Advances in Multimedia

Cross-Layer Optimized Wireless Multimedia Communications

Guest Editors: Haohong Wang, Jianwei Huang,
Mihaela Van Der Schaar, Dapeng Oliver Wu, and Zhu Han

Cross-Layer Optimized Wireless Multimedia Communications

Advances in Multimedia

Cross-Layer Optimized Wireless Multimedia Communications

Guest Editors: Haohong Wang, Jianwei Huang, Mihaela Van Der Schaar, Dapeng Oliver Wu, and Zhu Han

Contents

Hindawi Publishing Corporation
Advances in Multimedia
Volume 2007, Article ID 61391, 2 pages
doi:10.1155/2007/61391

Editorial

Cross-Layer Optimized Wireless Multimedia Communications

Zhu Han,[1] Haohong Wang,[2] D. Oliver Wu,[3] Jianwei Huang,[4] and M. Van Der Schaar[5]

[1] *Department of Electrical and Computer Engineering, Boise State University, 1910 University Drive, Boise, ID 83725, USA*
[2] *Marvell Semiconductor Incorporation, 5488 Marvell Lane, Santa Clara, CA 95054, USA*
[3] *Department of Electrical and Computer Engineering, University of Florida, P.O.Box 116130 , Gainesville, FL 32611-6130, USA*
[4] *Department of Information Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, China*
[5] *Electrical Engineering Department, University of California Los Angeles (UCLA), 66-147E Engineering IV Building, 420 Westwood Plaza, Los Angeles, CA 90095-1594, USA*

Received 26 August 2007; Accepted 26 August 2007

Recent advances in wireless and mobile communications provide ample opportunities for introducing new services. Supporting multimedia applications and services over wireless networks is challenging due to constraints and heterogeneities such as limited battery power, limited bandwidth, random time-varying fading effect, different protocols and standards, and stringent quality-of-service (QoS) requirements. Cross-layer design methodologies provide great promises for addressing these challenges and achieving reliable and high-quality end-to-end performance in wireless multimedia communications. In this special issue on cross-layer wireless multimedia communications, we have accepted a few papers that address such issues.

The first paper, "Quality-based backlight optimization for video playback on handheld devices," considers the problem caused by limited power consumption supply of mobile devices and the intensive battery resource request of the multimedia streaming applications. In this work, effective techniques for backlight energy saving are proposed to optimize the backlight dimming while maintaining acceptable service quality.

Both the second paper, "Asymptotic delay analysis for cross-layer delay-based routing in ad hoc networks," and the third paper, "A study on the usage of cross-layer power control and forward error correction for embedded video transmission over wireless links," provide analytical framework and evaluations for media delivery over wireless network. The former work considers multihop networks, where the accurate delay distribution estimation is used for QoS-based routing in multimedia applications. The latter paper considers single-hop network, where the source coding parameters, (e.g., packetization schemes and packet classification), medium access control procedures, (e.g., ARQ and forward error correction), and physical parameters (e.g., transmission power and channel sensing) are jointly optimized.

The forth paper, "Cross-layer path configuration for energy-efficient communication over wireless ad hoc networks," studies the energy-efficient configuration of multihop paths with ARQ mechanism in wireless ad hoc networks. The proposed approach jointly optimize the schedualing of the transmitting power of each transmitted node and the retransmission limit over each hop, under the constraints on maximal delay and minimal packet delivery ratio.

The fifth paper, "MOS-based multiuser multiapplication cross-layer optimization for mobile multimedia communication," proposes a cross-layer optimization strategy that jointly optimizes the application layer, the data-link layer, and the physical layer using a novel optimization scheme based on the mean opinion score as the unifying metric over different application classes.

The sixth paper, "Cross-layer perceptual ARQ for video communications over 802.11e wireless networks," presents an application-level perceptual ARQ algorithm for video streaming over 802.11e wireless networks. A simple and effective formula is proposed to combine the perceptual and temporal importance of each packet into a single priority value.

The seventh paper, "Cross-layer design of source rate control and congestion control for wireless video streaming," extends the QoS-aware congestion control mechanism to the wireless scenario, and provides a detailed discussion about how to enhance the overall performance in terms of rate smoothness and responsiveness of the transport protocol.

The eighth paper, "Identifying opportunities for exploiting cross-layer interactions in adaptive wireless systems,"

presents an analysis of the interaction of physical data link and network layer parameters with respect to throughput, bit-error rate, delay, and jitter. The goal of this analysis is to identify opportunities where systems designers might exploit cross-layer interactions to improve the performance of voice over IP, instant messaging, and file transfer applications.

The ninth paper, "MAC-layer QoS management for stream-ing rate-adaptive VBR video over IEEE 802.11e HCCA WLANs," proposes a cross-layer framework for efficiently delivering multiclass rate-adaptive variable bitrate 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.

The tenth paper, "A cross-layer optimization approach for energy efficient wireless sensor networks: coalition-aided data aggregation, cooperative communication, and energy balancing," proposes a cross-layer optimization approach to study energy-efficient data transport in coalition-based wireless sensor networks, where neighboring nodes are organized into groups to form coalitions, and sensor nodes within one coalition carry out cooperative communications.

In the final paper of this special issue, "Location-aware cross-layer design using overlay watermarks," location information of a mobile for efficient routing can be easily derived when a unique watermark is associated with each individual transceiver. In addition, cross-layer signaling and other interlayer interactive information can be exchanged with a new data pipe created by modulating the overlay watermarks.

The upcoming new applications for wireless multimedia communications open new opportunities for many interesting and comprehensive research topics targeting at concepts, methodologies, and techniques to support the future advanced mobile wireless applications. Clearly, the developments of the new schemes, mechanisms, and systems associated with the cross-layer designs and protocols will have a significant impact on the next generation of wireless communications and networks.

Zhu Han
Haohong Wang
Jianwei Huang
M. Van Der Schaar

Hindawi Publishing Corporation
Advances in Multimedia
Volume 2007, Article ID 83715, 10 pages
doi:10.1155/2007/83715

Research Article

Quality-Based Backlight Optimization for Video Playback on Handheld Devices

Liang Cheng,[1] Shivajit Mohapatra,[2] Magda El Zarki,[3] Nikil Dutt,[3] and Nalini Venkatasubramanian[3]

[1] *NVIDIA Corporation, Santa Clara, CA 95050, USA*
[2] *Motorola Labs, Schaumburg, IL 60196, USA*
[3] *Donald Bren School of Information and Computer Science, University of California, Irvine, CA 92697-3425, USA*

Received 18 December 2006; Accepted 23 April 2007

Recommended by Haohong Wang

For a typical handheld device, the backlight accounts for a significant percentage of the total energy consumption (e.g., around 30% for a Compaq iPAQ 3650). Substantial energy savings can be achieved by dynamically adapting backlight intensity levels on such low-power portable devices. In this paper, we analyze the characteristics of video streaming services and propose a cross-layer optimization scheme called quality adapted backlight scaling (QABS) to achieve backlight energy savings for video playback applications on handheld devices. Specifically, we present a fast algorithm to optimize backlight dimming while keeping the degradation in image quality to a minimum so that the overall service quality is close to a specified threshold. Additionally, we propose two effective techniques to prevent frequent backlight switching, which negatively affects user perception of video. Our initial experimental results indicate that the energy used for backlight is significantly reduced, while the desired quality is satisfied. The proposed algorithms can be realized in real time.

1. INTRODUCTION

With the widespread availability of 3G/4G cellular networks, mobile handheld devices are increasingly being designed to support streaming video content. These devices have stringent power constraints because they use batteries with finite lifetime. On the other hand, multimedia services are known to be very resource intensive and tend to exhaust battery resources quickly. Therefore, conserving power to prolong battery life is an important research problem that needs to be addressed, specifically for video streaming applications on mobile handheld devices.

Many handheld-device-based power saving techniques have been reported in the literature. They attempt to reduce power consumption at various computational levels. In [1], a number of architectural and software compiling strategies were proposed to optimize system cache and external memory access. Reference [2] specifically aims at MPEG-based applications. It proposes to scale the processor voltage and frequency to provide the necessary computing capability for decoding each video frame. In [3, 4], the power consumption of network interfaces (NICs) are optimized.

In this paper, we focus on the power consumption of the display unit in the handheld device. Most handheld devices are equipped with a thin-film transistor (TFT) liquid crystal display (LCD). For these devices, the display unit is driven by backlight illumination. The backlight consumes a considerable percentage of the total energy usage of the handheld device—for example, for a Compaq iPAQ device, it consumes 20%–40% of the total system power [5].

Dynamically dimming the backlight is considered an effective method to save energy [5–7]. The resultant reduced fidelity can be compensated for with scaling up of the pixel luminance. The luminance scaling, however, tends to saturate the bright part of the picture, thereby affecting the fidelity of the video quality.

In [6], a dynamic backlight luminance scaling (DLS) scheme is proposed. Based on different scenarios, three compensation strategies are discussed, that is, brightness compensation, image enhancement, and context processing. However, their calculation of the distortion does not consider the fact that the clipped pixel values do not contribute equally to the quality distortion. In [7], a similar method, namely, concurrent brightness and contrast scaling (CBCS),

is proposed. CBCS aims at conserving power by reducing the backlight illumination while retaining the image fidelity through preservation of the image contrast. Their distortion definition and proposed compensation technique may be good for static image-based applications, such as the graphic user interface (GUI) and maps, but might not be suitable for streaming video scenarios, because their contrast compensation further compromises the fidelity of the images. In addition, neither [6] nor [7] solves the problem associated with frequent backlight switching which can be quite distracting to the end user.

In this paper, we explicitly incorporate video quality into the backlight switching strategy and propose a quality adaptive backlight scaling (QABS) scheme. The backlight dimming affects the brightness of the video. Therefore, we only consider the luminance compensation such that the lost brightness can be restored. The luminance compensation, however, inevitably results in quality distortion. For the video streaming application, the quality is normally defined as the resemblance between the original and processed video. Hence, for the sake of simplicity and without loss of generality, we define the quality distortion function as the mean square error (MSE) (1) and the quality function as the peak signal to noise ratio (PSNR) (2), both of which are well-accepted objective video quality measurements,

$$\mathrm{MSE} = \frac{1}{M} \times \sum_{i=1}^{M} (x_i - y_i)^2, \tag{1}$$

$$\mathrm{PSNR(dB)} = 10 \log_{10} \sum_{i=1}^{M} \frac{255^2}{(x_i - y_i)^2}, \tag{2}$$

where x_i and y_i are the original pixel value and the reconstructed pixel value, respectively. M is the number of pixels per frame.

It is well known that MSE and PSNR are not the best measures to assess perceptual quality for most video sequences [8, 9]. However, they are widely used due to their simple implementation. A detailed discussion of the human visual system and the corresponding perceptual quality is beyond the scope of this paper. It is to be noted that any quality metric may be adopted to replace the used MSE and PSNR measures without affecting the validity of our proposed schemes.

As is mentioned in [7], for video applications, the continuous change in the backlight factor will introduce inter-frame brightness distortion to the observer. In our experiments, we find that the "unnecessary" backlight changes fall into two categories: (1) small continuous changes over adjacent frames; (2) abrupt huge changes over a short period. Therefore, we propose to quantize the calculated backlight to eliminate the small continuous change and use a low-pass digital filter to smooth the abrupt changes.

The rest of the paper is organized as follows. In Section 2, we introduce the principle of the LCD display—experimental results show that backlight dimming saves energy while the pixel luminance compensation results in minimal overhead. In Section 3, we present our QABS scheme, which includes determining the backlight dimming factor and two supplementary methods to avoid excessive backlight switching. Section 4 describes our prototype implementation, experimental methodology, and simulation results. We conclude our work in Section 5.

2. CHARACTERISTICS OF LCD

In this section, we outline the characteristics of the LCD unit from two perspectives, the LCD display mechanism and the LCD power consumption, both of which form the basis for our system design.

2.1. LCD display

The LCD panel does not illuminate itself; it displays by filtering the light source from the back of the LCD panel [6, 7]. There are three kinds of TFT LCD panels: transmissive LCD, reflective LCD, and transflective LCD. In transmissive LCD, the pixels are illuminated from behind (i.e., opposite the viewer) using a backlight. The transmissive LCD offers a high quality display with large power consumption, so it is widely used in laptop personal computers. The reflective LCD has a reflector on the back, which reflects the ambient environment light or a frontlight. Compared to a transmissive LCD, a reflective LCD uses modest amounts of power for illumination. Hence, most of the handheld devices use reflective LCD. Transflective LCD combines both transmissive and reflective mechanisms but is not as commonly used in handheld devices as the other two types.

In general, both transmissive and reflective mode LCD need artificial light source to illuminate the display. Hence, reducing the light power consumption is beneficial to all these three types of LCDs. For the sake of simplicity and without loss of generality, we use backlight to represent all types of light source—frontlight is also designated as backlight in the case of a reflective LCD. All of our proposed algorithms are applicable to both backlight and frontlight. Since the reflective mode LCD is more popular to the handheld devices, we base our algorithms and measurements on a Compaq iPAQ 3650—a PDA with reflective mode LCD. Figure 1 illustrates the schematic mechanism of the reflective mode LCD.

The perceptual luminance intensity of the LCD display is determined by two components: backlight brightness and the pixel luminance. The pixel luminance can be adjusted by controlling the light passing through the TFT array substrate. Users may detect a change in the display luminance intensity if either of these two components is adjusted. That is, the backlight brightness and the pixel luminance can compensate for each other. In Section 2.2, we will show that the pixel luminance does not have a noticeable impact on the energy consumption, whereas the backlight illumination results in high energy consumption. In general, dimming the backlight level while compensating for it with the pixel luminance is an effective way to conserve battery power in handheld devices.

We can therefore conclude that reducing backlight and increasing pixel luminance will result in power savings.

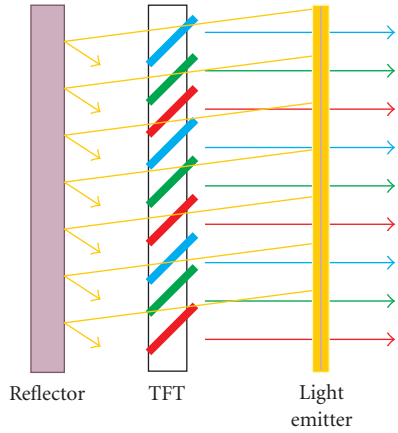


Figure 1: Reflective LCD.

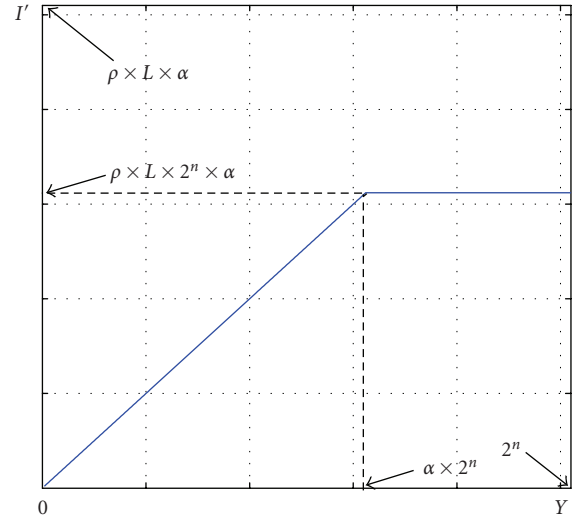


Figure 2: Clipping effect of LCD.

Let the backlight brightness level and the pixel luminance value be L and Y, respectively, and the perceived display luminance intensity be I. We may denote I using the following:,

$$I = \rho \times L \times Y, \tag{3}$$

where ρ is a constant ratio, denoting the transmittance attribute of the LCD panel, and as such $\rho \times Y$ is the transmittance of the pixel luminance.

We may reduce the backlight level to L' by multiplying L with a dimming factor α, that is, $L' = L \times \alpha$, $0 < \alpha < 1$. To maintain the overall display luminance I invariable, we need to boost the luminance of the pixel to Y'. Since the pixel luminance value is normally restricted by the number of bits that represent it (denoted as n), Y' may be clipped if the original value of Y is too high or the α is too low. The compensation of the backlight is described in (4),

$$Y' = \begin{cases} \dfrac{Y}{\alpha} & \text{if } Y < \alpha \times 2^n, \\ 2^n & \text{if } Y \geq (\alpha \times 2^n). \end{cases} \tag{4}$$

Combining (4) and (3), we have

$$I' = \begin{cases} I & \text{if } Y < \alpha \times 2^n, \\ \rho \times L \times \alpha \times 2^n & \text{if } Y \geq (\alpha \times 2^n). \end{cases} \tag{5}$$

Equation (5) clearly shows that the perceived display intensity may not be fully recovered, instead, it is clipped to $\rho \times L \times \alpha \times 2^n$ if $Y \geq (\alpha \times 2^n)$. In Figure 2, we illustrate the clipping effect of the display luminance.

In Figure 3, we show an image and its luminance histogram. This image is the first frame of a typical news video clip "ABC eye witness news" captured from a broadcast TV signal. Figure 4 illustrates the image and its luminance histogram after backlight dimming and pixel luminance compensation. Figure 4(b) shows that pixels with luminance higher than 156 are all clipped to 156. Compared with Figure 4(a), this clipping effect diminishes the differences in the brightness areas of the image, for example, the white caption and the face. This effect is subjectively perceived as luminance saturation and is objectively assessed as 30 dB using the PSNR quality metric with reference to the original image shown in Figure 3(a).

2.2. LCD power model

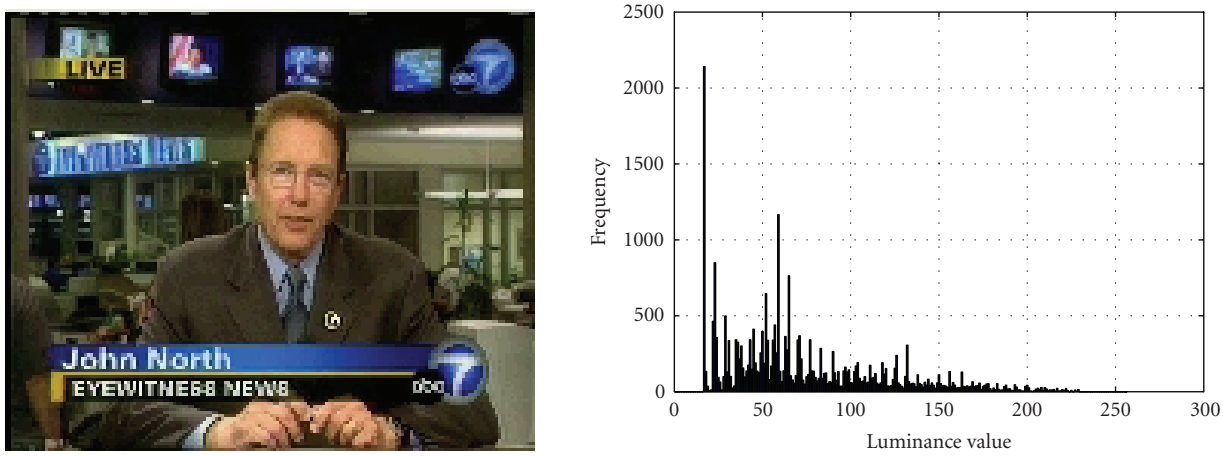
We ran several experiments and observed that dimming the backlight results in energy savings, whereas the compensation process, that is, scaling up the luminance of the pixel, has a negligible energy overhead. We measure the energy saving as the difference between the total system power consumption with the backlight set to different levels to that with the backlight turned to the maximum (brightest). In Figure 5, we plot the various backlight levels and their corresponding energy consumption for a Compaq iPAQ 3650 running Linux. A more detailed setup of our experiments is described in Section 4. It is noticed that the backlight energy saving is almost linear to the backlight level and can be estimated using (6),

$$y = a1 \times x + a2, \tag{6}$$

where y is the energy savings in watt; x denotes the backlight level; $a1$ and $a2$ are coefficients. We apply the curve fitting function of MATLAB and obtain $a1 = -0.0029567$ and $a2 = 0.73757$ with the largest residual fitting error as 0.085731.

Contrary to the backlight dimming, the pixel luminance scaling is uncorrelated to the energy consumption. In Figure 6, we show that for one specified backlight level (BL) the system energy consumption basically remains stable and is independent of the luminance scaling.

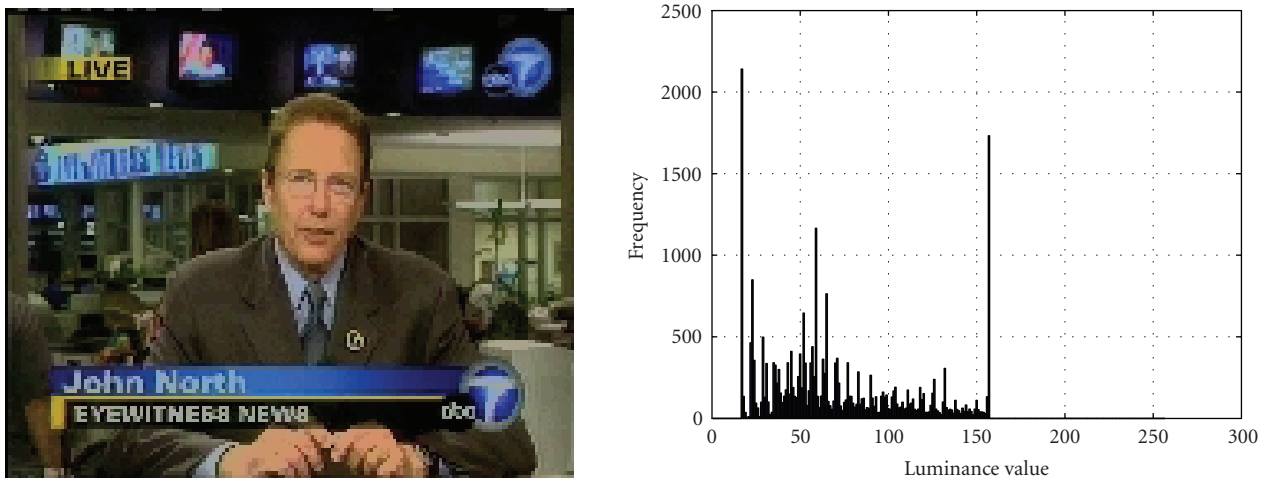
Figures 5 and 6 justify the validity of the proposed backlight power conservation approach, that is, dimming the backlight while enhancing the pixel luminance value. Note that in Figure 6,"BL" refers to the backlight level and



(a) Original image

(b) Histogram before clipping

Figure 3: Image and its luminance histogram before clipping.



(a) Compensated image

(b) Histogram after clipping

Figure 4: Image and its luminance histogram after clipping.

"luminosity scaling factor" refers to α. In the next section, we apply this method to the video streaming scenario, discussing a practical scheme to optimize the backlight dimming while taking into consideration the effect on video distortion.

3. ADAPTIVE BACKLIGHT SCALING

As explained in (5), the backlight scaling with the luminance compensation may result in quality distortion. The amount of backlight dimming, therefore, has to be restricted such that the video fidelity will not be seriously affected.

3.1. Optimized backlight dimming

We define the optimized backlight dimming factor as the one whose induced distortion is closest to a specified threshold. Henceforth, we replace the factor α with the real backlight level Alfa, Alfa $= N \times \alpha$ (N is the number of backlight levels (256 for Linux on iPAQ)), and the optimized backlight dimming is represented as Alfa*.

In Figure 7, we illustrate the image quality distortion in terms of MSE over different backlight levels. (Note that we use the image shown in Figure 3(a).) We see that as Alfa increases, the induced video quality distortion due to the brightness saturation monotonously decreases. Hence, for a given distortion threshold, we can find a unique Alfa(= Alfa*) for each image. In video applications, for a given distortion, different frames may have distinct Alfa*, depending on the luminance histogram of that frame. However, it is hard to have an accurate analytical representation of the quality distortion using Alfa as a parameter. We therefore adopt an optimized search-based-approach, where we calculate the MSE distortion with different Alfa until the specified distortion threshold is met. The results of our scheme

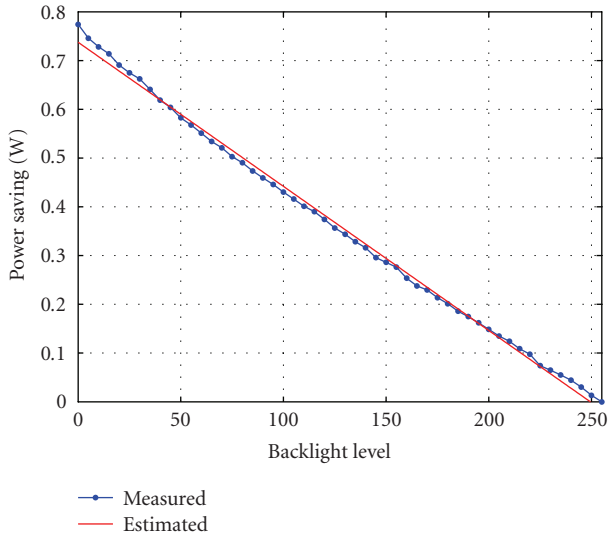


Figure 5: Power saving versus backlight level.

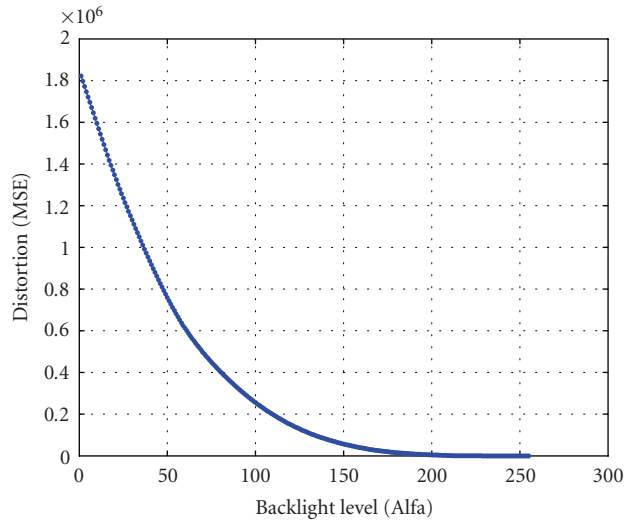


Figure 7: MSE with different Alfa.

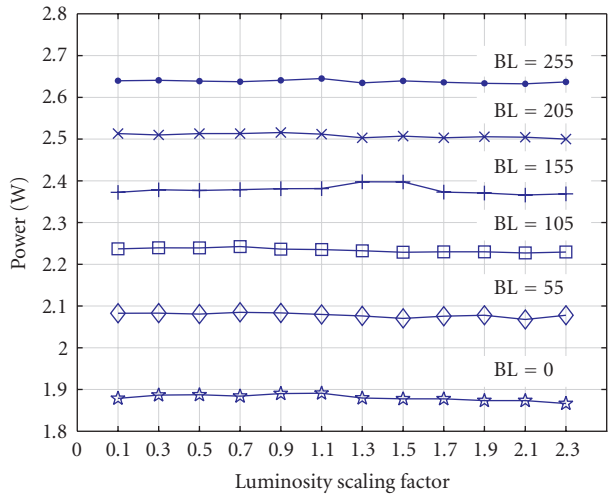


Figure 6: Energy overhead of luminosity scaling.

are accurate and can be used as the benchmark for the design of other analytical methods.

Algorithm 1 shows the exhaustive search algorithm for finding Alfa* for one image. FindAlfa(th) takes the distortion threshold (th) as input, and returns Alfa* as output. Note that MSE(Alfa) calculates the MSE with the specified Alfa for one frame.

However, the complexity of an exhaustive search shown in Algorithm 1 is too high. As shown in (2), the per-frame MSE calculation consists of M multiplications and $2M$ additions. M is the number of pixels in one frame, for example, $M = 25344$ for QCIF format video. We regard the per-frame MSE as the basic complexity measurement unit. We assume that the optimized backlight level is uniformly distributed in $[0, N]$, and thus the complexity of algorithm in Algorithm 1 is $O(N)$. In our test, $N = 256$. It is obvious that the optimized backlight dimming factor cannot be calculated in real time.

We therefore apply a faster bisection method [10] to improve the algorithm for finding Alfa*. Since we can easily find an upper bound (denoted as u) and a lower bound (denoted as d) on the backlight level, we obtain a good approximation using this method. We assume that $u > d$ and let ϵ be the desired precision. The algorithm based on the bisection method is illustrated in Algorithm 2.

By using the bisection method, we achieve a complexity of $O(\log_2 N)$ in the worst case. For instance, for $N = 256$ and $\epsilon = 1$, we only need to calculate the per-frame MSE at most eight times, which is relatively fast and can be realized in real time.

3.2. Smoothing the backlight switching

It has been discussed in [7] that the backlight dimming factor may change significantly across consecutive frames for most video applications. We call this abrupt backlight switching "flicker moment." The frequent switching of the backlight may also introduce an interframe brightness distortion to the observer due to the brightness compensation. Hence, it is necessary to reduce frequent backlight switching.

In our study, we observe that the calculated Alfa*, although based on an individual image, does not experience huge fluctuations during a video scene, that is, a group of frames that are characterized with similar content. Actually, the redundancy among adjacent frames constitutes the major difference between the video and the static image application and has long been utilized to achieve higher compression efficiency. Hence, the backlight switching should be smoothed out within the scene and most favorably only happen at the boundary of video scenes.

We propose two supplementary methods to smooth the acquired Alfa* in the same video scene. First, we apply

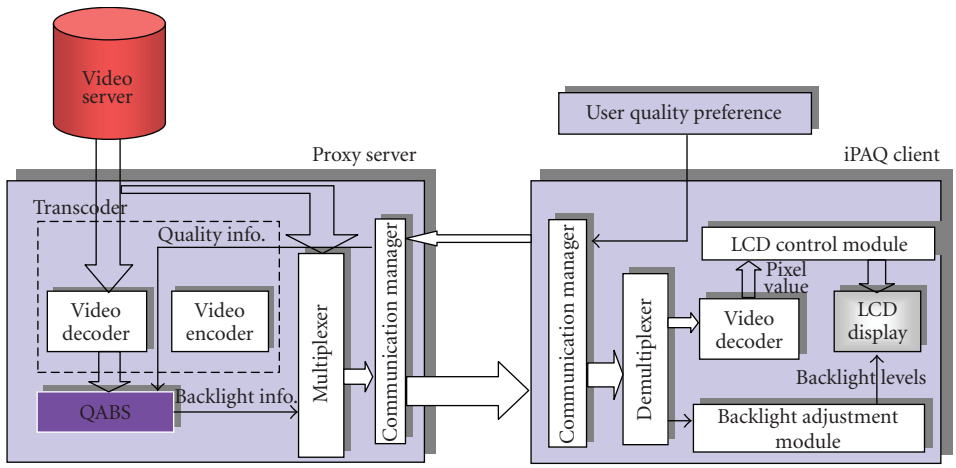


Figure 8: Prototype implementation.

Proc: FindAlfa(th)
(1) Alfa := 0;
(2) **while** Alfa ≤ N **do**
(3) **if** MSE(Alfa) > th **then**
(4) Alfa := Alfa + 1;
(5) **else**
(6) return(Alfa);
(7) **end if**
(8) **end while**

Algorithm 1: Exhaustive algorithm for finding Alfa*.

Proc: FastFindAlfa(th,ϵ)
(1) u := upper bound;
(2) d := lower bound;
(3) **while** $(u - d) > \epsilon$ **do**
(4) Alfa = round
$((d + u)/2)$;
(5) **if** (MSE(Alfa) > th) **then**
(6) u = Alfa;
(7) **else**
(8) d = Alfa;
(9) **end if**
(10) **end while**
(11) return(Alfa);

Algorithm 2: Fast algorithm for finding Alfa*.

a low-pass digital filter to eliminate any abrupt backlight switching that is caused by the unexpected sharp luminance change. The passband frequency is determined by the subjective perception of the "flicker moment" and the frame display rate. Second, we propose to quantize the number of backlight levels, that is, any backlight level between two quantization values can be quantized to the closest level, by which we prevent the needless backlight switching for small luminance fluctuations during one scene. In our experiments, we

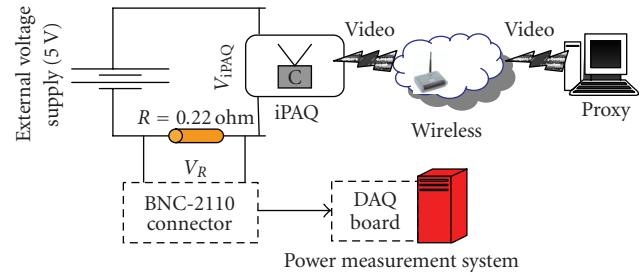


Figure 9: Setup for our measurements.

quantize all 256 levels to "N" levels (N = 5 in our study). We switch the backlight level only if the calculated Alfa* changes drastically enough, and falls into another quantized level.

4. PERFORMANCE EVALUATION

In this section, we introduce our prototype implementation, the methodology of our measurement, and the performance of the proposed algorithm.

4.1. Prototype implementation

Figure 8 shows a high level representation of our prototype system. Our implementation of the video streaming system consists of a video server, a proxy server, and a mobile client. We assume that all communication between the server and the mobile client is routed through a proxy server typically located in proximity to the client.

The video server is responsible for streaming compressed video to the client. The proxy server transcodes the received stream, adds the appropriate control information, and relays the newly formed stream to the mobile client (Compaq iPAQ 3650 in our case). In our initial implementation, for the sake of simplicity and without loss of generality, we use the proxy server to also double up as our video server.

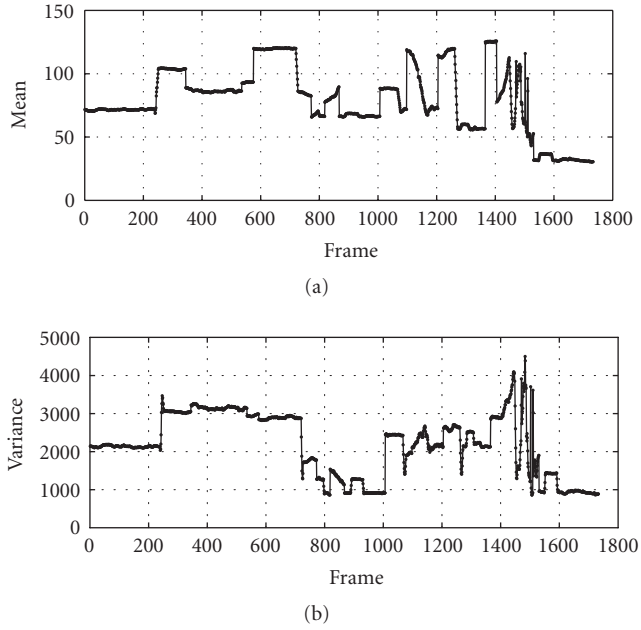


Figure 10: Basic statistics of *abc_news*.

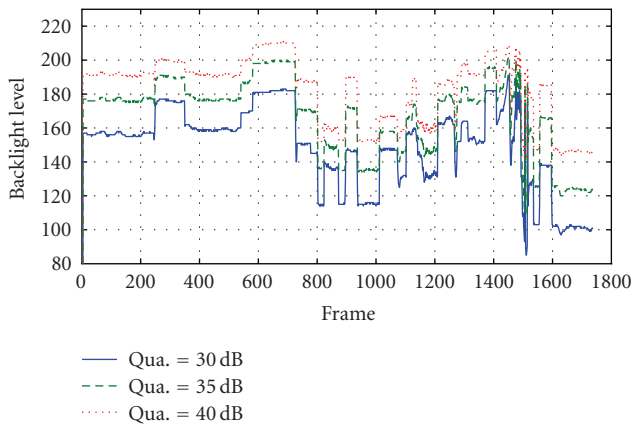


Figure 11: Alfa* adapted to three-given quality thresholds.

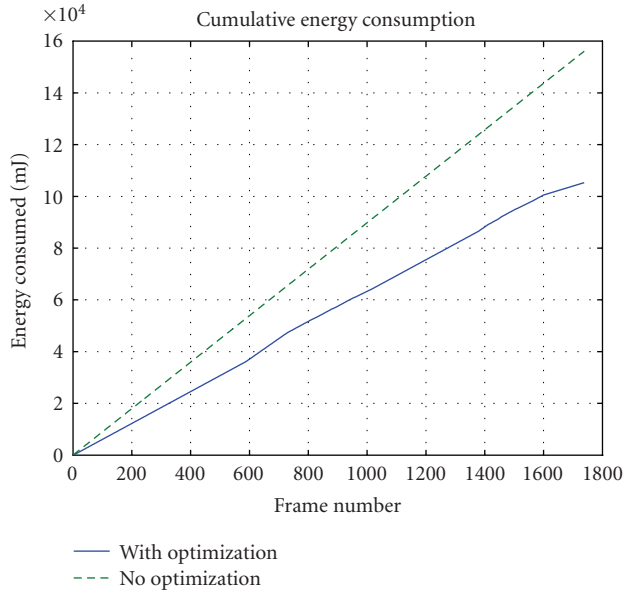


Figure 12: Energy consumption with and without optimization for *abc_news* video clip.

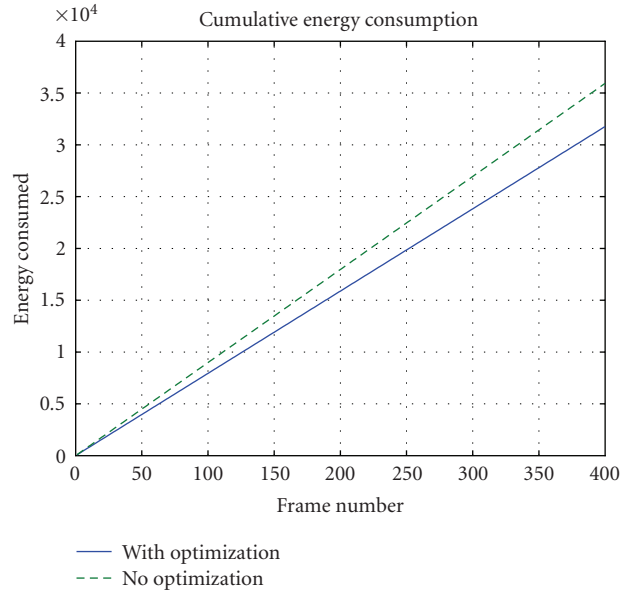


Figure 13: Energy consumption with and without optimization for *Foreman* video clip.

The proxy server includes four primary components—the video transcoder, the proposed QABS module, the signal multiplexer, and the communication manager. The transcoder uncompresses the original video stream and provides the pixel luminance information to the QABS module. The QABS module calculates the optimized backlight dimming factor based on the user quality preference feedback received from the client (user). The multiplexer is used to multiplex the optimized backlight dimming information with the video stream. The communication manager is used to send this aggregated stream to the client.

On the mobile client, the demultiplexer is used to recover the original video stream and the encoded backlight information from the received stream. The LCD control module renders the decoded image onto the LCD display. The backlight information is fed to the 'backlight adjustment module," which concurrently sets the backlight value for the LCD. In particular, users may send the quality request to the proxy when requesting a video sequence, based on his/her quality preference as well as concern for battery consumption.

Table 1: Results of QABS (G: good; F: fair; E: Excellent).

Alfa mean			Quality (dB)			Power saving(%)		
F	G	E	F	G	E	F	G	E
149	162	186	30.17	34.28	42.31	41.8%	36.7%	27.3%

4.2. Measurement methodology

For video quality and power measurements, we use the setup shown in Figure 9. The proxy in our experiments is a Linux desktop with a 1 GHz processor and 512 MB of RAM. All our measurements are made on a Compaq iPAQ 3650. In order to control the backlight and pixel luminance, we develop our own Linux-based API functions. We use a national instruments PCI DAQ board to sample voltage drops across a resistor and the iPAQ, and sample the voltage at 200 K samples/s. We calculate the instantaneous and average power consumption of the iPAQ using the formula $P_{\text{iPAQ}} = (V_R/R) \times V_{\text{iPAQ}}$.

4.3. Experimental results

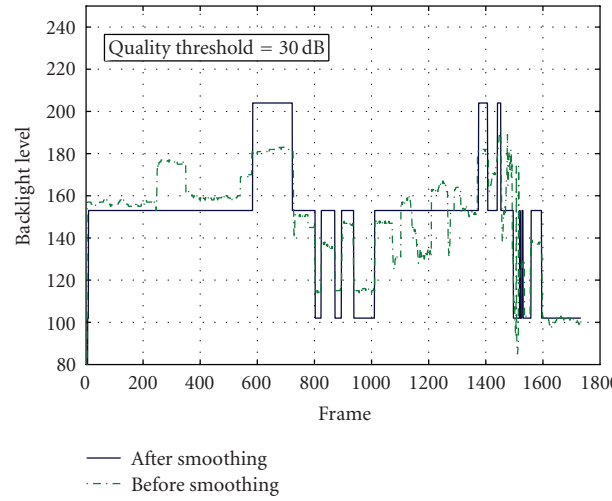
In our simulation, we use a video sequence captured from a broadcasted *ABC_news* program, whose first frame is shown in Figure 3(a). We chose this video as representative of a typical usage of a PDA—commuters watching the evening news on the way home. In Figure 10, we show the basic statistics (i.e., the mean and the variance of luminance per frame) of this video.

We assume that the users are given three quality options, fair, good, and excellent, which respectively correspond to the PSNR value of 30 dB, 35 dB, and 40 dB. After applying the algorithm "**Proc**: FastFindAlfa," we obtain the adapted Alfa* for these three quality preferences, as is shown in Figure 11. It can be seen that higher video quality needs higher backlight level on average.

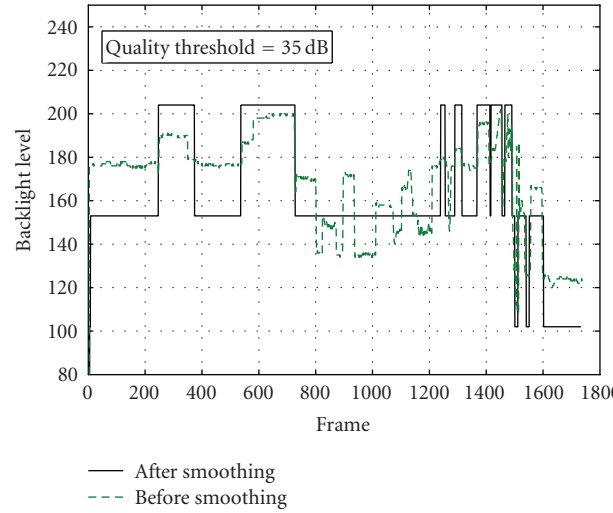
In Figure 14, we show Alfa* before and after the backlight smoothing process for different quality preferences. It is seen that the small variation and the abrupt change of the backlight switching are significantly eliminated after the filtering and quantization. In addition, as we expected, the backlight switching mostly happens at the boundary of major scenes.

In Table 1, we summarize the results of our QABS. The mean Alfa* of different quality preferences produces a quality on average very close to the predetermined quality threshold. It is noted that different quality requirements result in various power saving gains. Higher quality preference must be traded using more backlight energy. Nevertheless, we can still save 29% energy that is supposed to be consumed by the backlight unit if we set the quality preference to be "Excellent."

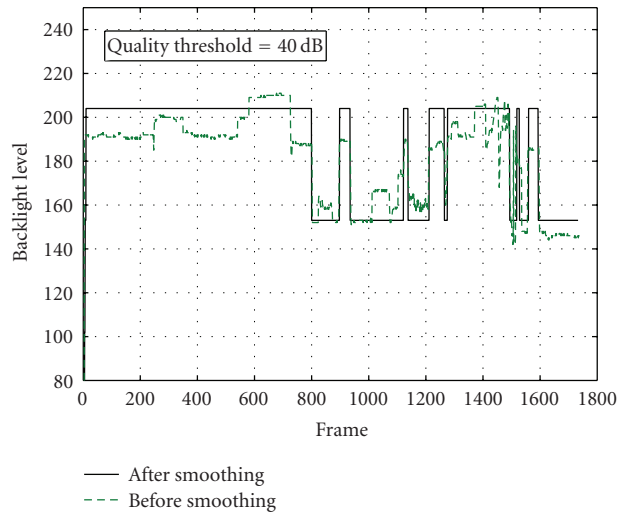
In Figure 15, we show that the filtering and quantization processes may lead to instantaneous quality fluctuation, which is contrasted to the consistent quality before backlight smoothing. Nevertheless, we observe that the quality fluctuation is around the designated quality threshold and mostly happens at scene changes.



(a) Alfa* for fair quality

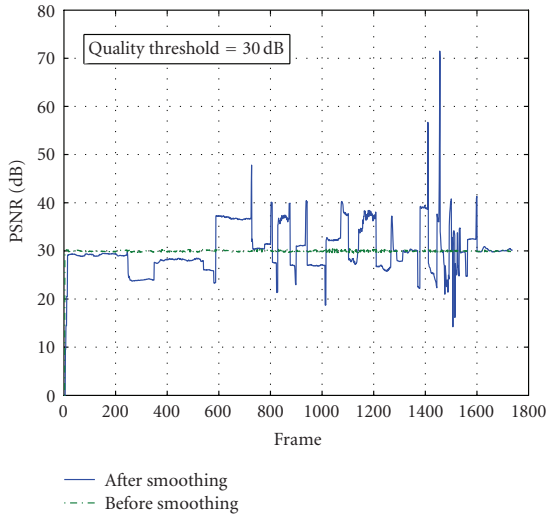


(b) Alfa* for good quality

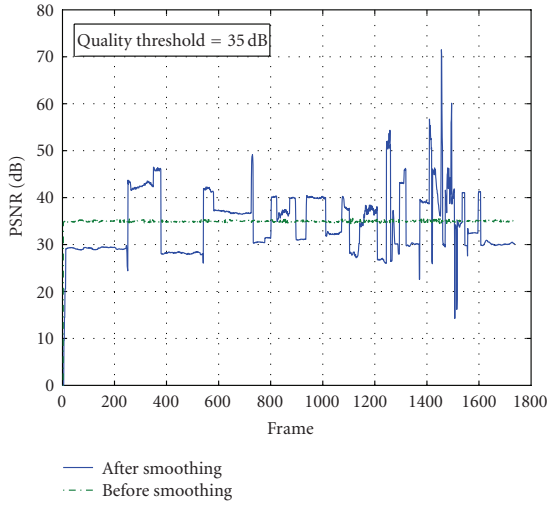


(c) Alfa* for excellent quality

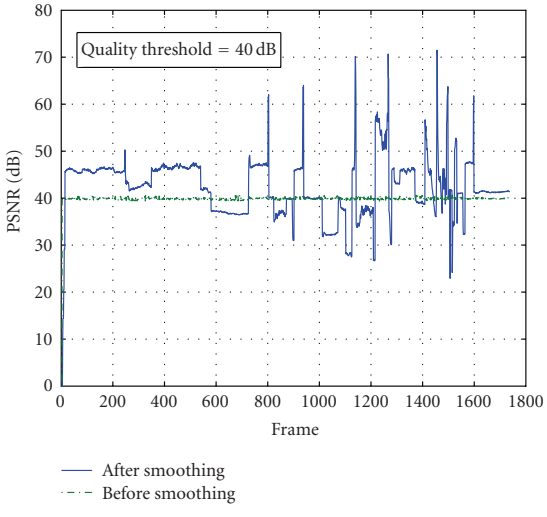
Figure 14: Optimized Backlight level before and after filtering and quantization.



(a) Fair quality



(b) Good quality



(c) Excellent quality

FIGURE 15: Actual PSNR quality before and after filtering and quantization.

In Figures 12 and 13, we compare the actual energy consumption on a Compaq ipaq with and without our quality aware backlight adaptation. As seen from the graph, the energy savings from our backlight adaptation for the ABCNews clip is 35%–40% of the total energy consumed due to backlight. Even for videos that offer very little opportunity to aggressively perform backlight adaptation (e.g., foreman video clip, which is simply a talking head), we can achieve energy savings as high as 14–20% with negligibal video quality sacrificing.

5. CONCLUSION

In this paper, we apply a backlight scaling technique to a proxy-based video streaming framework. We explicitly associate backlight switching to the perceptual video quality in terms of PSNR. The proposed adaptive algorithm is fast and effective for reducing the energy consumption while maintaining the designated video quality. To reduce the frequency of backlight switching, we also propose two supplementary schemes to smooth the backlight switch process such that the user perception of the video stream can be substantially improved. Our experiment shows that by applying our scheme, up to 40% power can be saved with negligibal video quality sacrificing.

ACKNOWLEDGMENTS

The authors would like to thank Stefano Bossi and Michael Philpott who helped them with the simulations and experimental system setup. This work was partially supported by National Science Foundation.

REFERENCES

[1] A. Azevedo, R. Cornea, I. Issenin, et al., "Architectural and compiler strategies for dynamic power management in COPPER project," in *Proceedings of International Workshop on Innovative Architecture (IWIA '01)*, Maui, Hawaii, USA, January 2001.

[2] K. Choi, W.-C. Cheng, and M. Pedram, "Frame-based dynamic voltage and frequency scaling for an MPEG player," *Journal of Low Power Electronics*, vol. 1, no. 1, pp. 27–43, 2005.

[3] S. Chandra, "Wireless network interface energy consumption: implications for popular streaming formats," *Multimedia Systems*, vol. 9, no. 2, pp. 185–201, 2003.

[4] M. Stemm, P. Gauthier, D. Harada, and R. H. Katz, "Reducing power consumption of network interfaces in hand-held devices," in *Proceedings of the 3rd International Workshop on Mobile Multimedia Communications (MoMuc '96)*, Princeton, NJ, USA, September 1996.

[5] S. Pasricha, M. Luthra, S. Mohapatra, N. Dutt, and N. Venkatasubramanian, "Dynamic backlight adaptation for low-power handheld devices," *IEEE Design and Test of Computers*, vol. 21, no. 5, pp. 398–405, 2004.

[6] N. Chang, I. Choi, and H. Shim, "DLS: dynamic backlight luminance scaling of liquid crystal display," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 8, pp. 837–846, 2004.

[7] W.-C. Cheng, Y. Hou, and M. Pedram, "Power minimization in a backlit TFT-LCD display by concurrent brightness and

contrast scaling," in *Proceedings of Design, Automation and Test in Europe Conference and Exhibition (DATE '04)*, vol. 1, pp. 252–257, Paris, France, February 2004.

[8] S. Wolf and M. H. Pinson, "Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system," in *Proceedings of the Multimedia Systems and Applications II*, vol. 3845 of *Proceedings of SPIE*, pp. 266–277, Boston, Mass, USA, September 1999.

[9] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[10] J. L. Zachary, *Introduction to Scientific Programming: Computational Problem Solving Using Maple and C.*, Telos Publishers, Surrey, UK, 1996.

Hindawi Publishing Corporation
Advances in Multimedia
Volume 2007, Article ID 90879, 11 pages
doi:10.1155/2007/90879

Research Article

Asymptotic Delay Analysis for Cross-Layer Delay-Based Routing in Ad Hoc Networks

Philippe Jacquet,[1,2] Amina Meraihi Naimi,[2] and Georgios Rodolakis[1,2]

[1] *Ecole Polytechnique, 91128 Palaiseau Cedex, France*
[2] *Institut National de Recherche en Informatique et en Automatique, Unité de recherche de Rocquencourt, 78153 Le Chesnay Cedex, France*

Received 21 December 2006; Revised 20 April 2007; Accepted 14 May 2007

Recommended by Haohong Wang

This paper addresses the problem of the evaluation of the delay distribution via analytical means in IEEE 802.11 wireless ad hoc networks. We show that the asymptotic delay distribution can be expressed as a power law. Based on the latter result, we present a cross-layer delay estimation protocol and we derive new delay-distribution-based routing algorithms, which are well adapted to the QoS requirements of real-time multimedia applications. In fact, multimedia services are not sensitive to average delays, but rather to the asymptotic delay distributions. Indeed, video streaming applications drop frames when they are received beyond a delay threshold, determined by the buffer size. Although delay-distribution-based routing is an NP-hard problem, we show that it can be solved in polynomial time when the delay threshold is large, because of the asymptotic power law distribution of the link delays.

1. INTRODUCTION

With the emergence of real-time applications in wireless networks, delay guarantees are increasingly required. In order to provide support for delay sensitive traffic in such networks, an accurate evaluation of the delay is a necessary first step. Moreover, in the case of mobile ad hoc networks, a detailed understanding of the impact of both the underlying MAC and routing protocols on the delay characteristics is crucial, since the traffic routes consist of several wireless hops. In this paper, we address the problem of the analytical evaluation of the delay distribution in a multihop wireless network. We consider a wireless network with IEEE 802.11 MAC protocol [1] under the optimized link state routing (OLSR) protocol [2]. The routing protocol is a table-driven protocol that operates under periodic broadcast control packets. IEEE 802.11 is the most popular MAC protocol in wireless LANs and mobile ad hoc networks. The main channel access mechanism is the distributed coordination function (DCF), which is a carrier sense multiple access with collision avoidance (CSMA/CA) scheme. DCF was designed initially for asynchronous traffic and since it is a random access protocol it does not provide any guarantee for delay sensitive applications.

The performance of the 802.11 protocol in single-hop wireless networks has been addressed in the past in several papers. A simple analytical model of the 802.11 DCF access mechanism was introduced in [3] and was used to analyze the saturation throughput performance. The MAC layer service time was studied in [4], by expanding the previous model. The delay in both saturated and unsaturated networks was also studied in [5], where each node was modelled as a discrete time queue. Multihop delay was studied in [6], in the context of mesh networks.

In this work, we evaluate the 802.11 protocol performance in terms of delay in the context of wireless ad hoc networks. We focus on the asymptotic properties of the delay and we obtain both one hop and multihop analytical delay estimates. We show that the asymptotic delay distribution can be expressed as a power law. Based on the latter result, we present a cross-layer delay estimation protocol and we derive new delay-distribution-based routing algorithms, which are well adapted to the QoS requirements of real-time multimedia applications.

We take the slotted time approach of [3]. We denote by W the end-to-end delivery delay of a packet. We analyze the delay distribution $P(W > T)$ and we show that in the case that

T is large (i.e., several times the average delay) the probability $P(W > T)$ decays as a power law, namely in $O(T^{-a})$, where a is a constant. In order to simplify the formula derivations, we perform the analysis under certain modelling assumptions. For instance, we use an M/G/1 queueing model for the nodes (a detailed description is provided in the methodology section), which cannot be considered a priori entirely realistic. However, we have verified via simulations that our model is pertinent and can accurately predict the shape of the node delay distributions in the domain of interest. Furthermore, the assumptions we make are not fundamental for our results, therefore we comment on plausible model generalizations whenever possible.

Based on the analysis, we present a cross-layer framework to evaluate the delay distribution $P(W > T)$ for any large T and use it in order to find routes that satisfy given delay requirements. A delay oriented quality of service for a connection is generally expressed via a maximum acceptable delay T and a maximum over-delay ratio ϵ, specified by the application, requiring that during the connection the constraint $P(W > T) < \epsilon$ is verified. In general finding the optimal route that minimizes an over-delay ratio is NP-hard [7]. Nevertheless, the fact that the delay distribution at every node router is in power law allows us to specify a polynomial approximation algorithm with an error factor of $1 + O(T^{-1})$.

The rest of the paper is organized as follows. In Section 2, we present an overview of the IEEE 802.11 DCF mode. In Section 3 we introduce the general model framework and we analyze the one hop delay. The multihop delay distribution is derived in Section 4. In Section 5 we describe a cross-layer delay estimation protocol based on OLSR. In Section 6 we present some simulation results. Finally, in Section 7, we indicate how the previous analysis can be used in delay-distribution-based routing.

2. 802.11 DISTRIBUTED COORDINATION FUNCTION OVERVIEW

The distributed coordination function (DCF) is the fundamental access method used in the IEEE 802.11 MAC protocol [1]. It is based on the carrier sense multiple access with collision avoidance (CSMA/CA) mechanism, which is designed to reduce the collisions due to multiple sources transmitting simultaneously on a shared channel. In the CSMA/CA protocol, a station transmits only if the medium is idle. The medium is considered as idle if it is sensed to be idle for a duration greater than the distributed inter-frame space (DIFS). If the medium is sensed as busy, the transmission is deferred until the end of the ongoing transmission. When the medium becomes idle after a busy period, the node does not transmit immediately, because multiple stations could have been waiting for the end of the transmission and may attempt to access the channel again at the same time. Therefore, the node starts a random wait by initializing its *backoff timer*. The backoff timer is randomly selected in an interval called the *contention window* and has the granularity of one slot. Every time the channel is sensed to be idle, the backoff counter is decremented. When the counter reaches zero, the node can start its transmission. If the channel is sensed as busy during the backoff procedure, the counter is frozen and then resumed when the channel becomes idle for a DIFS again. In spite of that, collisions can still occur. In order to reduce the probability of further collisions, the contention window is doubled after each collision to increase the random waiting time. The exponential backoff function is discussed in a more detailed manner in Section 3.2. To make sure that the transmitted frame has reached its destination, an acknowledgment frame is generated from the destination to the source.

The above carrier sense is called physical carrier sense because it is performed at the air interface. A virtual carrier sense is also possible in the DCF mode to resolve the problem of the hidden terminal. This problem occurs when two nodes that are not within hearing distance of each other create collisions at a third terminal that receives the transmission from both. The virtual carrier sense is performed at the MAC sub-layer. The channel is reserved before each transmission, so instead of transmitting the data frame after sensing that the channel is idle, the station sends an RTS (request to send) frame to the destination. The receiver replies by a CTS (clear to send) frame after which data transfer can start. However, the use of RTS/CTS frames imposes additional delay and bandwidth overhead. Therefore the RTS/CTS mechanism is recommended only for big packets.

3. ONE-HOP DELAY DISTRIBUTION ANALYSIS

3.1. Methodology overview

A wireless node can be seen as a buffer filled by incoming messages and with a single server that performs the CSMA/CA multiple access protocol. We model this system as an M/G/1 queue, that is, we assume

(1) the input packet flow in the buffer is Poisson of rate λ;

(2) service delays are independent.

In fact, the M/G/1 hypothesis is just a matter of simplifying approach. Since we are going to deal with heavy tailed distribution of service times, the consequence on queueing time distribution can be generalized to a much larger class of queueing models. For example, it is not necessary to assume independence between service times or to restrict to Poisson input in order to derive a power law queueing distribution, but in this case the coefficients change (see [8] for the case of a GI/GI/1 queue). Nonetheless, as we verify later in the simulations section, the M/G/1 hypothesis leads to satisfactory results.

3.2. Service delay determination

The IEEE 802.11 CSMA/CA protocol uses a rotating backoff where the nodes have to wait a random number of idle slots between transmission attempts. Let **C** be the random variable that expresses the number of busy slots between two

consecutive idle slots. Let $p(L)$ be the probability of collision that is experienced by packets, when the packet length is L (i.e., the physical transfer time of the packet in the channel, expressed in slots, which is proportional to its size in bits). The longer the packet is, the more likely it is to collide. We take the following assumptions.

(1) Durations between successive idle slots are independent and identically distributed.

(2) Collision events on successive transmissions are independent.

According to the CSMA protocol, the backoff counter is selected in an initial interval $\{1,\dots,W_{\min}\}$. If a collision occurs the nodes select a new backoff number in an enlarged interval $\{1,\dots,2W_{\min}\}$. The backoff interval length is multiplied by two after each collision. The backoff interval length is reset to $W_{\min}$ for the next packet. In practice there is a maximum number of retries after which the packet is discarded in case of permanent failure. The default maximum retry is 7 and can lead to a delay on the order of seconds. Since this delay is larger than the maximum acceptable delay we think of regarding connection QoS, it does not practically matter to set the maximum number of retries to infinity.

Let $C(z)$ be the probability generating function $\sum_n P(\mathbf{C} = n)z^{n+1}$, quantity C being expressed in slot duration. This generating function corresponds to the time needed for a backoff counter decrease, expressed by the random variable $C + 1$ (we add one slot to the quantity C for the decrease to be taken into account). Identity $C(z) = z$ would mean that $\mathbf{C} = 0$ always, that is, the channel is permanently sensed idle (note that in this case one slot is still needed for the counter decrease).

Let $\beta(z, L, p, k)$ be the probability generating function of the service delay when the packet length is L, the collision probability is p, and the initial backoff interval is k. The service delay of a packet corresponds to the time elapsed since it was extracted from the buffer until it is transmitted successfully. Therefore, it takes into account retransmissions due to collisions and it includes the time needed to access to the channel (corresponding to the rotating backoff decrementation) plus the fixed packet transmission length.

We will express all these quantities using generating functions, starting from the time needed to access the channel, or equivalently the backoff counter decrease. As discussed earlier, each backoff decrease is expressed by the random variable $C + 1$, with generating function $C(z)$. If the backoff counter is i, the total time to access the channel is the time needed for i counter decreases, or the sum of i times the random variable $C + 1$. From the independence assumption it comes that in this case the channel access time can be expressed by generating function $C(z)^i$. Since the initial backoff window is k, and the backoff counter value is selected uniformly at random in the interval $\{1,\dots,k\}$ (we also take here into account the DIFS interval), the generating function of the total channel access time can be written as $(1/k)\sum_{i=1,\dots,k} C(z)^i$, which results from the previous discussion by taking either possible value for i with probability $1/k$. Once the channel is accessed, the time needed to transmit the packet is fixed and equal to L,[1] therefore it can be expressed by generating function z^L. Hence the service time when no collision occurs comes from adding the previous two quantities, or equivalently the corresponding generating function is equal to the product of the above generating functions, that is,

$$\frac{z^L}{k}\sum_{i=1,\dots,k} C(z)^i = \frac{C(z)^{k+1} - C(z)}{C(z) - 1}\frac{z^L}{k}. \tag{1}$$

In order to account for packet collisions, we obtain the following recursion:

$$\beta(z, L, p, k) = \frac{C(z)^{k+1} - C(z)}{C(z) - 1}\frac{z^L}{k} \times (1 - p + p\beta(z, L, p, 2k)). \tag{2}$$

In case there is no collision (with probability $1 - p$), the service delay corresponds to our previous calculations. The term $\beta(z, L, p, 2k)$ is obtained from the case where there is a collision (with probability p), hence the procedure is repeated after doubling the backoff interval and this results in an additional service delay term.

The service delay probability generating function is

$$\beta(z) = E[\beta(z, L, p(L), CW_{\min})], \tag{3}$$

which is obtained by averaging on packet length L and collision probabilities $p(L)$.

Figure 1 shows the 200 first coefficients of $\beta(z)$ when $C(z) = 0.8z + 0.2z^4$, $L = 4$ and $p = 0.3$. In this theoretical example, the packet transferring time is 4 slots, and each decrementation of the backoff counter takes one slot with probability 80% and 4 slots with probability 20% (which means that the channel is busy with a packet transmission). The coefficients in this figure were obtained from numerical calculations using Maple, by iterating the recursive equation (2).

Remark 1. In case the RTS/CTS mechanism is used, the recursive equation (2) becomes

$$\beta(z, k) = \frac{1}{k}\frac{C(z)^{k+1} - C(z)}{1 - C(z)} \times ((1 - p_1 - p_2)z^{r+L} + (p_1 z^r + p_2 z^{r+L})\beta(z, 2k)), \tag{4}$$

where r is the RTS transmission time, p_1 and p_2 are the collision probabilities on RTS and data frames, respectively.

Note that in general $p_1 > p_2$, due to the channel reservation.

The modified recursive equation consists again of a backoff decrementation term $(1/k)((C(z)^{k+1} - C(z))/(1 - C(z)))$, and there are two cases for the additional delay:

(i) if there is no collision (with probability $1 - p_1 - p_2$) the delay is equal to the RTS plus the data frame transmission time (term z^{r+L});

[1] L can be adjusted to include *AckTimeouts* too.

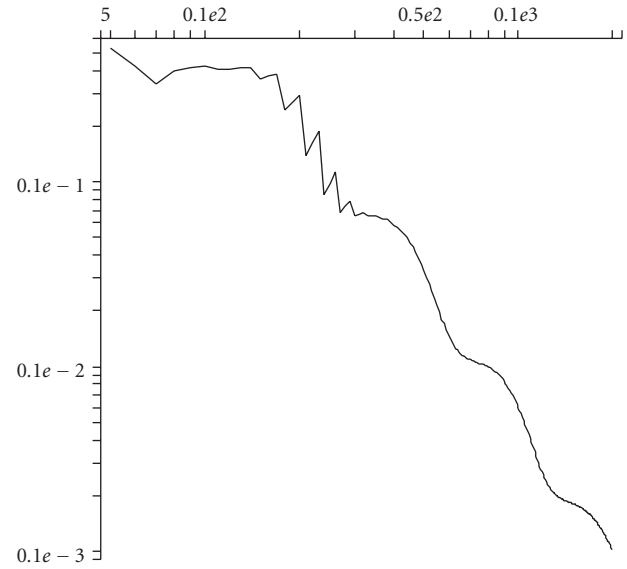


Figure 1: Coefficients of $\beta(z)$.

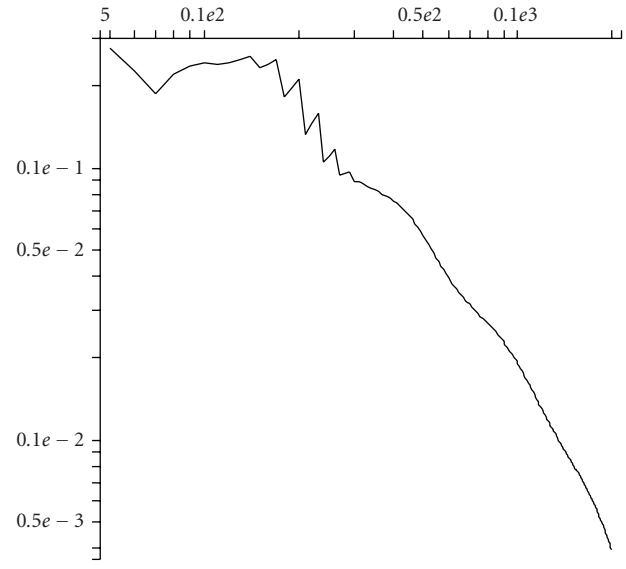


Figure 2: Coefficients of $w(z)$.

(ii) in case there is a collision on either an RTS frame (term $p_1 z^r$) or a data frame (term $p_2 z^{r+L}$) the procedure is repeated.

3.3. Delays including queueing

In order to compute the delay experienced by packets in the buffer, we take the formula for slotted M/G/1 for the queue delay probability generating function $q(z)$

$$q(z) = \exp\left((\beta(z)-1)\frac{\lambda}{2}\right)\frac{(1-\lambda\beta'(1))(1-z)}{1-z\exp\left(-(\beta(z)-1)\lambda\right)}. \tag{5}$$

This needs the provision that $\beta'(1)$ exists. We will see that this implies that $p < 1/2$. Similarly, for the existence of the kth moment of service time we need that $p < 2^{-k}$. If $\lambda \ll 1$ then we can replace (5) by

$$q(z) \approx \frac{(1-\lambda\beta'(1))}{1-(z/(1-z))(1-\beta(z))\lambda}. \tag{6}$$

The generating function of the overall delay, so called one hop delay (queueing + service), of a packet of length L with collision probability p, $w(z, L, p)$ satisfies the identity

$$w(z, L, p) = q(z)\beta(z, L, p, W_{\min}). \tag{7}$$

Figure 2 shows the coefficients of $w(z)$ for $\lambda = 0.02$. Notice that $\beta'(1) = 22.939\ldots$.

3.4. Asymptotic analysis

We denote by S the service time and W the overall delay in a router. In this section we derive asymptotic estimates for the distributions of the above quantities by applying Flajolet-Odlyzko theorems [9]. The proofs are given in the appendix.

Theorem 1. *The expansion for z around 1 holds:*

$$\begin{aligned}\beta(z, L, p, k) = {} & 1 + (1-z)v(1-z) \\ & + \left(kC'(1)(1-z)\right)^B \alpha\left(\log(1-z)\right) \\ & + O\left((1-z)^{B+1}\right),\end{aligned} \tag{8}$$

where $v(x)$ is a polynomial, $B = -\log_2 p$ assuming that B is not integer, and $\alpha(x)$ is a periodic function of period $\log 2$ *with small fluctuation.*

Theorem 2. *The probability that the service time is greater than T, for T large is*

$$P(S > T) = \left(W_{\min}C'(1)\right)^B \alpha^*(\log T)T^{-B} + O\left(T^{-B-1}\right), \tag{9}$$

where $\alpha^*(x)$ is also a periodic function of period $\log 2$ *with small fluctuation.*

Theorem 3. *The expansion for z around 1 holds:*

$$\begin{aligned}w(z) = {} & 1 + (1-z)u(1-z) \\ & + \frac{\lambda\left(W_{\min}C'(1)\right)^B}{1-\lambda\beta'(1)}\alpha\left(\log(1-z)\right)(1-z)^{B-1} \\ & + O\left((1-z)^B\right),\end{aligned} \tag{10}$$

where $u(x)$ is an analytic function.

Theorem 4. *The probability that the delay in a router is greater than T, for T large is*

$$P(W > T) = \frac{\lambda\left(W_{\min}C'(1)\right)^B}{1-\lambda\beta'(1)}\alpha^*(\log T)T^{1-B} + O\left(T^{-B}\right). \tag{11}$$

Notice that the delay distribution tail decays in power law. As a corollary it turns out that the existence of the kth moment of the delay needs $p < 2^{-k-1}$. Also, to obtain the asymptotic delay distribution estimate, only the average of the channel occupancy distribution $C'(1)$ is required, rather than the distribution $C(z)$.

Remark 2. Similarly, in case the RTS/CTS mechanism is used, we have the expansion $\beta(z,k) = 1 + zf(z) + c_2((1-z)k)^B + O((1-z)^{B+1})$ with $B = -\log \max_{i=1,2}\{p_i\} = -\log p_1$ and $f(z)$ an analytical function. This implies that both service time and delay distributions are asymptotically power laws.

4. MULTIHOP DELAY ANALYSIS

4.1. General case: correlated waiting times in queues

We now compute the end-to-end (multihop) delay distribution for any given route in the network, based on the one hop delay analysis discussed previously. For this purpose, we assume that when travelling on its route, the delay experienced by a packet on a router is independent of the delay experienced on other routers. This assumption makes the problem easier to handle mathematically. However, it is not a fundamental assumption for our result since it is known that the sum of several random variables in power law is still in power law whatever the dependence assumptions between them. The power law in the resulting distribution function will be the maximum of the respective power laws of the variables, except that the factor in front of it will depend on the dependence assumptions. To see this, consider n random variables $X_1, \ldots, X_n$, such that $P(X_i > T) \propto T^{-B_i}$ for all $1 \le i \le n$. We will show that $X_1 + \cdots + X_n$ is also in power law, and the value on the power is always the same regardless of any correlation between the random variables. Moreover, we will determine appropriate lower and upper bounds on the coefficients.

First of all an easy lower bound: we have $P(\sum_i X_i > T) \ge \max_i\{P(X_i > T)\}$.

Second, we can obtain an upper power law bound since

(1) $\sum_i X_i \le n \max_i\{X_i\}$,
(2) $P(\max_i\{X_i\} > T) = P(\exists i X_i > T) \le \sum_i P(X_i > T)$.

Therefore $P(\sum_i X_i > T) \le \sum_i P(X_i > T/n) \propto \sum_i n^{B_i} T^{-B_i}$, independently of any correlation between the variables, and derive appropriate upper bound delay routing in power law if the X_i's are the waiting delays in routers. Notice that the lower bound and upper bounds are asymptotically within a factor n^B of each other, with $B = \min_i\{B_i\}$.

4.2. Independence assumption

Assuming independence from now on, we can work on the exact asymptotic value instead of the bounds obtained in the general case, keeping in mind that these bounds are within a factor n^B of the exact value, where n is the route length and $B = \min_i\{B_i\}$.

Taking fully the independence assumption, when there are n routers in the route from the source to the destination, the probability generating function of the end-to-end delay is equal to the product $\prod_{i\in\text{route}} w_i(z)$ where $w_i(z)$ is the probability generating function of the delay at router number i and *route* is a set of router indices.

Still, according to Flajolet-Odlyzko results [9], if each $w_i(z)$ is of the form $1+(z-1)g_i(z)+c_i(z-1)^{B_i-1}+O((z-1)^{B_i})$, where c_i is a constant, then the leading term of $P(W(\text{route}) > T)$ is $\sum_{i\in\text{route}} c_i^* T^{1-B_i}$, with $c_i^* = c_i/\Gamma(2-B)$. Keeping only leading terms

$$P(W(\text{route}) > T) \approx c(\text{route}) T^{1-B(\text{route})}, \qquad (12)$$

where $B(\text{route}) = \min B_i$ and $c(\text{route}) = \sum_{B_j=B} c_j^*$.

An unexpected consequence of the above is that a good choice for the route should not be the shortest path in number of hops. In the shortest path the lap between two consecutive routers may be too large, leading to too large collision rates and therefore a too low value of $B(\text{route})$. If we take shorter hops between routers, then we will reduce the collision rate and get a larger value of $B(\text{route})$. Of course this would be done in the detriment of a larger number of hops and a larger value of $c(\text{route})$. But, since in $c(\text{route})T^{1-B(\text{route})}$ parameter T is supposed to be large, the reduction of $T^{1-B(\text{route})}$ would prevail in most cases on the $c(\text{route})$ increase.

Interestingly enough, increasing the number of hops and $c(\text{route})$ will in most cases increase the *average* end-to-end delay. Therefore we have the paradoxical case where increasing the average delay actually decreases the over-delay loss ratio. This is due to the fact that we expect the average delay to be much lower than the maximum acceptable delay T. Consequently, routing with respect to average delay as it is done in [10] may conflict with the minimization of the over-delay ratio. Conversely, the optimal route may be too long since it may have too short hops. In this case the connection may waste too many resources. Instead of choosing the route that minimizes $P(W > T)$ it is probably wiser to seek the shortest route that satisfies the requirement $P(W > T) \le \epsilon$.

5. CROSS-LAYER DELAY ESTIMATION PROTOCOL

In this section, we present a delay estimation protocol, used to obtain estimates of the delay distributions for all routes in the network in a proactive way. In Section 7 we show how this information can be used to optimize route computation. In fact, we propose an extension to the OLSR routing protocol to support delay estimation for any given route in a mobile ad hoc network.

Note that our delay distribution analysis holds regardless of the mobility of nodes. In fact, the end-to-end delay distribution analysis concerns a given route. Consequently, any changes in the path from the source to the destination result in a new power law for the end-to-end delay distribution. Our proposed delay estimation protocol is designed to support mobility in an ad hoc network. It is proactive and predictive, in the sense that, using the previously presented analytical model, it is able to estimate delays on links and routes even if no data packets have been transmitted yet. Consequently, we permanently have a prediction of the delay

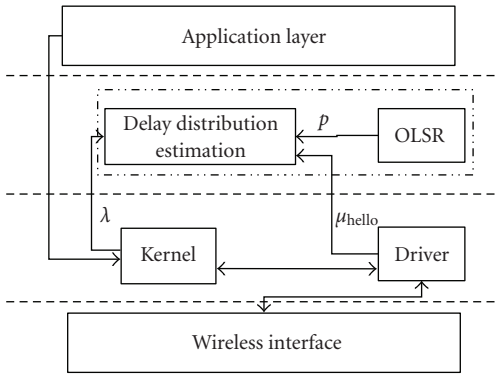


Figure 3: Delay estimation protocol framework.

distribution over all routes in the network, which can be used by a delay based routing protocol.

As mentioned previously, the single-hop delay distribution estimate is based on the knowledge of the collision probability and the average of the channel occupancy distribution C, which are basically MAC layer parameters. The multihop delay distribution is based on the knowledge of single hop characteristics along a given route, which clearly concerns the functioning of the routing protocol. Therefore, the extended protocol needs to interact with the MAC layer. In Figure 3, we depict the protocol framework.

5.1. Average of channel occupancy distribution C

The channel occupancy information concerns the internal functioning of wireless cards and is not actually known. However, the card acknowledges successful frame transmissions by sending special interrupts to the driver. This allows to measure the service time of transmitted packets. Knowing the service time, it is possible to deduce the access time in case of broadcast packets such as OLSR Hello messages (since they are not retransmitted when a collision occurs). Thus, based on (1), it is possible to derive the mean of the channel occupancy distribution C from the mean of the Hello access time distribution, noted by μ_{hello}. We have $C'(1) = 2\mu_{\text{hello}}/(W_{\min}+1)$.

5.2. Collision probability estimation

The collision probability is estimated by OLSR, since this information is not currently provided by wireless cards. OLSR uses Hellos in order to detect neighbors. A node is a neighbor if and only if the Hello collision rate is below a given threshold. Therefore OLSR has a procedure in the advanced neighbor sensing option that allows to compute the collision rate (link quality level parameter). It uses the Hello message sequence number in order to identify the missing Hellos. However there could be a difficulty in the fact that the collision probability $p(L)$ may depend strongly on packet length L. One may expect a dependence of the kind $-\log p(L) = aL+b$ where a and b are scalar coefficients. Since the neighbor has no idea of the size of missing Hellos, the transmitter should advertise the length distribution of its Hellos. Comparing with its received Hello distribution the neighbor would be able to determine the coefficients a and b. By default the neighbor assumes $a = 0$, that is, all packets have the same collision rate regardless of their length.

5.3. Advertising link quality

Multihop delay computation is based on the knowledge of the one hop delays of the route. Thus, each node must inform the entire network of its local information. For this purpose, a link quality advertisement (LQA) message is broadcasted in the network.

In OLSR, broadcast traffic is relayed via multipoint relay (MPR) nodes (nodes elected by their neighbors because they cover their two-hop neighborhood) to consume less resources. In order to save more on control traffic, OLSR offers the possibility for the nodes to advertise a small subset of their neighbor links. The advertised link set can be limited to MPR links, that is, the neighbors that have elected this node as an MPR. In this case the nodes have only a partial knowledge of the network topology, nevertheless, each node knows its own neighbor list and this guarantees that any given node can compute a shortest path to any arbitrary destination.

For our purpose, it is preferable to use the option full OLSR, that is, to advertize the whole neighbor set instead of the MPR selector set. In fact, in order to estimate the multihop delay distribution for all routes in the network, we need the complete network topology given by full OLSR. If the advertized links are limited to the MPR selector set, we have only a partial knowledge of the network topology. In the latter case, the nonadvertised links may offer better possibilities for delay-based routing. Consequently, the complete topology is necessary in order to find an optimal route with respect to the delay. The use of the full-OLSR option introduces an additional overhead due to larger link quality advertisement packets. However, broadcast traffic is still relayed with the optimized MPR-flooding mechanism, which significantly reduces the overhead.

The node advertises for each link ℓ the collision rate p_ℓ, and for itself it advertises the global λ, provided by the kernel as shown in Figure 3, and the value of $C'(1)$, or directly the tuple $(p, \lambda(W_{\min}C'(1))^B/(1-\lambda\beta'(1)))$.

6. SIMULATION RESULTS

We use the ns-2 [11] simulator to validate our delay modelling. We study various scenarios for different purposes. We compare the analytic service time distribution with the measured service time distribution (obtained by ns-2 simulations). We aim to show that the service time and the sojourn time (in other words the delay including queuing) are in power law. Furthermore, we investigate whether one-hop delays are independent within a route, and finally we show that the end-to-end delay is in power law too. Common simulation parameters are summarized in Table 1.

Table 1: Simulation settings.

MAC Parameters	$W_{\min}$ 32, slot 20 μs
Propagation model	Two-ray ground
Transmission range	250 m
Packet size	1000 bytes
Traffic type	Exponential (Poisson)
Simulation time	300 s
Simulation Area	800×800 m^2
Routing protocol	OLSR

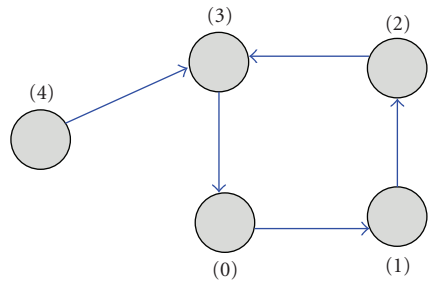


Figure 4: Topology 1.

Table 2: Measured parameters.

Channel occupancy	$C(z) \simeq 0.82 + 0.04z^{15} + 0.03z^{124} + 0.1z^{444}$
Packets per slot	$\lambda = 0.00024$ packets/slot (12 packets/s)
Collision probability	$p = 0.09$, $B = 3.45$
Packet length	$L = 229$ slots

One-hop delay measures

In the first scenario, we consider an ad hoc network with 5 nodes as shown in Figure 4. The 802.11 bandwidth is 1 Mb. Five exponential flows with 140 kbs data rate are launched between different pairs of nodes (represented by arrows in Figure 4). In order to study the cumulative delay distribution in node 2, we measure the main parameters in this node for the conducted simulation, as presented in Table 2.

Based on these parameters and (2), we compute analytically the service time distribution using Maple, which we draw in Figure 5. The service time distribution measured via ns-2 simulation is shown in Figure 6. To demonstrate that the service time distribution is in power law with $B = -\log_2(p)$ (here $B = 3.45$), as stated in Theorem 2, we draw the equation $Y = \alpha X^{-3.45}$, where α is a constant, and we compare the two plots. Figure 6 shows that, for T large enough, the service time distribution and Y have the same power law exponent.

In the same way, we measure the sojourn time distribution (also called node delay), which we present in Figure 7. We notice that for T between 4000 and 40000 slots (i.e., 80 milliseconds to 800 milliseconds), the node delay is in power law with exponent $1 - B = -2.45$, in accordance with Theorem 4.

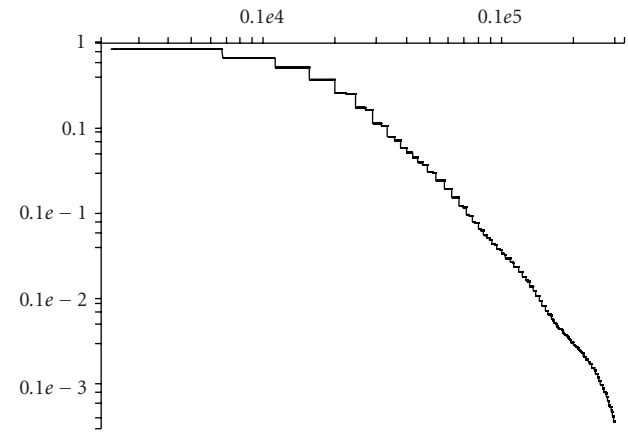


Figure 5: Analytic service time distribution.

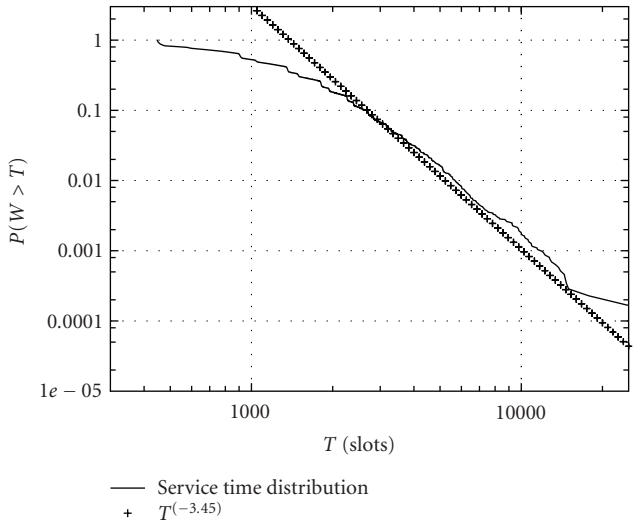


Figure 6: Measured service time distribution.

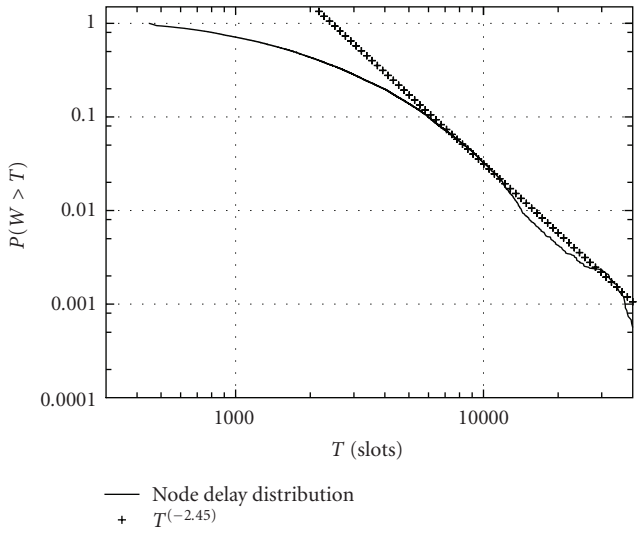


Figure 7: Measured node delay distribution.

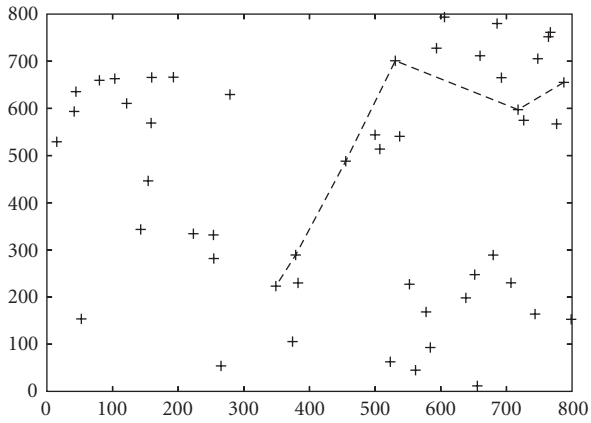


Figure 8: Topology 2.

Table 3: Collision probabilities for each hop of the path.

Throughput	Per hop collision probabilities %				
2 pkt/s	0.53	0.94	0.19	1.05	1.13
8 pkt/s	1.11	2.28	0.45	5.43	5.75

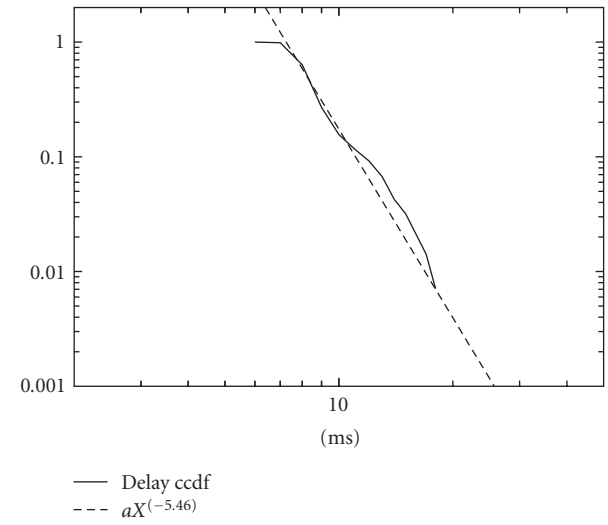
Multihop delay measures

Secondly, we consider a randomly generated topology of 50 nodes which is depicted in Figure 8. We launch 10 exponential flows in the network and we measure for each the end-to-end delay distribution. We run several simulations by varying the throughput from 2 to 8 packets per second. We consider the flow following the five-hop path shown in Figure 8. We measure the end-to-end delay of this flow as well as collision probabilities along the path (for each hop). Table 3 summarizes the obtained probabilities.

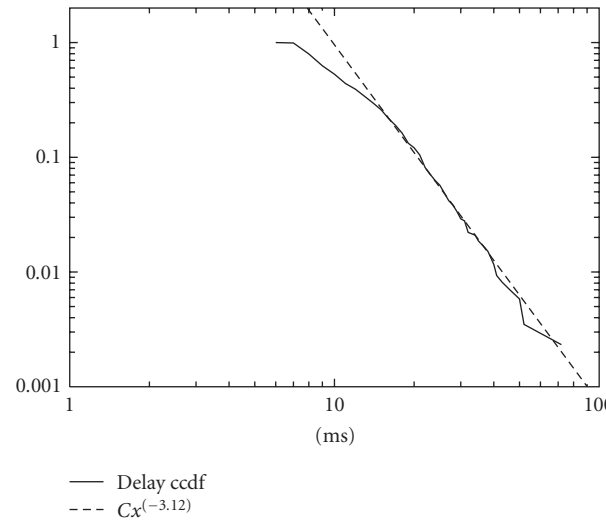
Figure 9 compares the measured end-to-end delay distribution with theoretical results for sending rates of 2 and 8 packets per second, respectively. According to the analysis of multihop delay distribution in Section 4, the power law exponent is equal to $1 + \log_2(p)$ such that p corresponds to the highest collision probability along the path. Referring to Table 3, the highest collision probability on this path is 0.0113 ($1 - B = -5.46$) and 0.0575 ($1 - B = -3.12$) for traffic rates of 2 and 8 pkt/s, respectively.

Impact of the packet size on the delay distribution

Considering the same scenario as previously, we run simulations by varying the packet size from 64 to 1000 bytes as shown in Figure 10. As we notice, for each size, the delay distribution is a power law with slightly different slopes, which correspond to slightly different collision probabilities. For bigger packets there is an increase in the collision probability. For example, for packets of 64 bytes the highest collision probability on the path is 5.5%, while it is equal to 8.2% for packets of 1000 bytes. In addition, we notice



(a) Sending rate of 2 pkt/s



(b) Sending rate of 8 pkt/s

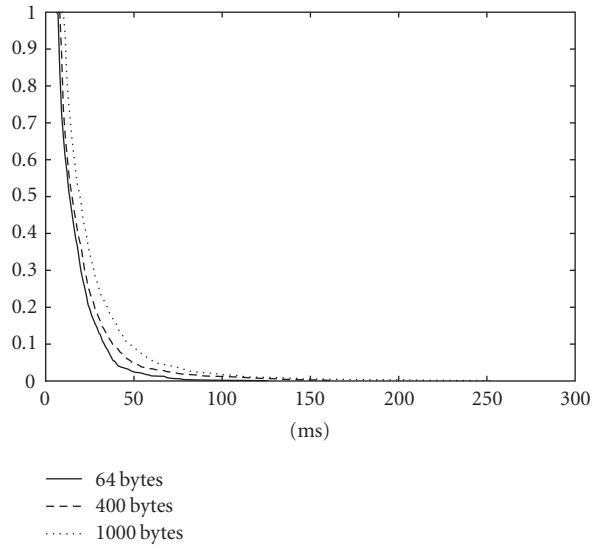
Figure 9: Comparison between end-to-end delay distribution and the corresponding power law.

that the curves are shifted to larger delay values when the packet size increases. The shift in the delay distribution corresponds to the increase of the factor c(route). In fact, bigger packets also require longer transmission times on the channel.

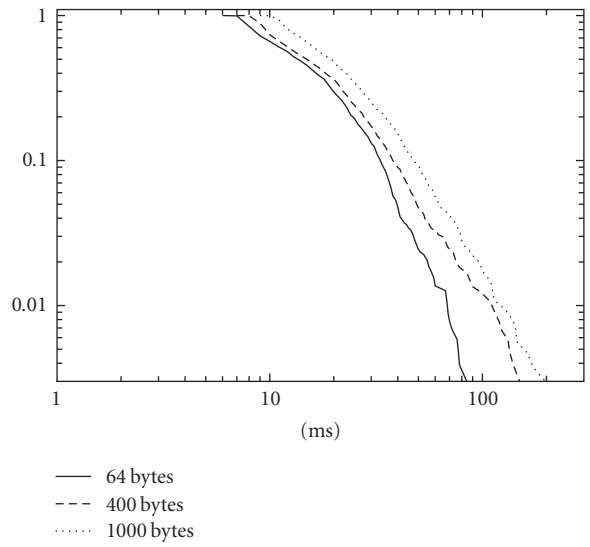
We note that, if the collision probability remains constant, according to (2), (7) bigger packets will only result in a shifted power law delay distribution with the same slope.

Dependence of the delays on the routers

We also measure the single-hop delay distributions along the path connecting the source to the destination. Let W_i, $i = 1 \cdots 5$, be the distribution generating function for each hop and W the end-to-end delay distribution generating function. We compute the product $\prod_{i=1\ldots5} W_i$ and we compare it

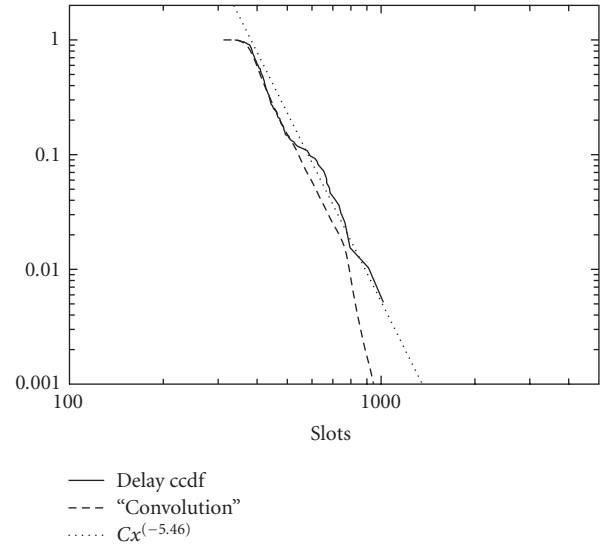


(a) Cumulative distribution function

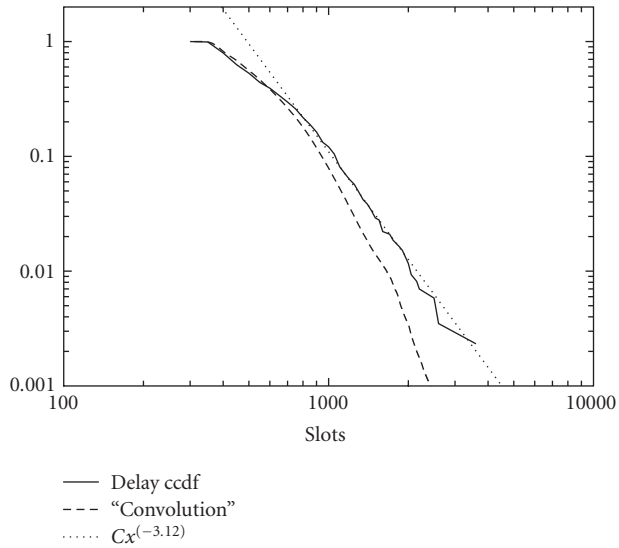


(b) Cumulative distribution function using log scale

Figure 10: Impact of packet size on the delay distribution.



(a) Sending rate of 2 pkt/s



(b) Sending rate of 8 pkt/s

Figure 11: Comparison between the end-to-end distribution and the convolution of single hop distributions.

to W. In case the delays are independent within a route, the above product corresponds to the end-to-end delay distribution. As shown in Figure 11, the curves representing W and $\prod_{i=1\cdots5} W_i$ are slightly different which means that there is a weak dependence between single-hop delays yet it is weaker when the network is lightly loaded. Notice that even when the independence assumption is not completely verified, the delay is still a power law as explained in Section 4.1.

7. DELAY-DISTRIBUTION-BASED ROUTING

As discussed earlier, routes with minimum average end-to-end delay are not necessarily those which satisfy a delay constraint, defined generally as the probability $P(W > T)$ that the delay W exceeds a maximum acceptable delay T, which is specified by the application layer. To compute such a probability, we need to know the delay distribution in every node, as presented in Section 5, instead of only the average delay. The problem of delay-distribution-based routing consists in finding a route that satisfies the application end-to-end delay requirement $P(W > T) < \epsilon$ for a given connection. Multiple routes satisfying such a delay constraint can be found in the network, hence a routing algorithm must select one among them. In this section, we explore two possible directions. The first direction consists in finding the optimal route that minimizes $P(W > T)$. The second direction consists in finding the shortest route (in hops) that satisfies the requirement $P(W > T) \leq \epsilon$.

Finding the optimal route

In general, finding the optimal route with respect to a delay distribution is NP hard [7]. But if we stick to the asymptotic expression, we can find a polynomial Dijkstra-like algorithm. The problem is to find the route that provides the best asymptotic expansion of the quantity $P(W(\text{route}) > T)$ when $T \to \infty$. By best asymptotic expansion we mean the one that provides asymptotically the lowest $P(W(\text{route}) > T)$. Since we expect that $P(W(\text{route}) > T)$ is asymptotically equivalent to $\sum_{i\in\text{route}} c_i^* T^{1-B_i}$ (cf. Section 4), the idea consists in minimizing the sum of the leading terms of the one-hop delay distributions along the route. Hence, the routing algorithm is effectively a Dijkstra algorithm, where the weights on the links are $c^* T^{1-B}$. Parameters c^* and B are calculated according to the analysis in Section 4. The weight of the route is the sum of the weights of the links, and the optimal route is the route that minimizes this sum.

When T is finite, the sum of the weights on a route gives an approximation of the end-to-end delay distribution within a factor $1 + O(T^{-B(\text{route})})$, according to the asymptotic analysis. Since $B(\text{route}) > 1$, the algorithm is optimal within a factor $1 + O(T^{-1})$.

Finding the shortest route satisfying the delay constraint

As discussed previously, the shortest route that satisfies the constraint $P(W > T) \leq \epsilon$ is generally preferable to a much longer route that minimizes the quantity $P(W > T)$. Moreover, a major problem due to the use of any dynamic metric is route fluctuation. However, the proposed routing on the shortest path that verifies an over-delay ratio constraint provides more stable routes. In the previous section we described a polynomial search algorithm which is optimal within a factor $1 + O(T^{-1})$ hence for T sufficiently large the search provides the optimal route. We showed that delay distribution routing can be reduced to polynomial routing based on an additive metric. In the present section we aim to find the shortest route according to a certain additive metric on links, that is, the number of hops, which satisfies a given constraint according to another additive metric, that is, the quantities $c^* T^{1-B}$.

In general such a multimetric optimization problem, also called multiconstrained path routing (MCP), is again NP-hard. In [12], the authors present an overview of some proposed polynomial time approximation algorithms in the case of 2 metrics, and study the general case of $K \geq 2$ metrics. However, in our particular context, since the first metric can only take integer values, we can easily solve the problem in polynomial time with dynamic programming.

We model the network as a weighted graph. We consider a source node s. We denote by v_j, $j = 1 \cdots n$, all the nodes in the network, where v_1 is the source s. Each link connecting two nodes (v_i, v_j) is associated to a weight w_{v_i,v_j} which corresponds to the asymptotic probability $c^* T^{1-B}$. For each node v_j we define $p(i, v_j)$ as the smallest known value according to the second metric (the sum of the weights w along the path) of all routes of length i according to the first metric (hop count) that connect the source node s to node v_j. Note that $i \leq n$, since the longest path in the network is at most n hops.

We describe the algorithm as follows. We initialize all values $p(0, v_j)$, for $j = 1 \cdots n$ to infinity except for $p(0, v_1) = 0$. We will compute $p(i, v_j)$ for all i, j. For each $i \geq 1$ in increasing order, we compute the values $p(i, v_j)$, $j = 1 \cdots n$, using equation

$$p(i, v_j) = \min_{c\in\mathcal{N}(v_j)} \left(p(i-1, c) + w_{c,v_j}\right), \tag{13}$$

where $\mathcal{N}(v_j)$ is the neighborhood of node v_j and w_{c,v_j} is the weight of the link (c, v_j). Notice that for all values of i that are smaller than the distance between s and v_j we have $p(i, v_j) = \infty$.

The aggregate computational cost of $p(i, v_j)$ for all nodes v_j and for a given i is $O(m)$, where m is the total number of links in the network. Hence, in the worst case, the total time needed to construct the table $p(i, v_j)$ for all possible values $i, j = 0 \cdots n$ is $O(mn)$. Once the table has been constructed, the shortest route to any destination d satisfying the required delay constraint, corresponds to the route of minimum i such that $p(i, d) < \epsilon$. In case the algorithm computes the route for one particular destination, the iteration on the route length i can stop as soon as a feasible route is found.

8. CONCLUSION AND PERSPECTIVES

In this paper, we analyze the delay distribution in 802.11 multihop networks. We demonstrate that for large values of T the cumulative delay distribution $P(W > T)$ is a power law. In practice, simulations show that this is true from T equal to approximately twice the average. The delay distribution for a specific route can be derived based on MAC-layer as well as network-layer parameters, hence we present a cross-layer solution for a delay estimation protocol as an extension to the OLSR routing protocol. Furthermore, the information from this protocol can be used to compute the route that satisfies the QoS delay requirements specified by a multimedia application. In fact, delay-distribution-based routing is known to be an NP-hard problem. However, the asymptotic analysis in power law makes it possible to obtain a polynomial, Dijkstra-like, algorithm.

It is important to note that the routing algorithm does not guarantee that the calculated route will satisfy the delay constraint after launching the new traffic. In case the new connection has a significant impact on the network conditions, it is necessary to dynamically control the delay, in order to check whether the constraint is still verified. Due to its proactive nature, the proposed delay estimation protocol allows to compute periodically the end-to-end delay, thus the routes can be readjusted. Therefore, it is interesting to combine the proposed delay routing algorithms with a mechanism providing dynamic delay control, as well as admission control when the connection delay requirements cannot be satisfied. In future work, we intend to implement such mechanisms, and subsequently to evaluate the performance of the proposed routing algorithms in this context.

Another direction for further research is to establish a link between bandwidth and delay requirements in the

context of wireless multihop networks. This would permit an adaptation of bandwidth-QoS routing solutions for delay-sensitive applications as well. Such a link can be established using the notion of equivalent bandwidth. In this particular case the equivalent bandwidth on a given path would correspond to the capacity of the network in transferring packets which satisfy this delay constraint. In [13], the authors show how to extend the effective bandwidth concept in order to provide a general model for a wireless channel in terms of connection level QoS metrics. It would be interesting to adapt the proposed model to the considered context of multihop IEEE 802.11 networks.

APPENDIX

Proof of Theorem 1. We fix L and p and set $e^{-\theta} = C(z)$ and denote $j(\theta,k) = \beta(z,k)$. We have $j(\theta,k) = ((1-e^{-k\theta})/k\theta) f(\theta)(1-p+pj(\theta,2k))$, with $f(\theta) = e^{\theta}(\theta/(1-e^{-\theta}))z^L$.

It is clear that $\theta = (1-z)C'(1) + O((1-z)^2)$. We define $g(\theta) = \prod_{i\geq 1}((1-e^{-\theta 2^{-i}})/\theta 2^{-i})$. Thus if

$$\nu(\theta,k) = g(k\theta)j(\theta,k), \tag{A.1}$$

then $\nu(\theta,k) = g(2k\theta)f(\theta)(1-p) + pf(\theta)\nu(\theta,2k)$. And $\nu(\theta,k) = ((1-p)/p)\sum_{i\geq 1}(f(\theta)p)^i g(2^i k\theta)$.

It can be proven that function $g(\theta)$ is analytical and behaves like $1+O(\theta)$ when $\theta \to 0$ and converges to zero faster than any power law when $\theta \to \infty$. Let $r_B(\theta)$ be the polynomial of degree $\lfloor B \rfloor$, which is the Taylor expansion of $g(\theta)e^{\theta}$ at $\theta = 0$. Recall that $B = -\log_2 p$.

Let $g_B(\theta) = g(\theta) - r_B(\theta)e^{-\theta}$. Clearly, $g_B(\theta) = O(\theta^{\lceil B \rceil})$ when $\theta \to 0$. We have $\nu(\theta,k) = u_B(\theta) + ((1-p)/p)\sum_{i\geq 1}(f(\theta)p)^i g_B(2^i k\theta)$ with

$$u_B(\theta) = \left(\frac{1-p}{p}\right)\sum_{i\geq 1}(f(\theta)p)^i r_B(2^i k\theta)e^{-2^i k\theta}. \tag{A.2}$$

Clearly, $u_B(\theta)$ is an analytical function with $u_B(\theta) = 1 + O(\theta)$. Let $\nu_B(\theta,k) = ((1-p)/p)\sum_{i\geq 1}(f(\theta)p)^i g_B(2^i k\theta)$. We will show that $\mu_B(\theta,k) = \theta^{-B}\nu_B(\theta,k)$ is bounded when $\theta \to 0$. Let $h_B(\theta) = \theta^{-B}g_B(\theta)$. We have $\mu_B(\theta,k) = ((1-p)/p)\sum_{i\geq 1}(f(\theta))^i h_B(2^i k\theta)k^B$.

Since $f(\theta) = 1 + O(\theta)$, when $\theta \to 0$, we have $\mu_B(\theta,k)$ which converges to $\alpha(\log\theta) = \sum_i h_B(2^i k\theta)k^B$, the sum being on all integers i, including the negative integers. The sum converges because $h_B(\theta) = O(\theta^{\epsilon})$ with $\epsilon = \lceil B \rceil - B$ and $h_B(\theta)$ decays faster than any power law. Notice that function $\alpha(x)$ is periodic of period $\log 2$.

Therefore $j(\theta,k) = \nu(\theta,k)/g(k\theta)$ has asymptotic expansion $u_B(\theta)/g(k\theta) + \alpha(\log\theta)\theta^B + O(\theta^{B+\epsilon})$. The theorem follows with a change of variable. □

Proof of Theorem 2. From the previous theorem it comes that $\beta(z)$ fits Flajolet-Odlyzko asymptotic conditions [9]. By writing the function $\alpha(\log(1-z))$ as a Fourier series $\alpha(\log(1-z)) = \sum_n \alpha_n(1-z)^{2in\pi/\log 2}$, and applying Flajolet-Odlyzko theorems, we have $P(S>T) = (W_{\min}C'(1))^B\alpha^*(\log T)T^{-B} + O(T^{-B-1})$, where α^* is periodic in $\log T$, of period $\log 2$: $\alpha^*(\log(T)) = \sum_n(\alpha_n/\Gamma(2-B-2in\pi/\log 2))T^{2in\pi/\log 2}$. □

Proof of Theorem 3. We substitute the expansion for $\beta(z)$ around $z=1$, derived in Theorem 1, in the formula for $q(z)$ given by (6). The theorem follows by using the expansion around $z=1$ in equation $w(z) = q(z)\beta(z)$. □

Proof of Theorem 4. We use the result of the previous theorem and we apply Flajolet-Odlyzko theorems on $w(z)$. □

ACKNOWLEDGMENT

Part of this work was presented at the 1st Workshop on resource Allocation in Wireless Networks, April 3rd, 2005, Riva Del Garda, Italy.

REFERENCES

[1] IEEE 802.11 Standard, "Wireless Lan Medium Access Control and Physical layer Specifications," June 1997.

[2] T. Clausen and P. Jacquet, "Optimised Link State Routing (OLSR)," IETF RFC 3626.

[3] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communcations*, vol. 18, no. 3, pp. 535–547, 2000.

[4] H. Zhai and Y. Fang, "Performance of wireless LANs based on IEEE 802.11 MAC protocols," in *Proceedings of the 14th IEEE on Personal, Indoor and Mobile Radio Communications (PIMRC '03)*, vol. 3, pp. 2586–2590, Beijing, China, September 2003.

[5] O. Tickoo and B. Sikdar, "Queueing analysis and delay mitigation in IEEE 802.11 random access MAC based wireless networks," in *Proceedings of the 23d Annual Joint Conference of IEEE Computer and Communications Societies (INFOCOM '04)*, vol. 2, pp. 1404–1413, Hong Kong, March 2004.

[6] J. Chen and Y. Yang, "Multi-hop delay performance in wireless mesh networks," in *Proceedings of IEEE Global Telecommunications Conference (Globecom '06)*, pp. 1–5, San Francisco, Calif, USA, November-December 2006.

[7] R. A. Guerin and A. Orda, "QoS routing in networks with inaccurate information: theory and algorithms," *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 350–364, 1999.

[8] J. W. Cohen, "A heavy-traffic theorem for the GI/G/1 queue with a Pareto-type service time distribution," *Journal of Applied Mathematics and Stochastic Analysis*, vol. 11, no. 3, pp. 247–254, 1998.

[9] P. Flajolet and A. M. Odlyzko, "Singularity analysis of generating functions," *SIAM Journal on Discrete Mathematics*, vol. 3, no. 3, pp. 216–240, 1990.

[10] H. Badis, A. Munaretto, K. Al Aghal, and G. Pujolle, "Optimal path selection in a link state QoS routing protocol," in *Proceedings of the 59th IEEE Vehicular Technology Conference (VTC '04)*, vol. 5, pp. 2570–2574, Milan, Italy, May 2004.

[11] K. Fall and K. Varadhan, *The ns manual*, 2002, http://www.isi.edu/nsnam/ns/ns-documentation.html.

[12] G. Xue, W. Zhang, J. Tang, and K. Thulasiraman, "Polynomial time approximation algorithms for multi-constrained QoS routing," to appear in *IEEE/ACM Transactions on Networking*.

[13] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, 2003.

Hindawi Publishing Corporation
Advances in Multimedia
Volume 2007, Article ID 95807, 14 pages
doi:10.1155/2007/95807

Research Article

A Study on the Usage of Cross-Layer Power Control and Forward Error Correction for Embedded Video Transmission over Wireless Links

Fabrizio Granelli,[1] Cristina E. Costa,[2] and Aggelos K. Katsaggelos[3]

[1] *Department of Information and Communication Technology, University of Trento, Via Sommarive 14, 38050 Povo (Trento), Italy*
[2] *CREATE-NET International Research Center, Via Solteri 38, 38100 Trento, Italy*
[3] *Electrical Engineering and Computer Science Department, Northwestern University, Evanston, IL 60208, USA*

Received 24 December 2006; Accepted 5 May 2007

Recommended by Haohong Wang

Cross-layering is a design paradigm for overcoming the limitations deriving from the ISO/OSI layering principle, thus improving the performance of communications in specific scenarios, such as wireless multimedia communications. However, most available solutions are based on empirical considerations, and do not provide a theoretical background supporting such approaches. The paper aims at providing an analytical framework for the study of single-hop video delivery over a wireless link, enabling cross-layer interactions for performance optimization using power control and FEC and providing a useful tool to determine the potential gain deriving from the employment of such design paradigm. The analysis is performed using rate-distortion information of an embedded video bitstream jointly with a Lagrangian power minimization approach. Simulation results underline that cross-layering can provide relevant improvement in specific environments and that the proposed approach is able to capitalize on the advantage deriving from its deployment.

1. INTRODUCTION

With the worldwide diffusion of wireless networks (WLANs, 3G cellular networks), the usage of wireless links for delivery of video streams is becoming a common practice. However, unlike wired links, wireless connections suffer from a number of specific drawbacks, including time-varying behavior, limited shared bandwidth, noisy medium, and user mobility. Moreover, the mobility of the user terminals implies the usage of batteries, and thus the need to limit transmission power and carefully utilize energy in order to maximize device lifetime. This aspect is also being considered by well-known and worldwide diffused standards like IEEE 802.11, which recently activated working group IEEE 802.11 h [1] in order to address the aspect of power saving.

In this scenario, optimized video transmission involves at least joint control of source coding parameters (encoded stream packetization, packet classification), medium access control procedures (ARQ, forward error correction), and physical parameters (transmission power, channel sensing). As a consequence, effective optimization can be achieved only by means of cross-layering solutions.

Indeed, cross-layering [2] is a well-known design principle which was recently proposed to overcome some limitations deriving from the standardized layering paradigm (such as OSI/OSI reference model) imposing independent design of protocols at different layers and standardized (and limited) interactions among adjacent layers. At this point, it is useful to mention that the authors share most of the cautionary perspective on cross-layering illustrated in [3], and—as it will be clearly discussed in the remainder of the paper—mainly aim at providing an analytical framework for evaluating the scenarios and the potential benefits deriving from the application of such a design principle.

The paper considers application of cross-layering paradigm to support unequal error protection (UEP) of video streams. UEP strategies for transmission of progressively coded images and videos have been implemented in various fashions, as some techniques designed for other contexts have been adapted to the transmission of embedded

bitstreams. Data is implicitly sorted by its importance, and this feature can be directly used for the implementation of error resilience techniques based on UEP. Traditional equal error protection (EEP) schemes consider all the data as having the same importance and assign the same degree of protection to the whole bitstream. On the other hand, UEP schemes give more importance, hence more protection, to the most critical parts of the coded image.

Unequal error protection techniques have been studied in the past [4] for differentiating protection between different layers in traditional layered scalable coding. These methods, based on techniques such as automatic repeat request (ARQ) and forward error correction (FEC), can be adapted to the transmission of embedded bitstreams. Reed-Solomon codes were used by Natu and Taubman [5] for the protection of JPEG2000 bitstreams during transmission over wireless channels. In [6], channel coding is used for implementing UEP of a JPEG2000 bitstream.

For video transmission, the application of UEP within a fine-grained scalable (FGS) bitstream was first considered by van der Schaar and Radha in [7], where the frame-grained loss protection (FGLP) framework was introduced. Based on it, Yang et al. [8] proposed a "degressive" protection algorithm (DEP) based on FEC for optimal assignment of protection redundancy among bit-planes. In [9], Wang et al. studied the problem of rate-distortion optimized UEP for progressive FGS (PFGS) over wireless channels using prioritized FEC for the base layer and enhancement layer. A similar problem was studied in [10], where the objective was to minimize the processing power for PFGS video given bandwidth and distortion constraints. In [11], a joint FEC and transmission power allocation scheme for layered video transmission over a multiple user CDMA networks was proposed. In that work, scalability was achieved using 3D SPIHT (wavelet based coding). The objective was to minimize the end-to-end distortion through optimal bit allocation among source layers and power allocation among different CDMA channels. The authors in [12] considered jointly adapting the source bit rate and the transmission power in order to maximize the performance of a CDMA system subject to a constraint on the equivalent bandwidth. In that work, an H.263+ codec was used to generate the layered bitstream. Relevant works on cross-layering between application and physical layers are provided by [13], where maximization of end-to-end quality of service is studied by enabling adaptive modulation and coding, and [14], where H.264/AVC video is protected by using turbo codes.

Recent works on the topic of optimization of video delivery over wireless channels are focused on the usage of embedded video streams (3D-ESCOT) [15, 16], and considered physical layer protocols are those employed in existing/next generation wireless networks (CDMA and OFDM). In [17], UEP, retransmissions, and interleaving are employed to reduce quality fluctuations in video quality of streaming scalable video, while in [18], motion-compensated temporal filtering (MCTF) scalable video coding (SVC) is supported by two-dimensional channel coding to improve resilience to channel errors. Preliminary works are also available, related to the problem of video transport over IP networks, like in [19], where UEP of video packets is proposed driven by a rate-distortion optimization principle.

However, a few works are available on video quality optimization using power control and FEC, such as [20], where joint coding-power control is employed for optimization of video delivery over CDMA cellular network, [21], where power control and channel coding are jointly employed to improve performance over wireless channels, and [22], where the authors introduced the application of cross-layer optimization between physical/link-layer functionalities (power, FEC) and application stream characteristics (embedded video stream).

In this paper, we analyze the problem of optimized video transmission over an uplink wireless channel, considering an architecture which enables to control forward error correction and transmission power on the basis of channel status and source information. Basically, the problem of providing the best possible quality is formulated and solved in order to derive useful guidelines on the effective benefits and usage scenarios of cross-layering. The reference coding algorithm will be MPEG-4 FGS [23], but the process can be applied to other embedded bitstreams as well. However, one of the main innovation points of the paper is the analysis of the potential benefits deriving from the implementation of cross-layering against standard "layered" approach.

In order to provide a virtual positioning of the work within the existing state-of-the-art, Table 1 classifies some of the relevant works presented above in terms of some relevant features. More in details, for each considered paper, the table illustrates whether it is based on scalable video, FEC, power control, ARQ (retransmission), rate-distortion model, and so forth. Only more relevant schemes are considered for sake of clarity.

More in details, the novel contributions of the paper can be outlined as follows:

(1) the definition of a suitable framework and an architecture for evaluating the potential benefits of cross-layer design in embedded video transmission over a wireless link under energy constraints,

(2) the analysis of the advantages and drawbacks deriving from cross-layer design rather than maintaining codec/link-layer independence in video transmission over wireless links,

(3) the flexibility of the proposed framework in terms of adaptability to different scenarios (i.e., different modulations, channel models, codec configurations),

(4) the usage of the rate-distortion characteristics of the embedded video bitstream, enabling to generalize the results to more than a specific video sequence (or codec).

The paper is organized as follows. Section 2 introduces the considered scenario and describes a possible architecture for enabling cross-layer interaction. Section 3 provides insight to the characterization of an embedded video stream, focusing on MPEG-4 FGS mode. Section 4 presents the statement of the problem and proposes an analytical method to

Table 1: Comparison among relevant state-of-the-art approaches and the work presented in the paper.

	Rate-distortion model	Video stream scalability	Power control	FEC	ARQ
Reference [8]	×	FGS	×	✓	×
Reference [9]	✓	FGS	×	✓	×
Reference [11]	×	Layered	✓	✓	×
Reference [17]	×	Multiple description	×	✓	✓
References [13, 18]	×	SVC/Layered	×	✓	×
Reference [19]	✓	Adaptive slicing	×	✓	×
Reference [14]	×	H.264/AVC	×	✓	×
Reference [20]	×	JSCC	✓	✓	×
Reference [21]	×	Layered	✓	✓	×
Proposed	✓	Embedded (MPEG-4 FGS for testing)	✓	✓	×

provide a numerical solution. Results are presented and analyzed in Section 5, while Section 6 draws conclusions and provides insights on future works on the topic.

2. THE CONSIDERED SCENARIO AND REFERENCE ARCHITECTURE

The considered scenario is a wireless video sensor network, where mobile devices equipped with video sensors (e.g., PDAs, smart phones) are moving within an area covered by wireless cells. Each mobile terminal within a wireless cell needs to transmit a video sequence to a remote control center or video server located on the Internet (see Figure 1). Mobility of the terminals implies the usage of batteries and thus constraints on power consumption in order to maximize the lifetime of the device.

We are interested in studying the optimal set of parameters able to provide satisfactory video quality and proper energy savings within the single-hop uplink transmission between the mobile node and the wireless access point. As mentioned in Section 1, power control and forward error correction are considered.

In order to provide the required functionality, an additional power control module needs to be implemented within the wireless interface of the mobile terminal, which controls the power allocated to each video packet on the basis of information provided by the MPEG-4 video encoder and radio MAC/physical modules.

Signal-to-noise ratio (SNR) information is required for effective power control since it provides an estimate of the status of the channel. Several works, like [24, 25], discussed how to properly estimate SNR on a wireless link, for example on an IEEE 802.11 link. A common assumption is to consider the link symmetric, implying that SNR observed from either station on the link is very similar, and thus allowing to use the SNR of the last ACK frame as an indication of the SNR at the other side [25]. Therefore, in the following paragraphs we will assume that SNR information can be directly calculated at the MAC level.

The resulting cross-layering reference architecture is presented in Figure 2, where the power control module drives

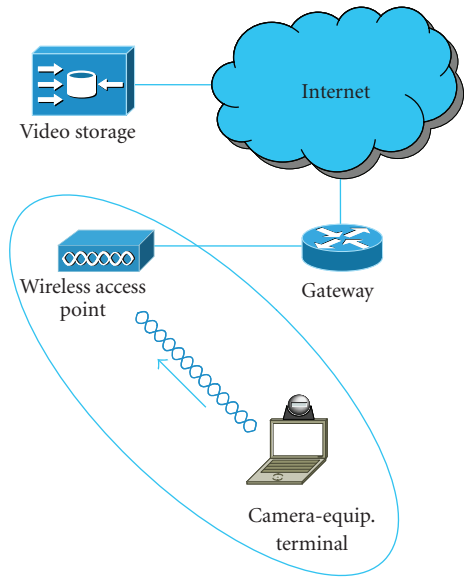


Figure 1: The scenario under consideration.

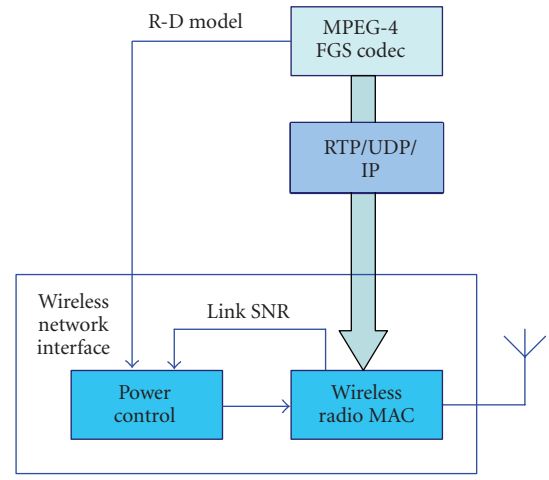


Figure 2: Block diagram of the main components at the mobile terminal side.

the power per frame allocated by the radio interface on the basis of the link state (Link SNR) and information related to the visual importance of each video packet in the form of rate-distortion characteristic of the compressed video (R-D model). In Figure 2, the wireless network interface represents the whole network adapter of the device, thus comprising the physical as well as the link layers—as often encountered in real implementations.

In the remainder of the paper, the encapsulation of video data by RTP, UDP, and IP protocols is not considered, since the modeling and analysis are oriented towards the MAC and physical level only. Furthermore, no fragmentation of the video packets produced by MPEG-4 FGS encoder is allowed at lower levels.

3. CHARACTERIZATION OF THE EMBEDDED VIDEO STREAM

3.1. Characteristics and model of a progressively coded bitstream

Multimedia data can be coded with progressive coding techniques. When decoded, the generated bitstreams progressively add enhancement information to the recovered data. The decoding process can be interrupted at any point, and the data decoded up to that point can be interpreted as a low-resolution or low-quality version of the fully decoded data. In progressively coded bitstreams, there are no distinct layers, as with traditional layered coding. Indeed, with the traditional approach, scalability is achieved by coding the data in different separate coding layers, starting with the *base layer* (BL), which contains basic information, and then generating one or more *enhancement layers* (ELs). For decoding, the BL is needed before the first EL can be decoded and so on. In progressive coding, scalability is achieved though direct truncation of the bitstream. This approach differs from traditional layered methods for video scalability because of its capability to achieve a smooth transition between different bit rates.

In the context of rate control, progressively coded bitstreams can be used for obtaining fine granularity data representations at lower bit rates, since such bitstreams have the property of allowing different spatial/quality resolutions depending on the amount of data being transmitted and decoded.

The most popular progressive coding implementations are based on wavelet transforms and/or bit-plane coding. These techniques enable the progressive coding of images, video, and even audio data. For example, for image coding, wavelet-based coding techniques, like those used in SPHIT [26] and EBCOT [27], can be used. These techniques differ on how the compression is achieved, but all of them can generate progressively coded bitstreams. In particular, wavelet transform is used by the newest image compression standard, JPEG2000 [28, 29], which is based on the EBCOT paradigm. JPEG2000 not only delivers a state-of-the-art compression performance, but is also flexible to accommodate tools for the implementation of region of interest (RoI) coding, perception-based quality optimization, and quality layers.

Wavelets can be further used in video compression. For example, 3D wavelet coding schemes, such as 3D SPHIT [30], can be used in obtaining embedded bitstreams of video data. These techniques group together a sequence of frames and apply the 3D wavelet transform to them, eventually allowing both temporal and quality scalabilities.

Another important approach is represented by the fine granular scalability (FGS), as included in the streaming profile of the MPEG-4 standard [23, 31], that uses a mixed implementation of layered scalability and bit-plane coding for obtaining two layers, a BL with essential information about the sequence, and a progressively coded EL that adds information and detail to the BL.

Due to its structure, the EL can be truncated at any point and still be used to add information to the decoded BL. The inherent scalability and flexibility of FGS enable complexity scalability and easy resource adaptation depending on the capabilities of video devices. Thus, FGS is suitable for video conferencing and video multicasting. An interesting overview of applications enabled by FGS technology is given in [32].

Due to the many applications of the various forms of scalability, the joint video team (JVT) composed by ITU-T and ISO/IEC experts groups is currently working towards a scalable extension of the H.264/MPEG4-AVC [33]. The current reference model, commonly known as scalable video coding (SVC), includes both fine grain SNR (quality) scalability (FGS) and coarse grain SNR (CGS) scalability modes. SVC has obtained the important result of successfully addressing the problem of coding efficiency reduction, a typical issue of previous scalability schemes such as MPEG-4. Even if scalable codes have been heavily criticized in the past for such reason, the advent of SVC gives to scalable coding a new perspective, and it is not unreasonable to expect an increase of efficiency from future developments.

For the purpose of our discussion, we consider a generic embedded coded bitstream with a set of truncation points. We separately consider the data added between a given truncation point and the subsequent truncation point, and define in this way different coding layers. Embedded bitstreams are constructed in such a way that data in one point of the bitstream is strictly dependent on preceding data. As a consequence, the incorrect reception of one layer (preceding data) affects the decoding of all the information added by the subsequent ones.

For sake of simplicity, we assume that layer zero or base layer (see Section 3.2), which contains the most important information (such as header information for images or basic frame information for video), together with a small amount of data, is always received correctly. This can be possibly achieved by strong protection of such layer, for example, implementing an automatic repeat request (ARQ) scheme, or by transmitting BL data during the negotiation phase at the beginning of a video transmission transaction.

Under this hypothesis, we can compute the average distortion in the reconstructed data by considering the contributions of each single layer. If we consider an additive distortion metric, such as the mean-squared error (MSE), the

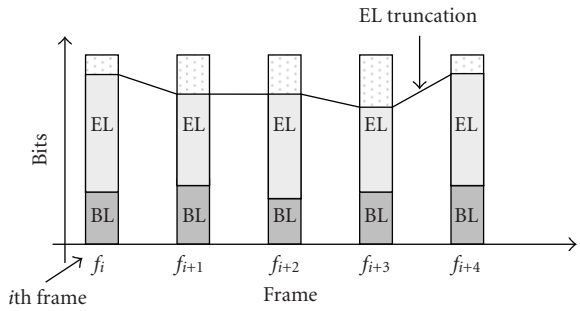


Figure 3: Example of rate control applied to FGS bitstream.

distortion associated with the reconstructed data can be expressed as follows:

$$D = D_0 - E[\Delta] = D_0 - \sum_{l=1}^{L} \prod_{i=1}^{l} (1 - \rho_i) \cdot \Delta_l, \qquad (1)$$

where D_0 is the distortion incurred if only layer zero is received, $E[\Delta]$ is the mean distortion reduction due to the reception of L layers, ρ_l is the probability of loss for the lth layer, and Δ_l is the distortion reduction due to the reception of the lth layer.

From (1), we notice that lower-numbered layers have greater importance than higher ones, thus the transmission of each layer critically depends on its position in the bitstream. Stronger protection of the lower layers becomes therefore essential, and as a natural consequence it is straightforward to combine this multilayered approach with a UEP scheme for achieving improved quality of the received video.

3.2. Progressive scalability in MPEG4

In MPEG-4 FGS, the BL behaves as a standard baseline MPEG-4 compressed bitstream, while the EL is obtained by encoding the difference between the BL and the original sequence. Fine granular scalability is provided by the EL, since residual data is block-encoded and the DCT coefficients are bit-plane coded, thus generating an embedded bitstream. Since the EL encoded data is transmitted starting from the most significant bit-plane (MSBP) to the least significant bit-plane (LSBP), a truncated EL bitstream (Figure 3) can still be used for improving, together with the BL, the reconstruction quality of the video sequence.

If the individual bit-planes are considered, the contribution given to the quality of the decoded sequence by the data encoded in the EL bitstream decreases as we move from the most significant bit-plane (MSBP) to the least significant one (LSBP). At the same time, data from MSBP is easier to compress, since it is more correlated than data from the LSBP. This aspect is shown in the characteristic rate-distortion (R-D) curve (Figure 5). Typically, the most significant bit-plane (BP1) is the smallest in size (i.e., it requires the smallest number of bits), and bit-plane size significantly increases from the most significant to the least significant bit-plane. In Figure 4, the average size of the different bit-planes

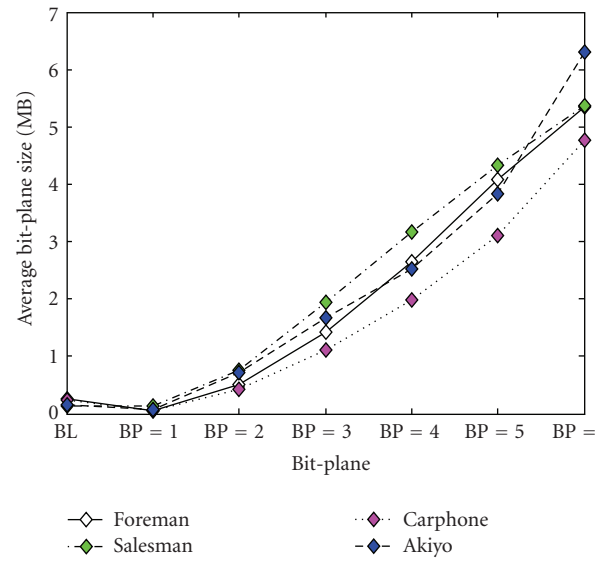


Figure 4: Average size of the bit-planes for the various QCIF reference sequences.

is plot for various reference sequences (BL is encoded with a bit rate of 14 Kbps). More details on the R-D curve are given in Section 3.3.

By taking advantage of the structure of MPEG-4 FGS coding, we can implement a prioritized UEP scheme for the EL packet [7]. This approach is possible mainly for two reasons. First, it is not possible to decode the data of a bit-plane without decoding the preceding bit-planes. Second, the data of the MSBP also carries more (perceptually relevant) information with respect to the LSBP.

The highest level of protection can be given to the MSBP, gradually reducing the protection as the bit-planes become (perceptually) less significant. This coding approach can simplify rate control algorithms implementation and can be used in combination with unequal error protection policies [7, 34].

3.3. Rate-distortion model of the FGS bitstream

Working with an operational rate-distortion (ORD) model, that is, measuring the distortion for each packet, is typically a computationally intensive task. A workaround to this problem is to use a rate-distortion (R-D) model of the EL based on the statistics collected during the encoding phase. The rate-distortion model can be derived either from empirical considerations or from analytical calculations. An interesting analysis of the FGS EL layer is given by Loguinov and Radha in [35], where also a distortion model is defined.

In this work, we utilize experimental measurements to construct an R-D curve. Based on them, the R-D model is built using a piecewise linear curve (linear within each bit-plane) [36, 37]. This is reasonable if we assume the statistical properties uniform within a single bit-plane and consider the fact that inside a bit-plane the distortion improves gradually by adding bit-plane information one MB at a time. R-D

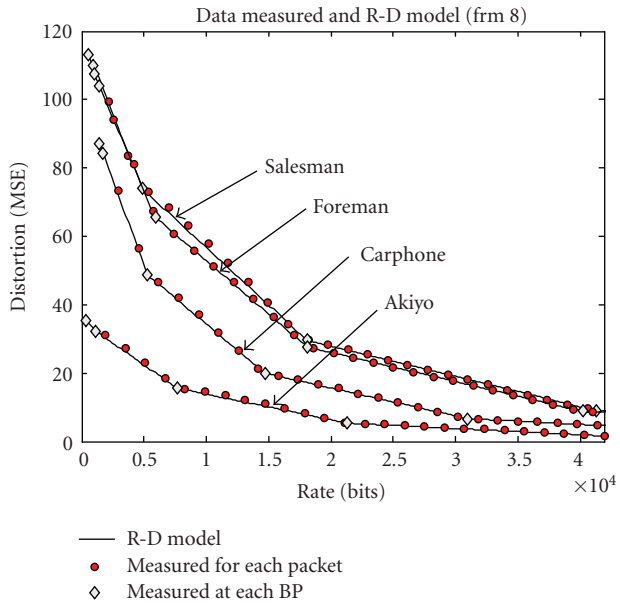


Figure 5: Comparison between the measured data and the R-D curve calculated from the BP data.

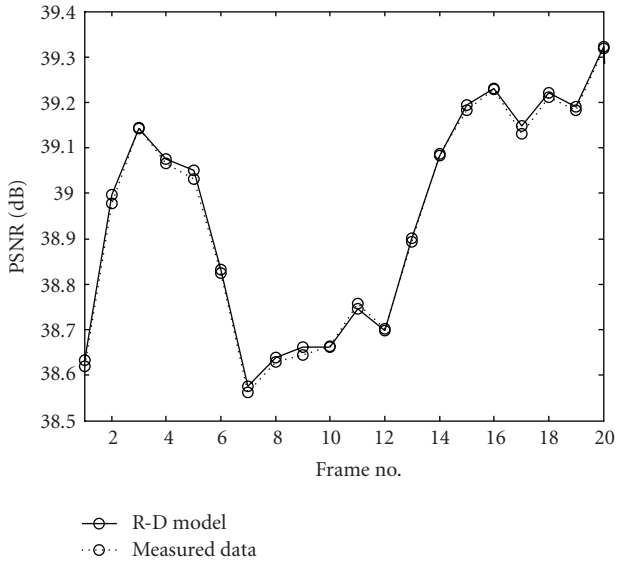


Figure 6: Comparison between the PSNR obtained using measured data and the R-D model.

data measured for each BP is shown in Figure 5 for different video sequences. As it can be seen, they validate quite well the linear-within-a-bit-plane model. The actual R-D data and the linear R-D model are both utilized by the energy allocation algorithm, to be described later, resulting in the reconstructed video of the quality shown in Figure 6. As it can be observed, the quality resulting from the two R-D models is almost indistinguishable. It is worthwhile to be mentioned here that the R-D data can be easily calculated also in the frequency domain, as highlighted in [38].

4. MODELING AND ANALYZING THE PROBLEM

4.1. The general framework

For simplicity, we consider transmitting each layer in a single link-layer packet. This assumption is not limiting, as it is possible to consider layer fragmentation in multiple link-layer packets just evaluating the distortion contribution of each portion of data in a way similar to the one described in Section 3.3. Since the design objective is to limit the energy consumption within a predetermined budget, we assume that the transmission power can be controlled at packet level. Unequal error protection can then be achieved by adjusting the transmission power used by each packet.

Link-layer retransmissions, which are a common strategy to improve reliability of communication on wireless links, are not considered, mainly for two reasons.

(1) Retransmissions introduce additional delay in data delivery, which is not controllable. As a consequence, delayed packets could be received "too late" and therefore become unusable, thus introducing serious power waste.

(2) The cost for retransmitting a packet multiplies the power consumption related to each packet. Since the goal is to minimize power consumption (or maximize battery lifetime), a single packet retransmission would double power consumption, while a lower increase in the transmission power could bring more benefits in terms of probability of correct reception.

Based on the hypothesis described above, the problem of transmitting a bitstream of compressed multimedia data with minimum end-to-end distortion D, given a limited energy budget, can be formulated as follows:

$$\min_{P_l} D, \text{ s.t. } E_{\text{tot}} = \sum_{l=1}^{L} \frac{B_l \cdot P_l}{R}, \tag{2}$$

where R is the channel rate, L the total number of transmitted packets (or layers), B_l and P_l are the size in bits and the power assigned to the lth packet, respectively, and E_{tot} the available total energy. The number of packets L is related to the bit budget. Indeed, given a bit rate R and the desired maximum transmission time T, the bit budget must equal $T \cdot R$. In case of equal size packets, L is given by $T \cdot R$ divided by the packet size B_l.

The problem can be solved using the Lagrangian relaxation method. By introducing the Lagrange multiplier, λ, (2) can be converted into the following unconstrained problem:

$$\min_{P_l} J = \min_{P_l} \left\{ D_0 - \sum_{l=1}^{L} \prod_{i=1}^{l} (1 - \rho_i) \cdot \Delta_l + \lambda \left(\sum_{l=1}^{L} \frac{B_l \cdot P_l}{R} - E_{\text{tot}} \right) \right\}, \tag{3}$$

where J is the resulting cost function.

Since the average transmission power used by a modulation scheme directly affects the probability of packet loss, we can represent the relationship between the loss probability ρ_l of the lth packet and the transmission power P_l with a function g, such that $\rho_l = g(P_l)$. We assume that such relationship is known at the transmitter, defined using an analytical model of the wireless channel or through actual measurements.

The necessary condition for an optimum point of the cost function J is that its first derivative with respect to P_j, with $j = 1, \ldots, L$, is null. The first derivative of the cost function J with respect to P_L can be written as

$$\frac{\partial J}{\partial P_L} = \frac{\partial \rho_L}{\partial P_L} \cdot \prod_{i=1}^{L-1} (1-\rho_i) \cdot \Delta_L + \lambda \frac{B_L}{R}. \tag{4}$$

Setting it equal to zero and rearranging terms, it is possible to derive the following expressions, for $\rho_i \neq 1$:

$$\prod_{i=1}^{L-1} (1-\rho_i) = \lambda \cdot f(P_L, B_L, \Delta_L), \tag{5}$$

where

$$f(P_L, B_L, \Delta_L) = -\frac{B_L}{R \cdot \Delta_L} \cdot \left(\frac{\partial \rho_L}{\partial P_L}\right)^{-1}. \tag{6}$$

From (5), we can obtain the expression

$$\prod_{i=1}^{j} (1-\rho_i) = \lambda \cdot f(P_L, B_L, \Delta_L) \cdot \prod_{h=j+1}^{L-1} (1-\rho_h)^{-1}, \quad \text{for } j = 1, \ldots, L-2. \tag{7}$$

The first derivative of the cost function J with respect to P_j can be written, for $\rho_j \neq 1$ and $i < L$, as

$$\frac{\partial J}{\partial P_j} = \frac{\partial \rho_j}{\partial P_j} \cdot (1-\rho_j)^{-1} \cdot \sum_{l=j}^{L} \left(\prod_{i=1}^{l} (1-\rho_i) \right) \cdot \Delta_l + \lambda \frac{B_j}{R} = 0. \tag{8}$$

Substituting expressions (7) into (8), we obtain, for $j < L$,

$$\frac{\partial J}{\partial P_j} = \frac{\partial \rho_j}{\partial P_j} \cdot (1-\rho_j)^{-1} \cdot \left[\sum_{l=j}^{L-1} \left[\lambda \cdot f(P_L, B_L, \Delta_L) \cdot \prod_{h=j+1}^{L-1} (1-\rho_h)^{-1} \right] \cdot \Delta_l + \lambda \cdot (1-\rho_L) \cdot f(P_L) \cdot \Delta_L \right] + \lambda \cdot \frac{B_j}{R} = 0, \tag{9}$$

or

$$\frac{\partial J}{\partial P_j} = \lambda \cdot \left[\frac{\partial \rho_j}{\partial P_j} \cdot \frac{(1-\rho_L)}{(1-\rho_j)} \cdot f(P_L) \cdot \left[\sum_{l=j+1}^{L} \prod_{h=j+1}^{L} (1-\rho_h)^{-1} \cdot \Delta_{l-1} + \cdot \Delta_L \right] + \frac{B_j}{R} \right] = 0. \tag{10}$$

By substituting $f(P_L, B_L, \Delta_L)$ from (6) into (9) and eliminating λ, we obtain

$$-\frac{\partial \rho_j}{\partial P_j} \cdot (1-\rho_j)^{-1} \cdot \frac{B_L}{R \cdot \Delta_L} \cdot \left(\frac{\partial \rho_L}{\partial P_L}\right)^{-1} \cdot (1-\rho_L) \cdot \left[\sum_{l=j+1}^{L} \prod_{h=j+1}^{L} (1-\rho_h)^{-1} \cdot \Delta_{l-1} + \cdot \Delta_L \right] + \frac{B_j}{R} = 0. \tag{11}$$

After some simple manipulations, we obtain, for $j < L$,

$$\left(\frac{\partial \rho_j}{\partial P_j}\right)^{-1} \cdot (1-\rho_j) \cdot B_j = \left(\frac{\partial \rho_L}{\partial P_L}\right)^{-1} \cdot (1-\rho_L) \cdot B_L \cdot \left[\sum_{l=j+1}^{L} \left[\prod_{h=l}^{L} (1-\rho_h)^{-1} \right] \cdot \frac{\Delta_{l-1}}{\Delta_L} + 1 \right]. \tag{12}$$

In the last expression, the left-hand side represents the information related to the jth packet, and it depends on the power of its subsequent packets, $(j+1)$th to Lth. This closed form expression will be used later, in combination with the channel model, for calculating the optimal energy distribution.

Following a similar approach, it is possible to formulate and solve the dual problem, that is, how to transmit the bitstream by minimizing the transmission energy E, subject to a distortion constraint. This problem can be formulated as follows:

$$\min_{P_l} \underbrace{\left\{ \sum_{l=1}^{L} \frac{B_l \cdot P_l}{R} \right\}}_{E}, \qquad \text{s.t. } D_{\text{tot}} = D_0 - E[\Delta], \tag{13}$$

where D_{tot} is the maximum acceptable distortion for the frame. Again, the optimal power assignment must satisfy the same relationship given by (12). The dual problem presented here is useful for applications in which a desired level of visual quality must be maintained using the least amount of energy.

4.2. AWGN channel model

In this section, we consider the transmission of the embedded bitstream through a channel with additive white Gaussian noise (AWGN) by using different digital modulation schemes. Such channel model is considered for its simplicity, even if any channel (multipath, indoor, outdoor) can be considered—given that it is possible to derive the relationship between energy (or power) and probability of error per bit. Moreover, it was demonstrated that AWGN model can be used also in case of CDMA modulation, under specific hypothesis [39].

Since most link-layer protocols discard packets containing errors, from the application point-of-view we can assume that generally the probability of packet loss can be written as the probability of losing at least one bit:

$$\rho_j = 1 - (1 - \varepsilon_j)^{B_j}, \tag{14}$$

where ε is the bit-error probability, which depends on the adopted modulation scheme.

Well-known results from the communication theory [40] provide a relationship between ε and the transmission parameters for different modulations. For the BPSK modulation, for example, ε is given by

$$\varepsilon = Q\left(\sqrt{2 \cdot \gamma_b}\right), \quad \text{with} \gamma_b = \frac{E_b}{N_0}, \tag{15}$$

where E_b is the bit energy, N_0 the noise power per Hz; and the function $Q(x)$ is given by:

$$Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-u^2/2} \cdot du. \tag{16}$$

Generalizing, for several digital modulation schemes, the bit-error probability can be written as

$$\varepsilon = a \cdot Q\left(\sqrt{\alpha \cdot \gamma_b}\right), \tag{17}$$

where the values of a and α for different modulations are summarized in Table 2.

Equation (12) can then be expressed as

$$\left(1 - a \cdot Q\left(\sqrt{\alpha \cdot \frac{E_b^j}{N_0}}\right)\right) \cdot \sqrt{E_b^j} \cdot e^{(\alpha \cdot E_b^j)/(2 \cdot N_0)}$$

$$= \left(1 - a \cdot Q\left(\sqrt{\alpha \cdot \frac{E_b^L}{N_0}}\right)\right) \cdot \sqrt{E_b^L} \cdot e^{(\alpha \cdot E_b^L)/(2 \cdot N_0)}$$

$$\cdot \left[1 + \sum_{l=j+1}^{L} \prod_{h=l}^{L} \left(1 - a \cdot Q\left(\sqrt{\alpha \cdot \frac{E_b^h}{N_0}}\right)\right)^{-B_h} \cdot \frac{\Delta_{l-1}}{\Delta_L}\right]. \tag{18}$$

The minimization problem can be solved by finding the value of E_b^L that satisfies the energy constraint. Since it is not possible to provide an analytical closed form solution to the problem, a numerical method is used.

Table 2: Parameters a, α, and the spectral efficiency r_b/B_T for different modulations.

Modulation	a	α	r_b/B_T
FSK	1	1	1
BPSK; PSK	1	2	1
MSK; QAM; QPSK	1	2	2

Once the energy levels for each packet are known, the packet transmission power for packet j can be calculated using the following relationship:

$$P_j = E_b^j \cdot R. \tag{19}$$

4.3. Power control with forward error correction

During video transmission, for increasing the protection of the information stored in the packets, it is possible to add redundancy bits by employing a forward error correction (FEC) code. This operation, usually performed at MAC level in building the MAC frame, enables to recover a given percentage of transmission errors without requiring interaction between the sender and the receiver.

In this section, we address the problem of power control in presence of FEC for optimizing video streams delivery, and analyze the performance of the system in different scenarios. Optimal energy distribution is jointly employed with error correction schemes in order to achieve optimal nonuniform error protection using different modulation schemes. Moreover, the analytical formulation of the proposed framework can be used for estimating the degree of protection offered by power control and FEC, as well as the system sensitivity to the parameters setup. In the remainder of the paper, we will assume that FEC code rate is fixed for all the EL packets—modeling a fixed frame check sequence (FCS) field which is quite common in wireless networks.

A well-known FEC code is the Reed-Solomon code [41], which divides the codeword in m-bit symbols. By its definition, an RS(n, k) code with a codeword of n symbols and k symbols of data can correct up to $t = (n - k)/2$ symbols errors. A symbol represents a sequence of bits, considered as a single "block of information" for the purpose of application of the FEC code.

The probability of packet loss deriving from the employment of an FEC strategy can then be written as the probability of receiving more than t uncorrect symbols, that is,

$$\rho = 1 - \sum_{i=0}^{t} \binom{n}{i} \cdot P_{\text{symbol}}^i \cdot (1 - P_{\text{symbol}})^{n-i}, \tag{20}$$

where P_{symbol} is the error probability for a symbol of m bits, and depends on the bit-error rate ε in the following way:

$$P_{\text{symbol}} = 1 - (1 - \varepsilon)^m. \tag{21}$$

The term ε is the bit-error probability, and depends on the employed modulation parameters and the characteristics of the transmission channel, as described in Section 4.2.

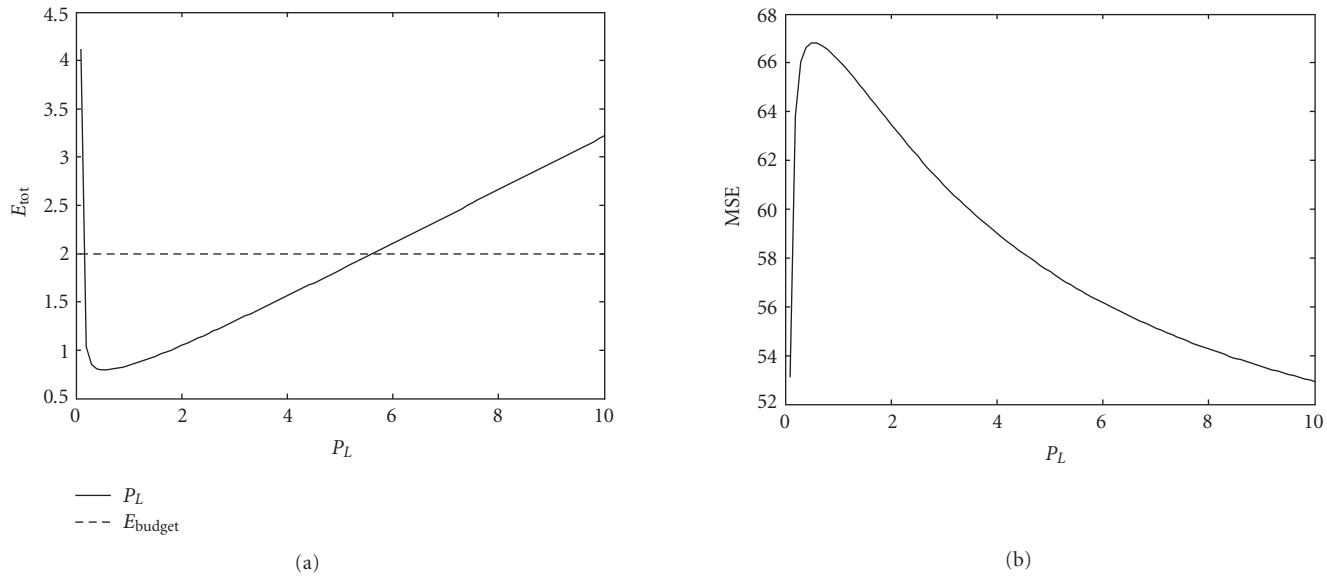


FIGURE 7: Foreman QCIF sequence, frame 10, equal-size packets: (a) total energy E_{tot} (J) and (b) frame distortion (MSE) (dB) versus the power of the last packet, P_L (W). In (a), the dotted line represents the energy budget.

4.4. Solution for AWGN channel model

As a consequence of the introduction of the FEC schemes, (8) can then be expressed as in the following formula:

$$\frac{\sqrt{E_b^j}\cdot e^{(E_b^j\cdot\alpha)/(2\cdot N_0)}\cdot(1-\varepsilon_j)^{1-m}\cdot(1-\rho_j)}{\sum_{i=0}^{t}\binom{n}{i}\cdot P_{\text{symbol},j}^{i-1}\cdot(1-P_{\text{symbol},j})^{n-i-1}\cdot(i-n\cdot P_{\text{symbol},j})} = \frac{\sqrt{E_b^L}\cdot e^{(E_b^L\cdot\alpha)/(2\cdot N_0)}\cdot(1-\varepsilon_L)^{1-m}\cdot(1-\rho_L)}{\sum_{i=0}^{t}\binom{n}{i}\cdot P_{\text{symbol},L}^{i-1}\cdot(1-P_{\text{symbol},L})^{n-i-1}\cdot(i-n\cdot P_{\text{symbol},L})} \cdot\left[\sum_{l=j+1}^{L}\prod_{h=l}^{L}(1-\rho_h)^{-1}\cdot\frac{\Delta_{l-1}}{\Delta_L}+1\right]. \tag{22}$$

The minimization problem can be solved by finding the value of E_b^L that satisfies the energy constraint. Indeed, (22) allows to compute the energy to be assigned to the packet j, E_b^j, from those assigned to the subsequent packets $\{E_b^{j+1},\ldots,E_b^L\}$. Starting from a given E_b^L, we can then compute the optimal energy distribution $\{E_b^1,\ldots,E_b^L\}$ to be assigned to the packets and the energy budget required for transmitting them.

It is possible then to use (12) to construct a curve $E_{\text{budget}} = f(E_b^L)$ that extrapolates the relationship between the energy budget and the optimal choice of E_b^L.

A numerical algorithm (e.g., the bisection method) can be used to find the value of E_b^L, and thus the optimal energy distribution $\{E_b^1,\ldots,E_b^L\}$, that satisfies the energy constraint. In the case that a solution with E_b^L greater than zero does not exist, the solution of the minimization problem must be searched using $L' = L - 1$ packets.

Once the energy levels for each packet are known, the transmission power for the jth packet can be calculated as $P_j = E_b^j \cdot R$. An example of the numerical approach described above is presented in Figure 7.

5. RESULTS

A number of experiments have been run simulating the transmission of various sequences. Results presented in this section are obtained by simulating the transmission of the QCIF test sequence *foreman*, with a frame rate of 10 fps, noise power N_0 of 10^{-5} W/Hz, and $B_T = 200$ KHz, unless otherwise stated. The sequence is encoded with the MPEG-4 FGS algorithm, setting a fixed bit rate of 14 Kbps for the BL. Packet size is fixed (100 bytes). The maximum number of transmitted EL packets, L, is calculated based on the video frame rate and the available bit rate r_b. The value of r_b depends on the transmission bandwidth B_T and the spectral efficiency of the adopted modulation scheme (Table 2), according to the equation

$$\sigma_e = \frac{r_b}{B_T} \quad (\text{bps/Hz}). \tag{23}$$

Figure 8 presents the results achieved by the proposed optimization approach as compared with equal energy distribution among the packets, in the case of BPSK modulation, for four different test sequences: *foreman*, *carphone*, *salesman*, and *akiyo*. The employed performance metric is the *peak signal-to-noise ratio* (PSNR) at the video decoder, as is usual in video transmission schemes. The proposed solution provides a relevant advantage in error-prone situations, while for an energy value E_{tot} higher than (approximately) 0.5 Joule the performance is similar to equal energy distribution—due to the good performance of the employed

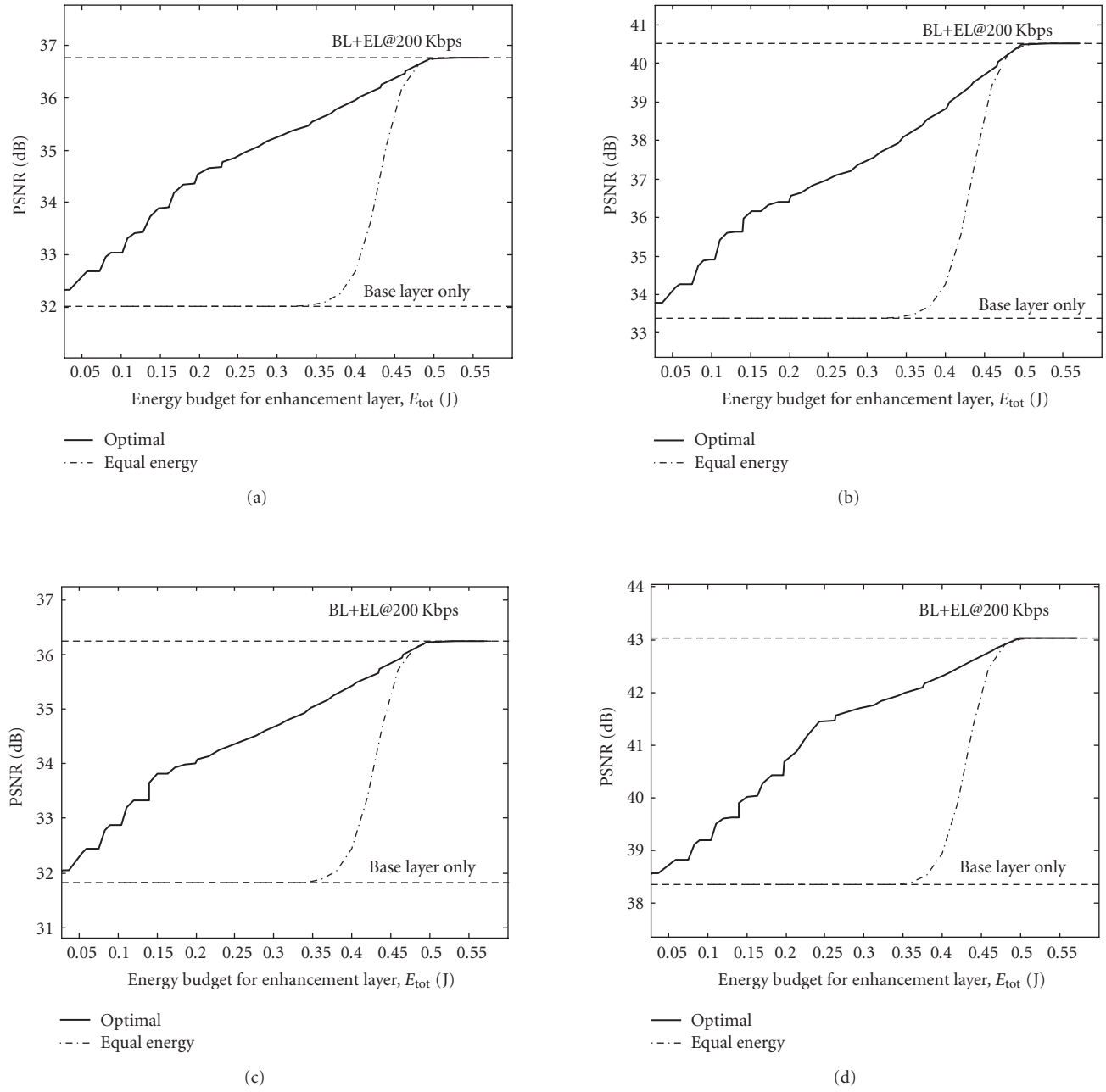


Figure 8: Performance comparison between the proposed approach and equal energy distribution (BPSK modulation, RS(160,100), B_T = 200 KHz). Test sequences are: (a) *foreman*, (b) *carphone*, (c) *salesman*, (d) *akiyo*.

modulation and FEC code. In error-prone situations, the optimal algorithm prefers not to send the less important packets allowing a higher energy protection of the remaining bitstream. This consideration remains valid for all considered test sequences.

This behavior is highlighted for the *foreman* sequence in Figure 9, where a detail on the performance contribution in terms of visual quality is shown for different choices of L.

The impact of different levels of error protection (different FEC code rates) is presented in Figure 10, where the PSNR is plot against the total available energy for two codes with different correction capabilities. Clearly, it is possible to achieve better performance in error-prone scenarios by using a stronger code, at the expense of a lower performance if the channel conditions improve.

Figure 11 summarizes the obtained results for the transmission of a sequence using different modulation schemes (BPSK, MSK, FSK). BPSK achieves better performance than FSK, while MSK provides a higher spectral efficiency, thus allowing transmission of a higher portion of the EL bitstream. Figure 12 demonstrates that the joint employment of FEC codes and power control improves the robustness of the

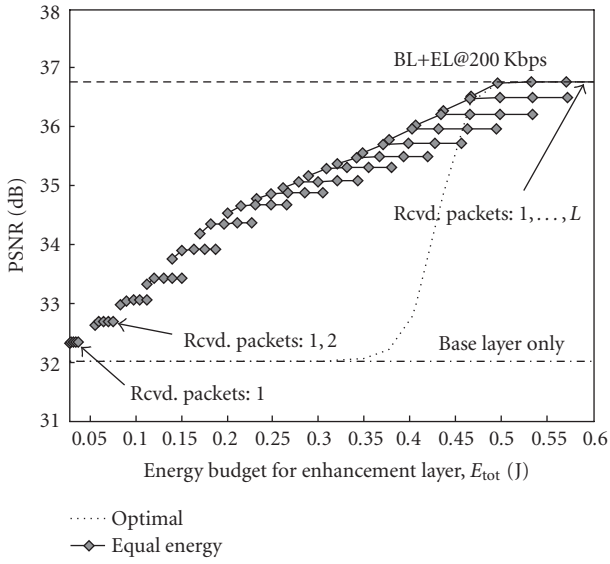


Figure 9: Performance comparison between the proposed approach and equal energy distribution (BPSK modulation, RS(160,100), $B_T = 200$ KHz). Incremental PSNR contribution for each received packet beyond the BL is represented.

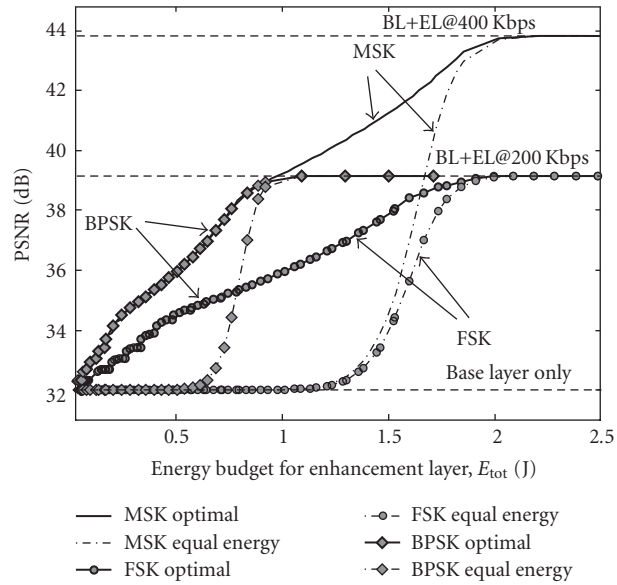


Figure 11: Comparison among different modulation schemes (BPSK, FSK, and MSK) in terms of PSNR at the receiver (RS(110,100), $B_T = 200$ KHz).

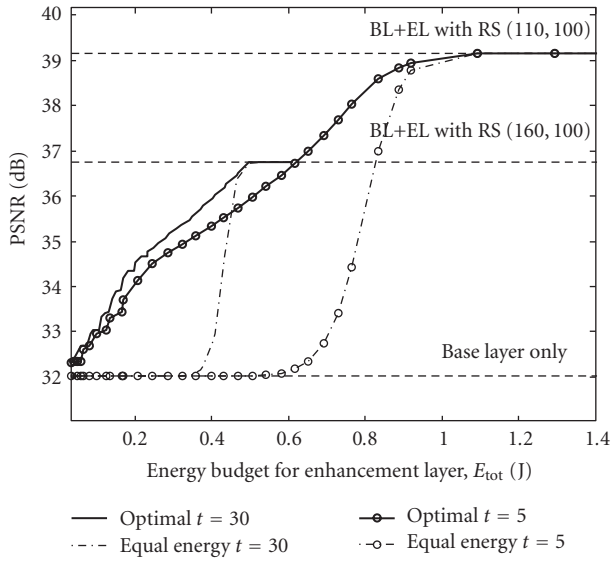


Figure 10: Performance of the proposed approach and equal energy distribution for different RS codes: RS(110,100) and RS(160,100) (BPSK, $B_T = 200$ KHz).

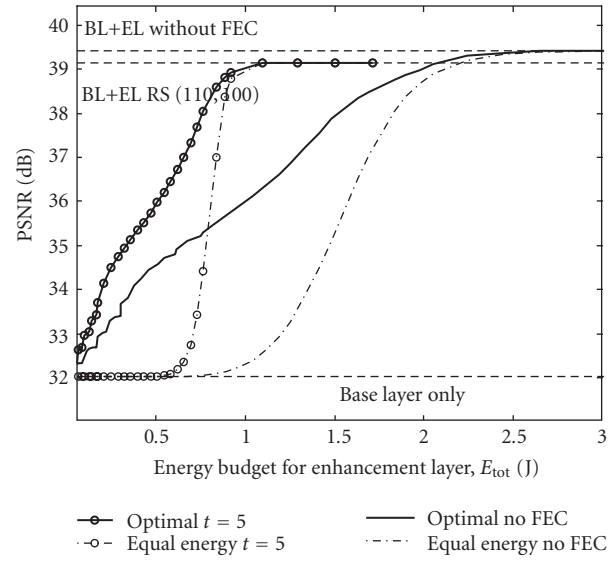


Figure 12: Performance improvement deriving by the introduction of RS(110,100) code (BPSK, $B_T = 200$ KHz).

video stream. On the other hand, equal energy distribution is characterized by a step-like behavior, making it very sensitive to changing channel conditions—especially if FEC is employed.

Finally, Figure 13 summarizes the obtained results, comparing the considered cross-layering approach (UEP) with equal energy distribution (EEP). This allows us to conclude that there is a relevant advantage in the employment of optimized power control and forward error correction—especially when the available energy budget is limited, while above a certain level of available transmission resources transmission is reliable and thus unequal protection (and cross-layering) is less valuable. Moreover, the introduction of FEC makes the performance more sensitive to the available power, and as a consequence requires reliable estimation of the state of the channel.

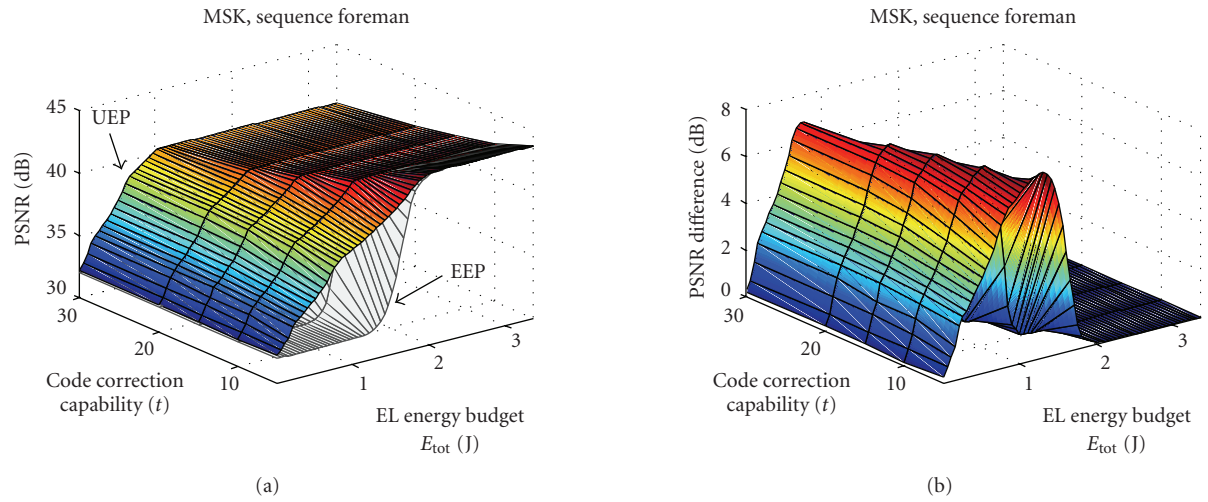


Figure 13: PSNR performance of cross-layer UEP (a) and improvement against EEP (b) versus energy budget and code correction capability of FEC.

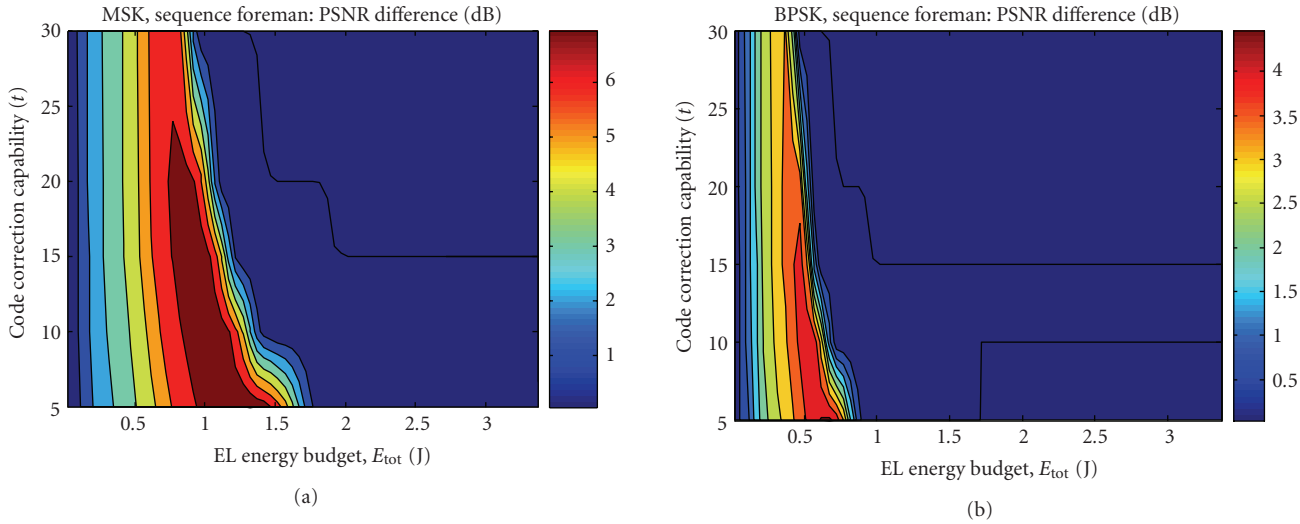


Figure 14: Performance gain of cross-layer UEP against EEP for MSK (a) and BPSK (b) modulations.

Summarizing, from the analysis of our model we conclude that cross-layering oriented towards unequal error protection can provide a relevant performance gain, especially for limited availability of transmission resources (energy budget lower than 2 Joules in Figure 13) or in the case of a very noisy channel. In particular, interaction between the video encoder and power control is advisable for power-constrained devices, while FEC should be considered more carefully since it is making the system more sensitive to channel conditions, even if it could provide additional power saving. This aspect is further illustrated in Figure 14(a), which represents a different view of the graph in Figure 13(b). Figure 14(b) underlines that similar behavior holds for other modulations as well (BPSK in particular).

Finally, Figure 15 provides a comparison among different modulation schemes (FSK, MSK, and BPSK) in terms of energy spent against maximum achievable PSNR gain. The curves are plot for different correction capabilities of the code (t). The graph clearly outlines that maximum power savings is achieved by using BPSK (providing the optimal parameter settings, as well), while the highest gain (more than 7 dB against EEP, more than 3 dB against BPSK) derives by the employment of MSK at the expense of a slightly higher energy consumption.

As a final remark regarding the possible implementation of a system supporting the proposed cross-layered architecture, the complexity of the devices will need to be increased (with reference to the fully layered architecture) mainly to support interaction between the source encoder and the link and physical layer protocols, in order to enable information exchange. The process of optimization does not severely impact on the complexity of the resource-constrained video

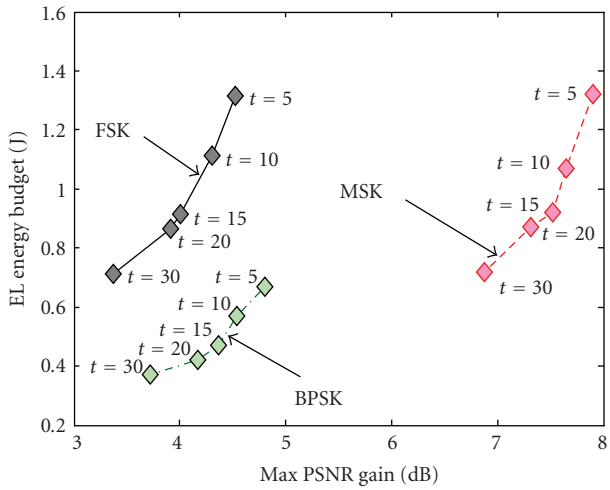


FIGURE 15: Relationship between PSNR gain of cross-layer UEP strategy against EEP and available energy budget.

device, since having a suitable R-D model of the video source (which is likely, i.e., in the case of surveillance applications) the whole process can be performed offline and then allow adaptive power and FEC allocation on the basis of a simple lookup table for each EL packet.

6. CONCLUSIONS AND FUTURE WORK

Cross-layering is an interesting design paradigm, recently proposed to overcome the limitations deriving from the ISO/OSI layering principle in order to improve performance of communications in specific scenarios, such as wireless multimedia communications. However, most available solutions are based on empirical reasoning, and do not provide a theoretic background supporting such approaches. The paper provides an analytical framework for the study of single-hop embedded video delivery over a wireless link, enabling the study of cross-layer interactions for performance optimization using power control and FEC and providing a useful tool for determining the potential benefits of such architecture. From analysis of the achieved results, the following remarks can be derived: (i) cross-layering is necessary to allow video transmission in presence of limited transmission resources (bandwidth, power); (ii) introduction of FEC codes brings relevant performance gains, but performance becomes more sensitive to available energy budget; (iii) adaptation of transmission power and FEC based on R-D model provides more benefits than independent usage of power control or FEC, supporting a wider range in terms of energy budget.

Future work will deal with the analysis of more complex scenarios (such as in presence of link-layer retransmissions, multihop communication) and with the investigation on the definition of an end-to-end R-D model for multihop video transmission.

REFERENCES

[1] IEEE standard 802.11h supplement, "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. Spectrum and transmit power management extensions in the 5GHz band in Europe," 2003.

[2] Q. Wang and M. A. Abu-Rgheff, "Cross-layer signalling for next-generation wireless systems," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '03)*, vol. 2, pp. 1084–1089, New Orleans, La, USA, March 2003.

[3] V. Kawadia and P. R. Kumar, "A cautionary perspective on cross-layer design," *IEEE Wireless Communications*, vol. 12, no. 1, pp. 3–11, 2005.

[4] U. Horn, K. Stuhlmüller, M. Link, and B. Girod, "Robust Internet video transmission based on scalable coding and unequal error protection," *Signal Processing: Image Communication*, vol. 15, no. 1, pp. 77–94, 1999.

[5] A. Natu and D. Taubman, "Unequal protection of JPEG2000 code-streams in wireless channels," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '02)*, vol. 1, pp. 534–538, Taipei, Taiwan, November 2002.

[6] Z. Wu, A. Bilgin, and M. W. Marcellin, "Unequal error protection for transmission of JPEG2000 codestreams over noisy channels," in *Proceedings of IEEE International Conference on Image Processing (ICIP '02)*, vol. 1, pp. 213–216, Rochester, NY, USA, September 2002.

[7] M. van der Schaar and H. Radha, "Unequal packet loss resilience for fine-granular-scalability video," *IEEE Transactions on Multimedia*, vol. 3, no. 4, pp. 381–394, 2001.

[8] X. K. Yang, C. Zhu, Z. G. Li, G. N. Feng, S. Wu, and N. Ling, "A degressive error protection algorithm for MPEG-4 FGS video streaming," in *Proceedings of IEEE International Conference on Image Processing (ICIP '02)*, vol. 3, pp. 737–740, Rochester, NY, USA, September 2002.

[9] G. Wang, Q. Zhang, W. Zhu, and Y.-Q. Zhang, "Channel-adaptive unequal error protection for scalable video transmission over wireless channel," in *Visual Communications and Image Processing*, vol. 4310 of *Proceedings of SPIE*, pp. 648–655, San Jose, Calif, USA, January 2001.

[10] Q. Zhang, W. Zhu, and Y.-Q. Zhang, "Network-adaptive scalable video streaming over 3G wireless network," in *Proceedings of IEEE International Conference on Image Processing (ICIP '01)*, vol. 3, pp. 579–582, Thessaloniki, Greece, October 2001.

[11] S. Zhao, Z. Xiong, and X. Wang, "Joint error control and power allocation for video transmission over CDMA networks with multiuser detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 6, pp. 425–437, 2002.

[12] Y. S. Chan and J. W. Modestino, "Transport of scalable video over CDMA wireless networks: a joint source coding and power control approach," in *Proceedings of IEEE International Conference on Image Processing (ICIP '01)*, vol. 2, pp. 973–976, Thessaloniki, Greece, October 2001.

[13] Y. Pei and J. W. Modestino, "Multi-layered video transmission over wireless channels using an adaptive modulation and coding scheme," in *Proceedings of IEEE International Conference on Image Processing (ICIP '01)*, vol. 2, pp. 1009–1012, Thessaloniki, Greece, October 2001.

[14] P. Raibroycharoen, M. M. Ghandi, E. V. Jones, and M. Ghanbari, "Performance analysis of H.264/AVC video transmission with unequal error protected turbo codes," in *Proceedings of the 61st IEEE Vehicular Technology Conference (VTC '05)*, vol. 3, pp. 1580–1584, Stockholm, Sweden, May-June 2005.

[15] S. Zhao, Z. Xiong, X. Wang, and J. Hua, "Progressive video delivery over wideband wireless channels using space-time differentially coded OFDM systems," *IEEE Transactions on Mobile Computing*, vol. 5, no. 4, pp. 303–316, 2006.

[16] S. Zhao, Z. Xiong, and X. Wang, "Optimal resource allocation for wireless video over CDMA networks," *IEEE Transactions on Mobile Computing*, vol. 4, no. 1, pp. 56–67, 2005.

[17] T. Gan, L. Gan, and K.-K. Ma, "Reducing video-quality fluctuations for streaming scalable video using unequal error protection, retransmission, and interleaving," *IEEE Transactions on Image Processing*, vol. 15, no. 4, pp. 819–832, 2006.

[18] Y. Wang, T. Fang, L.-P. Chau, and K.-H. Yap, "Two-dimensional channel coding scheme for MCTF-based scalable video coding," *IEEE Transactions on Multimedia*, vol. 9, no. 1, pp. 37–45, 2007.

[19] O. Harmanci and A. M. Tekalp, "Rate-distortion optimal video transport over IP with bit errors," in *Proceedings of IEEE International Conference on Image Processing (ICIP '06)*, pp. 1305–1308, Atlanta, Ga, USA, October 2006.

[20] Y. S. Chan and J. W. Modestino, "Video delivery over CDMA cellular networks using rate-compatible punctured turbo (RCPT) codes combined with joint source coding-power control," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '03)*, vol. 6, pp. 3585–3589, San Francisco, Calif, USA, December 2003.

[21] H. Zheng and K. J. R. Liu, "Power optimized space-time code for layer coded multimedia over wireless channels," in *Proceedings of IEEE International Conference on Image Processing (ICIP '99)*, vol. 3, pp. 95–99, Kobe, Japan, October 1999.

[22] C. E. Costa, F. G. B. de Natale, and F. Granelli, "Embedded packet video transmission over wireless channels using power control and forward error correction," in *Proceedings of IEEE International Conference on Communications (ICC '05)*, vol. 3, pp. 1433–1437, Seoul, Korea, May 2005.

[23] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 301–317, 2001.

[24] 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.

[25] J. del Prado Pavon and S. Choi, "Link adaptation strategy for IEEE 802.11 WLAN via received signal strength measurement," in *Proceedings of IEEE International Conference on Communications (ICC '03)*, vol. 2, pp. 1108–1113, Anchorage, Alaska, USA, May 2003.

[26] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 3, pp. 243–250, 1996.

[27] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Transactions on Image Processing*, vol. 9, no. 7, pp. 1158–1170, 2000.

[28] M. Boliek, C. Christopoulos, and E. Majani, Eds., *JPEG2000—Part I Final Draft International Standard (ISO/IEC FDIS15444-1)*, ISO/IEC JTC1/SC29/WG1 N1855, August 2000.

[29] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Practice and Standards*, Kluwer Academic Publishers, Boston, Mass, USA, 2002.

[30] B.-J. Kim, Z. Xiong, and W. A. Pearlman, "Low bit-rate scalable video coding with 3-D set partitioning in hierarchical trees (3-D SPIHT)," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 8, pp. 1374–1387, 2000.

[31] ISO/IEC MPEG-4, "Information Technology—Coding of Audio Video Object—Part 2: Visual—Amendment 4: Streaming Video Profile," MPEG 2000/N3518, July 2000.

[32] M. van der Schaar, L. G. Boland, and Q. Li, "Novel applications of fine-granular-scalability: Internet & wireless video, scalable storage, personalized TV, universal media coding," in *Proceedings of the 5th World Multi-Conference on Systemics, Cybernetics and Informatics and the 7th International Conference on Information Systems Analysis and Synthesis (SCI/ISAS '01)*, Orlando, Fla, USA, July 2001.

[33] "Joint Scalable Video Model JSVM-4, JVT Q202, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG," Nice, France, October 2005.

[34] 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.

[35] M. Dai, D. Loguinov, and H. Radha, "Statistical analysis and distortion modeling of MPEG-4 FGS," in *Proceedings of IEEE International Conference on Image Processing (ICIP '03)*, vol. 3, pp. 301–304, Barcelona, Spain, September 2003.

[36] L. Zhao, J. W. Kim, and C.-C. J. Kuo, "MPEG-4 FGS video streaming with constant-quality rate control and differentiated forwarding," in *Viual Communications and Image Processing*, vol. 4671 of *Proceedings of SPIE*, pp. 230–241, San Jose, Calif, USA, January 2002.

[37] L. Zhao, J. W. Kim, and C.-C. J. Kuo, "Constant quality rate control for streaming MPEG-4 FGS video," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '02)*, vol. 4, pp. 544–547, Phoenix, Ariz, USA, May 2002.

[38] M. Dai and D. Loguinov, "Analysis of rate-distortion functions and congestion control in scalable Internet video streaming," in *Proceedings of the 13th International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV '03)*, pp. 60–69, Monterey, Calif, USA, June 2003.

[39] G. L. Turin, "Effects of multipath and fading on the performance of direct-sequence CDMA systems," *IEEE Journal on Selected Areas in Communications*, vol. 2, no. 4, pp. 597–603, 1984.

[40] A. B. Carlson, *Communication Systems*, McGraw-Hill, New York, NY, USA, 2001.

[41] T. S. Rappaport, *Wireless Communications—Principles and Practice*, Prentice-Hall, Upper Saddle River, NJ, USA, 1999.

Hindawi Publishing Corporation
Advances in Multimedia
Volume 2007, Article ID 19860, 9 pages
doi:10.1155/2007/19860

Research Article

Cross-Layer Path Configuration for Energy-Efficient Communication over Wireless Ad Hoc Networks

Hong-Chuan Yang,[1] Kui Wu,[2] and Wu-Sheng Lu[1]

[1] *Department of Electrical and Computer Engineering, University of Victoria, BC, Canada V8W 3P6*
[2] *Department of Computer Science, University of Victoria, BC, Canada V8W 3P6*

Received 27 December 2006; Revised 12 April 2007; Accepted 12 April 2007

Recommended by Jianwei Huang

We study the energy-efficient configuration of multihop paths with automatic repeat request (ARQ) mechanism in wireless ad hoc networks. We adopt a cross-layer design approach and take both the quality of each radio hop and the battery capacity of each transmitting node into consideration. Under certain constraints on the maximum tolerable transmission delay and the required packet delivery ratio, we solve optimization problems to jointly schedule the transmitting power of each transmitting node and the retransmission limit over each hop. Numerical results demonstrate that the path configuration methods can either significantly reduce the average energy consumption per packet delivery or considerably extend the average lifetime of the multihop route.

1. INTRODUCTION

There have been considerable interests in selforganizing, fast deployable wireless ad hoc networks within both academic society [1, 2] and industry [3]. Since these networks consist of a group of battery operated wireless devices, they are ideal for providing instantaneous wireless services without deploying access points or wired infrastructure. On the other hand, limited battery lifetime greatly affects the usefulness of wireless ad hoc networks. Therefore, it is of great importance to develop energy-efficient communication techniques for such networks. Recently, there have been substantial research efforts in developing energy-efficient routing protocols for wireless ad hoc networks (see, e.g., [4–9]). The basic idea of these energy-efficient routing protocols is to integrate energy metrics into the route search or maintenance process. While saving considerable amount of energy compared to traditional routing protocols, these energy-aware routing algorithms become more complex and very difficult to implement. *Decoupling routing algorithms and other add-on features, for example, energy saving in our case, is of great importance from the point of view of protocol engineering and has become a broadly accepted industrial practice.*

In this paper, we follow the above philosophy to improve energy efficiency of wireless ad hoc networks. Our objective is to develop energy-efficient configuration algorithms for a multihop path that has been obtained through traditional routing protocols. The rationale behind this approach is that if the path obtained is not properly configured, the battery energy at some intermediate nodes may be quickly depleted and the whole path becomes unusable. The resulting route recovery operations [10, 11] will lead to extra energy cost for the whole network. In addition, since the path configuration is decoupled with any routing protocols or transport layer protocols, *it can serve as an add-on feature to existing routing or transport protocols with low implementation complexity*. Specifically, we consider a multihop path with hop-by-hop automatic repeat request (ARQ) mechanism. With hop-by-hop ARQ, a data packet must be acknowledged in the current hop before it could be transmitted over the next hop. Otherwise, packet retransmission occurs. Traditionally, this transmission/retransmission process is continued until either the packet arrives at the destination node correctly or the packet is dropped because the maximum number of allowed retransmissions is exceeded for that packet. This retransmission limit usually stems from the delay constraint of the data traffic, especially those generated by voice and/or video applications (in this work, we focus on the transmission delay while assume queuing delay due to multiple flow has been subtracted from the total delay budget). Obviously, this best-effort transmission strategy will lead to the best end-to-end path reliability, that is, the lowest packet loss rate. The energy

cost for packet delivery with this strategy, however, will be large. Moreover, we can intuitively expect that if the number of retransmissions performed over an intermediate hop is large, the probability that the packet can successfully reach its final destination within the delay constraint will be small. It will be more energy efficient if we drop the packets immediately when the probability for the packet to reach its ultimate destination within a given delay constraint becomes very small.

With this observation in mind, we propose an energy-efficient hop-by-hop retransmission strategy for multiple-hop transmission. In particular, we allocate each hop along the path of a number of permitted retransmissions in advance.[1] If the number of performed retransmissions over a hop reaches its prespecified limit, then the transmitting node of that hop will drop the packet. To determine the number of allowed retransmissions for each hop as well as select the transmitting power for each transmitting node, we formulate optimization problems which take into account the delay constraint of the data packets, the channel quality of each hop, and the available energy supply of each transmitting node. These optimization problems are solved at the destination node where the channel quality and energy resource information of each hop have been collected during the route discovery process and the solution are then used to configure the multihop path.

Specifically, we develop two path configuration algorithms. The first algorithm, termed as *minimum-energy configuration*, targets at reducing the average energy consumption per packet delivery over the multihop path. We show through numerical examples that the minimum-energy algorithm can save considerable energy for packet delivery, compared with the traditional best-effort retransmission strategy, while guaranteeing a given quality of service (QoS) level in terms of the packet delivery ratio within a given delay constraint. While the minimum-energy configuration can reduce the average energy cost per packet transmission, it does not take into account the available energy resources of intermediate nodes along the path. We then develop another path configuration algorithm, termed as *maximum-lifetime configuration*, that tries to extends the lifetime of the multihop path by taking into consideration both the link quality of each hop and the battery resource of the transmitting nodes. Numerical examples also show that the maximum-lifetime configuration algorithm can prolong the lifetime of the multihop path at the cost of slightly increased average power consumption per packet delivery, compared to minimum-energy path configuration.

The rest of the paper is organized as follows. Section 2 introduces the system and channel model under consideration. In Section 3, we study the packet delivery ratio and average energy consumption with the best-effort transmission strategy as a benchmark. In Section 4, the minimum-energy configuration problem is formulated and solved. The optimization problem for the maximum-lifetime configuration is then presented in Section 5. Selected numerical example is presented and discussed in Section 6. In Section 7, we explain in detail how to incorporate our path configuration algorithms with existing routing/transport protocols. Finally, we conclude the paper in Section 8.

2. SYSTEM AND CHANNEL MODELS

2.1. Multihop path with fading

We consider a multihop path obtained via a certain routing protocol, where there are L hops between the source node, S, and the destination node, D. Let R_k denote the kth intermediate node for $k = 1, \ldots, L-1$. We can represent the ith hop as $R_{i-1}R_i$, $1 \le i \le L$, with the notation $R_0 = S$ and $R_L = D$. The radio link for each hop is assumed to be subject to independent Rayleigh block fading. In particular, the amplitude of the fading signal during a packet transmission can be considered constant and varies independently for the next transmission. The cumulative distribution function (CDF) $P_{\gamma_i}(x)$ of the instantaneous received signal-to-noise ratio (SNR) γ_i at R_i for the ith hop is given by

$$P_{\gamma_i}(x) = 1 - \exp\left(-\frac{x}{\overline{\gamma}_i}\right), \tag{1}$$

where $\overline{\gamma}_i$ is the average received SNR of the ith hop, which is proportional to the transmitting power of the transmitting node R_{i-1}, denoted by p_i. Specifically, we have $\overline{\gamma}_i = G_i \cdot p_i$, where G_i is a parameter depending on the antenna gain, the distance between the two nodes, and the shadowing effect, and so forth. We assume that G_i remains constant for the time duration of interest. We also assume that each transmitting node can select its transmitting power within the range of $(0, p_{\max}]$, where $p_{\max}$ is the common maximum transmitting power for all transmitting nodes.

The packet error rate over a radio hop is in general a complex function of the instantaneous received SNR of that hop. Simultaneous transmission over other hops will also cause interference to current hop. Note that since nodes cannot simultaneously transmit and receive packets, interference will only come from nonneighboring hops and therefore is small. In this paper, we treat the interference from other hops as background noise and approximate the packet error rate for the ith hop with the probability that the instantaneous received SNR γ_i is smaller than a fixed threshold γ_T [12–14]. Mathematically, the packet error probability of the ith hop $R_{i-1}R_i$, $1 \le i \le L$, denoted by P_i, is approximated by

$$P_i = P_{\gamma_i}(\gamma_T) = 1 - \exp\left(-\frac{\gamma_T}{G_i \cdot p_i}\right). \tag{2}$$

Note that the above equation associates the packet error rate for the ith hop with the transmitting power of its transmitting node R_{i-1}.

[1] Alternatively, we can set a limit for the total number of allowed retransmissions up to the current hop. In this case, the packet dropping decision will depend on the number of retransmissions performed in the previous hops, which will lead to a more complicated configuration algorithm and will be addressed in a different paper.

2.2. Hop-by-hop ARQ for delay sensitive traffics

We assume that the multihop path employs hop-by-hop ARQ mechanism. With hop-by-hop ARQ, the transmitting node of a certain hop waits for a positive acknowledgment before advancing to the transmission of the next data packet. If the positive acknowledgment is not received within a given threshold time, the transmitting node will retransmit the packet until the packet is positively acknowledged. Then the next node along the path will transmit the packet to the subsequent nodes in the same fashion. Traditionally, this process is continued until either the packet arrives at the destination correctly or the packet is dropped because the maximum number of allowed retransmissions is exceeded for that packet. This retransmission limit usually stems from the delay constraint of the data traffic, especially those generated by voice and/or video applications. In this paper, we propose to optimally select the retransmission limit as well as the transmitting power for each hop for energy saving purpose.

We consider the transmission of delay sensitive traffic over the multihop radio path. More specifically, the traffic has the QoS requirement that packets must be delivered to the destination node without error within T_D seconds with a required probability P_{req}. Note that we focus on the allowed transmission delay while assume queuing delay due to multiple flow has been taken into account during the routing process and subtracted from the total delay budget. The common round-trip time of each individual hop is assumed to be T_R and, as such, the total number of allowed transmissions/retransmissions[2] is $N = \lfloor T_D/T_R \rfloor$. The QoS requirement for the traffic can then be rephrased as follows: the packets must arrive at the destination correctly within N total transmission/retransmissions, or equivalently within $N - L$ retransmissions, with the probability of P_{req}. Finally, while there may be multiple packet traveling along the path at the same time, we ignore the interference between different packet transmissions. Note that if a node cannot transmit and receive at the same time, simultaneous transmission on adjacent hops will not occur.

3. ANALYSIS ON UNCONFIGURED BEST-EFFORT TRANSMISSION

In this section, we consider the best-effort transmission strategy for packet transmission over a multihop path. With best-effort transmission, every node along the path tries to deliver the packet to the next node without error by performing as many retransmissions as necessary with maximum transmitting power $p_{\max}$, that is, $p_i = p_{\max}$ for $i = 1, 2, \ldots, L$. A packet is dropped only if the maximum number of allowed retransmissions is exceeded. We derive closed-form expressions for the packet delivery ratio and average energy consumption for a single packet delivery with best-effort transmission.

Let x_i denote the number of transmissions and retransmissions that are actually performed over the ith hop. We note that with the best-effort strategy, a packet can arrive at the destination without error after $k = \sum_{i=1}^{L} x_i$ transmissions/retransmissions, where $L \le k \le N$. The probability of each realization of vector $\mathbf{x} = [x_1, x_2, \ldots, x_L]$, satisfying (i) $k = \sum_{i=1}^{L} x_i$, (ii) $1 \le x_i < k$, and (iii) $L \le k \le N$, can be calculated as

$$P_{\text{succ}}(\mathbf{x}) = \prod_{l=1}^{L} P_l^{x_l - 1}(1 - P_l), \tag{3}$$

where P_l is the packet error probability for the lth hop. Note that P_l was given in (2) with p_l now equal to $p_{\max}$ for all l. Summing up the probabilities for all possible vectors, we obtain the packet delivery ratio P_{succ} with best-effort transmission strategy as

$$P_{\text{succ}} = \sum_{k=L}^{N} \left[\sum_{\substack{\sum_{i=1}^{L} x_i = k \\ 1 \le x_i < k}} \left(\prod_{l=1}^{L} P_l^{x_l - 1}(1 - P_l) \right) \right]. \tag{4}$$

We now determine the average energy consumption for a single data packet delivery, regardless of whether the packet arrives at the destination node correctly within the delay constraint. Note that if a packet fails to arrive at the destination within the maximum number of retransmissions, it may be dropped on any one of the L hops. In this case, all $N - L$ allowed retransmissions must have been performed. Note that if the packet is dropped on the jth hop, then the vector $\mathbf{x}$ satisfies (i) $x_i = 0$ for $j < i \le L$; (ii) $\sum_{i=1}^{j} x_i = N - (L - j)$; and (iii) $1 \le x_i \le N - (L - j)$ for $1 \le i \le j$ and the probability for each such vector is equal to

$$P_{\text{drop}}^{(j)}(\mathbf{x}) = \left(\prod_{l=1}^{j-1} P_l^{x_l - 1}(1 - P_l) \right) P_j^{x_j}. \tag{5}$$

Therefore, the probability that the packet is dropped on the jth hop $P_{\text{drop}}^{(j)}$ is obtained as

$$P_{\text{drop}}^{(j)} = \sum_{\substack{\sum_{i=1}^{j} x_i = N - (L-j) \\ 1 \le x_i \le N - (L-j)}} \left(\prod_{l=1}^{j-1} P_l^{x_l - 1}(1 - P_l) \right) P_j^{x_j}. \tag{6}$$

For a particular realization of vector $[x_1, x_2, \ldots, x_L]$, the corresponding energy consumption $\mathcal{E}$ is equal to $T \cdot \sum_{i=1}^{L} x_i p_{\max}$, where T is the time duration required for transmitting a data packet. For simplicity, in the rest of the paper, we set $T = 1$ without loss of generality. Therefore, we obtain the following

[2] Since the receiving node may transmit to the next node once it correctly receives the packet, without waiting for the positive acknowledgment to reach the transmitting node, the actual value of N may be slightly greater than $\lfloor T_D/T_R \rfloor$. We ignore those extra transmissions for the sake of brevity here.

analytical expression for the average energy consumption per packet delivery with best-effort transmission strategy as

$$\mathbf{E}[\mathcal{E}] = \sum_{k=L}^{N} \left[\sum_{\substack{\sum_{i=1}^{L} x_i = k \\ 1 \le x_i < k}} \left(\sum_{i=1}^{L} x_i p_{\max} \right) \left(\prod_{l=1}^{L} P_l^{x_l - 1} (1 - P_l) \right) \right] + \sum_{j=1}^{L} \left[\sum_{\substack{\sum_{i=1}^{j} x_i = N-(L-j) \\ 1 \le x_i \le N-(L-j)}} \left(\sum_{i=1}^{j} x_i p_{\max} \right) \times \left(\prod_{l=1}^{j-1} P_l^{x_l - 1} (1 - P_l) \right) P_j^{x_j} \right], \tag{7}$$

where $\mathbf{E}[\cdot]$ denotes the statistical expectation.

4. MINIMUM-ENERGY CONFIGURATION

In this section, we consider the minimum-energy configuration of a multihop link for energy-efficient packet delivery. We assign a maximum retransmission limit to each individual hop, denoted by $\hat{x}_i$, in advance. As such, packet drop may occur in any hop when the retransmission limit for that hop is reached. We first derive closed-form expressions for the message delivery ratio and average energy consumption with an arbitrary transmitting power and retransmission limit configuration. Then, we formulate and solve an optimization problem to configure the path through jointly setting the transmitting power for each transmitting node and the number of allowed transmissions/retransmissions over each hop.

4.1. Packet delivery ratio and energy consumption analysis

Let $\hat{\mathbf{x}}$ and $\mathbf{p}$ denote the vector of the number of permitted transmissions/retransmissions and the transmitting power over the ith hop, respectively, that is, $\hat{\mathbf{x}} = [\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_L]$ and $\mathbf{p} = [p_1, p_2, \ldots, p_L]$. We first determine the probability of successful packet transmission over the multihop link for a particular choice of the vectors $\hat{\mathbf{x}}$ and $\mathbf{p}$. Note that node R_{i-1} will drop a packet if the data packet has been transmitted $\hat{x}_i$ times over the ith hop $R_{i-1}R_i$ without being correctly received by R_i. It can be shown that the packet delivery ratio $P_{\text{succ}}(\hat{\mathbf{x}}, \mathbf{p})$ is given by

$$P_{\text{succ}}(\hat{\mathbf{x}}, \mathbf{p}) = \prod_{i=1}^{L} \left(1 - P_i^{\hat{x}_i}\right), \tag{8}$$

where P_i is the packet error probability for the ith hop, which is given in (2) as a function of p_i.

We now calculate the average energy consumption for a single packet transmission regardless of whether it is successfully delivered to the destination within the delay constraint. For a particular realization of $\hat{\mathbf{x}}$ and $\mathbf{p}$, the average power consumption per packet delivery over the configured multihop link is given by

$$\mathbf{E}[\mathcal{E}(\hat{\mathbf{x}}, \mathbf{p})] = \sum_{i=1}^{L} \mathbf{E}[x_i] p_i, \tag{9}$$

where x_i denotes the actual number of transmissions and retransmissions performed over the ith hop, which becomes a discrete random variable (RV) taking integer values from 0 to $\hat{x}$. Note that the distribution of x_i depends on the values of $\hat{x}_j$ and p_j for $1 \le j \le i$. For the first hop, the source node R_0 would repeatedly transmit a data packet until either it is successfully received by R_1 or the number of maximum retransmissions for the first hop $\hat{x}_i$ is exceeded. Conditioning on the number of retransmissions used in a successful delivery and applying the total probability theorem, it can be shown that the probability that a data packet is correctly received by R_1 is $(1 - P_1) \cdot \sum_{k=1}^{\hat{x}_1} {P_1}^{k-1}$. Moreover, we can easily obtain the probability that a packet is dropped in the first hop is ${P_1}^{\hat{x}_1}$. Combining the two mutually exclusive cases, we can write $\mathbf{E}[x_1]$ as

$$\mathbf{E}[x_1] = (1 - P_1) \sum_{k=1}^{\hat{x}_1} {P_1}^{k-1} \cdot k + P_1^{\hat{x}_1} \cdot \hat{x}_1. \tag{10}$$

After similar algebraic manipulations as in [15, page 36], we have

$$\mathbf{E}[x_1] = \frac{1 - P_1^{\hat{x}_1}}{1 - P_1}. \tag{11}$$

For the second hop R_1R_2, x_2 may be either zero or a positive integer depending on whether the packet can reach R_1 or not. If the packet is successfully delivered to R_1, we can follow the similar approach for deriving (11) to calculate the average number of transmission/retransmissions performed by R_1. Therefore, noting that the probability that a data packet can reach R_1 correctly is equal to $1 - P_1^{\hat{x}_1}$, it can be shown that

$$\begin{aligned} \mathbf{E}[x_2] &= P_1^{\hat{x}_1} \times 0 + \left(1 - P_1^{\hat{x}_1}\right) \\ &\quad \times \left[(1 - P_2) \sum_{k=1}^{\hat{x}_2} {P_2}^{k-1} \cdot k + P_2^{\hat{x}_2} \cdot \hat{x}_2 \right] \\ &= \left(1 - P_1^{\hat{x}_1}\right) \cdot \frac{1 - P_2^{\hat{x}_2}}{1 - P_2}. \end{aligned} \tag{12}$$

With the above derivation in mind, we now develop a general expression for $\mathbf{E}[x_i]$, $i \ge 2$. Note that x_i is nonzero if and only if the packet is successfully delivered over the first $i - 1$ hops and finally received by R_{i-1}, the probability of which is given by $\prod_{j=1}^{i-1}(1 - P_j^{\hat{x}_j})$. Also note that the average number of transmissions/retransmissions conducted in the ith hop is $(1 - P_i^{\hat{x}_i})/(1 - P_i)$, after the packet arrives at R_{i-1} correctly. Therefore, we have

$$\mathbf{E}[x_i] = \prod_{j=1}^{i-1} \left(1 - P_j^{\hat{x}_j}\right) \cdot \frac{1 - P_i^{\hat{x}_i}}{1 - P_i}, \quad i \ge 2. \tag{13}$$

Combining (9), (10), and (13), we obtain a closed-form expression for the average energy consumption per packet delivery over a configured L-hop path as

$$\mathbf{E}[\mathcal{E}(\hat{\mathbf{x}},\mathbf{p})] = \frac{1-P_1^{\hat{x}_1}}{1-P_1}\cdot p_1 + \sum_{i=2}^{L}\prod_{j=1}^{i-1}\left(1-P_j^{\hat{x}_j}\right)\cdot\frac{1-P_i^{\hat{x}_i}}{1-P_i}\cdot p_i. \tag{14}$$

4.2. Minimum-energy optimization

Based on the closed-form expressions for the packet delivery ratio and average energy consumption of a multihop wireless path, we are now in a position to formulate an optimization problem for the multihop route configuration. In particular, we seek to select vectors $\hat{\mathbf{x}}$ and $\mathbf{p}$ so that the average energy consumption for packet delivery is minimized and the packet can arrive at the destination node within $N-L$ retransmissions with probability at least P_{req}. This leads to the following optimization problem:

$$\underset{\hat{\mathbf{x}},\mathbf{p}}{\text{minimize}}\ \mathbf{E}[\mathcal{E}(\hat{\mathbf{x}},\mathbf{p})] \tag{15a}$$

$$\text{subject to } P_{\text{succ}}(\hat{\mathbf{x}},\mathbf{p}) \geq P_{\text{req}}, \tag{15b}$$

$$p_{\max} > p_i > 0 \quad \text{for } 1 \leq i \leq L, \tag{15c}$$

$$\sum_{i=1}^{L}\hat{x}_i = N, \quad \hat{x}_i \in \{1,2,\ldots,N\}, \tag{15d}$$

where in this case the packet delivery ratio $P_{\text{succ}}(\hat{\mathbf{x}},\mathbf{p})$ becomes a function of both power configuration vector $\mathbf{p}$ and retransmission configuration vector $\hat{\mathbf{x}}$.

Note that in the optimization problem (15), $\hat{x}_i$ can only take integer values whereas p_i are continuous variables, and that both the objective function and the constraints given in (19b) are nonlinear functions of $\hat{x}_i$ and p_i. Therefore, the optimal configuration problem of a multihop link is actually a mixed integer nonlinear programming (MINP) problem [16]. In general, optimization problems of this kind are NP-hard and few algorithms guarantee to find the global minimum. However, in practical systems, the number of hops in a multihop wireless link is usually small. In this case, (15) can be efficiently solved by using small scale MINP algorithms such as the branch-and-bound algorithm [17]. Because of space limitation, we omit the details of applying the branch-and-bound algorithm to solve the optimization problem. Obviously, the calculation of the solution to the optimization problem will incur additional energy consumption to the destination node. However, as we will observe in the later numerical examples, the average energy saving for packet transmissions with route configuration based on the possibly local-minimum solution is significant compared to the unconfigured best-effort approach, which justifies the energy cost spent in solving the optimization problem.

5. MAXIMUM LIFETIME CONFIGURATION

While the minimum-energy configuration in the previous section can reduce the average energy cost per packet transmission, it does not take into account the available energy resources of intermediate nodes along the path. Consider, as an example, a transmitting node with low battery supply and sending data packets over an unfavorable radio hop. With the minimum-energy configuration, this node will be configured with high transmitting power and a large number of retransmissions. This configuration will quickly deplete the battery resource of this node and leave the whole path unusable, which will not only cause the interruption of the data transmission but also lead to extra route recovery operations. In this section, we develop a maximum-lifetime configuration algorithm for multihop paths in wireless ad hoc networks. In particular, we take into consideration both the link quality of each hop and the battery capacity of the transmitting nodes in determining the transmitting power for each transmitting node and the maximum number of allowed retransmissions for each hop to extend the lifetime of the multihop, in terms of the average number of packets that can be transmitted.

Note that for a particular pair of path configuration vectors $\hat{\mathbf{x}} = [\hat{x}_1,\hat{x}_2,\ldots,\hat{x}_L]$ and $\mathbf{p} = [p_1,p_2,\ldots,p_L]$, the average packet delivery ratio of that multihop path is given in (8). We assume that the multihop path reaches its lifetime when the battery supply of any intermediate node becomes too little to support a single packet transmission. Let B_i, $1 \leq i \leq L$, denote the remaining battery resource of the ith node for packet transmission. The average path lifetime is defined as

$$T(\hat{\mathbf{x}},\mathbf{p}) = \min_i\left\{\frac{B_i}{\mathbf{E}[x_i]p_i}\right\}, \tag{16}$$

where $\mathbf{E}[x_i]$ is the average energy consumption per packet delivery over the ith hop, which is given in (13).

Based on the closed-form expressions for the packet delivery ratio and the average path lifetime, we can formulate another optimization problem to configure the multihop path. In particular, we seek to select vectors $\hat{\mathbf{x}}$ and $\mathbf{p}$ so that the average path lifetime is maximized under the constraint that the packet can arrive at the destination node within $N-L$ retransmissions with probability at least P_{req}. This leads to the following optimization problem:

$$\underset{\hat{\mathbf{x}},\mathbf{p}}{\text{maximize}}\ T(\hat{\mathbf{x}},\mathbf{p}) \tag{17a}$$

$$\text{subject to } P_{\text{succ}}(\hat{\mathbf{x}},\mathbf{p}) \geq P_{\text{req}}, \tag{17b}$$

$$p_{\max} > p_i > 0 \quad \text{for } 1 \leq i \leq L, \tag{17c}$$

$$\sum_{i=1}^{L}\hat{x}_i = N, \quad \hat{x}_i \in \{1,2,\ldots,N\}. \tag{17d}$$

From (16) and (17), we see that (17) is a constrained optimization problem with a minimax-type objective function to which few optimization algorithms are directly applicable. To deal with this problem, let δ be a lower bound of $B_i/\mathbf{E}[x_i]p_i$, $i = 1,\ldots,L$, for vectors $\hat{\mathbf{x}}$ and $\mathbf{p}$ satisfying constraints in (17), that is,

$$\frac{B_i}{\mathbf{E}[x_i]p_i} \geq \delta \quad \text{for } 1 \leq i \leq L. \tag{18}$$

It follows from (16) and (18) that $T(\hat{\mathbf{x}},\mathbf{p}) \geq \delta$. Hence, maximizing $T(\hat{\mathbf{x}},\mathbf{p})$ subject to the constraints in (17) amounts to

Table 1: Results of minimum-energy path configuration.

$\mathbf{c}$	$\mathbf{p}$	$\hat{\mathbf{x}}$
[0.158, 0.06, 0.158]	[0.338, 0.214, 0.354]	[4, 3, 4]
[0.158, 0.06, 0.05]	[0.319, 0.154, 0.175]	[4, 4, 3]
[0.05, 0.06, 0.158]	[0.166, 0.152, 0.326]	[3, 4, 4]

maximizing the lower bound δ subject to the constraints in (17) and (18). In this way, the optimization problem in (17) is reformulated as

$$\underset{\hat{\mathbf{x}},\mathbf{p},\delta}{\text{maximize}}\ \delta \tag{19a}$$

$$\text{subject to}\ \frac{B_i}{\mathbf{E}[x_i]p_i} \geq \delta, \quad 1 \leq i \leq L, \tag{19b}$$

$$P_{\text{succ}}(\hat{\mathbf{x}}, \mathbf{p}) \geq P_{\text{req}}, \tag{19c}$$

$$p_{\max} > p_i > 0, \quad 1 \leq i \leq L, \tag{19d}$$

$$\sum_{i=1}^{L} \hat{x}_i = N, \quad \hat{x}_i \in \{1, 2, \ldots, N\}, \tag{19e}$$

where lower bound δ is treated as an additional variable. Note that the maximum-lifetime configuration of a multihop path is again an MINP problem [16]. Since the number of hops in a multihop wireless link is usually not large, (19) can also be efficiently solved by using small scale MINP algorithms such as the branch-and-bound algorithm [17]. As we will see in the next section, even with possibly local-minimum solution, the maximum-life configuration can extend path lifetime and save considerable energy, compared to the unconfigured best-effort case.

6. NUMERICAL EXAMPLES AND DISCUSSION

In this section, we illustrate the effectiveness of the path configuration algorithms over multihop paths through numerical examples. In particular, we consider a 3-hop path, that is, $L = 3$, while noting that most of observations hold for paths with a larger number of hops. To simplify the following presentation, we define a channel coefficient vector $\mathbf{c}$, whose ith entry c_i is given by

$$c_i = \frac{\gamma_T}{G_i}, \quad i = 1, 2, 3. \tag{20}$$

It can be seen that the channel condition of the ith hop becomes worse as the corresponding coefficient c_i increases. The QoS requirement of the traffic is assumed to be that packets should reach its destination after $N = 11$ transmission/retransmission with probability of at least $P_{\text{req}} = 0.95$. The maximum transmitting power $p_{\max}$ is set to 0.56 W. For the maximum-lifetime configuration, we assume that the battery capacities of the three transmitting nodes are set as $B_1 = 600$ J, $B_2 = 500$ J, and $B_3 = 400$ J, respectively.

6.1. Minimum-energy configuration

In Table 1, we present the solutions of the minimum-energy configuration problem given in (15) for three different choices of channel coefficient vector $\mathbf{c}$. As we can see, the path configuration algorithm selects a higher transmission power and allocates a larger number of retransmissions to a hop experiencing poor channel condition, as one can expect by intuition. We also notice that this bias in route configuration towards poorer hops is not inversely proportional to the channel quality. In particular, we note that $p_i/p_j < c_i/c_j$ for $c_i > c_j$. Finally, we observe from the first choice of vector $\mathbf{c}$ that although the first and the last hops experience the same poor channel condition, the configuration algorithm allocates more power to the last hop and the same retransmission limit for both hops. This is because once a packet arrives at the last hop, less energy will be wasted if the packet is successfully transmitted to the destination than if the packet is lost eventually.

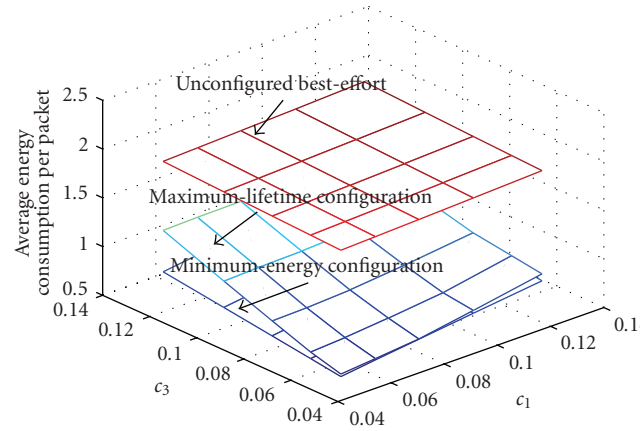


Figure 1: Average power consumption per packet with the maximum-lifetime configuration, minimum-energy configuration, and unconfigured best-effort strategies ($P_{\text{req}} = 0.95$, $N = 11$, $p_{\max} = 0.56$ W).

The energy saving offered by the minimum-energy configuration algorithm is illustrated in Figure 1. In generating the numerical results, we fix the channel coefficient of the second hop c_2 to be 0.06 while varying c_1 and c_3 from 0.05 to 0.158.[3] We first compare the average energy consumption for a single packet delivery in the 3-hop wireless link with minimum-energy configuration and unconfigured best-effort case (i.e., each node always uses the maximum transmitting power $p_{\max}$ for each transmission/retransmission). It can be observed that route configuration can save considerable amount of energy compared to the traditional best-effort strategy. For example, when c_1 and c_3 are equal to 0.0998 and 0.0792, respectively, the average power consumption required for a packet delivery with minimum-energy configuration is only 36.38% of that of the unconfigured best-effort case. It can also be seen that both strategies consume less energy on average as the channel

[3] This range for the channel coefficients and the choice of 0.7 W for $p_{\max}$ guarantee that even the worst hop, that is, the hop with channel coefficient 0.158, has a packet loss rate less than 20%.

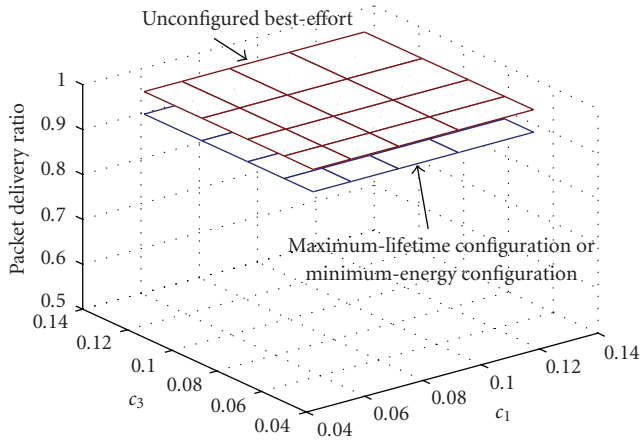


FIGURE 2: Packet delivery ratio with the maximum-lifetime configuration, minimum-energy configuration, and unconfigured best-effort strategies ($P_{\text{req}} = 0.95$, $N = 11$, $p_{\max} = 0.56$ W).

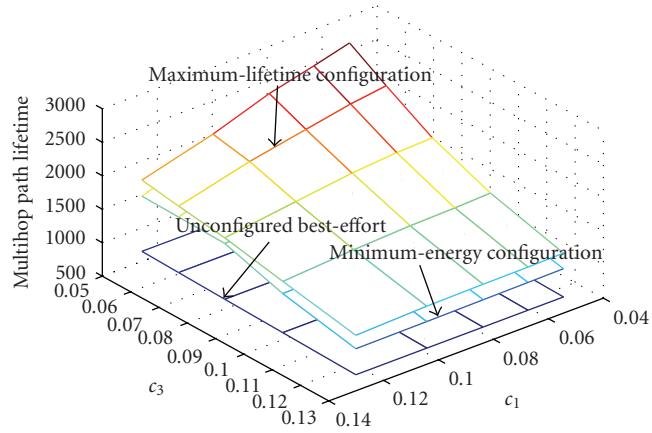


FIGURE 3: Path lifetime with the maximum-lifetime configuration, minimum-energy configuration, and unconfigured best-effort strategies for a 3-hop path ($P_{\text{req}} = 0.95$, $N = 11$, $p_{\max} = 0.56$ W).

TABLE 2: Results of maximum-lifetime path configuration.

c	p	$\hat{\mathbf{x}}$
[0.05, 0.05, 0.126]	[0.443, 0.367, 0.143]	[2, 2, 7]
Minimum-energy case	[0.159, 0.128, 0.247]	[3, 4, 4]
[0.126, 0.05, 0.05]	[0.180, 0.248, 0.188]	[5, 3, 3]
Minimum-energy case	[0.254, 0.165, 0.132]	[4, 3, 4]

coefficients c_1 and c_3 decrease. That is because smaller values of the channel coefficients represent better channel conditions.

Figure 2 plots the packet delivery ratio with the minimum-energy configuration and unconfigured best-effort case as the functions of c_1 and c_3. It can be observed that for all value pairs of c_1 and c_3, the minimum-energy configured path can always provide a packet delivery ratio greater or equal to P_{req}, which satisfies the QoS requirement. Note also that the application of traditional best-effort strategy leads to a slightly higher packet delivery ratio compared to the case with minimum-energy configuration, as expected. However, considering Figures 1 and 2 together, we can observe that the minimum-energy path configuration achieves the appealing property of maintaining acceptable path reliability while significantly reducing the average energy consumption.

6.2. *Maximum lifetime configuration*

In Table 2, we present the solutions of the maximum-lifetime path configuration problem given in (19) for two different choices of channel coefficient vector. For comparison, we also present the results of the same path with minimum-energy configuration in the italic format. As we can see, the maximum-lifetime configuration algorithm selects a smaller transmitting power and allocates a larger number of retransmissions to transmitting node with less battery resource, compared with the minimum-energy configuration. As such, the average energy consumption over the corresponding hop decreases to achieve a longer path lifetime. For example, for the first choice of channel coefficient vector, the average energy consumption of the third hop with maximum-lifetime configuration is 0.327, while a value of 0.41 is observed with minimum-energy configuration.

We now compare the maximum-lifetime configuration with the minimum-energy configuration and the unconfigured best-effort case. In Figure 3, we plot the path lifetime with the three strategies. It can be seen that with maximum-lifetime configuration, the multihop route achieves the largest lifetime. For example, when $c_1 = 0.063$ and $c_3 = 0.0998$, there is a 23.1% and a 150% increase in the path lifetime with the maximum-lifetime configuration, compared with minimum-energy configuration and traditional best-effort case, respectively. It can also be seen that as c_1 and c_3 decrease, that is, the channel conditions improve, all three schemes would lead to an increased path lifetime, as expected.

For comparison purpose, we have also plotted the average energy consumption with maximum-lifetime configuration in Figure 1 and the corresponding packet delivery ratio in Figure 2. As we can see, the maximum-lifetime configuration can also save considerable amount of energy per packet delivery compared to the unconfigured best-effort case. We also notice that the maximum-lifetime configuration will lead to a slightly larger average power consumption than minimum-energy configuration for the same selection of channel coefficients. From Figure 2, we observe that, similar to the minimum-energy configuration, the maximum-lifetime configuration can always provide a packet delivery ratio greater or equal to 0.95 for all value pairs of c_1 and c_3. Considering Figures 1, 2, and 3 together, we can observe that the maximum-lifetime configuration achieves the property of maintaining acceptable path reliability and considerably low energy consumption while significantly improving the lifetime of an existing path.

7. IMPLEMENTATION CONSIDERATION

The route configuration algorithm presented in this paper targets at the efficient usage of existing paths obtained by routing protocols. As such, our route configuration algorithm could become an optional but a desired feature of any existing routing protocol of wireless ad hoc network for improved energy efficiency. In this section, we adopt a commonly used routing protocol, dynamic source routing (DSR) [10], as an example to illustrate how our route configuration algorithm can be utilized. The principle discussed in this section is applicable to most routing protocols.

The DSR protocol [10] uses the source routing approach (i.e., every data packet carries the whole path information in its header) to forward packets. When a source node wants to send messages to a destination node but does not know a path to the destination, the source node initiates the route discovery process by broadcasting a Route REQuest (RREQ) message. Each node, once receiving the RREQ message, puts its node ID in the RREQ message and rebroadcasts the message. When the RREQ message reaches the destination node, the destination node replies with a Route REPly (RREP) message to the source node, using the reversed path that the RREQ message just traversed. The source node can obtain the complete path information after it receives the RREP message.

Our path configuration algorithms could be applied to the route discovery process of DSR as follows. From implementation point of view, decoupling the routing protocol and energy saving means that the routing algorithm is irrelevant to energy metrics. Nevertheless, data structure of control messages may need to change to facilitate path configuration. Each RREQ message should piggyback parameters of the link status of intermediate hops, that is, the parameters G_i in Section 2, and the remaining battery resource, that is, B_i in Section 5. After receiving RREQ, the destination node uses the collected information to solve the optimization problem and determine the retransmission limit and the power level for each intermediate hop. Then, it piggybacks the configuration information in the RREP packet. Each intermediate node, once receiving the RREP packet, will configure itself accordingly.

Since our path configuration algorithms are essentially independent of routing protocols or transport layer protocols, our path configuration algorithms could also be used for an end-to-end data flow. For instance, when a source node wants to establish a TCP connection with a destination node, the TCP SYN message can piggyback the link status information to the destination node. The destination node, after it calculates the path configuration with our algorithms, can use the TCP ACK message to notify each intermediate node the path configuration instruction. Similar operations could be performed even during an on-going data flow, by piggybacking the control information in the end-to-end data and acknowledgment messages.

We stress that route configuration is not performed on a packet-by-packet basis. When the network topology and network link quality are stable, the frequency of path reconfiguration could be quite small. In mobile ad hoc networks, however, route reconfiguration should be performed when the network topology changes and new paths are searched for. Although the destination node may consume extra energy in calculating optimal path configuration, such cost is not prohibitive as long as the number of nodes remains small. On the other hand, the path configuration process has substantial benefits due to the fact that a properly configured path will last longer and also the fact that energy cost on calculation is negligible compared to that on message transmission. For instance, the energy cost for transmitting 1 bit could be equivalent to the energy cost of executing up to 800 instructions [18].

8. CONCLUSION

In this paper, we have proposed the minimum-energy and maximum lifetime configuration algorithms to configure an existing multihop path with ARQ mechanism under a given QoS requirement and delay constraint. Our algorithms could work as an add-on function with most existing routing and transport protocols. Numerical results clearly illustrate the benefit of the proposed methods. We observed that the new algorithms can prolong the lifetime of the multihop path while maintaining an acceptable packet delivery ratio and considerably low overall average energy consumption. We have also investigated the tradeoff of path lifetime versus average energy consumption between the minimum-energy and maximum-lifetime configuration schemes.

ACKNOWLEDGMENTS

The authors would like to thank Mr. Le Yang for his help with the preparation of the figures and for some insightful discussions regarding these figures. This work was supported in part by startup funds from the University of Victoria and in part by Discovery Grant from NSERC. This is an expanded version of work which was presented in part at the 25th IEEE International Performance Computing and Communications Conference (IPCCC '06), Phoenix, Arizona, USA, April 2006.

REFERENCES

[1] N. Vaidya, "Open problems in mobile ad hoc networking," in *Proceedings of the 26th Annual IEEE Conference on Local Computer Networks (LCN '01)*, p. 516, Tampa, Fla, USA, November 2001.

[2] A. Bicket, D. Aguayo, S. Biswas, and R. Morris, "Architecture and evaluation of an unplanned 802.11b mesh network," in *Proceedings of the 11th Annual International Conference on Mobile Computing and Networking (MOBICOM '05)*, pp. 31–42, Cologne, Germany, August-September 2005.

[3] R. Draves, J. Padhye, and B. Zill, "Comparison of routing metrics for static multi-hop wireless networks," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM '04)*, pp. 133–144, Portland, Ore, USA, August 2004.

[4] J. Gomez and A. Campbell, "Power-aware routing optimization for wireless ad hoc networks," in *Proceedings of High Speed*

Networks Workshop (HSN '01), Balatonfured, Hungary, June 2001.

[5] S. Doshi, S. Bhandare, and T. X. Brown, "An on-demand minimum energy routing protocol for a wireless ad hoc network," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 6, no. 3, pp. 50–66, 2002.

[6] M. Agarwal, J. H. Cho, L. Gao, and J. Wu, "Energy efficient broadcast in wireless ad hoc networks with Hitch-hiking," in *Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '04)*, vol. 3, pp. 2096–2107, Hong Kong, March 2004.

[7] J.-H. Chang and L. Tassiulas, "Energy conserving routing in wireless ad-hoc networks," in *Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '00)*, vol. 1, pp. 22–31, Tel Aviv, Israel, March 2000.

[8] C. E. Jones, K. M. Sivalingam, P. Agrawal, and J. C. Chen, "A survey of energy efficient network protocols for wireless networks," *Wireless Networks*, vol. 7, no. 4, pp. 343–358, 2001.

[9] G. Zussman and A. Segall, "Energy efficient routing in ad hoc disaster recovery networks," in *Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '03)*, vol. 1, pp. 682–691, San Francisco, Calif, USA, March-April 2003.

[10] D. A. Maltz, J. Broch, J. Jetcheva, and D. B. Johnson, "The effects of on-demand behavior in routing protocols for multihop wireless ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 8, pp. 1439–1453, 1999.

[11] C. E. Perkins and P. Bhagwat, "Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers," in *Proceedings of the Conference on Communications Architectures, Protocols and Applications (SIGCOMM '94)*, pp. 234–244, London, UK, August 1994.

[12] M. Zorzi and S. Pupolin, "Optimum transmission ranges in multihop packet radio networks in the presence of fading," *IEEE Transactions on Communications*, vol. 43, no. 7, pp. 2201–2205, 1995.

[13] M. Zorzi, R. R. Rao, and L. B. Milstein, "Error statistics in data transmission over fading channels," *IEEE Transactions on Communications*, vol. 46, no. 11, pp. 1468–1477, 1998.

[14] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, pp. 1746–1755, 2004.

[15] D. Zwillinger, *Standard Mathematical Tables and Formulae*, CRC Press, Boca Raton, Fla, USA, 30th edition, 1996.

[16] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*, John Wiley & Sons, New York, NY, USA, 1999.

[17] S. Leyffer, "Integrating SQP and branch-and-bound for mixed integer nonlinear programming," *Computational Optimization and Applications*, vol. 18, no. 3, pp. 295–309, 2001.

[18] K. Sohrabi, J. Gao, V. Ailawadhi, and G. J. Pottie, "Protocols for self-organization of a wireless sensor network," *IEEE Personal Communications*, vol. 7, no. 5, pp. 16–27, 2000.

Hindawi Publishing Corporation
Advances in Multimedia
Volume 2007, Article ID 94918, 11 pages
doi:10.1155/2007/94918

Research Article

MOS-Based Multiuser Multiapplication Cross-Layer Optimization for Mobile Multimedia Communication

Shoaib Khan,[1] Svetoslav Duhovnikov,[1] Eckehard Steinbach,[1] and Wolfgang Kellerer[2]

[1] *Media Technology Group, Technische Universität München, 80687 München, Munich, Germany*
[2] *DoCoMo Communications Laboratories Europe GmbH, Future Networking Lab, 80687 2 Munich, Germany*

Received 7 January 2007; Revised 30 April 2007; Accepted 5 June 2007

Recommended by Jianwei Huang

We propose a cross-layer optimization strategy that jointly optimizes the application layer, the data-link layer, and the physical layer of a wireless protocol stack using an application-oriented objective function. The cross-layer optimization framework provides efficient allocation of wireless network resources across multiple types of applications run by different users to maximize network resource usage and user perceived quality of service. We define a novel optimization scheme based on the mean opinion score (MOS) as the unifying metric over different application classes. Our experiments, applied to scenarios where users simultaneously run three types of applications, namely voice communication, streaming video and file download, confirm that MOS-based optimization leads to significant improvement in terms of user perceived quality when compared to conventional throughput-based optimization.

1. INTRODUCTION

In order to achieve efficient resource usage in a wireless network and to provide high quality of services to the largest possible number of users, it is necessary to obtain an optimal configuration of the wireless transmission system. Dynamic changes of transmission conditions and concurrently running applications by different users make dynamic optimization of resources a complex task. In realistic scenarios, multiple users share the wireless medium and run rather diverse applications such as video streaming/conferencing, voice telephony, and file download. Dynamic allocation of resources across all the users and all the applications provides an opportunity to achieve increasing network resource usage and to maximize the user satisfaction at the same time.

Application-driven cross-layer optimization (CLO) has been studied for systems supporting single applications [1–4]. However, in reality, the users sharing the wireless medium, for example, in a cell, usually run different applications. User satisfaction translates into a different set of requirements for each type of application. Furthermore, the impact of losses on the user-perceived quality is also very much application-dependent. Jointly optimizing the system for different users and applications requires (1) defining a common metric that quantifies the user perceived quality of service for the service delivery and (2) mapping network and application parameters onto this metric.

The challenge of optimization across multiple applications has been treated mainly in the form of throughput maximization [5, 6]. Maximizing throughput leads to optimum performance only for applications which are insensitive to delay and packet loss. Multimedia applications such as video streaming and voice telephony are highly sensitive to changes in data rate, delay, and packet losses. Even the importance of a packet changes dynamically depending on the transmission history of previous packets. Due to these reasons, throughput maximization leads to performance which is usually not optimal with respect to user perceived quality for multimedia applications.

A possible metric to capture user satisfaction is the mean opinion score (MOS). MOS was originally proposed for voice quality assessment and provides a numerical measure of the quality of human speech at the destination. The scheme uses subjective tests (opinionated scores) that are mathematically averaged to obtain a quantitative indicator of the system performance. To determine MOS, a number of listeners rate the quality of test sentences read aloud over the communication circuit by a speaker. A listener gives each sentence a rating as follows: (1) bad; (2) poor; (3) fair; (4) good; (5) excellent. The MOS is the arithmetic mean of all the individual scores.

The multiapplication CLO approach proposed in this paper extends the use of MOS as a user-perceived quality metric to other applications, such as video streaming, web browsing and file download. This enables us to optimize across applications using a common optimization metric. The objective function to be maximized can be chosen, for example, to be the average MOS of all the users competing for the resources of the wireless communication system:

$$F(\tilde{\mathbf{x}}) = \frac{1}{K}\sum_{k=1}^{K} \lambda_k \cdot \mathrm{MOS}_k(\tilde{\mathbf{x}}), \tag{1}$$

where $F(\tilde{\mathbf{x}})$ is the objective function with the cross-layer parameter tuple. $\tilde{\mathbf{x}} \in \tilde{X} \cdot \tilde{X}$ is the set of all possible parameter tuples abstracted from the protocol layers representing a set of candidate operation modes. λ_k are free parameters which can be chosen in two different ways. For a priority-based scheme, they can be chosen to provide different relative importance of the user as determined by the service agreement between the user and the service provider. For an equal-priority system, λ_k can be chosen to ensure fairness among the users. In this paper, we take the second approach. Although the MOS functions for different applications can be different, a linear combination, as in (1), can be used because the range of the functions is the same, that is, from 1 to 5. The decision of the optimizer can be expressed as

$$\tilde{\mathbf{x}}_{\mathrm{opt}} = \arg\max_{\tilde{\mathbf{x}}\in\tilde{X}} F(\tilde{\mathbf{x}}), \tag{2}$$

where $\tilde{\mathbf{x}}_{\mathrm{opt}}$ is the parameter tuple which maximizes the objective function. Once the optimizer has selected the optimal values of the parameters, it distributes them to all the individual layers which are responsible for translating them back into actual layer-specific modes of operation.

In this work, the abstracted parameters for the physical and data link layers are transmission rate R and packet error probability (PEP) for all users for all candidate modes of operation. For a detailed description of the principle of parameter abstraction and the formulation of objective functions for multiuser cross-layer optimization, please refer to [1–3, 7].

The proposed MOS-based optimization approach has several advantages with respect to previous work. First, compared to traditional techniques for multiuser diversity [8], it allows us to directly relate network parameters, such as rate (R) and packet error probability (PEP) to a user-perceived application quality metric such as MOS. Second, compared to the application-driven cross-layer optimization described in [2, 3], it allows us to further maximize the optimization gain by taking advantage of the diversity not only across multiple users running the same application, but also across users running different applications. Our experiments applied to scenarios including multiple concurrent video streaming, voice telephony, and file download applications show that MOS-based optimization significantly outperforms throughput-based optimization.

This paper is arranged as follows. In Section 2, we describe MOS functions for three different applications, namely voice telephony, file download, and video streaming.

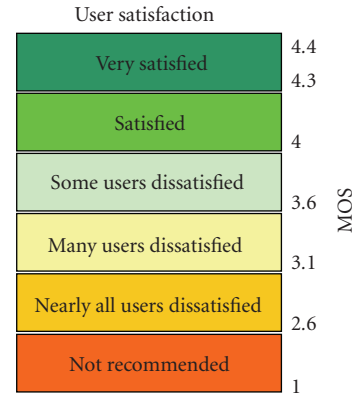


Figure 1: Relation between MOS and user satisfaction [9].

In Section 3, we give a detailed description of our multiapplication cross-layer optimization framework. Section 4 gives an overview of our simulation setup that is used to compare our approach with throughput maximization. Section 5 presents our experimental results and Section 6 concludes the paper.

2. MEAN OPINION SCORE (MOS)

The objective function of (1) requires the mapping of transmission characteristics (in our case transmission rate and packet error probability) to MOS for different applications. We now describe this mapping for voice communication, file download, and video streaming applications.

2.1. Voice communication

The traditional method of determining voice quality is to conduct subjective tests with panels of human listeners. The results of these tests are averaged to give MOS but such tests are expensive and are not feasible for online voice quality assessment. For this reason, the ITU-T has standardized a model, perceptual evaluation of speech quality (PESQ) [10], an algorithm that predicts with high correlation the quality scores that would be given in a typical subjective test. This is done by making an intrusive test and processing the test signals through PESQ.

PESQ measures one-way voice quality: a signal is injected into the system under test and the degraded output is compared by PESQ with the input (reference) signal. The output of the PESQ algorithm is a numerical value that corresponds to MOS. The mapping between MOS and user satisfaction is presented in Figure 1.

The PESQ algorithm is computationally too expensive to be used in real-time scenarios. To solve this problem, we propose a model to estimate MOS as a function of the transmission rate R and the packet error probability (PEP). The available rate determines the voice codec that can be used. In Figure 2 we show experimental curves for MOS estimation as a function of PEP for different voice codecs. The curves are drawn using an average over a large number of voice samples and channel realizations (packet loss patterns).

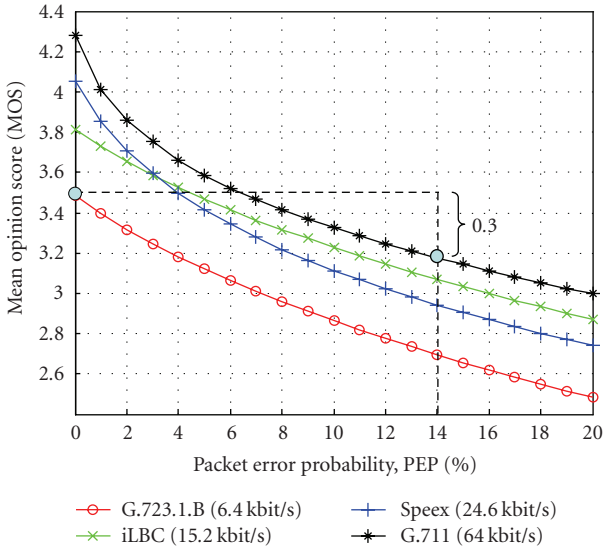


Figure 2: PESQ-based MOS versus packet error probabilities for different voice codecs.

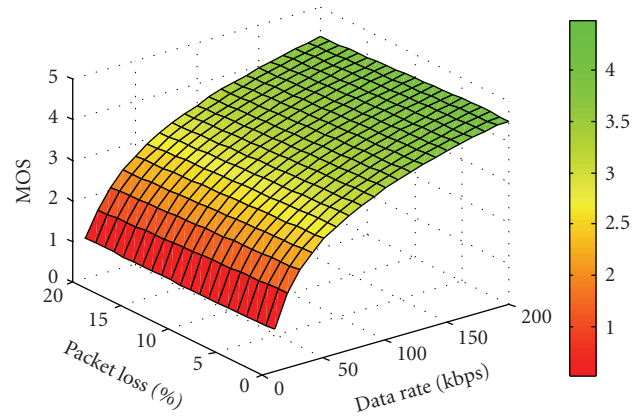


Figure 3: MOS as a function of transmission rate and packet error probability for file download applications.

These curves can be stored at the optimizer for every codec that is supported. If transcoding from an unsupported codec is required, the corresponding curve has to be signaled to the optimizer as side information. It should be noted that the average MOS (averaged over a large number of packet loss patterns for a fixed PEP) of individual voice samples may differ as much as 10% for the highest considered PEP, but the deviation from the mean values (averaged over a large number of voice samples) as shown in Figure 2 is found to be less than 7%.

Depending on the distortion imposed by the source codec, every voice codec leads to a different MOS value in the case of error-free transmission. Also the codecs exhibit different sensitivities to packet losses. As an example, let us consider two lower layer parameter tuples (R = 64 kbps, PEP = 14%), and (R = 6.4 kbps, PEP = 0%) and assume these two represent possible operating modes of the lower layers for a particular user. In this example, the second parameter tuple (R = 6.4 kbps, PEP = 0%) leads to a gain of 0.3 on the MOS scale and the cross-layer optimizer would select it as its outcome.

2.2. File download

To estimate user satisfaction for file download applications, we use the logarithmic MOS-throughput relationship introduced in [11] which results from the assumption that the utility of an elastic traffic (e.g., FTP service) is an increasing, strictly concave, and continuously differentiable function of throughput. We assume that every user has subscribed for a given data rate and user satisfaction is characterized by the actual rate the user receives. The MOS is estimated based on the current rate R offered to the user by the system and packet error probability PEP:

$$\text{MOS} = a * \log_{10}[b * R * (1 - \text{PEP})], \tag{3}$$

where a and b are determined from the maximum and minimum user perceived quality. If a user has subscribed for a specific rate R_{service} and receives $R = R_{\text{service}}$, then in case of no packet loss user satisfaction on the MOS scale should be maximum, that is, 4.5. On the other hand, we define a minimum transmission rate (e.g., 10 kbps in Figure 3) and assign to it a MOS value of 1. Using the parameters a and b, we fit the logarithmic curve in (3) for the estimated MOS. Varying the actual transmission rate R and packet error probability, PEP, this model results in the MOS surface shown in Figure 3.

2.3. Streaming video

Assessment of video quality is addressed in the literature with a wide variety of techniques. References [12, 13] are ITU recommendations to perform subjective assessment of TV and multimedia quality, respectively. Reference [14] gives a perceptual quality metric with respect to blockiness in compressed video. In [15], authors propose a reference-free method to estimate subjective quality using blurriness of the reconstructed video. Assuming that human visual perception is highly adapted for extracting structural information from a scene, [16] proposes a method of image quality assessment using degradation of structural information and develops a structural similarity index (SSIM). Reference [17] gives a comparison of different computational models of video quality, carried out by the video quality experts group (VQEG) of ITU.

Peak signal-to-noise ratio (PSNR) is an objective measurement of video quality which is widely used due to its simplicity and high degree of correlation with subjective quality [17]. PSNR is based on mean square error (MSE) as follows:

$$\text{PSNR} = 10 * \log_{10} \frac{255^2}{\text{MSE}}. \tag{4}$$

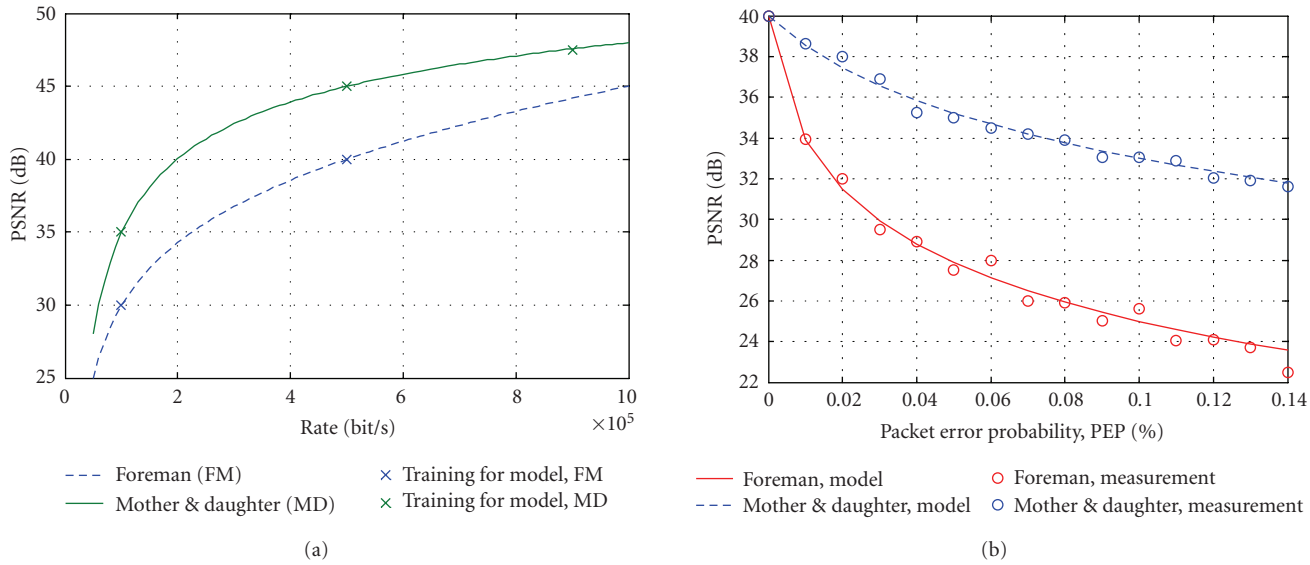


Figure 4: Illustration of the source distortion model (left), and the loss distortion model (right) using two test video sequences "Foreman" and "Mother and Daughter."

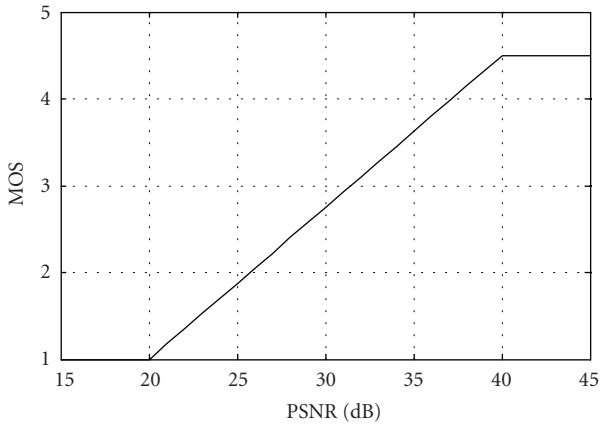


Figure 5: MOS versus PSNR.

The distortion of a video sequence can be expressed in terms of MSE. Distortion is assumed to be composed of two components, namely the *source* distortion D_S and the *loss* distortion D_L:

$$D = D_S + D_L, \tag{5}$$

D_S is due to the compression of the video sequence, while D_L is due to the losses generated in the network. Consequently, D_S depends on the video source rate, R and D_L is a function of packet error probability PEP. In this paper, we apply the source distortion model as proposed in [18], and the loss distortion is assumed to be a linear function of PEP [19],

$$D = D_S + D_L = \frac{a}{\exp(R/b) - 1} + \beta \cdot \text{PEP}, \tag{6}$$

where a, b, and β are model parameters. The source distortion model requires three pairs of rate and distortion measurements, as illustrated in Figure 4(a) for two test video sequences. The loss distortion model requires measuring distortion for different PEP and uses best-fit to compute $\beta \cdot \beta$ is assumed to be independent of the video encoding rate. The validation of the loss distortion model is shown in Figure 4(b). Encoding is done with the H.264 reference encoder, with 30 frames per second in QCIF format. Each packet is assumed to have a fixed size of 125 bytes. Each video frame is encapsulated into one or more such packets.

In this work, we assume a simple linear mapping between PSNR and MOS. We assume that the maximum user satisfaction is achieved for a PSNR of 40 dB and the minimum user satisfaction results for PSNR values below 20 dB. The upper limit comes from the fact that reconstructed video sequences with 40 dB PSNR are almost indistinguishable from the original and below 20 dB very severe degradations distort the video. Figure 5 shows our assumed relationship between PSNR and MOS.

3. MULTIAPPLICATION CROSS-LAYER OPTIMIZATION

Based on the MOS framework described above, we are able to optimize the system taking actual user perceived quality of service into account. Our optimization scheme is not only applicable to the application types described in Section 2, but to any general mix of applications.

3.1. Architecture

In [1, 2], we have proposed a cross-layer optimization architecture (Figure 6) with a component, called cross-layer optimizer (CLO), that periodically selects the optimal parameter settings of the different layers. This architecture is inspired by the CLO approach presented in [7]. Our CLO uses abstractions of different layers and optimizes the assignment of resources to each user. In our work, the abstracted parameters from the lower layers are rate R and packet er-

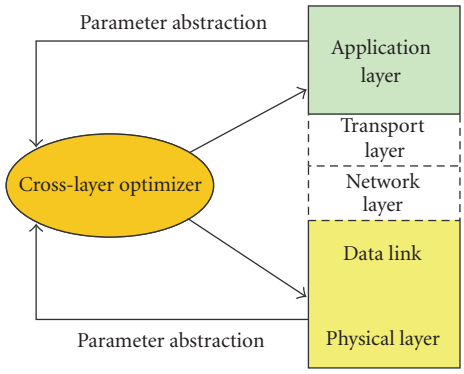


Figure 6: CLO architecture.

ror probability PEP for every user for all possible modes of operation. From the application layer we extract the utility functions (MOS versus PEP). We assume that the optimizer is located at the base station, the utility functions are generated at the sender and are sent as side information along with the media bitstream.

3.2. Optimization policy

As an example, we consider three types of users: U—requesting voice service, V—file download, and W—video streaming. Depending on the type of application, the mobile users require different resources over the wireless channel. The available transmission rate for each user depends on the modulation scheme, the channel code rate, and the assigned share of the medium access. In our example, a user requesting voice service may be served with different voice codecs (G.711, Speex, iLBC, or G.723.1.B), his data may be encoded with different channel code rates 1/2, 1/3, 1/4, or 1 (uncoded) and DBPSK or DQPSK modulation can be used. Every transmission policy gives different quality of service to the user and requires different amount of channel resources.

We create sets of transmission policies for every service. T_U is the set of transmission policies for voice service, T_V is the set of transmission policies for the file download service, and T_W is the set of transmission policies for the video service.

3.3. Mean opinion score maximization

The goal of this optimization is to achieve maximum user satisfaction and fairness among the users. For every user, depending on the service, we define a decision variable for every transmission policy—whether this user is served with a given transmission policy or not. Consequently, these decision variables are of boolean type, that is, either the user transmits its information using this policy or not. For the voice users, we have decision variables u_{ij}, where "i" denotes the ith user and "j" refers to the jth transmission policy available for the voice users.

Mobile users in the wireless network have time-varying position, which results in variable SNR at the receiver. Based on the SNR, we compute an estimate of the PEP [20] for different modulation schemes (DBPSK and DQPSK) and different channel code rates, that is, for all candidate transmission policies. A channel realization is generated and the estimation of the PEP is performed for all the transmission policies given the particular SNR at the receiver.

Our objective function for multiuser multiapplication cross-layer optimization is defined in (7). A maximization of the sum of the MOS perceived by every user in our multimedia wireless network has to be achieved. The parameter λ is used to ensure fairness among the users.

Maximize

$$\sum_{i\in U}\sum_{j\in T_U}\lambda_{ui}u_{ij}E[\mathrm{MOS}_{ij}]+\sum_{i\in V}\sum_{j\in T_V}\lambda_{vi}v_{ij}E[\mathrm{MOS}_{ij}]+\sum_{i\in W}\sum_{j\in T_W}\lambda_{wi}w_{ij}E[\mathrm{MOS}_{ij}] \tag{7}$$

subject to

$$\begin{aligned}\sum_{j\in T_U}u_{ij}&=1,\quad \forall i\in U,\\ \sum_{j\in T_V}v_{ij}&=1,\quad \forall i\in V,\\ \sum_{j\in T_W}w_{ij}&=1,\quad \forall i\in W,\end{aligned} \tag{8}$$

$$\sum_{i\in U}\sum_{j\in T_U}r_{ij}u_{ij}+\sum_{i\in V}\sum_{j\in T_V}r_{ij}v_{ij}+\sum_{i\in W}\sum_{j\in T_W}r_{ij}w_{ij}\le \text{total symbol rate}. \tag{9}$$

In our example, every user must be associated with only one transmission rate, channel code rate, and modulation scheme. The decision variables u_{ij}, v_{ij}, and w_{ij} are of boolean type which leads to the constraints (8). The total available symbol rate for all the users is constrained to be less than the total symbol rate of the system. Every transmission policy has an associated symbol rate r_{ij} and the sum of all the chosen symbol rates of all the users must be less than or equal to the total symbol rate. The above problem can be solved with a full search through the possible parameter space which has the worst case number of searches of $|T_U|^{K_U}\cdot|T_V|^{K_V}\cdot|T_W|^{K_W}$ where $|T_U|$, $|T_V|$ and $|T_W|$ are the numbers of transmission policies and K_U, K_V, K_W are the numbers of users of user classes U, V, and W, respectively.

The parameters λ_{ui}, λ_{vi}, λ_{wi} in (7) are inserted to ensure a fair allocation of resources. The optimizer finds a resource allocation which maximizes the user satisfaction based on MOS. In this case, there is a possibility that even though the system performance is maximized, a given user is not satisfied. This could be caused by low receiver SNR and the optimizer can decide to allocate the resources to the other users. This contradicts with the fairness we are trying to offer to the users independent of their location. To solve this problem, we propose to select the scaling coefficients λ_{ui}, λ_{vi}, λ_{wi} based on the history of the user estimated MOS. On every rate allocation procedure, we find the user with the maximum average of the estimated MOS from the previous steps. Let us assume that we are at rate allocation step "N" and we have K users in

the system. The value of the maximum perceived MOS by a single user is found by

$$\text{MaxMOS}_N = \frac{1}{N-1}\max\left(\sum_{n=1}^{N-1}\text{MOS}_{1n}; \sum_{n=1}^{N-1}\text{MOS}_{2n}; \ldots; \sum_{n=1}^{N-1}\text{MOS}_{Kn}\right). \tag{10}$$

The scaling coefficient for every user is calculated with

$$\lambda_{kN} = \frac{\text{MaxMOS}_N}{(1/(N-1))\sum_{n=1}^{N-1}\text{MOS}_{kn}}, \quad k = 1 \cdots K. \tag{11}$$

The user with the maximum perceived MOS has a scaling coefficient of one. The other users have scaling coefficients in the range [1; 4.5], because the denominator is also bounded in the interval [1; MaxMOS_N]. Since these λ values scale the estimated MOS for every transmission policy and we maximize the sum of the MOS of all the users, the optimizer assigns transmission policies with high estimated MOS to the users with higher λ. This ensures fairness by providing higher resources to the users which have been receiving lower MOS up to the time of the current optimization step.

3.4. Throughput maximization

A common network performance metric is the throughput of the system. Traditionally, the goal of the network operator is to maximize the network throughput. By throughput we consider the effective rate (goodput) G_{ij} of a given user i at time j:

$$G_{ij} = R_{ij} * (1 - \text{PEP}_{ij}) \tag{12}$$

with R_{ij} is the actual transmission rate and PEP_{ij} is the packet error probability. The objective function for such an optimization model is to maximize the sum of the goodput allocated to all the users in the system and is given with (13). The assumption is that a higher goodput will result in a higher user satisfaction regardless of the application type.

For throughput maximization, we have the same set of decision variables as in (7)–(9). The difference is the absence of the scaling parameter λ. Here we do not need scaling of the allocated transmission rate, because the transmission rates required by different applications are not comparable. Additionally, in order to make a fair comparison with our MOS-based optimization, we include a constraint on the packet error probability, $\text{PEP}_{\max}$, for each application type, so that the real-time applications are assigned a *sensible* share of the resources.

Maximize

$$\sum_{i\in U}\sum_{j\in T_U} u_{ij}G_{ij} + \sum_{i\in V}\sum_{j\in T_V} v_{ij}G_{ij} + \sum_{i\in W}\sum_{j\in T_W} w_{ij}G_{ij} \tag{13}$$

subject to

$$\sum_{j\in T_U} u_{ij} = 1, \quad \forall i \in U, \tag{14}$$

$$\sum_{j\in T_V} v_{ij} = 1, \quad \forall i \in V, \tag{15}$$

$$\sum_{j\in T_W} w_{ij} = 1, \quad \forall i \in W, \tag{16}$$

$$\sum_{i\in U}\sum_{j\in T_U} r_{ij}u_{ij} + \sum_{i\in V}\sum_{j\in T_V} r_{ij}v_{ij} + \sum_{i\in W}\sum_{j\in T_W} r_{ij}w_{ij} \leq \text{total symbol rate}, \tag{17}$$

$$\text{PEP}_i \leq \text{PEP}_{\max,i}. \tag{18}$$

3.5. Greedy resource allocation algorithm

The full-search resource allocation described in Sections 3.3 and 3.4 becomes computationally infeasible as the number of users in the system grows. For example, with three voice users, two ftp users and two video users, the number of resource allocations that have to be considered is $4.845 \cdot 10^{12}$.

The greedy allocation algorithm used in this work is similar to the work in [19]. It is initialized by assigning equal amount of resources to every user. In each subsequent step, a small amount of resources is taken from the user with the lowest sensitivity to a decrease of resources and assigned to the user that receives the maximum benefit. This is repeated until there is no further improvement in the objective function. The greedy algorithm for the MOS-maximization is described below. The throughput maximization is performed in a similar way.

Let Θ_i denote the utility function, and α_i the share of resource (symbol rate) of user i. Then, $\Delta\Theta_i$ denotes the change of utility for user i due to a change of its resource share, $\Delta\alpha_i$, where $\sum_{i=1}^{K}\alpha_i = 1$, that is, the sum of resource share over all the users in the system equals unity. The greedy allocation can be expressed as an iterative maximization of the incremental utility values of two users i and j:

$$\max_{(i,j)\in\{1,\ldots,K\}} \frac{\Delta\Theta_i}{\Delta\Theta_j}, \quad i \neq j, \tag{19}$$

where $\Delta\Theta_i$ and $\Delta\Theta_j$ are changes of utility due to an increase, $\Delta\alpha_i$, and decrease, $\Delta\alpha_j$, of resource share for user i and j, respectively, and K is the total number of users.

3.6. Generalization

Our cross-layer optimization scheme is not limited to a certain mix of only those application types we described in Section 2, but it is applicable to any general scenario. In Section 3.3, we propose the MOS-based optimization scheme where it is assumed that the application-layer side information (SI) is provided to the cross-layer optimizer in the form of MOS-functions. In a more general scenario, we need to consider the case when some of the streams provide SI and some do not. For this purpose, we classify the streams into two categories: SI and non-SI streams. Our strategy for these

two groups is going to be as follows: MOS-maximization for the SI streams, and throughput maximization for the non-SI streams. An initial resource allocation (e.g., symbol rate) among the two classes would be necessary. This can be assigned with the information on QoS requirement for each class. We regard this as a separate optimization issue. In our simulations, the resources (symbol rates) among the two classes are assigned in proportion to the number of streams belonging to each class. Let K_M and K_T be the number of SI and non-SI streams, respectively. Then, for the purpose of our simulation, optimization problems of Sections 3.3 and 3.4 are modified only in (9) and (17) to incorporate the symbol rate constraints of (20) and (21) respectively, as follows:

$$\sum_{i\in U}\sum_{j\in T_U} r_{ij}u_{ij} + \sum_{i\in V}\sum_{j\in T_V} r_{ij}v_{ij} + \sum_{i\in W}\sum_{j\in T_W} r_{ij}w_{ij} \le \text{Total Symbol Rate} \cdot \frac{K_M}{K}, \tag{20}$$

$$\sum_{i\in U}\sum_{j\in T_U} r_{ij}u_{ij} + \sum_{i\in V}\sum_{j\in T_V} r_{ij}v_{ij} + \sum_{i\in W}\sum_{j\in T_W} r_{ij}w_{ij} \le \text{Total Symbol Rate} \cdot \frac{K_T}{K}, \tag{21}$$

where K is the total number of users.

4. SIMULATION

The simulations shown in this paper are performed with the following parameter settings. We assume a total of seven simultaneous users in the wireless network. Three voice users, one male and two female voices, are used. The voice samples are 60 seconds long. The voice signal comes from the backbone network encoded with G.711 voice codec at 64 kbps. In the base station, following the optimization output, the signal could be transcoded to 6.4 kbps with G.723.1 codec, 15.2 kbps with iLBC codec, 24.6 kbps with Speex, or it can be transmitted without transcoding at 64 kbps.

Two users perform a file download using FTP. Both of them have subscribed for a service with maximum offered transmission rate of 192 kbps.

Two users are using video streaming service. The video sequences used for our simulation are "foreman" and "mother and daughter," encoded with the H.264 reference software encoder. The GOP structure is I-P-P-..., encoded at 30 frames per second in QCIF resolution (176×144 pixels).

The λ values in (7) are all initialized to 1 in our experiments. The total available symbol rate is constant and we have examined three different cases: 500 Ksymbol/s, 1000 Ksymbol/s, and 1500 Ksymbol/s. The supported modulation schemes are DBPSK and DQPSK. Channel code rates of 1/2, 1/3, 1/4, and 1 (uncoded) are supported, using convolutional code.

To reflect user mobility, the receiver SNR for every optimization step is drawn randomly for every user from a uniform distribution from 5 dB to 25 dB. The system is active for 60 seconds and we assume that the average channel characteristics remain constant for 1.2 second periods, which results in 50 optimization loops. $\text{PEP}_{\max}$ is set to be 0.1, 0.2, and 0.3 for video, voice, and ftp services, respectively.

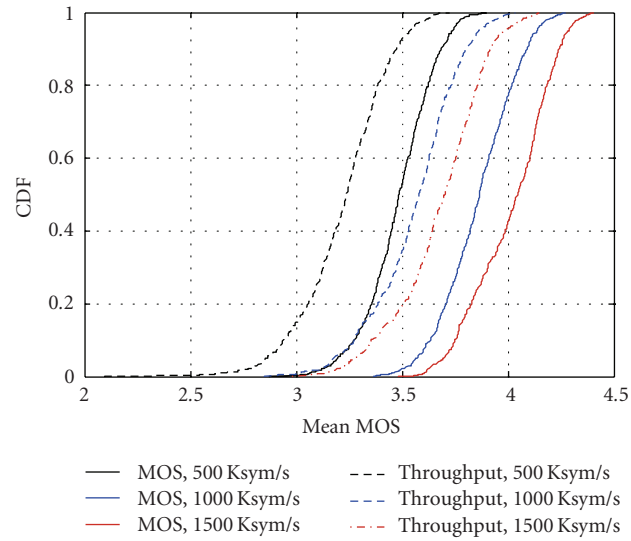


Figure 7: Mean opinion score of all seven users for three different total symbol rates (500, 1000, and 1500 Ksymbol/s) and two different optimization techniques, MOS maximization and throughput maximization.

The wireless system we have implemented in this work does not refer to any particular physical layer interface. We kept it intentionally simple, as the main goal of our work is to demonstrate the potential gain for any wireless system considering joint optimization across multiple different applications.

For the voice users, the signal samples are partitioned into 1.2 seconds and every sample is encoded using the voice codec determined by the optimization algorithm. At the end of the optimization loops, these voice samples are assembled into a single file and the perceived quality (MOS) is computed by comparing the original signal and the distorted one using PESQ.

For the video user, if a slice (packet) is lost, it is not written in the bit stream, which tells the decoder to invoke the error concealment algorithm. The PSNR of every frame and the resulting average PSNR are computed. The average PSNR is converted to an MOS value using the relationship shown in Figure 5. For file download we compute the MOS using the relationship given in (3).

5. RESULTS

5.1. Comparison between MOS-based and throughput-based optimization

In this section, a comparison between the two investigated optimization approaches (MOS maximization and throughput maximization) is performed. We use the setup described in the previous section and the results are based on 600 runs.

Figure 7 shows the cumulative density function (CDF) of mean opinion score over all the users for the two optimization approaches and three different total system rates: 500 Ksymbol/s (overloaded system), 1000 Ksymbol/s (moderately loaded system), and 1500 Ksymbol/s (lightly

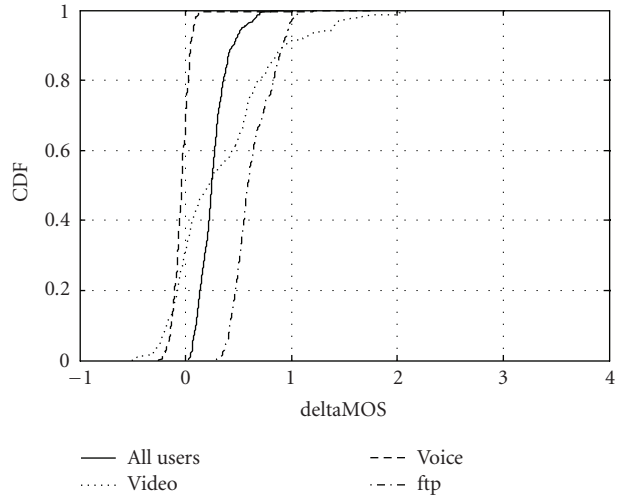


Figure 8: MOS gain per user, system symbol rate of 500 Ksymbol/s.

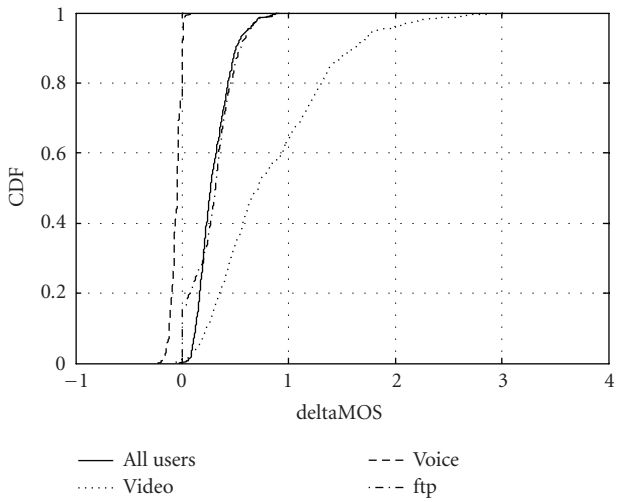


Figure 9: MOS gain per user, system symbol rate of 1000 Ksymbol/s.

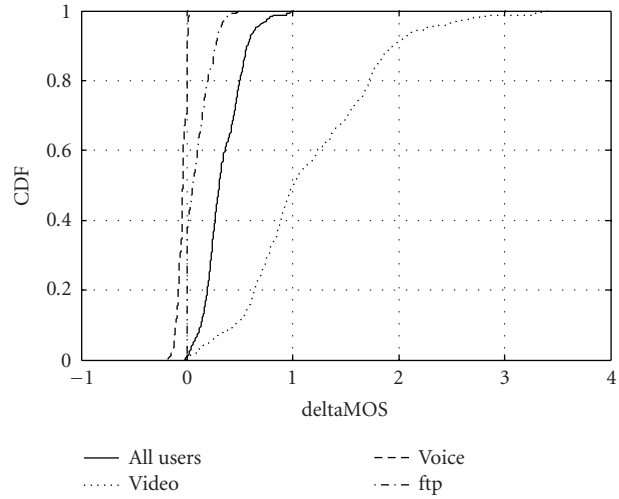


Figure 10: MOS gain per user, system symbol rate of 1500 Ksymbol/s.

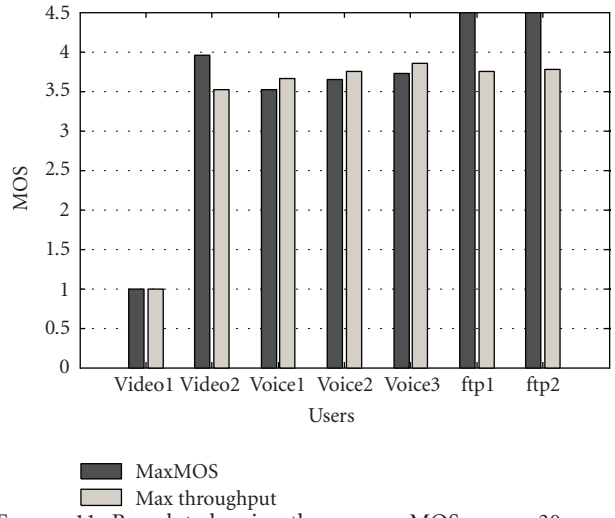
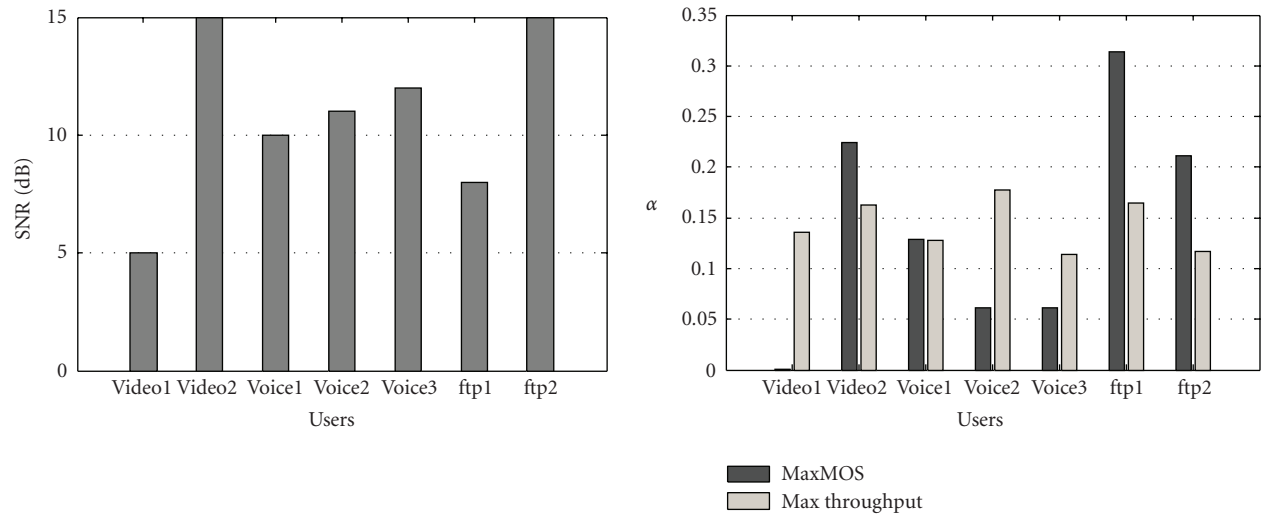


Figure 11: Bar plot showing the average MOS over a 30-second simulation run.

loaded system). The average gain in MOS for the three systems is 0.26, 0.30, and 0.34, respectively.

Figures 8–10 present the gain per user in the system. The curves are produced as a difference between the mean MOS computed with MOS maximization and throughput maximization. Starting with a system symbol rate of 500 Ksymbol/s (Figure 8), in 50% of the simulations, the average gain for all users is 0.26. The video and FTP users are benefited with a little penalty on the voice users. In Figures 9 and 10, we observe increasingly higher gain for the video users, with little noticeable loss of quality for the voice users. The quality of voice and video services are very sensitive to packet losses. In our throughput maximization approach we set a maximum allowable packet error probability, PEP_{max} of 0.2 for voice and 0.1 for video service. This turns out to be a reasonable choice for the voice users and the second video user "(mother and daughter)," but is too high for the first video user "(foreman)," because of its dynamic content which is highly sensitive to packet loss. In our MOS-based cross-layer optimization approach, application requirements are taken care of individually for each user, which results in optimum allocation of resources in terms of user perceived quality.

Figure 11 shows the average MOS of the seven users over a 30-second simulation run. Figure 12(a) shows the receiver SNR which was fixed during the simulation and Figure 12(b) shows the resource shares, α. Video1 receives a very low SNR, which results in poor received video quality for both optimization approaches. However, our MOS maximization approach, being aware of the utility function of the applications, does not assign any resource to this user, and distributes the *saved* resources to other users which results in higher mean MOS.



(a) (b)

Figure 12: (a) Mean receiver SNR of seven users, (b) resource shares.

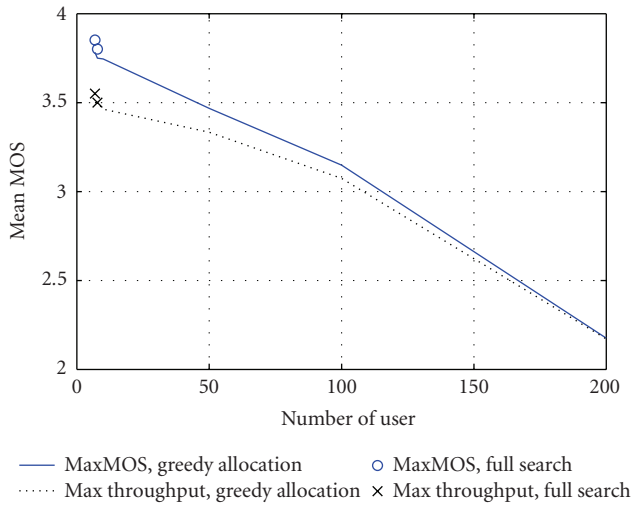


Figure 13: Mean MOS versus the number of users for the greedy algorithm and the full search algorithm.

5.2. Performance of the greedy search algorithm

The full-search algorithm described in Sections 3.3 and 3.4 is computationally too expensive to be implemented. In order to use our cross-layer optimization scheme in a real-time scenario, we also developed a greedy search algorithm, as described in Section 3.5. The resource that is allocated among the users is the time-share, which translates to a particular symbol-rate for a user, as the total symbol-rate of the system is fixed. Figure 13 shows the mean MOS versus the number of users using the greedy search approach for $K = 7$ to 200 while the total symbol rate of the system is fixed at 1000 Ksymbol/s. In the simulations when we use varying number of user, we keep the number of video and ftp user fixed at two and two, respectively, and increase only the number of voice user. At $K = 7$ and $K = 8$, we also compute the mean MOS using the full-search approach, and we observe little difference between the two approaches. For $K > 8$, the computation for the full-search approach becomes infeasible, while the greedy search remains fast enough to be used in online optimizations. As the number of user increases, the gap between the MOS-based and throughput-based approaches gradually decreases. Please note that for satisfied users, the MOS should stay above 3.5. At this level, we see significant improvements when using the MOS-based optimization.

The convergence speed of the greedy algorithm can be measured in terms of the number of iterations. The number of iterations tends to be dependent on the resource allocation step size $\Delta\alpha_i$ and a minimum threshold of utility improvement at each iteration, $\Delta\Theta_{th}$. The improvement of utility at each iteration, $\Delta\Theta_{\text{incr}} = \Delta\Theta_i - \Delta\Theta_j$ is compared with the threshold, $\Delta\Theta_{th}$. The algorithm is assumed to converge when $\Delta\Theta_{\text{incr}} \leq \Delta\Theta_{th}$. For a comparison of the number of iterations required for different number of users with a wide range of channel conditions, we keep these two parameters fixed, $\Delta\alpha_i = 0.0001$ and $\Delta\Theta_{th} = 0.00005$ (MOS).

Figure 14 shows the CDF of the number of iterations for 5, 10, and 50 users. The worst-case number of iterations is found to be in the range of 3000 to 4000 iterations. It is interesting to find that the 5-user case may take more iterations to converge than what we observe for the case of 50 users, the reason being the use of equal step size for both cases. Further fine-tuning is possible by choosing a step-size that is a function of the number of users. Also, for those applications, which have a limit on the rate (e.g., voice communication applications with at most 64 kbps), we can speed up the greedy algorithm by using this fact during initialization.

The time to complete each iteration, however, increases with the number of users. The convergence speed of the greedy algorithm in terms of time is shown in Figure 15. The measurements are taken from the Matlab-based simulation environment, with an Intel dual-core T2300 1.66 GHz

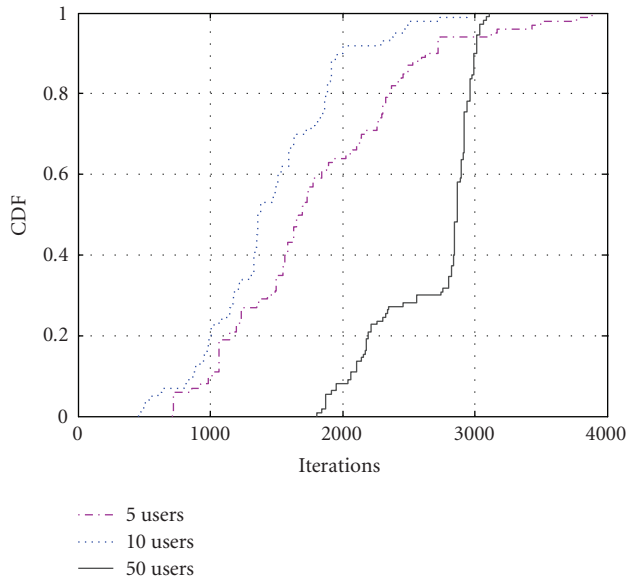


Figure 14: CDF of number of iterations with greedy algorithm.

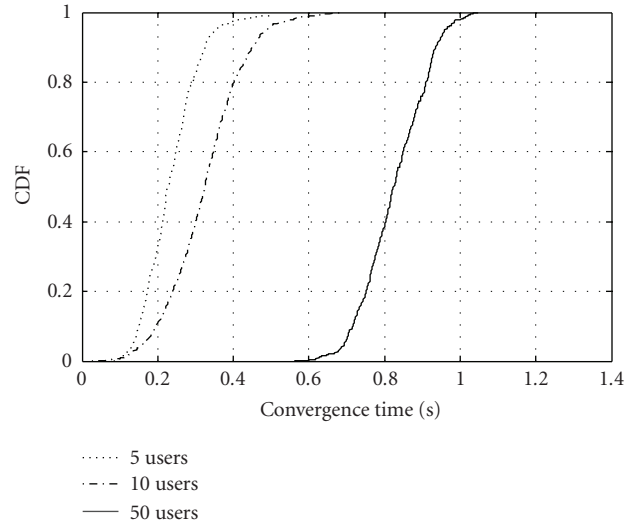


Figure 15: CDF of convergence time with greedy algorithm.

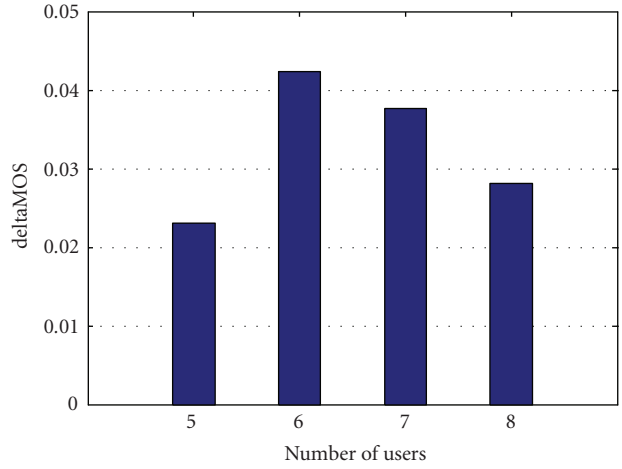


Figure 16: Worst-case difference in MOS (deltaMOS) between full-search and greedy algorithm.

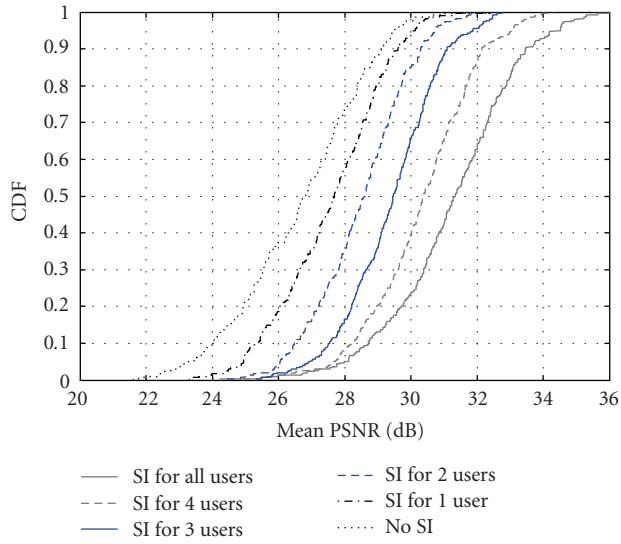


Figure 17: CDF of mean PSNR (in dB) for a five-user video streaming scenario.

processor. Using dedicated software and hardware environments, the convergence speed is expected to be much faster.

Figure 16 shows the worst-case performance gap between full-search and greedy algorithm. The performance gap is computed as the difference between the MOS values obtained by using the full-search and the greedy algorithm. We find that the gap is reasonable.

5.3. Optimization with and without side information

In this section, we consider the more general case when some of the streams provide application-layer side information (MOS functions) and some do not. As discussed in Section 3.6, we perform MOS-based optimization for the SI streams, and throughput-based optimization for the non-SI streams. In this section, we consider a five-user video streaming scenario. Figure 17 shows the CDF of mean PSNR over five users for all possible cases of with and without SI. For this scenario, we use the same video sequence ("foreman" sequence) for all five users with a wide variety of channel conditions. For the case "SI for all user" MOS-based optimization is used, while for the case of "no SI," throughput-based optimization is performed. For the other cases, both approaches are used in combination. Figure 17 shows that we have an average gain of 1 dB PSNR for each additional stream with SI. It is easy to extend this strategy to a system having different application types, although the results will be more involved due to the different quality metrics for different applications.

6. CONCLUSION

In this paper, we propose a novel multiuser cross-layer optimization approach across multiple applications using MOS as a common application layer performance metric. With this approach we are able to dynamically optimize the wireless transmission system resource usage and the user perceived quality of service in a multiuser environment. We compare our approach to a traditional approach where allocation is done with the goal of maximizing overall throughput. Our simulation results show significant improvements in terms of user perceived quality for a variety of circumstances.

REFERENCES

[1] W. Kellerer, L.-U. Choi, and E. Steinbach, "Cross-layer adaptation for optimized B3G service provisioning," in *Proceedings of the 6th International Symposium on Wireless Personal Multimedia Communications (WPMC '03)*, pp. 57–61, Yokosuka, Japan, October 2003.

[2] L.-U. Choi, W. Kellerer, and E. Steinbach, "Cross layer optimization for wireless multi-user video streaming," in *Proceedings of the International Conference on Image Processing (ICIP '04)*, vol. 3, pp. 2047–2050, Singapore, October 2004.

[3] S. Khan, Y. Peng, E. Steinbach, M. Sgroi, and W. Kellerer, "Application-driven cross-layer optimization for video streaming over wireless networks," *IEEE Communications Magazine*, vol. 44, no. 1, pp. 122–130, 2006.

[4] 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.

[5] V. Tsibonis, L. Georgiadis, and L. Tassiulas, "Exploiting wireless channel state information for throughput maximization," in *Proceedings of the 22nd IEEE Annual Joint Conference of Computer and Communications Societies (INFOCOM '03)*, vol. 1, pp. 301–310, San Francisco, Calif, USA, March-April 2003.

[6] X. Liu, E. K. P. Chong, and N. B. Shroff, "Transmission scheduling for efficient wireless utilization," in *Proceedings of the 20th IEEE Annual Joint Conference of Computer and Communications Societies (INFOCOM '01)*, vol. 2, pp. 776–785, Anchorage, Alaska, USA, April 2001.

[7] M. T. Ivrlač and J. A. Nossek, "Cross layer optimization-an equivalence class approach," in *Proceedings of the International Symposium on Signals, Systems, and Electronics (ISSSE '04)*, Linz, Austria, August 2004.

[8] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Communications Magazine*, vol. 41, no. 10, pp. 74–80, 2003.

[9] ITU-T G.107, "The E-model, a computational model for use in transmission planning".

[10] ITU-T P.862, "PESQ: an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs".

[11] F. P. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[12] ITU-R Recommendation BT.500, "Methodology for the subjective assessment of the quality of television pictures".

[13] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications".

[14] S. Winkler, A. Sharma, and D. McNally, "Perceptual video quality and blockiness metrics for multimedia streaming applications," in *Proceedings of the 4th International Symposium on Wireless Personal Multimedia Communications (WPMC '01)*, pp. 547–552, Aalborg, Denmark, September 2001.

[15] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proceedings of IEEE International Conference on Image Processing (ICIP '02)*, vol. 3, pp. 57–60, Rochester, NY, USA, September 2002.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[17] Final report from the Video Quality Experts Group, "On the validation of objective models of video quality assessment," October 2003.

[18] L. U. Choi, M. T. Ivrlač, E. Steinbach, and J. A. Nossek, "Sequence-level models for distortion-rate behaviour of compressed video," in *Proceedings of IEEE International Conference on Image Processing (ICIP '05)*, vol. 2, pp. 486–489, Genova, Italy, September 2005.

[19] D. Jurca and P. Frossard, "Media-specific rate allocation in multipath networks," *IEEE Transactions on Multimedia*, vol. 9, no. 5, 2007.

[20] M. T. Ivrlač, "Parameter selection for the Gilbert-Elliott model," Tech. Rep. TUM-LNS-TR-03-05, Institute for Circuit Theory and Signal Processing, Munich University of Technology, Munich, Germany, May 2003.

Hindawi Publishing Corporation
Advances in Multimedia
Volume 2007, Article ID 13969, 12 pages
doi:10.1155/2007/13969

Research Article

Cross-Layer Perceptual ARQ for Video Communications over 802.11e Wireless Networks

P. Bucciol,[1] E. Masala,[1] E. Filippi,[2] and J. C. De Martin[1]

[1] *Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy*
[2] *Advanced System Technologies, STMicroelectronics S.r.l., Cornaredo, 20010 Milano, Italy*

Received 29 December 2006; Revised 13 April 2007; Accepted 2 July 2007

Recommended by Zhu Han

This work presents an application-level perceptual ARQ algorithm for video streaming over 802.11e wireless networks. A simple and effective formula is proposed to combine the perceptual and temporal importance of each packet into a single priority value, which is then used to drive the packet-selection process at each retransmission opportunity. Compared to the standard 802.11 MAC-layer ARQ scheme, the proposed technique delivers higher perceptual quality because it can retransmit only the most perceptually important packets reducing retransmission bandwidth waste. Video streaming of H.264 test sequences has been simulated with *ns* in a realistic 802.11e home scenario, in which the various kinds of traffic flows have been assigned to different 802.11e access categories according to the Wi-Fi alliance WMM specification. Extensive simulations show that the proposed method consistently outperforms the standard link-layer 802.11 retransmission scheme, delivering PSNR gains up to 12 dB while achieving low transmission delay and limited impact on concurrent traffic. Moreover, comparisons with a MAC-level ARQ scheme which adapts the retry limit to the type of frame contained in packets and with an application-level deadline-based priority retransmission scheme show that the PSNR gain offered by the proposed algorithm is significant, up to 5 dB. Additional results obtained in a scenario in which the transmission relies on an intermediate node (i.e., the access point) further confirms the consistency of the perceptual ARQ performance. Finally, results obtained by varying network conditions such as congestion and channel noise levels show the consistency of the improvements achieved by the proposed algorithm.

1. INTRODUCTION AND MOTIVATIONS

The IEEE 802.11 wireless local area networking standard [1] provides network access capabilities to an ever expanding array of devices, including multimedia-enabled devices. The 802.11 standard addresses channel noise and MAC-level collision issues by means of a link-layer automatic repeat request (ARQ) scheme. This mechanism is well suited for generic data transmission, because it is fast and simple to implement. However, for the specific—and increasingly important—case of multimedia traffic, more advanced ARQ techniques could deliver higher perceptual quality as well as use network resources more efficiently.

Multimedia communications exhibit peculiar features if compared with conventional data transmissions. Two of the most important characteristics of multimedia streams, in fact, are the highly nonuniform perceptual importance of data and the strong time sensitivity. Most ARQ techniques designed for multimedia communications often consider one or both characteristics. For instance, the *Soft ARQ* proposal [2] avoids retransmitting late data that would not be useful at the decoder, thus saving bandwidth. This technique has also been adapted to deal with layered encoded streams.

Other works focused on optimizing prioritization mechanisms in order to take advantage of the different perceptual importance of the syntax elements contained in a compressed multimedia bitstream. For instance, video packets can be protected by different error correcting codes depending on the type of frame to which the packets belong, as in [3], in which an additional ARQ scheme that privileges the most important classes of data is also implemented by means of different retry limit values. The work in [4] proposes to schedule video frames according to the priority given by their position inside the group of pictures (GOP), and at the same time, it assigns different priorities to motion and texture information contained in each packet. A retry limit adaptation scheme for layered video has been proposed in [5]. That work presents an algorithm which can dynamically determine the

best retry limit value for each layer, depending on channel error and MAC-level buffer overflow probabilities, in a priority queueing transmission system.

Further improvements can be achieved by optimizing the transmission policy for each single packet rather than relying on a priori determination of the average importance of the elements contained in the compressed bitstream [6, 7]. In the low-delay wireless video transmission system presented in [8], for instance, packets are retransmitted or not depending on whether the distortion caused by their loss is above a given threshold. However, it is not clear how to optimally determine such threshold. Given a way to associate distortion values to each packet, rate-distortion optimization of the transmission policies has also been proposed [9–11].

This work addresses the specific case of video streaming over a congested 802.11 network. In order to overcome the limitations of the standard 802.11 MAC-level ARQ which retransmits all packets regardless of their importance, we propose a cross-layer perceptual ARQ scheme which exploits information about the *perceptual* and the *temporal* importance of each packet. The proposed ARQ scheme is composed of three parts: an algorithm which determines retransmission opportunities, a retransmission scheduling algorithm, and a formula to compute a priority value for each packet. More specifically, first a set of retransmission opportunities is determined at the beginning of each GOP, then the scheduling algorithm retransmits packets according to their priority and on the basis of the receiver feedback. The priority of each packet is computed using a simple and flexible formula that combines perceptual importance and maximum delay constraint. Perceptual importance is evaluated using the analysis-by-synthesis technique [10], which is explained in Section 3.

This paper presents a detailed analysis of the cross-layer perceptual ARQ scheme first presented in [12]. Extensive simulation results quantify the impact of the main algorithm parameters and illustrate how varying levels of congestion or channel noise affect the performance of the proposed ARQ scheme. This work focuses on a congested 802.11e home network scenario in which the access point represents the home access gateway (HAG). Test H.264 video transmissions in presence of several concurrent interfering flows are simulated. Two scenarios have been considered: direct transmission, from the access point to a PC, and indirect transmission from the PC to a TV set, relaying on the access point. Both perceptual video quality (as measured by PSNR) and network performance metrics are obtained in different conditions and for different values of the main algorithm parameters, such as the maximum retransmission bandwidth. Moreover, the sensitivity of the proposed technique to variations in the scenario (i.e., the amount of concurrent traffic and channel noise) has also been evaluated. Besides the standard MAC-level ARQ scheme, two reference techniques have been implemented and studied for comparison purposes. The first technique is a deadline-driven application-level ARQ in which the highest retransmission priority is given to the packet whose playout deadline is the nearest, similarly to [2]. The second one is a MAC-level ARQ technique which imposes different retry limits value for each type of packet, as proposed in [3], to give unequal protection to the various elements of the video sequence.

The remainder of the paper is organized as follows. Section 2 briefly reviews the H.264 standard focusing on communication issues. Then, Section 3 provides details on the analysis-by-synthesis distortion estimation technique, which is used compute perceptual importance values. Section 4 describes the cross-layer ARQ technique studied in this work. Simulation setup and an extensive discussion of results are presented in Sections 5 and 6, respectively. Finally, conclusions are drawn in Section 7.

2. H.264 VIDEO COMMUNICATIONS

In this work, we consider video communications based on the state-of-the-art H.264 video codec [13, 14]. This codec is particularly suitable for transmission over packet networks. In fact, one of the most interesting characteristics of the H.264 standard is the attempt to decouple the coding aspects from the bitstream adaptation needed to transmit it over a particular channel. The part of the standard that deals with the coding aspects is called Video Coding Layer (VCL), while the other is the network adaptation layer (NAL).

As in previous video coding standards, the H.264 VCL groups consecutive macroblocks into *slices*, that are the smallest independently decodable units. Slices are important because they allow to subdivide the coded bitstream into independent packets, so that the loss of a packet does not affect the ability of the receiver to decode the bitstream of others.

Differently from other video coding standards, the H.264 provides a NAL which aims to efficiently support transmission over IP networks [15]. In particular, it relies on the use of the real-time transport protocol (RTP), which is well suited for real-time wired and wireless multimedia transmissions. The implementation of our proposed algorithm is compliant with this NAL specification.

However, some dependencies exist between the VCL and the NAL. For instance, the packetization process is improved if the VCL is instructed to create slices of about the same size of the packets and the NAL told to put only one slice per packet, thus creating independently decodable packets. The packetization strategy, as the frame subdivision into slices, is not standardized and the encoder has the possibility to vary both of them for each frame. Usually, however, the maximum packet size (hence slice size) is limited and slices cannot be too short due to the resulting overhead that would reduce coding efficiency.

3. DISTORTION ESTIMATION

The quality of multimedia communications over packet networks may be impaired in case of packet loss. The amount of quality degradation strongly vary depending on the importance of the lost data. In order to design efficient loss protection mechanisms, a reliable importance estimation method for multimedia data is needed. Such importance is often defined a priori, based on the average importance of the elements of the compressed bitstream, as with the data partitioning approach.

In order to provide a quantitative importance estimation method at a finer level of granularity, we define the importance of a video coding element, such as a macroblock or a packet, as a value proportional to the distortion that would be introduced at the decoder by the loss of that specific element. The potential distortion of each element, could, therefore, be computed using the analysis-by-synthesis technique [10]. The conceptual scheme is depicted in Figure 1. In this work, we apply the analysis-by-synthesis technique on a packet basis. Hence, the video sequence has to be coded and packetized before the activation of the algorithm. The analysis-by-synthesis distortion estimation algorithm performs, for each packet, the following steps:

(1) decoding, including concealment, of the bitstream simulating the loss of the packet being analyzed (synthesis stage);

(2) quality evaluation, that is, computation of the distortion caused by the loss of the packet. The original and the reconstructed picture after concealment are compared using, for example, MSE;

(3) storage of the obtained value as an indication of the perceptual importance of the analyzed video packet.

The previous operations can be implemented with small modifications of the standard encoding process. The encoder, in fact, usually reconstructs the coded pictures simulating the decoder operations, since this is needed for motion-compensated prediction. If step (1) of the analysis-by-synthesis algorithm exploits the operations of the encoding software, complexity is only due to the simulation of the concealment algorithm. In case of simple temporal concealment techniques, this is trivial and the task is reduced to provide the data to the quality evaluation algorithm.

The analysis-by-synthesis technique, as a principle, can be applied to any video coding standard. In fact, it is based on repeating the same steps that a standard decoder would perform, including error concealment. Obviously, the importance values computed with the analysis-by-synthesis algorithm are dependent on a particular encoding, that is, if the video sequence is compressed with a different encoder or using a different packetization, values will be different. Note, however, that in principle the analysis-by-synthesis scheme does not impose any particular restriction on encoding parameters or packetization.

Due to the interdependencies usually present between data units, the simulation of the loss of an isolated data unit is not completely realistic, particularly for high packet loss rates. Every possible combination of events should ideally be considered, weighted by its probability, and its distortion computed by the analysis-by-synthesis technique, obtaining the expected distortion value. For simplicity, however, we assume that all preceding data units have been correctly received and decoded. Nevertheless, this leads to a useful approximation as demonstrated by some applications of the analysis-by-synthesis approach to MPEG-coded video [6, 7, 10]. The results section will show the effectiveness of the proposed video transmission algorithm which relies on these distortion values.

The application of the analysis-by-synthesis method is straightforward when considering elements of the video stream which do not contribute to later referenced frames, since the mismatch due to concealment does not propagate. If propagation is possible, the distortion caused in subsequent frames should be evaluated until it becomes negligible, for instance, at the beginning of the next group of pictures (GOP) for MPEG video, or until its value falls below a given threshold. In this case, the complexity of the proposed approach could be high, but it is still suitable for stored-video scenarios that allow precomputation. In order to reduce complexity, statistical studies on many different video sequences have been conducted and a model-based approach [16] has been developed. According to that model, the encoder computes the distortion that would be caused by the loss of the packet into the current frame and then, using a simple formula, it computes an estimation of the total distortion which includes future frames. The reader is referred to the work in [16] for further details.

4. IMPORTANCE-BASED CROSS-LAYER ARQ

4.1. Overview

This work proposes an application-level end-to-end ARQ technique which relies on the perceptual and temporal importance of each multimedia packet in order to optimize the usage of retransmission bandwidth. The technique has been designed to work with the IP/UDP/RTP protocol stack [17]. RTCP packets are used to provide feedback information.

According to the proposed technique, every packet is transmitted once, then it is stored in a retransmission buffer $\mathrm{RTX}_{\mathrm{buf}}$ waiting for its acknowledgment. The receiver periodically generates RTCP receiver reports (RR) containing an ACK or a NACK for each transmitted packet. A NACK is generated when the receiver detects a missing packet by means of the RTP sequence number. When a retransmission opportunity is available, packets in the retransmission buffer are sent in the order given by their combined temporal-perceptual priority, as defined in Section 4.4.

A few key parameters can be used to tune the performance of the proposed technique, for instance, the peak transmission bandwidth B_{peak} granted to retransmissions and the relative weight of temporal with respect to perceptual importance.

4.2. Retransmission opportunities

The first step of the scheduling algorithm consists in determining the transmission time of each packet. The task is carried out, at the beginning of each GOP, by equispacing the packets of each frame inside their respective frame interval.

Let B_{GOP} be the bandwidth needed to transmit the current GOP and let B_{peak} be the peak bandwidth granted to the transmission, including retransmissions. Retransmission opportunities are identified using the following algorithm.

(1) Determine the number of retransmission opportunities N_{rtx}, for the current GOP, as $N_{rtx} = \lfloor B_{\mathrm{peak}} - B_{\mathrm{GOP}} \rfloor / \overline{S}_{pck}$, where $\overline{S}_{pck}$ is the average size of all the packets belonging to the original video sequence.

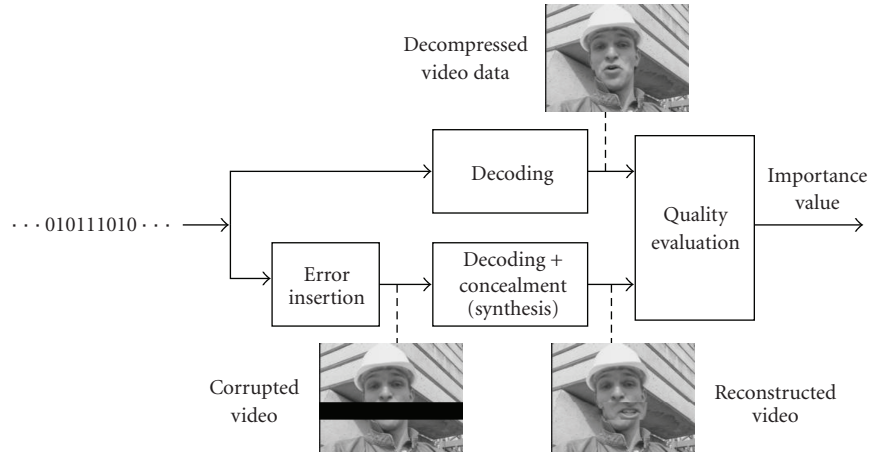


Figure 1: Conceptual scheme of the analysis-by-synthesis technique.

(2) Determine the time instants corresponding to the retransmission opportunities.

(2.1) Compute the total size, including retransmissions, of each frame. Identify the smallest one.

(2.2) Set the time of a retransmission opportunity as the midway between the time instant of the first packet of the smallest frame (including retransmissions) and the last packet of the previous frame.

(2.3) Repeat steps (2.1) and (2.2) until N_{rtx} opportunities have been determined, considering at each step the opportunities filled by packets of size $\overline{S}_{pck}$ and including them in the total frame size.

This procedure may create retransmission bursts between each frame, but has the advantage to be simple to implement; if desired, a more uniform distribution of the retransmission opportunities can be designed. Note also that the opportunities will not be necessarily used completely.

4.3. Scheduling algorithm

The retransmission policy, illustrated in Figure 2, is implemented by means of a retransmission buffer RTX_{buf}. After that a packet is sent for the first time, it is placed in the RTX_{buf}, waiting for its acknowledgment, and marked as *unavailable* for retransmission. When an ACK is received, the corresponding packet in the RTX_{buf} is discarded because it has been successfully transmitted. If a NACK is received, the corresponding packet is marked as *available* for retransmission. Packets belonging to the RTX_{buf} that will never arrive at the decoder in time for playback (considering the estimated FTT) are discarded. To limit the impact of receiver report losses, the sender piggybacks the highest sequence number for which it received an ACK or a NACK. The receiver always repeats in the receiver reports the status information for all the packets whose sequence number is less than the piggybacked one.

A priority function (see Section 4.4) is computed for each packet marked as *available* in the RTX_{buf} each time a new

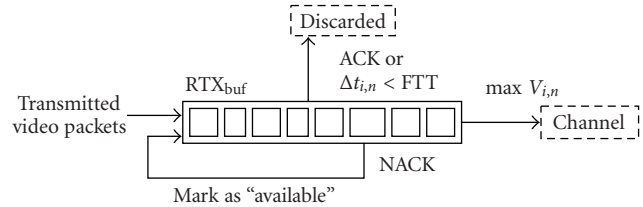


Figure 2: Diagram of the scheduling algorithm for packet retransmission.

retransmission opportunity approaches. The packet with the highest priority is then transmitted.

It is important to stress that the retransmission opportunities computed according to B_{peak} *not necessarily* will be actually used by the algorithm, leading to an actual bandwidth usage which can be considerably lower than B_{peak}.

4.4. Packet priority function

In a generic multimedia streaming scenario, each packet must be available at the decoder a certain amount of time before it is played back to allow the decoder to process it. Let t_n be the time the nth frame is played back. All packets containing data needed to synthesize the nth frame must be available at the decoder at time $t_n - T_P$, where T_P is the decoder processing time. Note that data dependencies in the coded video (e.g., due to reference to future frames) must also be considered.

We define *deadline* of a packet as the time instant at which that packet must be available at the decoder to be played back correctly. Let $t_{i,n}$ be the deadline of a packet i belonging to the nth frame. From the definition, it is clear that $t_{i,n} = t_n - T_P$. If a packet never arrives or it arrives after $t_{i,n}$, it will cause a distortion increase $D_{i,n}$ that can be evaluated using the analysis-by-synthesis technique.

Obviously, the sender should always select a packet for transmission only among the ones that can arrive before their deadline, that is, $t_{i,n} > t_s + \text{FTT}$, where t_s is the instant of the next retransmission opportunity and FTT (forward trip

time) is the time needed to transmit the packet, which is typically time-varying depending on the network state. Defining the distance from the deadline as $\Delta t_{i,n} = t_{i,n} - t_s$, the previous condition can be rewritten as $\Delta t_{i,n} > \text{FTT}$.

A policy is needed to choose which packet must be retransmitted and in which order, because at any given time several packets satisfy the condition $\Delta t_{i,n} > \text{FTT}$. Consider, for instance, the packets containing the video data of a certain frame, each packet has the same $\Delta t_{i,n}$. Within a frame, the sender should transmit, or retransmit, the packet with the highest $D_{i,n}$ that has not been yet successfully received. The decision is not as clear when choosing between sending an element A with low distortion $D_{A,n-1}$ in an older frame and an element B with high distortion $D_{B,n}$ in a newer frame. In other words, there is a tradeoff between the importance of the video data and its distance from the deadline (which can be seen as a sort of temporal importance.) A reason in favor of sending A is because its playback time is nearer ($\Delta t_{A,n-1} < \Delta t_{B,n}$), that reduces the number of opportunities to send it. On the other hand, if B arrives at the decoder, it will reduce the potential distortion of a value greater than A (because $D_{B,n} > D_{A,n-1}$.) A detailed study of the problem can be found in [2].

A criterion is needed to select, at each retransmission opportunity, the video packet which maximizes the expected quality of the transmission. We propose to compute, for each packet, a priority function of both its potential distortion and its distance from the deadline:

$$V_{i,n} = f(D_{i,n}; \Delta t_{i,n}). \tag{1}$$

Packets will then be sorted by $V_{i,n}$ and the one with the highest priority value is sent. The issue is to find an effective, and, if possible, simple function that combines the distortion value with the distance from the deadline. We propose to use the following function:

$$V_{i,n} = D_{i,n} + wC\frac{1}{\Delta t_{i,n}}. \tag{2}$$

The normalization factor C is computed as

$$C = \overline{D_{i,n}} \cdot T_B, \tag{3}$$

where T_B is the receiver buffer length, in seconds, and $\overline{D_{i,n}}$ is the average packet distortion. The normalization factor, C, is designed to balance the perceptual and temporal importance of the packet for the average case. The size of the receiver buffer T_B is, in fact, approximately equal to the mean value of the distance from the deadline, assuming that the receiver buffer is almost full. The weighting factor w in (2) is introduced to control the relative importance of the perceptual and temporal terms of the formula.

5. SIMULATION SCENARIO

The network simulator *ns* [18] has been used to assess the performance of the proposed technique. An 802.11e MAC layer [19] has been configured to operate over an 802.11a physical layer with a channel bandwidth of 36 Mbit/s. A packet error model has been implemented in *ns* based on BER curves obtained from 802.11 channel measurements, with different noise levels and packet sizes.

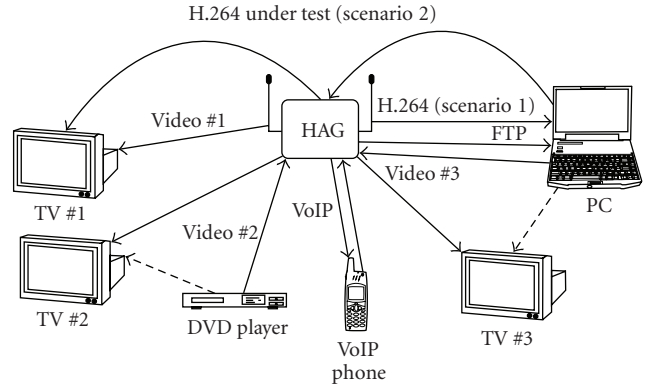


Figure 3: Home network scenario used in the experiments. The tested H.264 video stream is transmitted from the home access gateway (HAG), that is, the access point, to the PC (scenario 1) or from the PC to the TV set (scenario 2). The solid lines show the actual path of the transmitted packets, while the dashed lines indicate logical connections.

We simulated an H.264 video streaming transmission in a realistic home network scenario, in which many wireless devices, that is, three TV sets, a DVD player, a PC, and a VoIP terminal, all share the same physical bandwidth. The home access gateway is represented by the access point. Three concurrent video transmissions, a VoIP call, an FTP transfer as well as the H.264 video transmission under test, are active at the same time. Two scenarios have been analyzed. In scenario 1, the H.264 packets originate from the access point and they are directly sent to the PC, while in scenario 2, the H.264 packets are sent from the PC to the TV #1 by means of the access point which acts as a relay node. Both scenarios are depicted in Figure 3.

Tests have been performed using three different standard sequences (*paris*, *tempete*, *bus*) at CIF (352×288) resolution. They were encoded using version 6.1e of the H.264 test model software [13] with a fixed quantization parameter. The GOP encoding scheme is IBBPBBPBBPBB. The characteristics of the tested video sequences are shown in Table 1. Each sequence is concatenated with itself to reach a length of approximately 500 seconds. The video encoder is instructed to make RTP packets whose size is approximately constant. Unless otherwise noted, the playout buffer size is 1-second long. The decoder implements a simple temporal concealment technique that replaces a corrupted or missing macroblock with the macroblock in the same position in the previous frame.

The assignment of the various kinds of traffic to the 802.11e access categories has been based on the Wi-Fi alliance WMM specification [20]. The FTP stream is assigned to the lowest priority class, that is, access category (AC) 0. The tested H.264 stream is assigned to AC1, while all the remaining video flows are sent as AC2. The VoIP and the receiver reports flows are assigned to AC3, which provides the highest available QoS. This assignment provides

Table 1: Characteristics of the sequences used as H.264 streams.

Sequence	Avg. bitrate (kbit/s)	Encoding distortion (dB)	Resolution	Frame rate (fps)	Maximum packet size (bytes)	Avg. number of packets/s
Paris	765	35.68	352×288	30	750	146
Tempete	1205	34.23	352×288	30	1200	141
Bus	1304	34.25	352×288	30	750	235

Table 2: Access category assignment for all traffic.

Access category	Stream	Bandwidth
AC0	FTP	Variable
AC1	Tested H.264	765–1304 kbit/s
AC2	Video #1	1.5 Mbit/s
	Video #2	3 Mbit/s
	Video #3	6 Mbit/s
AC3	VoIP	72 kbit/s
	RTP receiver reports	3–6 kbit/s

the maximum protection against receiver report losses. The maximum number of MAC retransmissions is seven for all the classes except AC1, for which no MAC-level retransmissions are used unless MAC-level ARQ techniques are simulated. We assigned the tested H.264 video stream and the other video flows to different access categories because the retry limit can be specified only for each access category and not for each flow. To ensure fairness in the comparisons, however, the tested H.264 stream flow has been assigned to an access category whose priority is lower than the other video streams. Table 2 summarizes the assignments and the bandwidth of each flow. Note that the rate of the RTCP flow due to the receiver reports is very modest. It ranges from 3 to 6 kbit/s for a 100 milliseconds receiver report interval, and, if needed, could be further improved by packing ACK and NACK information more efficiently than the current implementation.

6. RESULTS

6.1. MAC-level ARQ techniques

First, the performance of the MAC-level ARQ techniques is presented. Two techniques have been studied. The first one is the current 802.11 standard ARQ implementation without any modification. The second technique employs MAC-level retransmissions with a different retry limit depending on the characteristics of the data contained in the packet, which allows MAC-level unequal protection as done in [3]. We divided the video flow into two classes, the first one containing packets belonging to I and P frames, and the second one for the rest of the packets (B frames), and we tested several retry limit values ($RL_{I,P}$, RL_B) for these two classes. We refer to this technique as the class-based retry-limit (CBRL) ARQ technique.

As a reference for comparisons, the performance of both techniques as a function of the retry limit has been studied for both the *paris* and the *tempete* sequences. For the MAC-level ARQ scheme, we varied the retry limit of the AC1 class, while for the CBRL ARQ technique, various combinations of retry limit values ($RL_{I,P}, RL_B$) have been used, in particular, for the case $RL_{I,P} = RL_B + 1$ and $RL_{I,P} = RL_B + 2$. The PSNR performance of the MAC-level and the CBRL ARQ techniques is nearly equivalent and the maximum is achieved when $RL_{I,P}$ is equal to four. For higher retry limit values, the performance slightly decreases due to the higher packet delay caused by the increased network congestion. The higher packet delay, in fact, causes the expiration of the MAC-level timeout of many packets. The used bandwidth, expressed as a percentage of the average bitrate of the original sequence, shows a saturation effect when the retry limit is increased over a certain threshold, that is, about four in our simulations. However, note that the number of packets in queues is limited, although infinite size queues are assumed due to the MAC-level timeout. The corresponding used bandwidth is about 120%, that is, 20% of the bandwidth is used for retransmissions. The best results obtained with the MAC-level and the CBRL ARQ techniques will be used as reference values in the rest of the paper.

6.2. Application-level ARQ techniques: direct transmission (scenario 1)

This section is aimed at investigating the performance of the proposed perceptual ARQ algorithm in scenario 1 by means of performance indicators such as the PSNR value and the used bandwidth. Moreover, the impact of the two main parameters of the algorithm, namely, the peak bandwidth (B_{peak}) and the weighting parameter (w), will also be examined. In addition to the MAC-level and CBRL ARQ techniques, another retransmission scheme, that is, the *Soft ARQ*, will be used as reference to evaluate the performance of the proposed perceptual ARQ algorithm. The *Soft ARQ* proposal [2] assigns the highest retransmission priority to the packet whose playout deadline is the nearest. Packets whose deadline is the same (belonging to the same frame) are retransmitted sequentially. The algorithm implements the same strategy of our proposed application-level ARQ technique to compute retransmission opportunities, but the distortion term in (2) is not considered.

Figure 4 shows the performance in terms of PSNR of the proposed ARQ technique as a function of the peak bandwidth parameter, which is expressed as a percentage of the sequence average bitrate, for different values of the w parameter when the *tempete* sequence is transmitted. The graph also includes the three reference ARQ techniques. The dashed curve shows the performance of the *Soft ARQ* scheme, while the best performance achieved by the

MAC-level and the CBRL ARQ techniques, as determined in Section 6.1 by varying the MAC-level retry limit, is represented by two horizontal lines. In fact, for these two reference techniques, it is not possible to impose a peak bandwidth parameter B_{peak} constraint and the average used bandwidth is only indirectly controlled by means of the retry limit.

The proposed perceptual ARQ technique achieves a consistent performance gain, up to 0.8 dB, with respect to the best performance achieved by the standard MAC-level ARQ technique, while the gain with respect to the best performance of the CBRL ARQ technique is up to 0.5 dB, provided that the B_{peak} parameter is set higher than about 127%. Since this percentage value is computed with reference to the average bitrate of the video sequence, a value less than 127% may impose a bandwidth constraint which is sometimes lower than the instantaneous bitrate of the video sequence. This is the main reason of the performance decrease under the B_{peak} value. Hence, it is recommended that the B_{peak} parameter is set to about 130% (or higher values), as done for a large part of the results presented in this paper. With this B_{peak} setting, the bandwidth usage of the proposed ARQ technique is equal or slightly higher than the one used by the two reference techniques and the impact on concurrent traffic is very limited, as it will be shown later in this section. Hence, the 130% value represents the best tradeoff for video streaming applications in scenario 1.

The *Soft ARQ* performance in Figure 4 shows that for the same B_{peak} parameter the distortion term in (2) plays an important role. In fact, using the perceptual importance of each packet when selecting the packet to retransmit allows to achieve the error-free performance for a smaller peak bandwidth value. Note that the saturation value in the graph is equal to the encoding distortion. Different values of the w parameter have a significant impact on the performance, especially when the B_{peak} parameter is low. In this situation, the best value for the w parameter is zero, thus the packets should be retransmitted based on the perceptual importance only, that is, the distortion term in (2). If the B_{peak} parameter is increased, the best PSNR performance of the proposed technique is achieved for progressively increasing values of the w parameter. This effect can be observed in Figure 5, which presents the PSNR values as a function of the w parameter for different B_{peak} values. Each curve presents a maximum for progressively increasing values of w until flatness when the values reach the error-free PSNR performance. For the recommended B_{peak} value of approximately 130%, performance maximization is achieved using a w value equal to about one.

The results of the same set of experiments for the *paris* sequence are shown in Figure 6. The behavior is similar to the case of the *tempete* sequence. The performance gain is up to 0.5 dB when considering the best performance of the standard MAC-level ARQ technique, and up to 0.8 dB for the case of the CBRL ARQ technique.

Figure 7 shows the results achieved with the *bus* sequence. For this sequence, the performance of the standard MAC-level and the CBRL ARQ techniques are not shown because they are very low compared to the other results. The best performance of the standard MAC-level ARQ technique is 21.21 dB. The CBRL ARQ technique also provides similar

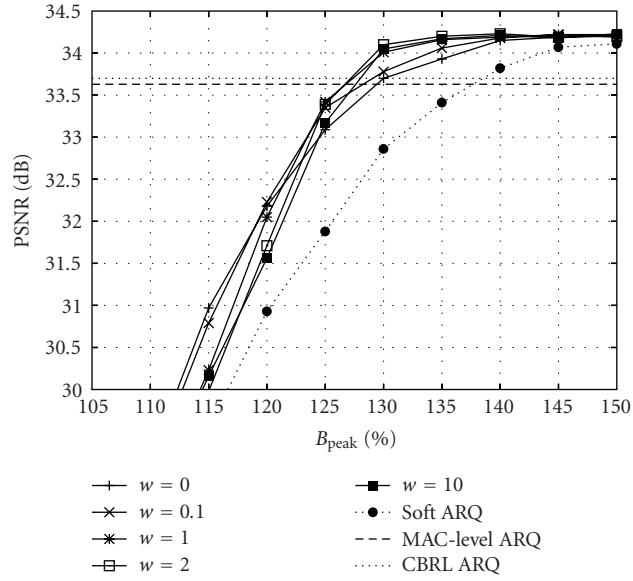


Figure 4: PSNR as a function of the peak bandwidth for the proposed ARQ scheme for different values of the w parameter, compared to the Soft ARQ technique and to the best performance of the MAC-level and CBRL ARQ schemes (*tempete* sequence).

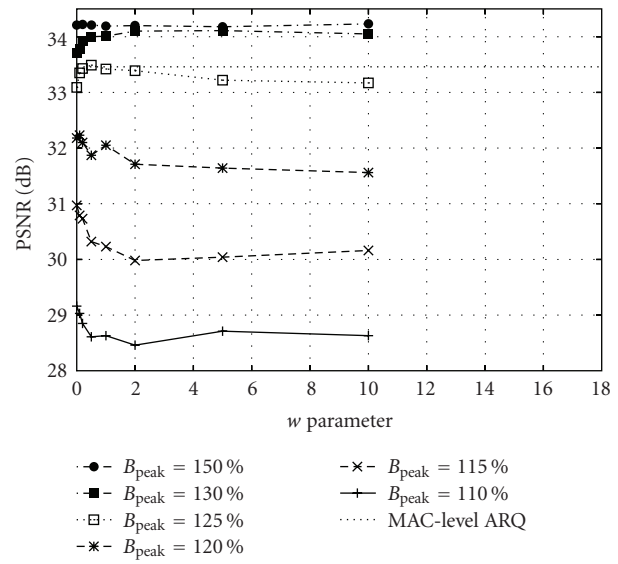


Figure 5: PSNR as a function of w parameter for the proposed perceptual ARQ technique (*tempete* sequence).

performance (23.72 dB), which is well below the acceptable quality threshold. Hence, the gain of the proposed algorithm with respect to the MAC-level ARQ technique is up to 12 dB.

Note that the performance of the MAC-level ARQ schemes for the case of the *bus* and *tempete* sequences is strongly different despite the similar sequence average bitrate. The fact can be explained noting that the *bus* sequence is packetized into about 66% more packets per second compared to the *tempete* sequence (please refer to the last column of Table 1). The higher number of packets of the bus

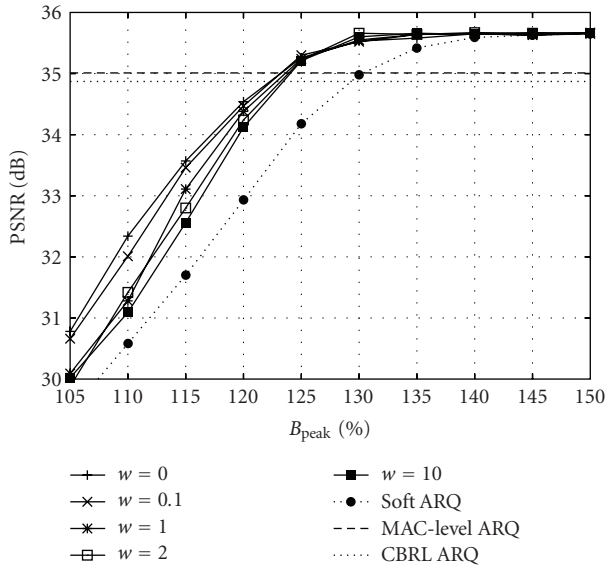


Figure 6: PSNR as a function of the peak bandwidth for the proposed ARQ scheme for different values of the w parameter, compared to the Soft ARQ technique and to the best performance of the MAC-level and CBRL ARQ schemes (*paris* sequence).

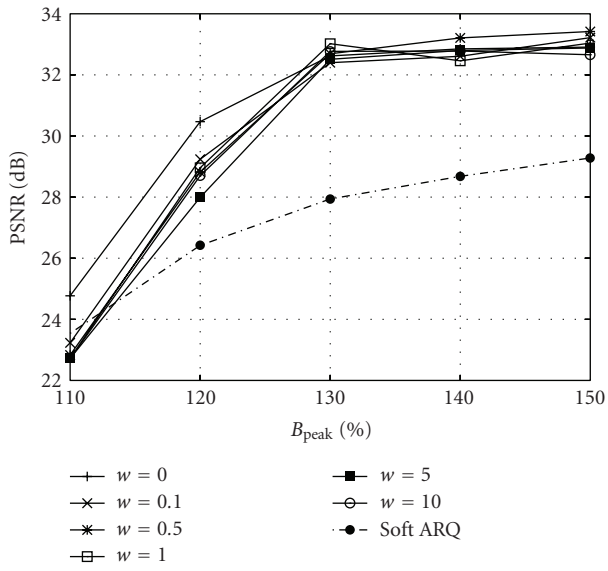


Figure 7: PSNR as a function of the peak bandwidth for the proposed ARQ scheme for different values of the w parameter, compared to the MAC-level ARQ performance (*bus* sequence). The maximum PSNR values achieved by the MAC-level ARQ and the CBRL ARQ techniques are not reported because they are very low.

sequence causes a saturation effect in the 802.11 network. Consequently, the performance drastically decreases. The application-level ARQ algorithms, instead, cause a lower number of network access attempts because packets are never retransmitted at the MAC level, hence good performance can be achieved in this congested scenarios as shown by the results. In this situation, it is very important to use retransmission opportunities for more perceptually important packets, as shown by the curve with w equal to zero. Moreover, when packets have the same playout deadline, the proposed perceptual ARQ technique selects them for retransmission in decreasing order of perceptual importance because in this case the total importance value is influenced only by the left term in (2). The *Soft ARQ* technique, instead, does not exploit the different perceptual importance of the packets and retransmits packets with the same deadline sequentially. This behavior leads to lower performance compared to the perceptual ARQ algorithm.

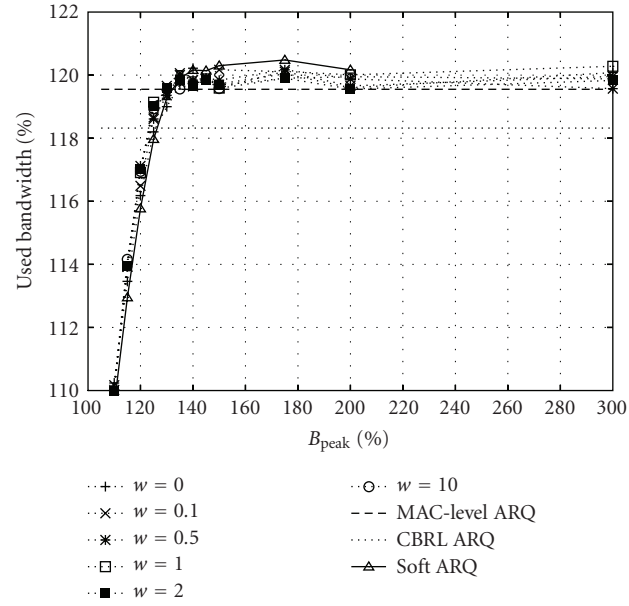


Figure 8: Used bandwidth as a function of the peak bandwidth for the proposed ARQ scheme, compared to the Soft ARQ technique and to the value corresponding to the best PSNR performance of the MAC-level and CBRL ARQ schemes (*paris* sequence).

Figure 8 shows the average bandwidth used by the various algorithms, expressed as a percentage of the sequence average bitrate. Note that the value is much lower than the corresponding peak bandwidth parameter (B_{peak}). The peak transmission bandwidth, in fact, is fully used only when a GOP has much higher bandwidth than the average. Therefore, if B_{peak} is increased, the PSNR gain comes from allowing the algorithm to timely retransmit a higher number of packets when it is more needed. The two horizontal lines in the graph represent the bandwidth usage of the MAC-level and the CBRL ARQ techniques corresponding to their best PSNR performance. On average, the perceptual ARQ algorithm presents bandwidth usage similar to the MAC-level ARQ and *Soft ARQ* techniques and slightly higher than CBRL ARQ technique, up to 2% for the *paris* and *tempete* (not shown) sequences. However, the PSNR performance of the perceptual ARQ algorithm is consistently better than the other algorithms.

Figure 9 shows the average used bandwidth for the *bus* sequence. In this case, the difference is higher in comparison with the MAC-level ARQ (up to 5%) and the CBRL ARQ (up to 9%), but these algorithms are unable to provide an

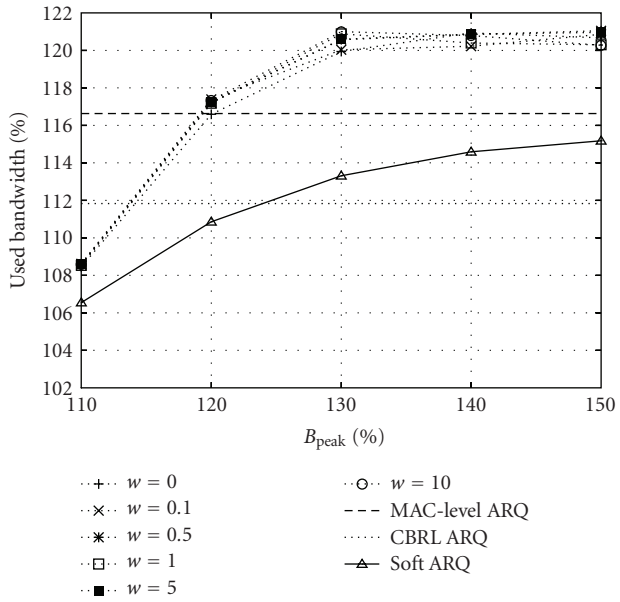


Figure 9: Used bandwidth as a function of the peak bandwidth for the proposed ARQ scheme, compared to the Soft ARQ technique and to the value corresponding to the best PSNR performance of the MAC-level ARQ scheme (*bus* sequence).

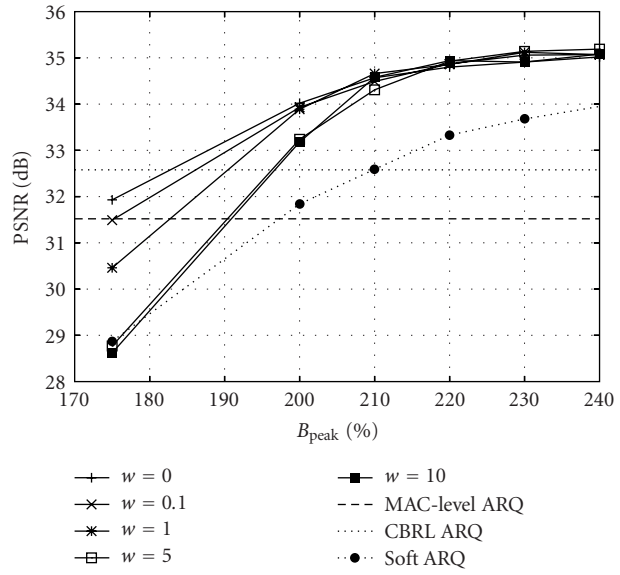


Figure 10: PSNR as a function of the peak bandwidth for the proposed ARQ scheme for different values of the w parameter, compared to the Soft ARQ technique and to the best performance of the MAC-level and CBRL ARQ schemes (*paris* sequence.) Transmission through the relay node.

acceptable video quality. The increase with respect to the *Soft ARQ* technique is up to 8%, but the corresponding increase of the proposed perceptual ARQ algorithm in terms of PSNR gain is significant, up to 5 dB.

6.3. Application-level ARQ techniques: transmission through a relay node (scenario 2)

The performance of the proposed perceptual ARQ algorithm has also been evaluated when a node transmits the H.264 video sequence to another node by means of the access point (scenario 2). In this case, the transmission originates from the PC and is directed to a TV set, as shown in Figure 3. Note that, in case of packet losses, the application-level ARQ techniques (i.e., the proposed perceptual ARQ and the *Soft ARQ*) employ an end-to-end retransmission approach, hence packets are sent again from the PC even if they are lost during transmission from the AP to the TV set. In this scenario, we also simulated the MAC-level and the CBRL ARQ techniques for comparison purposes. However, note that to implement the CBRL ARQ technique in this scenario, some mechanism has to be designed to let the access point know the classification of each video packet so that the correct retry limit for each packet can be used, otherwise the retry limit information would not be available at the access point.

Figures 10 and 11 show the PSNR performance of the proposed technique as a function of the B_{peak} parameter and for different w values. The performance is compared with the other three algorithms. The PSNR gain with respect to the MAC-level ARQ is much more pronounced than in scenario 1, up to 6 dB. With respect to the CBRL ARQ technique, the gain is still significant, up to 4 dB in the *tempete* case.

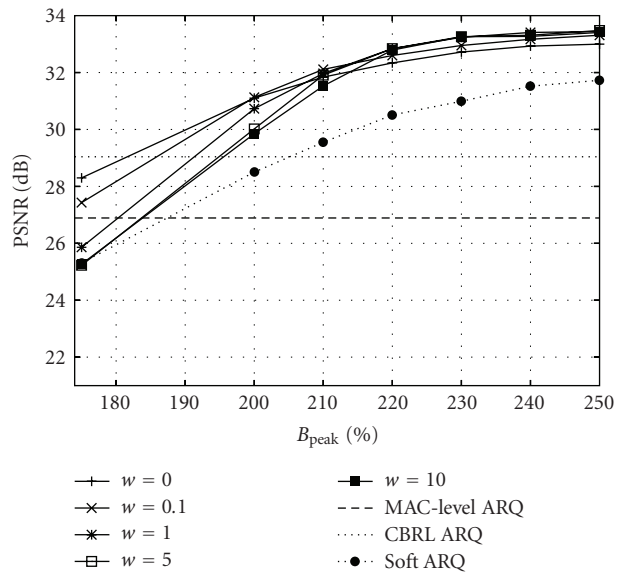


Figure 11: PSNR as a function of the peak bandwidth for the proposed ARQ scheme for different values of the w parameter, compared to the Soft ARQ technique and to the best performance of the MAC-level and CBRL ARQ schemes (*tempete* sequence.) Transmission through the relay node.

The comparison with the *Soft ARQ* algorithm shows that the PSNR gain is up to 3 dB if the value of the w which leads to the best performance is chosen. Note also that the performance of the proposed perceptual ARQ technique is more sensible to the value of the w parameter than in scenario 1. When the B_{peak} parameter is low, the w parameter should

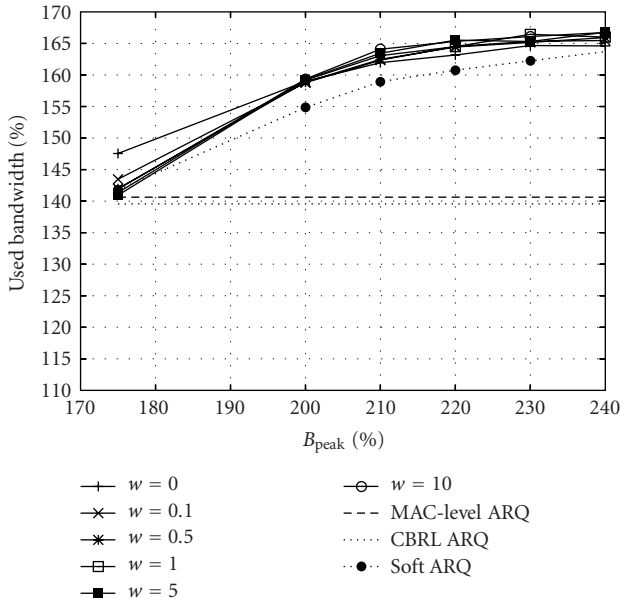


Figure 12: Used bandwidth as a function of the peak bandwidth for the proposed ARQ scheme, compared to the Soft ARQ technique and to the value corresponding to the best PSNR performance of the MAC-level and CBRL ARQ schemes (*paris* sequence). Transmission through the relay node.

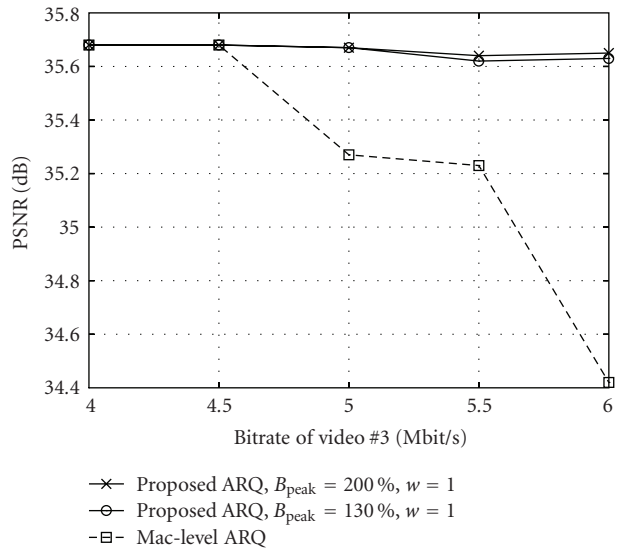


Figure 13: PSNR as a function of the Video #3 bandwidth (varied from 4 to 6 Mbps) for both the MAC-level ARQ and the proposed ARQ scheme; *paris* sequence.

be equal to zero, so that the most perceptually important packets are privileged when a retransmission opportunity is available. The case of the *bus* sequence is not shown because none of the considered techniques is able to achieve an acceptable performance. In this case, all packets in the network experience long access delays due to the large number of packets offered to the network, and this causes a generalized performance decrease.

In Figure 12, the average used bandwidth for the *paris* sequence in scenario 2 is reported. The perceptual ARQ technique shows a used bandwidth increase with respect to the MAC-level and the CBRL ARQ techniques up to about 25% of the sequence average bitrate, however a considerable PSNR gain is provided by the perceptual ARQ technique. Moreover, the difference in used bandwidth is limited, about 5%, while the perceptual ARQ technique can achieve a PSNR gain up to 2 dB (as seen in the previous graph) with respect to the *Soft ARQ* technique. Simulations results thus indicate that the proposed perceptual ARQ technique can be effectively used in an infrastructured scenario to perform video communication between nodes.

6.4. Delays

We now present further results obtained in scenario 1 in which the video sequence under test is transmitted from the access point to the destination node without using any intermediate hop. The perceptual ARQ technique has also been evaluated in terms of the average delay, which is reported in Table 3. The results show that the MAC-level ARQ scheme causes a relatively high delay for the *tempete* and *bus* sequences, about one second or more, which might be annoying in real-time applications. Note also that, for the *bus* sequence in the MAC-level ARQ case, the playout buffer had to be increased to 2 seconds. On the contrary, the perceptual ARQ technique achieves a very low transmission delay, especially in the *paris* and *tempete* cases (50–80 milliseconds). The average delay for the *bus* sequence is slightly higher, about 250 milliseconds, which however greatly improves with respect to the 1.3 seconds average delay of the MAC-level ARQ technique and it allows to use a 1 second playout buffer as in the other cases. Moreover, increasing the peak bandwidth parameter further reduces the transmission delay. Hence, the proposed perceptual ARQ algorithm can be very interesting in scenarios with very strict delay requirements.

Table 3: Average delay (millisecond) for the standard MAC-level ARQ and the proposed perceptual ARQ technique.

Sequence	MAC-level ARQ	Perceptual ARQ	
	retry limit = 4	B_{peak}=130%	B_{peak} = 200%
Paris	152	81	61
Tempete	870	76	49
Bus	1315	258	235

6.5. Influence of the scenario

This section assesses the impact of variations in scenario 1 on the performance of the proposed perceptual ARQ technique. First, the effect of different network congestion levels is evaluated. Figures 13 and 14 show the PSNR performance as a function of the bandwidth of the heaviest video flow (Video #3), varied from 4 Mbit/s to 6 Mbit/s. As illustrated in the graphs, the proposed ARQ technique is only

Table 4: Packet loss rate (%) of concurrent traffic for different retransmission schemes. B_{peak} is equal to 130% for the application-level ARQ schemes. The table shows the highest observed value.

Sequence	Traffic flow	MAC-level ARQ	Soft ARQ	Perceptual ARQ
Paris	Video #1	17.31	17.40	15.64
	Video #2	18.51	18.64	17.40
	Video #3	17.52	17.72	15.69
	VoIP	0.002	0.00	0.253
Tempete	Video #1	19.98	19.93	20.24
	Video #2	21.39	21.27	21.69
	Video #3	20.46	20.59	20.80
	VoIP	0.002	0.00	0.004
Bus	Video #1	19.45	22.16	23.06
	Video #2	21.05	23.74	24.71
	Video #3	20.39	23.60	24.55
	VoIP	0.003	0.00	0.00

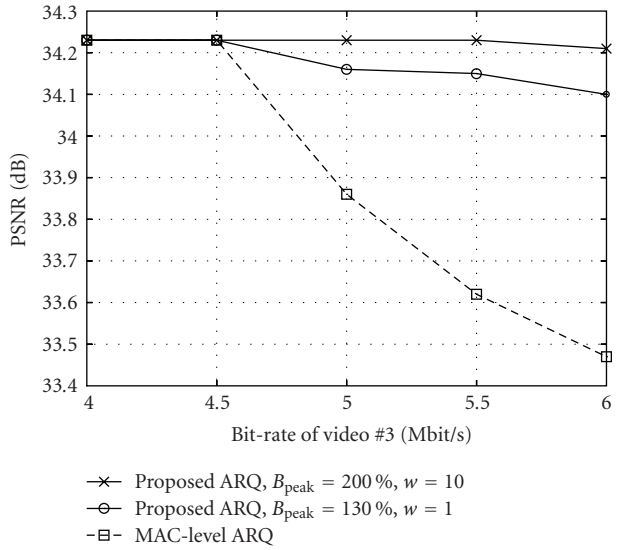


Figure 14: PSNR as a function of the Video #3 bandwidth (varied from 4 to 6 Mbps) for both the MAC-level ARQ and the proposed ARQ scheme; *tempete* sequence.

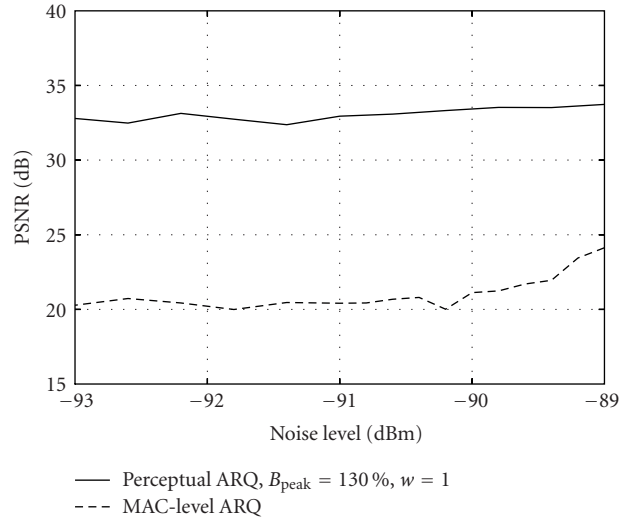


Figure 15: PSNR as a function of channel noise for both the MAC-level ARQ and the proposed ARQ scheme; *bus* sequence.

minimally affected by the augmented network load, while the MAC-level ARQ suffers a sharp decrease in terms of PSNR performance. Note also that the Video #3 transmission is relayed by the access point, hence the actual traffic offered to the network is doubled compared to the nominal bitrate.

Figure 15 compares the PSNR performance of the proposed perceptual ARQ and the MAC-level ARQ techniques as a function of the average channel noise level, for the *bus* sequence. Besides the performance gap between the perceptual and MAC-level ARQ, it is interesting to note that the perceptual ARQ performance is almost constant over the full range of the considered noise level.

The last set of results investigates the impact of the various ARQ techniques on the concurrent traffic. Results are shown in Table 4, in terms of the packet loss rate experienced by the various traffic flows, except the FTP flow which is not included because the throughput it can deliver is very limited and not significant due to the high network congestion. The results compare three techniques, namely, the MAC-level ARQ, Soft ARQ, and the proposed perceptual ARQ, for the three considered video sequences. The shown values represent the highest packet loss rate measured in the simulations. The B_{peak} parameter is set equal to 130%, which is the saturation point for the PSNR performance of the proposed perceptual ARQ technique in scenario 1. For all the three concurrent high-bandwidth video flows, the packet loss rate increase is less than 0.34% when using the perceptual ARQ instead of the MAC-level ARQ technique for the *tempete* sequence. For the *paris* video sequence, the perceptual ARQ technique causes a lower packet loss rate compared to the MAC-level ARQ. The packet loss rate slightly increases in the *bus* case (up to 4.16%); however, the proposed perceptual ARQ technique is able to deliver an acceptable quality while in the same conditions, the degradation using the

MAC-level ARQ is intolerable. Similar considerations hold when the perceptual ARQ technique is compared with the *Soft ARQ* technique, but in this case the packet loss difference is smaller. Finally, the results show that the impact on the VoIP transmission, which is assigned to AC3, that is, the highest-QoS access category, is negligible in all cases.

7. CONCLUSIONS

In this paper, we presented and investigated the performance of a perceptual ARQ algorithm for video streaming over 802.11e wireless networks. The algorithm uses a simple and effective formula in order to combine the perceptual and temporal importance of each packet into a single priority value, which is then used to drive the packet selection process at each retransmission opportunity. Extensive simulations of H.264 video streaming in a heavily congested 802.11e home scenario have been carried out by means of *ns*. The results show that the proposed method consistently outperforms the standard link-layer 802.11 retransmission scheme, with PSNR gains up to 12 dB. Comparisons with a MAC-level ARQ scheme which adjusts the retry limit of each packet based on the frame type and with an application-level deadline-based priority retransmission scheme show that the PSNR gain offered by the proposed perceptual ARQ algorithm is significant, up to 5 dB. Further results indicate that the proposed algorithm presents a very low transmission delay and a limited impact on concurrent traffic. Finally, consistent performance is achieved with various network congestion and channel noise levels.

ACKNOWLEDGMENTS

This work was supported in part by STMicroelectronics. The authors would like to thank the anonymous reviewers for their insightful and constructive comments.

REFERENCES

[1] "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," *ISO/IEC 8802-11, ANSI/IEEE Std 802.11*, 1999.

[2] M. Podolsky, S. McCanne, and M. Vetterli, "Soft ARQ for layered streaming media," Tech. Rep. UCB/CSD-98-1024, Computer Science Division, University of California, Berkeley, Calif, USA, November 1998.

[3] Y. Shan and A. Zakhor, "Cross layer techniques for adaptive video streaming over wireless networks," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '02)*, vol. 1, pp. 277–280, Lausanne, Switzerland, August 2002.

[4] S. H. Kang and A. Zakhor, "Packet scheduling algorithm for wireless video streaming," in *Proceedings of International Packetvideo Workshop (PV '02)*, Pittsburgh, Pa, USA, April 2002.

[5] 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.

[6] E. Masala, D. Quaglia, and J. C. De Martin, "Adaptive picture slicing for distortion-based classification of video packets," in *Proceedings of the 4th IEEE Workshop on Multimedia Signal Processing*, pp. 111–116, Cannes, France, October 2001.

[7] F. De Vito, L. Farinetti, and J. C. De Martin, "Perceptual classification of MPEG video for Differentiated-Services communications," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '02)*, vol. 1, pp. 141–144, Lausanne, Switzerland, August 2002.

[8] S. Aramvith, C.-W. Lin, S. Roy, and M.-T. Sun, "Wireless video transport using conditional retransmission and low-delay interleaving," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 6, pp. 558–565, 2002.

[9] J. Chakareski, P. A. Chou, and B. Aazhang, "Computing rate-distortion optimized policies for streaming media to wireless clients," in *Proceedings of the Data Compression Conference (DCC '02)*, pp. 53–62, Snowbird, Utah, USA, April 2002.

[10] E. Masala and J. C. De Martin, "Analysis-by-synthesis distortion computation for rate-distortion optimized multimedia streaming," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '03)*, vol. 3, pp. 345–348, Baltimore, Md, USA, July 2003.

[11] J. Chakareski and B. Girod, "Rate-distortion optimized packet scheduling and routing for media streaming with path diversity," in *Proceedings of Data Compression Conference (DCC '03)*, pp. 203–212, Snowbird, Utah, USA, March 2003.

[12] P. Bucciol, G. Davini, E. Masala, E. Filippi, and J. C. De Martin, "Cross-layer perceptual ARQ for H.264 video streaming over 802.11 wireless networks," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '04)*, vol. 5, pp. 3027–3031, Dallas, Tex, USA, November-December 2004.

[13] ITU-T Rec. H.264 & ISO/IEC 14496-10 AVC, "Advanced video coding for generic audiovisual services," *ITU-T*, May 2003.

[14] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[15] S. Wenger, "H.264/AVC over IP," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 645–656, 2003.

[16] F. De Vito, D. Quaglia, and J. C. De Martin, "Model-based distortion estimation for perceptual classification of video packets," in *Proceedings of the 6th IEEE Workshop on Multimedia Signal Processing (MMSP '04)*, pp. 79–82, Siena, Italy, September-October 2004.

[17] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: a transport protocol for real-time applications," *RFC 3550*, July 2003.

[18] UCB/LBNL/VINT, "Network Simulator—ns-version 2," http://www.isi.edu/nsnam/ns/, 1997.

[19] IEEE 802 Committee, "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*, September 2005.

[20] Wi-Fi alliance, "WMM specification, v.1.1," http://www.wi-fi.org/, 2006.

Hindawi Publishing Corporation
Advances in Multimedia
Volume 2007, Article ID 68502, 13 pages
doi:10.1155/2007/68502

Research Article

Cross-Layer Design of Source Rate Control and Congestion Control for Wireless Video Streaming

Peng Zhu,[1] Wenjun Zeng,[2] and Chunwen Li[3]

[1] *Hitachi (China) Research and Development Corporation, Beijing 100004, China*
[2] *Department of Computer Science, University of Missouri-Columbia, MO 65211, USA*
[3] *Department of Automation, Tsinghua University, Beijing 100084, China*

Received 30 December 2006; Revised 22 May 2007; Accepted 11 July 2007

Recommended by Zhu Han

Cross-layer design has been used in streaming video over the wireless channels to optimize the overall system performance. In this paper, we extend our previous work on joint design of source rate control and congestion control for video streaming over the wired channel, and propose a cross-layer design approach for wireless video streaming. First, we extend the QoS-aware congestion control mechanism (TFRCC) proposed in our previous work to the wireless scenario, and provide a detailed discussion about how to enhance the overall performance in terms of rate smoothness and responsiveness of the transport protocol. Then, we extend our previous joint design work to the wireless scenario, and a thorough performance evaluation is conducted to investigate its performance. Simulation results show that by cross-layer design of source rate control at application layer and congestion control at transport layer, and by taking advantage of the MAC layer information, our approach can avoid the throughput degradation caused by wireless link error, and better support the QoS requirements of the application. Thus, the playback quality is significantly improved, while good performance of the transport protocol is still preserved.

1. INTRODUCTION

Multimedia streaming over the wireless channels has been a very challenging issue due to the dynamic uncertain nature (e.g., variable available bandwidths and random packet losses) of the wireless channels [1]. To address this problem, many solutions have been proposed, of which congestion control for streaming media at the transport layer and source rate control at the application layer are two basic components.

At the transport layer, congestion control for streaming multimedia is adopted to make sure the users have a fair share of the network resources. Congestion control for the wireless scenario has to differentiate packet loss due to wireless link error from packet loss due to congestion, as the transport protocol needs to decrease the sending rate only when there is congestion in the network [2]. Source rate control is typically adopted at the application layer to optimize the playback quality, subject to the bandwidth constraint provided by the congestion control mechanism and the QoS requirements of the multimedia application (e.g., the end-to-end delay constraint).

However with the traditional layered design principle, source rate control and congestion control are usually designed separately without sufficient communication with each other, which imposes a limitation on the overall system performance. For example, traditional congestion control for streaming multimedia usually needs to smooth their sending rate to help the application achieve smooth playback quality. But this does not work all the time, because the coding complexity of the video frames may change abruptly. Moreover, source rate control alone cannot guarantee the end-to-end delay constraint at any time due to the minimum bandwidth requirement and the quality smoothness constraint of the video source. Actually, the end-to-end delay constraint also imposes constraints on the sending rate, which cannot be well supported in the layered design mode.

The cross-layer design approach, on the other hand, can achieve the better overall system performance, and there have been many cross-layer design solutions proposed for wireless video streaming [3]. However, most of them mainly concern about how to utilize the information from the MAC/physical layer. In this paper, we extend our work in [4] to the wireless scenario, and propose cross-layer design of source rate

control and congestion control for wireless video streaming. Our main contributions are as follows.

(1) We first extend the *QoS-aware* congestion control mechanism-TFRCC (TCP friendly rate control with compensation) that we proposed in [4] to better support the QoS requirements of multimedia applications, to the wireless scenario. We provide a detailed discussion about how to obtain a smooth measurement of the parameters, and how to enhance the overall performance of the transport protocol in terms of rate smoothness and responsiveness.

(2) Based on the above work, we extend our previous work, joint design of source rate control and QoS-aware congestion control, to wireless video streaming. How to enhance the cross-layer design mechanism is discussed and a thorough performance evaluation is conducted under different wireless scenarios. Simulation results show that our approach can significantly improve the playback quality of the application and maintain good long-term TCP-friendliness of the transport protocol, thus optimizing the overall performance.

The remainder of this paper is organized as follows. Section 2 discusses the related work. In Section 3, we briefly introduce our joint design work in [4]. Section 4 describes how to extend the QoS-aware congestion control algorithm to the wireless scenario. Simulation results are presented in Section 5. Section 6 gives the concluding remarks.

2. RELATED WORK

2.1. Congestion control for streaming multimedia

Congestion control for streaming multimedia has to take into account not only the fairness and responsiveness of the transport protocol, but also the smoothness of the sending rate to help the multimedia application achieve better playback quality [5]. A number of TCP-friendly congestion control schemes for the wired channels have been proposed to provide smooth sending rates. These include the window-based schemes [6, 7] and the rate-based schemes which can be further classified into the probe-based [8, 9] and equation-based schemes [10, 11].

The above approaches all assume that every packet loss is an indication of congestion, which is not held for the wireless channels. As in the wireless scenario, packet losses can also be attributed to link error [12]. So these mechanisms cannot be directly applied to the wireless scenario. To overcome this problem, several mechanisms have been proposed to distinguish packet losses due to link errors from those due to congestion [13]. For example, an agent is proposed to be installed at the edge of wired and wireless networks to measure the conditions of these two types of networks separately, thus the wireless loss can be differentiated from the packet loss due to congestion [14, 15]. In [12, 16], the proposed methods focus on differentiating the congestion loss from the erroneous packet loss by adopting some heuristic methods such as interarrival time or packet pair. Such solutions expect a packet to exhibit a certain behavior under network congestion or wireless errors. However, a specific behavior of a packet in the Internet reflects the joint effect of several factors, and it is hard to predict the behaviors of the packets by using a simple pattern. Akan and Akyildiz proposed an equation-based approach—the analytical rate control scheme (ARC) for multimedia traffic in wireless networks [2], which only requires the statistical information of the wireless losses, thus avoiding precise differentiation between the congestion loss and the erroneous packet loss. Chen and Zakhor have proposed a pure end-to-end approach in [17, 18], which creates multiple simultaneous TFRC connections on the same wireless path instead of distinguishing packet losses due to link errors from those due to congestion to fully utilize the bandwidth.

2.2. Source rate control for streaming multimedia

Source rate control at the application layer is to make the source rate match the channel condition to achieve better video quality. At the receiver side, adaptive media playout mechanism is proposed in [19, 20] to make sure the end-to-end delay constraint is met by adaptively varying the playout speed at the receiver. At the sender side, adaptive adjustment of the source rate is proposed based on the channel condition and the QoS requirements of the application. For example, the proportional plus derivative (PD) controller is used in [1] to determine the source rate according to the encoder buffer state. In [21], both the encoder buffer state and the end-to-end delay constraint are considered using a virtual network buffer management algorithm for bitstream switching applications.[1] In [22], a global rate control model is adopted to take into account the encoder buffer state as well as the end-to-end delay constraint of the application. A new transport protocol building upon the TFRC protocol is also proposed to take into account the characteristics (e.g., variable packet size) of the multimedia flows and achieve better performance.

Compared to traditional congestion control and source rate control, our approach is unique in that it provides a more flexible framework to allow a joint decision of source rate and sending rate to optimize the overall system performance.

2.3. Cross-layer design for streaming multimedia

Cross-layer design has been proposed for wireless video streaming to improve the overall system performance [23, 24]. However most of these schemes mainly concern about how to utilize the information from the MAC/physical layer.

Recently, the cross-layer design principle is also introduced to make the congestion control take into account both TCP-friendliness and "multimedia-friendliness." For example, constrained TCP-friendly adaptation framework (CTFAF) is proposed in [9], where the design of the congestion

[1] Bitstream switching can be thought of as a kind of source rate control. In bitstream switching applications, there are several streams with different bit rate for the same content. The sender can intelligently switch to the stream with the appropriate bit rate according to the bandwidth constraint.

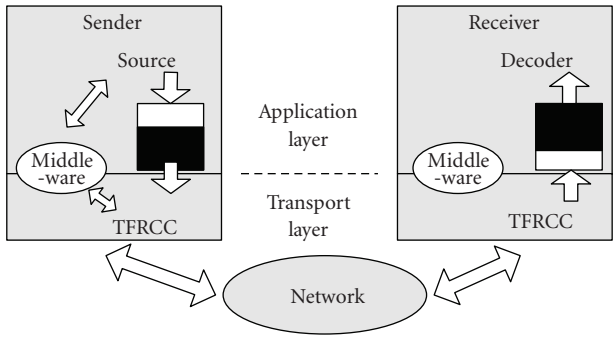


Figure 1: The system architecture.

control algorithm takes into account some QoS constraints (e.g., maximum and minimum bandwidth, rate adaptation granularity) of multimedia applications. In [25], a media-friendly congestion control mechanism is proposed to optimize the overall video quality by using a utility-based model. It adopts a two-timescale approach, which can optimize the video quality in short term (multimedia-friendliness) and meet the TCP-friendliness in long term. In [26], the bandwidth requirement of the multimedia application can be met by achieving proportional fairness with TCP, and multimedia-friendliness is regarded as TCP-friendliness with a weighting factor. The congestion control can adaptively adjust the weighting factor to get the necessary bandwidth and meet the QoS requirements of the multimedia application.

In [4], we propose a joint design algorithm of source rate control at application layer and congestion control at transport layer for streaming video over the Internet. By allowing the sending rate to temporarily violate TCP-friendliness to better support the QoS requirements of the application, the video quality is significantly improved. The long-term TCP-friendliness can be preserved by adopting the rate compensation algorithm. However, the proposed congestion control mechanism has the same assumption that packet loss is the sign of congestion, thus cannot be directly applied to the wireless scenario.

In this paper, we focus on cross-layer design of source rate control and congestion control for wireless video streaming, which can make transport protocol better support the QoS requirements (e.g., end-to-end delay constraint) of the application, and achieve the better overall performance.

3. JOINT DESIGN FOR THE WIRED SCENARIO

3.1. The system architecture

The system architecture is illustrated in Figure 1. At the transport layer, we proposed a QoS-aware congestion control mechanism, named TFRCC (TCP-friendly rate control with compensation), based on TFRC [10]. TFRCC can provide better support for the QoS requirements of the application by allowing temporal violation of TCP-friendliness, while the long-term TCP-friendliness can be preserved by introducing a rate compensation algorithm. At the application layer, the virtual network buffer management mechanism, denoted as VB (as described below), is used to translate the QoS requirements of the application to the constraint of source and sending rates. There is a middleware component located between the application layer and the transport layer. The joint decision of the source rate and the sending rate is done within the middleware at the sender.

3.2. The joint design algorithm

Next we will briefly introduce the joint design algorithm. One can refer to [4] for more details. From the application layer perspective, let us assume a virtual network buffer located between the sender and the receiver that abstracts the potentially complex network topology, and accounts for the delay and loss of packets introduced in the network. Denote $Be(k)$, $Bd(k)$, and $Bv(k)$,[2] respectively, as the encoder buffer, the decoder buffer, and the virtual network buffer occupancies at time k (when frame k is to be placed into the encoder buffer). Let $R(k)$, $Rs(k)$, and $C(k)$, respectively, be the kth video frame size, the amount of data sent by the sender, and the amount of data actually received by the receiver at time k. Denote BE and BD, respectively, as the encoder and decoder buffer sizes, and suppose that N is the end-to-end startup delay (in terms of frame number). Then it can be easily derived that if we can maintain the encoder buffer to meet (1) by selecting appropriate source and sending rates, the overflow and underflow of the encoder and decoder buffers can be avoided:

$$\max\left(0, \sum_{i=k+1}^{k+N} C(i) - Bv(k) - BD\right) \leq Be(k) \leq \min\left(BE, \sum_{i=k+1}^{k+N} C(i) - Bv(k)\right). \tag{1}$$

Let us count the feedback intervals of TFRCC as K. At time k, by using the nominal sending rate of current feedback interval $Ri(K)$ (bytes/frame) to estimate the receive rates of the future N frame periods in (1), we can derive the following two bounds for $Be(k)$ according to (1):

$$B_u = \min\left(N * Ri(K) - Bv(K), BE\right), \quad B_l = \max\left(0, N * Ri(K) - Bv(K) - BD\right). \tag{2}$$

Note that we will place extra safety margins for the bounds in the implementation to deal with the possible estimation error of $Bv(K)$ caused by the possible feedback loss or other factors. The above constraints are derived by VB at application layer.

[2] $Bv(k)$ can be estimated according to the difference between the amount of data sent at the sender and the amount of data received at the receiver.

At transport layer, TFRCC first uses the same algorithm as TFRC to calculate the TCP-friendly sending rate $B(K)$ (bytes/s). Note that the actual sending rate of TFRCC $Rs(k)$ is allowed to temporally violate TCP-friendliness, so TFRCC uses a rate compensation algorithm, based on the TCP-friendly sending rate $B(K)$ and the accumulated difference between the amount of data actually sent and the ideal TCP-friendly value, to determine the nominal sending rate $Ri(K)$ so as to preserve long-term TCP-friendliness.

Then with the encoder buffer constraint of (2) provided by VB and the long-term TCP-friendliness constraint of $Ri(K)$ provided by TFRCC, the source rate and sending rate are jointly determined in the middleware component of the sender.

3.2.1. Decision of the source rate and the sending rate

The actual sending rate $Rs(k)$ is usually set to $Ri(K)$ for good TCP-friendliness. Then the source rate is determined to maintain the encoder buffer within the bounds of (2), together with the consideration of the video quality smoothness constraint and the minimum acceptable/maximum necessary video quality of the video source.

3.2.2. Adaptation at the beginning of a new feedback interval

Suppose at time k_1, the sender receives a new feedback from the receiver, then the sending rate is updated as $Ri(K+1)$. Note that there exists the maximum admissible sending rate constraint, which is imposed by the encoder and decoder buffer sizes. If $Ri(K+1)$ is too large so that $N * Ri(K+1) > BE+BD+Bv(K+1)$, we can find that it will lead to $B_u < BE < B_l$ according to (2), which consequently causes the decoder buffer to overflow. In this case, we need to decrease the sending rate to make sure the maximum sending rate constraint is met (i.e., making $B_l < BE$). The corresponding change of the amount of sent data caused by the sending rate adjustment will be recorded and compensated later.

Moreover, at times $k_1-N,\ldots,k_1-1$, the estimation of the future receive rates using $Ri(K)$ might not have been accurate and the constraints of (1) might not actually be met, since the sending rate after time k_1 has been changed to $Ri(K+1)$. So if necessary, the readjustment of the size of the encoded frame $k_1-N,\ldots,k_1-1$ (if still available in the encoder buffer), subject to the quality smoothness constraint, is used to make sure that the decoder buffer at times $k_1,\ldots,k_1+N-1$ will not underflow and overflow. If this cannot prevent the decoder buffer from underflow or overflow, we will have to adjust the sending rate to pull back the decoder buffer fullness to within the safety region. For example, if the decoder buffer will underflow, we will temporarily increase the sending rate (i.e., making $Rs(k)$ larger than $Ri(K)$) to meet the end-to-end delay constraint of the application. This temporal adjustment of the sending rate will lead to un-TCP-friendliness, and the corresponding change of the amount of sent data will be recorded and compensated later.

4. CROSS-LAYER DESIGN FOR THE WIRELESS SCENARIO

In this section, we discuss how to extend our joint design mechanism for the wired scenario to the wireless scenario. In our previous work, TFRCC uses the same algorithm as TFRC to calculate the TCP-friendly sending rate $B(K)$ and has the same assumption that packet loss is the sign of congestion, so it cannot be directly applied to the wireless scenario. To overcome this problem, we incorporate ARC [2]—an equation-based congestion control mechanism for the wireless scenario—into our framework, which means that we use the same algorithm as ARC to calculate the TCP-friendly rate $B(K)$. Note that any other equation-based congestion control mechanism for the wireless scenario or any work that can differentiate the congestion loss from the erroneous packet loss can also be incorporated into our framework. To differentiate the extended work from the original work of [4], we denote the modified congestion control mechanism as TFRCC-W.

4.1. ARC

ARC is an equation-based mechanism. It models the ideal behavior of the TCP source over lossy links (i.e., reducing the send rate if packet loss is due to congestion, while performing no rate change if packet loss is due to wireless link error), and derives the following throughput formula:

$$B = \frac{s}{4 * \mathrm{RTT}}\left(3+\sqrt{25+24\left(\frac{1-\omega}{\pi-\omega}\right)}\right), \tag{3}$$

where B is the sending rate in bytes/s, s is the packet size, RTT is the round-trip time, ω is the packet loss ratio due to wireless link error, and π is the overall packet loss ratio (including packet losses due to congestion and wireless link error). Then the sender will perform rate control according to (3) to avoid the unnecessary rate reduction due to wireless link error. Note that the overall loss ratio π can be measured at the receiver, and the wireless loss ratio ω can be retrieved from the underlying MAC layer at the sender if the first link is wireless link. For the case in which the sender is not a mobile station, the information regarding the wireless portion of the end-to-end path, that is, the wireless loss ratio ω, should be conveyed to the sender through the feedback.

4.2. Details of TCP-friendly rate calculation

To make the sending rate change smoothly, we need to perform a smooth measurement of the parameters used in (3), which is not discussed in [2]. Here we propose to use the weighted average value over the last N feedback interval to obtain a smooth estimation of the loss ratio. Instead of directly smoothing ω and π, we define the "loss interval" l as

$$l = \frac{1-\omega}{\pi-\omega}, \tag{4}$$

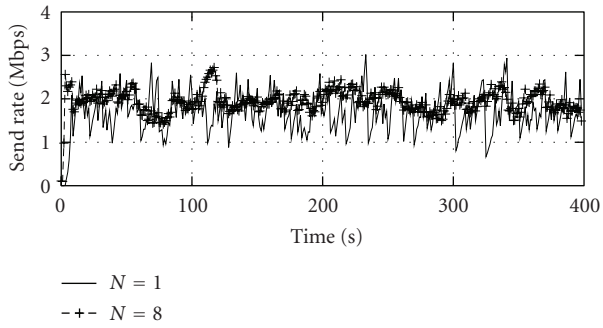


Figure 2: The sending rate of TFRCC-W.

and then compute the average "loss interval" $\hat{l}$[3] as

$$\hat{l} = \sum_{i=1}^{N} m_i * l_i, \tag{5}$$

where m_i is the weight assigned to the ith previous feedback interval,[4] π_i and ω_i are, respectively, the measured overall loss ratio and wireless loss ratio of the ith previous feedback interval, and l_i is computed according to π_i, ω_i, and (4). We first set N to 8, and use the following weights: $m_1, \ldots, m_4 = 1/6$; $m_5 = 2/15$; $m_6 = 0.1$; $m_7 = 1/15$; $m_8 = 1/30$. Then the TCP-friendly sending rate B can be calculated according to (3) and $\hat{l}$, which is computed according to (5) and (4).

We need to deal with the situation where the newest measurement ω^0 is no less than π^0 to avoid "divided-by-zero" problem. This means that there is no packet loss due to congestion within current feedback interval or there is a measurement error. In this case, we cannot directly use (4) any more. So we first let $\omega^0 = \pi^0$, then compute the "loss interval" by combining current interval and last interval. Denote the number of packets sent within current interval and last interval, respectively, as Num^0 and Num_1. Then we update ω_1 and π_1 as follows:

$$\begin{aligned} \omega_1 &= \frac{\omega^0 * \text{Num}^0 + \omega_1 * \text{Num}_1}{\text{Num}^0 + \text{Num}_1}, \\ \pi_1 &= \frac{\pi^0 * \text{Num}^0 + \pi_1 * \text{Num}_1}{\text{Num}^0 + \text{Num}_1}; \end{aligned} \tag{6}$$

l_1 can be updated according to the updated ω_1 and π_1.

Figure 2 shows the sending rate curves of one TFRCC-W flow without parameter smoothing (i.e., $N = 1$) and one flow with parameter smoothing ($N = 8$) when they compete for a bottleneck. It can be found that the protocol has satisfactory performance in terms of rate smoothness by using the proposed measurement mechanism.

[3] Note that the reason of not directly smoothing ω and π is that even with smoothed ω and π, l might sometimes be still not smooth enough to make the sending rate change smoothly, especially when the packet loss ratio due to congestion is very small.

[4] Note that we use the same weighing factors as TFRC [10], which have been proven to have good smoothing effect.

4.3. Performance enhancement of TFRCC-W

The responsiveness and rate smoothness of TFRCC-W depend largely on the sample count N. It is easily understood that with a large value of N, better rate smoothness can be achieved, but the responsiveness of the protocol will deteriorate; while with a small value of N, the protocol will have better responsiveness and worse rate smoothness. Figure 3(a) depicts the sending rate of one TFRCC-W flow with setting N to 3 and one with setting N to 16 when they compete for a bottleneck. It can be easily seen that compared to the flow with setting N to 3, the flow with setting N to 16 has better rate smoothness when the network is underloaded, but shows worse responsiveness when there is a sudden decrease of the available bandwidth.

So we need to investigate how to choose an appropriate value of N to achieve the reasonable tradeoff between the responsiveness and the rate smoothness. Obviously, a constant value of N is not the optimal choice. When the available bandwidth is high, the packet loss ratio is so low that the time interval between two consecutive packet losses is very long. So at this time, N should be sufficiently large to allow sufficient samples of packet loss for effective smoothing. On the other hand, when the available bandwidth is low, the packet loss ratio increases. A large value of N will make the packet loss samples too large to reflect the recent network condition, thus deteriorate the responsiveness of the protocol. At this time, the value of N is expected to be small.

Let us first see how TFRC [10] resolve this problem. TFRC uses the following formula to calculate the TCP-friendly sending rate:

$$B = \frac{s}{R\sqrt{2p/3} + 4R \min\left(1, 3\sqrt{3p/8}\right) p(1 + 32p^2)}, \tag{7}$$

where B is the sending rate in bytes/s, s is the packet size, R is the round-trip time, p is the steady-state loss event rate.

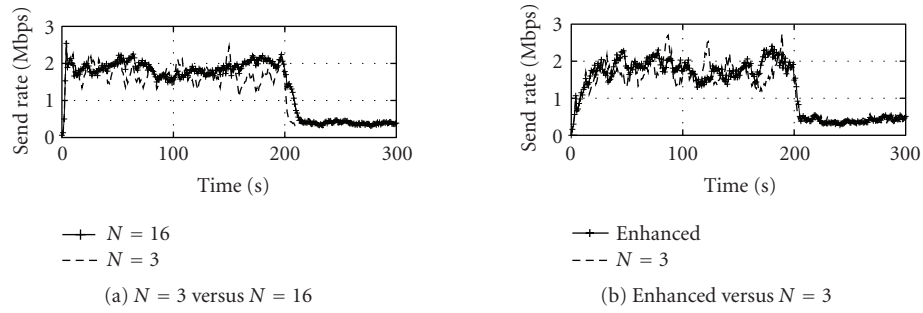
A loss event is defined as the first packet loss every one RTT, and p is measured in terms of loss intervals, spanning the number of packets between consecutive loss events. The most recent M (usually set to 8) loss intervals are averaged, using decaying weights to get the smoothed value $\hat{l}_t$. The loss event rate p is calculated as the inverse of the average loss interval $\hat{l}_t$.

The average time interval between two consecutive sending packets, t_p, is the inverse of the packet rate (packet/s), that is, B/s. So we have

$$t_p = \frac{1}{B/s} = R\sqrt{\frac{2p}{3}} + 4R \min\left(1, 3\sqrt{\frac{3p}{8}}\right) p(1 + 32p^2). \tag{8}$$

Let us denote the average time interval between two consecutive loss events as t_l. Then we can get

$$\begin{aligned} t_l &= \hat{l}_t * t_p = t_p / p \\ &= R\sqrt{\frac{2}{3p}} + 4R \min\left(1, 3\sqrt{\frac{3p}{8}}\right)(1 + 32p^2). \end{aligned} \tag{9}$$



(a) $N = 3$ versus $N = 16$ (b) Enhanced versus $N = 3$

Figure 3: Performance comparison between TFRCC-W flows with different values of N.

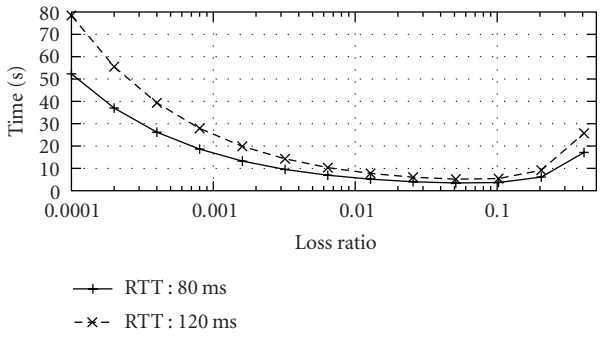


Figure 4: The sampling time length of TFRC.

So we can get the sampling time length of TFRC as

$$L_{\text{TFRC}} = 8 * \left(R\sqrt{\frac{2}{3p}} + 4R \min\left(1, 3\sqrt{\frac{3p}{8}}\right)(1 + 32p^2)\right). \tag{10}$$

Then we plot the $L_{\text{TFRC}} - p$ curve (see Figure 4) when RTT is, respectively, set to 80 milliseconds and 120 milliseconds. From Figure 4, we can find that the sampling time length of TFRC is variable and depend on the network condition, which helps TFRC achieve reasonable tradeoff between the rate smoothness and responsiveness.

So in the enhanced version of TFRCC-W, we adopt the same sampling time length as TFRC, that is, first calculate the sample time length of TFRC according to (10),[5] then determine N according to the derived time length value. Figure 3(b) depicts the sending rate of one enhanced TFRCC-W flow and one with N of 3 when they compete for a bottleneck in the same scenario as in Figure 3(a). It can be easily seen that the enhanced TFRCC-W has better overall performance in terms of rate smoothness and responsiveness.

[5] Note that in (10), p means the packet loss ratio only due to congestion. So in the computation, we do not use the overall packet loss ratio π, but $1/l$, which means the packet loss ratio due to congestion (see details in [2]).

5. PERFORMANCE EVALUATION AND ENHANCEMENT FOR OUR PROPOSED CROSS-LAYER MECHANISM

5.1. Performance evaluation of TFRCC-W

We use NS-2 simulator [27] to investigate the performance of TFRCC-W. Note that in this subsection, we mainly concern about the performance of the transport layer. So TFRCC-W adapts its sending rate only according to (3), and does not consider how to support the QoS requirements of the application (i.e., as FTP applications run). We use the well-known dumbbell topology, where one TFRC or TFRCC-W flow and several TCP flows compete for one bottleneck link with a capacity of 15 Mbps and a transmission delay of τ millisecond. We assume that the senders of TFRC or TFRCC-W flows connect to the bottleneck via wireless links with the wireless loss ratio p_w, that is, the first link of every TFRC/TFRCC-W flow is wireless. The feedback interval of TFRCC-W is fixed to 1 second. We set the packet drop ratio in the bottleneck caused by congestion to be 0.001, and the transmission delay τ is, respectively, set to 50 milliseconds and 100 milliseconds. Then we record the average throughput of one TFRCC-W and one TFRC flows when they, respectively, run through the bottleneck with the wireless loss ratio p_w increasing from $1e-6$ to 0.01. Simulation results are depicted in Figure 5, and every point in the figure is the average value of ten runs.

The results show that TFRC and TFRCC-W have similar throughput when the wireless loss ratio is very low, which means that TFRCC-W has good TCP-friendliness in the wired scenario. With the increasing of the wireless loss ratio, TFRCC-W shows much better performance than TFRC, and can effectively avoid the throughput degradation caused by the wireless link error.

5.2. Performance evaluation of the proposed cross-layer mechanism

In this subsection, we investigate the performance of our proposed cross-layer design mechanism using NS-2 simulation. The simulation topology is depicted in Figure 6, where m multimedia mobile terminals are connected to the IP backbone via wireless access point. In the backbone, there are three links (R1-R2, R2-R3, and R3-R4), each of which has a capacity of R Mbps and a transmission delay of τ millisecond. All the wireless links have the same loss ratio of 0.5%. To

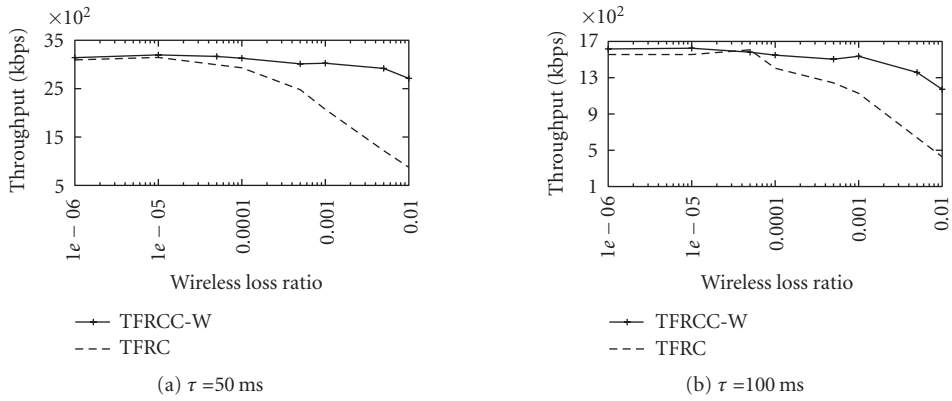


Figure 5: Performance comparison between TFRCC-W and TFRC in the wireless scenario.

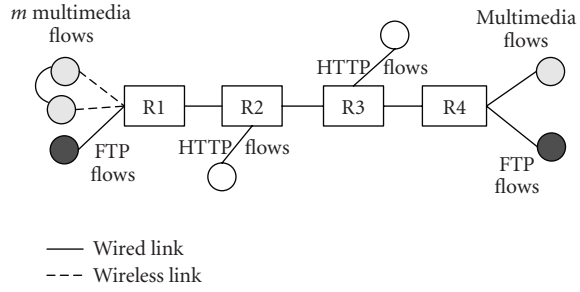


Figure 6: The simulation topology.

simulate the real Internet environment, various cross-traffic (FTP flows and WWW flows) combinations have been used. The FTP flows are modeled as greedy sources which can always send data at the available bandwidth. The WWW flows are modeled as on-off processes.

The standard video sequence of "Foreman" (300 frames) in QCIF format is circularly used as the video source, and is encoded using an MPEG-4 fine granularity scalable (FGS) coder [28]. The encoder uses interframe coding (with the GOP size of 10 and the frame type of I and P) and quantization step size of 31 to generate the base layer, which provides the minimum video quality. Then the FGS coder generates the embedded enhancement layer bitstream, which can be cut off at any bit to adapt the source rate with fine granularity. The frame rate is set to 25 frames per second (fps). The maximum necessary PSNR is set to 40 dB. We packetize the base layer and enhancement layer separately, and the MSS (maximum segment size) is set to 1000 bytes. In this paper, we use a simple error resilience algorithm. If the base layer of some frame is lost or late, the base layer of the previous frame will be used in decoding. If there is a packet loss in the enhancement layer, all less important packets in that frame will be discarded as they all depend on the lost, more important packet. Note that in MPEG-4 FGS, loss of enhancement layer packets is not going to cause distortion propagation to subsequent frames.

We compare the performance of three source rate/congestion control algorithms. One uses the global rate control model proposed in [22] with TFRC as the congestion control mechanism, denoted as GM-TFRC, and one uses the global model and ARC [2], denoted as GM-ARC. The other is our proposed algorithm, denoted as VB-TFRCC-W. Note that GM-TFRC and GM-ARC belong to traditional separate design approaches. For fair comparisons, all of the congestion control mechanisms have the same feedback interval of 1 second. In Section 5.2.1, we mainly intend to show that VB-TFRCC-W can provide better QoS support for the application. The overall system performance will be evaluated in Section 5.2.2.

5.2.1. Support of the QoS requirements of the application

In this scenario, each of the three links (R1-R2, R2-R3 and R3-R4) has a capacity of 14 Mbps, a transmission delay of 40 milliseconds, and an RED queue with the maximum threshold of 120 packets. Mobile terminals are supposed to be the receivers of the multimedia flows, that is, the last links of all the multimedia flows are wireless. The startup delay is set to a large value, that is, 125 frames (5 seconds), and the encoder and decoder buffer sizes are both set to 400 kB. We adopt a dynamic scenario, which lasts 600 seconds. There are 3 GM-TFRC flows, 3 GM-ARC flows, 3 VB-TFRCC-W flows, and 3 FTP flows running throughout the entire simulation. As the background flows, 100 FTP flows join at around 350 seconds, and 10 WWW flows join at 300 seconds. Here we use the average PSNR and PSNR deviation to evaluate the video quality, where the average PSNR deviation of one video sequence is calculated by averaging the PSNR difference between every two adjacent frames.

Because TFRC is mainly designed for the wired channels, it will reduce the sending rate as long as there is one packet loss. As a result, it shows poor TCP-friendliness when there exist wireless packet losses (see Figure 8). Consequently, the low throughput leads to poor video quality (see Table 1[6]). ARC and TFRCC-W, on the other hand, can take

[6] Note that all the entries in Table 1 are the average value of all the flows using the same source rate/congestion control algorithm.

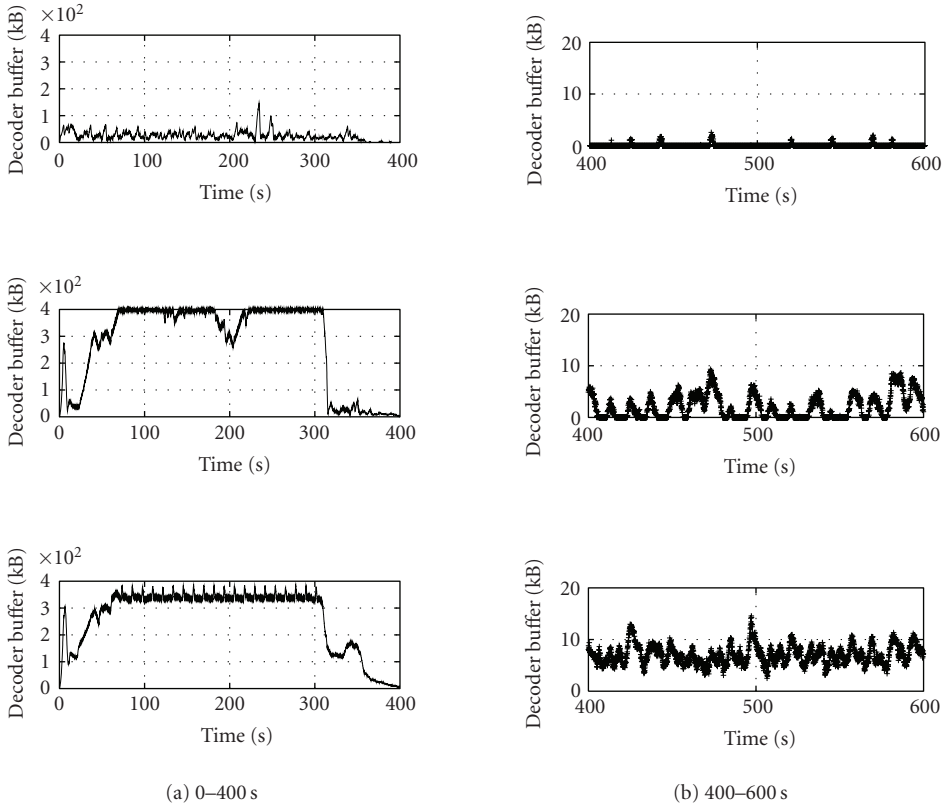


Figure 7: The decoder buffer occupancies of one GM-TFRC flow, one GM-ARC flow, and one VB-TFRCC-W flow; top: GM-TFRC, middle: GM-ARC, bottom: VB-TFRCC-W.

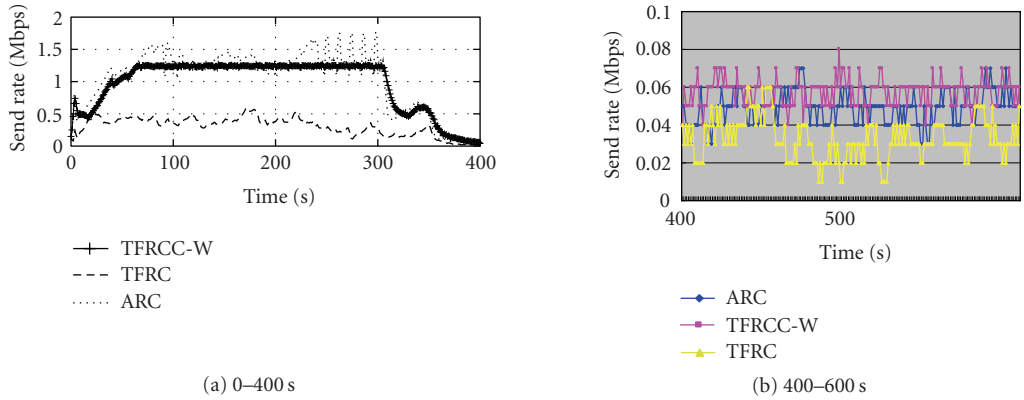


Figure 8: The sending rate of one GM-TFRC flow, one GM-ARC flow, and one VB-TFRCC-W flow.

into account the effect of wireless losses, and achieve higher throughput.

Furthermore, our cross-layer design approach can provide better support for the QoS requirement of the application than GM-ARC. Within the first 300 seconds, the network is underloaded, and the available bandwidth may occasionally be higher than the maximum admissible sending rate constrained by buffer sizes (see the discussion in Section 3.2.2). So from Figure 7, we can find that the decoder buffer of GM-ARC overflows. VB-TFRCC-W, on the other hand, takes into account this sending rate constraint (see Figure 8), and successfully avoids the decoder buffer overflow. With the joining of 100 FTP flows around 350 seconds, the available bandwidth becomes so low that source rate control alone cannot guarantee the end-to-end delay constraint being met because of the minimum bandwidth requirement and quality smoothness constraint of the video source. So the decoder buffer underflow occurs for GM-ARC (see Figure 7). However VB-TFRCC-W can meet the end-to-end delay constraint by making the sending rate temporarily larger than

Table 1: PSNR of GM-TFRC, GM-ARC, and VB-TFRCC-W.

	Sender side		Receiver side	
	PSNR	Variation	PSNR	Variation
GM-TFRC	29.67	0.26	24.70	0.35
GM-ARC	33.75	0.2	30.05	0.97
VB-TFRCC-W	33.51	0.17	32.74	0.57

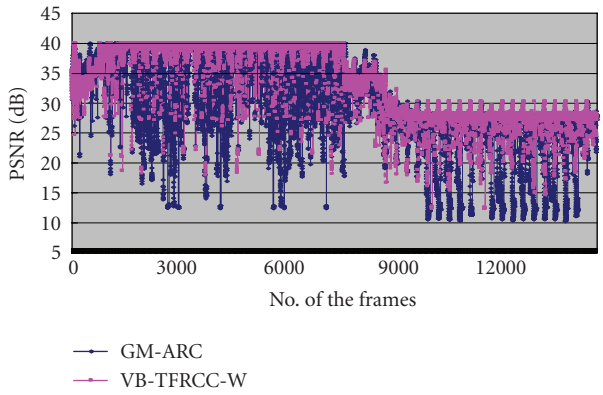


Figure 9: The PSNR curve of the sequences decoded by the decoder of one GM-ARC flow and one VB-TFRCC-W flow.

the TCP-friendly value when necessary (see Figure 8). So VB-TFRCC-W can almost avoid the decoder buffer underflow. As a result, VB-TFRCC-W can significantly reduce the video quality degradation due to lost/late packets between the sender and the receiver, and achieve better playback quality (higher average PSNR and smoother PSNR variation) than GM-ARC (see Table 1 and Figure 9). Note that very low PSNR values (e.g., less than 25 dB) in Figure 9 typically indicate an effective loss of base layer packet for a frame, which introduces significant quality degradation for the lost frame and the subsequent frames.

5.2.2. *Overall system performance evaluation*

We make a comparison for the overall system performance of GM-ARC and VB-TFRCC-W. Each of the three links (R1-R2, R2-R3, and R3-R4) has a capacity of 10 Mbps. We would like to test the overall system performance, respectively, under RED queues and under Drop Tail queues. When the queues of the three links (R1-R2, R2-R3 and R3-R4) are RED, the startup delay is set to a small value, that is, 15 frames, and the encoder and decoder buffer sizes are both set to 100 kB. When the queues are Drop Tail, the startup delay is set to 125 frames (5 seconds), and the encoder and decoder buffer sizes are both set to 500 kB. Mobile terminals are supposed to be the senders of the multimedia flows, that is, the first links of all the multimedia flows are wireless. In this scenario, we would like to simulate the scenario where the available bandwidth is very low. The simulation lasts 600 seconds. There are 5 GM-ARC flows, 5 VB-TFRCC-W flows, and 5 FTP flows running throughout the entire simulation. As the background flows, 70 FTP flows join at 50 seconds, and depart at 300 seconds.

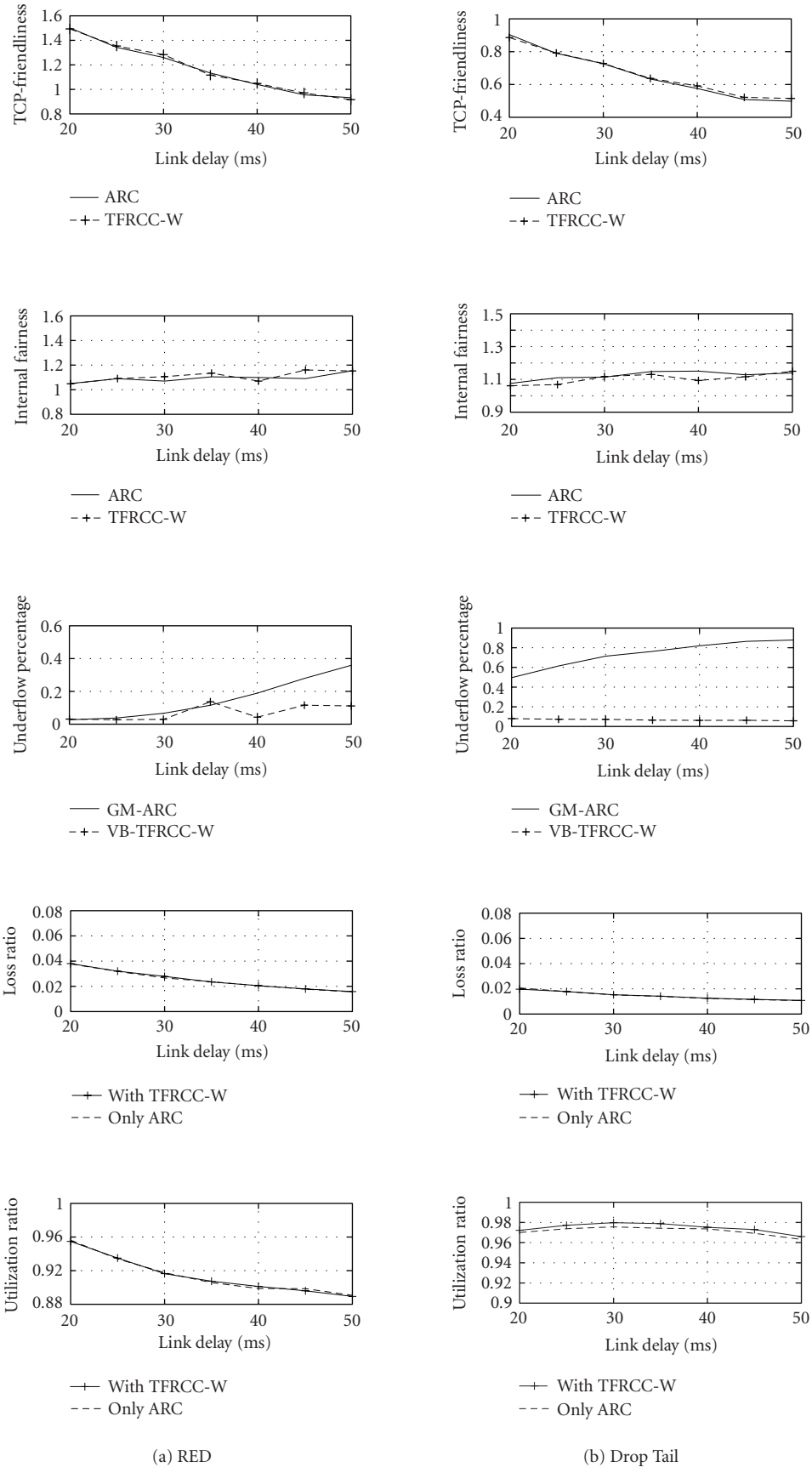
To evaluate the long-term TCP-friendliness and internal fairness (i.e., the fairness among the flows using the same congestion control mechanism) of the transport protocol, we adopt the metrics defined in [29, Chapter 4], where a value close to 1 indicates a good TCP-friendliness or internal fairness. The underflow percentage of the decoder buffer between 50 seconds to 300 seconds (i.e., when the available bandwidth is low), is used to evaluate the support for the end-to-end delay constraint of the application. The underflow percentage of the decoder buffer is computed as the percentage of the frames which are lost or arrive at the decoder buffer later than the prescribed time. We run the simulations under different link transmission delay τ (varying from 20 milliseconds to 50 milliseconds), and simulation results are depicted in Figure 10. Note that every point in Figure 10 is the average value of 5 runs. From the simulation results, it can be found that with the increase of the link delay τ, the TCP-friendliness (i.e., the throughtput) of GM-ARC decreases for both the RED case and the DropTail case. So the underflow percentage of the decoder buffer increases for GM-ARC. VB-TFRCC-W, on the other hand, can provide better QoS support for the application and maintain the underflow percentage of the decoder buffer at a low level, although its throughput also decreases. For the performance of the transport protocol, VB-TFRCC-W shows similar long-term TCP-friendliness and internal fairness as GM-ARC.

We also concern about how the overall network performance is affected with different mechanisms. Here we use the overall packet loss ratio introduced in *the wired channels*, and the utilization ratio of the bottleneck bandwidth to evaluate the overall network performance. We replace 5 VB-TFRCC-W flows with 5 GM-ARC flows (which means that there are 10 GM-ARC flows in the simulations) and repeat the above simulations. Then we compare the results to the previous simulation results when there exist VB-TFRCC-W flows. From Figure 10, we can find that there is almost no difference in the overall network performance between using GM-ARC and VB-TFRCC-W. So our proposed algorithm will not deteriorate the overall network performance although it sometimes exhibits temporal un-TCP-friendly behaviors.

5.3. *Enhancement of the cross-layer mechanism*

Our cross-layer mechanism allows the sending rate to temporarily violate TCP-friendliness to support the QoS requirements of the application, as described in Section 3. When the TCP-friendly bandwidth is too low to make sure the end-to-end delay constraint of the application is met, the sending rate can be temporarily larger than the TCP-friendly value to help the application meet the end-to-end delay constraint, which effectively prevent the decoder buffer from underflow.

However this does not work well under all the network scenarios. Let us suppose that the network is congested and the TCP-friendly bandwidth is B. The minimum bandwidth which can help the application to meet the end-to-end delay constraint is assumed to be RI ($RI > B$). Suppose that the application adopts our cross-layer mechanism and sends the data with the rate of RI. Obviously,



(a) RED

(b) Drop Tail

Figure 10: Overall system performance evaluation between GM-ARC and VB-TFRCC-W.

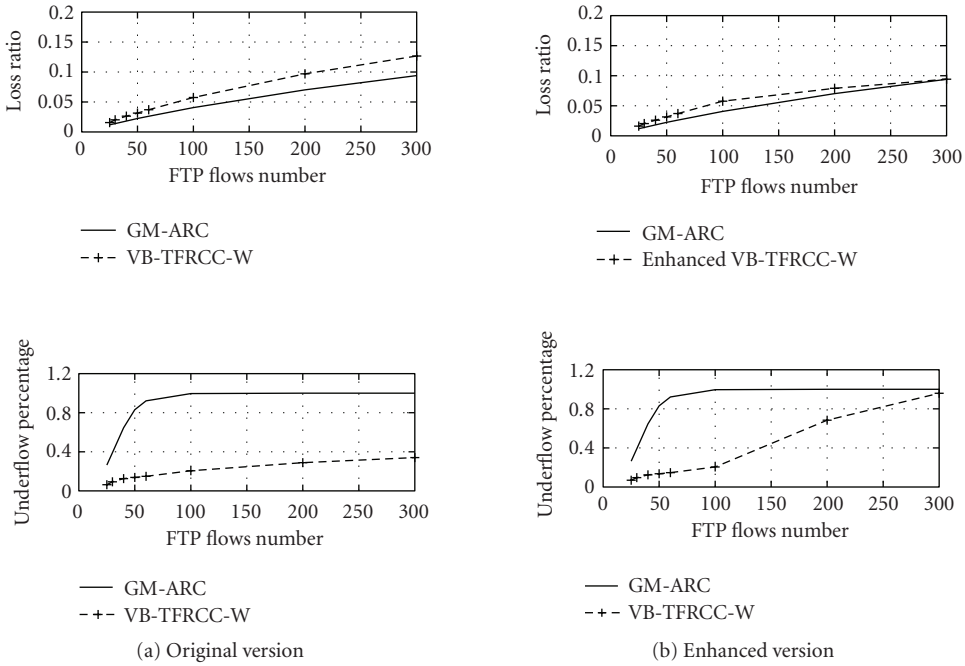


Figure 11: Performance evaluation for the enhanced VB-TFRCC-W.

our cross-layer mechanism can achieve good performance only when the additional sent data beyond the TCP-friendly value (i.e., $RI - B$) can go through the network and be received by the receiver. However, when the network is heavily congested, its effective throughput may not afford such additional data going through the network, which cannot prevent the decoder buffer from underflow. On the other hand, such additional data will deteriorate network congestion. We use *ns*-2 simulation to evaluate the above analysis. The simulation environment is the same as in Section 5.2. The capacity R of the three links (R1-R2, R2-R3, and R3-R4) is set to 15 Mbps, and the Drop Tail queue is used. Ten GM-ARC or VB-TFRCC-W flows compete the bottleneck link with N FTP flows (N increases from 25 to 300), and we record the underflow percentage of decoder buffer and the overall packet loss ratio in the wired links, respectively, when GM-ARC flows run and when VB-TFRCC-W flows run. The simulation results are depicted in Figure 11(a), and note that every point in the figure is the average value of three runs. We can find that when the network is slightly or moderately congested (i.e., N is not very large), our cross-layer mechanism can maintain the underflow percentage of the decoder buffer at a low level, as opposed to GM-ARC. But with the increasing of N, the network becomes extremely congested and cannot meet the minimum bandwidth requirement of multimedia flows. Thus the decoder buffer underflow percent can no longer be maintained at an acceptable level even with our mechanism. Meanwhile, additional data (i.e., $RI - B$) sent by our mechanism will deteriorate network congestion and lead to the significant increasing of the overall packet loss ratio in the network.

The above analysis and simulation results show that our cross-layer design is not suitable for such extreme congestion scenario. To solve this problem, we add an "intelligent switching" function, which can switch between the cross-layer design mode and layered design mode (i.e., the TCP-friendly state) according to the network congestion level. Here a good metric to evaluate the network congestion level is the underflow percentage of the decoder buffer, which actually indicates how much the network can meet the bandwidth requirement of the multimedia application. If the recent decoder buffer underflow percentage is larger than P_H,[7] it means that the network might be too congested to support the minimum bandwidth requirement of the application, and the sender will be switched from the cross-layer mode to the layered design mode (i.e., setting the sending rate to the TCP-friendly value). Then if the recent decoder buffer underflow percentage is below P_L (set to 0.1 in the simulations), it means that the network has recovered from the serious congestion, and the sender will return back to the cross-layer design mode. We repeat the simulation by using the enhanced version of VB-TFRCC-W, and the results are shown in Figure 11(b). We can find that the enhanced version can make sure the network performance (e.g., the overall packet loss ratio in the wired links) is not degraded by intelligent switch to the TCP-friendly state during the serious network congestion.

[7] Note that P_H indicates the allowed maximum value of the decoder buffer underflow percentage, which is determined by the minimum acceptable video quality of the application. In the simulations, P_H is set to 0.4.

6. CONCLUSION AND FUTURE WORK

This paper proposes cross-layer design of source rate control and congestion control for wireless video streaming. With a joint decision of the source rate and sending rate by taking into account the information from the application layer, the transport layer, and the MAC layer, the proposed cross-layer design approach can effectively avoid throughput degradation caused by wireless link error, and help the multimedia application achieve better playback quality, while maintaining good performance of the transport protocol.

In this paper, we mainly use the simulation to evaluate the performance of the proposed mechanism. Next we will try to implement our mechanism in the real wireless network to evaluate its practical performance. Although we incorporate ARC into our framework for extension to wireless, it should be noted that other rate control schemes for wireless can also be incorporated, for example, MULTFRC [17] and AIO-TFRC [18], which belong to pure end-to-end approaches and do not need the cross-layer information. It is a very interesting topic to study how to combine MULTFRC or AIO-TFRC, and our proposed framework.

REFERENCES

[1] S. Jacobs and A. Eleftheriadis, "Streaming video using TCP flow control and dynamic rate shaping," *Journal of Visual Communication and Image Representation*, vol. 9, no. 3, pp. 211–222, 1998.

[2] Ö. B. Akan and I. F. Akyildiz, "ARC: the analytical rate control scheme for real-time traffic in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 4, pp. 634–644, 2004.

[3] M. van der Schaar and S. Shankar, "Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms," *IEEE Wireless Communications*, vol. 12, no. 4, pp. 50–58, 2005.

[4] P. Zhu, W. Zeng, and C. Li, "Joint design of source rate control and QoS-aware congestion control for video streaming over the Internet," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 366–376, 2007.

[5] J. Widmer, R. Denda, and M. Mauve, "A survey on TCP-friendly congestion control," *IEEE Network*, vol. 15, no. 3, pp. 28–37, 2001.

[6] D. Bansal and H. Balakrishnan, "Binomial congestion control algorithms," in *Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '01)*, vol. 2, pp. 631–640, Anchorage, Alaska, USA, April 2001.

[7] N. R. Sastry and S. S. Lam, "CYRF: a theory of window-based unicast congestion control," *IEEE/ACM Transactions on Networking*, vol. 13, no. 2, pp. 330–342, 2005.

[8] R. Rejaie, M. Handley, and D. Estrin, "RAP: an end-to-end rate-based congestion control mechanism for realtime streams in the Internet," in *Proceedings of the 18th Annual Joint Conference of the IEEE Computer and Communications Societie (INFOCOM '99)*, vol. 3, pp. 1337–1345, New York, NY, USA, March 1999.

[9] D. Sisalem, *TCP-friendly congestion control for multimedia communication in the Internet*, Ph.D. thesis, Technical University of Berlin, Berlin, Germany, 2000.

[10] M. Handley, S. Floyd, J. Padhye, and J. Widmer, "TCP friendly rate control (TFRC): protocol specification," IETF RFC 3448, January 2003.

[11] Y.-G. Kim, J. Kim, and C.-C. Jay Kuo, "TCP-friendly Internet video with smooth and fast rate adaptation and network-aware error control," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 2, pp. 256–268, 2004.

[12] S. Cen, P. C. Cosman, and G. M. Voelker, "End-to-end differentiation of congestion and wireless losses," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 703–717, 2003.

[13] F. Yang, Q. Zhang, W. Zhu, and Y.-Q. Zhang, "End-to-end TCP-friendly streaming protocol and bit allocation for scalable video over wireless Internet," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 4, pp. 777–790, 2004.

[14] G. Cheung and T. Yoshimura, "Streaming agent: a network proxy for media streaming in 3G wireless networks," in *IEEE International Packet Video Workshop*, Pittsburgh, Pa, USA, April 2002.

[15] H. Balakrishnan, V. Padmanabhan, S. Seshan, and R. Katz, "A comparison of mechanisms for improving TCP performance over wireless links," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 756–769, 1997.

[16] D. Barman and I. Matta, "Effectiveness of loss labeling in improving TCP performance in wired/wireless network," in *Proceedings of the 10th IEEE International Conference Network Protocols (ICNP '02)*, pp. 2–11, Paris, France, November 2002.

[17] M. Chen and A. Zakhor, "Multiple TFRC connections based rate control for wireless networks," *IEEE Transactions on Multimedia*, vol. 8, no. 5, pp. 1045–1062, 2006.

[18] M. Chen and A. Zakhor, "AIO-TFRC: a light-weight rate control scheme for streaming over wireless," in *Proceedings of the International Conference on Wireless Networks, Communications and Mobile Computing (WirelessCom '05)*, vol. 2, pp. 1124–1129, Maui, Hawaii, USA, June 2005.

[19] M. Kalman, E. Steinbach, and B. Girod, "Adaptive media playout for low-delay video streaming over error-prone channels," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 6, pp. 841–851, 2004.

[20] E. G. Steinbach, N. Färber, and B. Girod, "Adaptive playout for low latency video streaming," in *Proceedings of IEEE International Conference on Image Processing (ICIP '01)*, vol. 1, pp. 962–965, Thessaloniki, Greece, October 2001.

[21] B. Xie and W. Zeng, "Rate-distortion optimized dynamic bitstream switching for scalable video streaming," in *IEEE International Conference on Multimedia and Expo (ICME '04)*, vol. 2, pp. 1327–1330, Taipei, Taiwan, June 2004.

[22] J. Viéron and C. Guillemot, "Real-time constrained TCP-compatible rate control for video over the Internet," *IEEE Transactions on Multimedia*, vol. 6, no. 4, pp. 634–646, 2004.

[23] C. E. Luna, Y. Eisenberg, R. Berry, T. N. Pappas, and A. K. Katsaggelos, "Joint source coding and data rate adaptation for energy efficient wireless video streaming," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 10, pp. 1710–1720, 2003.

[24] T. Ahmed, A. Mehaoua, R. Boutaba, and Y. Iraqi, "Adaptive packet video streaming over IP networks: a cross-layer approach," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 385–401, 2005.

[25] J. Yan, K. Katrinis, M. May, and B. Plattner, "Media- and TCP-friendly congestion control for scalable video streams," *IEEE Transactions on Multimedia*, vol. 8, no. 2, pp. 196–206, 2006.

[26] C. Chen, Z.-G. Li, and Y.-C. Soh, "TCP-friendly source adaptation for multimedia applications over the Internet," in *Proceedings of the 15th International Packet Video Workshop (PV '06)*, Hangzhou, China, April 2006.

[27] S. Floyd and S. McCanne, "Network Simulator, LBNL public domain software," http://www.isi.edu/nsnam/ns/.

[28] H. M. Radha, M. van der Schaar, and Y. Chen, "The MPEG-4 fine-grained scalable video coding method for multimedia streaming over IP," *IEEE Transactions on Multimedia*, vol. 3, no. 1, pp. 53–68, 2001.

[29] J. Padhye, *Towards a comprehensive congestion control framework for continuous media flows in best effort networks*, Ph.D. thesis, University of Massachusetts Amherst, Amherst, Mass, USA, 2000.

Hindawi Publishing Corporation
Advances in Multimedia
Volume 2007, Article ID 49604, 11 pages
doi:10.1155/2007/49604

Research Article

Identifying Opportunities for Exploiting Cross-Layer Interactions in Adaptive Wireless Systems

Troy Weingart, Douglas C. Sicker, and Dirk Grunwald

Department of Computer Science, University of Colorado, 430 UCB, Boulder, CO 80309-0430, USA

Received 31 December 2006; Revised 15 April 2007; Accepted 2 July 2007

Recommended by Zhu Han

The flexibility of cognitive and software-defined radio heralds an opportunity for researchers to reexamine how network protocol layers operate with respect to providing quality of service aware transmission among wireless nodes. This opportunity is enhanced by the continued development of spectrally responsive devices–ones that can detect and respond to changes in the radio frequency environment. Present wireless network protocols define reliability and other performance-related tasks narrowly within layers. For example, the frame size employed on 802.11 can substantially influence the throughput, delay, and jitter experienced by an application, but there is no simple way to adapt this parameter. Furthermore, while the data link layer of 802.11 provides error detection capabilities across a link, it does not specify additional features, such as forward error correction schemes, nor does it provide a means for throttling retransmissions at the transport layer (currently, the data link and transport layer can function counterproductively with respect to reliability). This paper presents an analysis of the interaction of physical, data link, and network layer parameters with respect to throughput, bit error rate, delay, and jitter. The goal of this analysis is to identify opportunities where system designers might exploit cross-layer interactions to improve the performance of Voice over IP (VoIP), instant messaging (IM), and file transfer applications.

1. INTRODUCTION

The flexibility of cognitive and software-defined radios presents an opportunity for researchers to reexamine how network protocol layers operate with respect to providing quality of service aware transmission among wireless nodes. This opportunity is enhanced by the continued development of spectrally responsive devices—ones that can detect and respond to changes in the radio frequency environment. Present wireless network protocols define reliability and other performance-related tasks narrowly within layers. For example, the frame size employed on 802.11 can substantially influence the throughput, delay, and jitter experienced by an application, but there is no simple way to adapt this parameter. Furthermore, while the data link layer of 802.11 provides error detection capabilities across a link, it does not specify additional features, such as forward error correction schemes, nor does it provide a means for throttling retransmissions at the transport layer (currently the data link and transport layer can function counterproductively with respect to reliability).

This paper presents an analysis of the interaction of physical, data link and network layer parameters with respect to throughput, bit error rate, delay, and jitter. We look specifically at (1) power, (2) bitrate, (3) forward error correction, (4) frame size, (5) automatic repeat request, and (6) selective queueing. The goal of this analysis is to identify key opportunities for system designers to exploit cross-layer interactions improving the performance of Voice over IP (VoIP), instant messaging (IM), or file transfer applications. This research utilizes both simulation and system platforms in its analysis. For simulation purposes, we employed the OPNET network simulation environment. In developing an actual radio, we reverse engineered a commercial off-the-shelf (COTS) product to create adaptive data link and routing layers. The extensions to the COTS platform, what we refer to as SoftMAC and MultiMAC, served as an experimental framework for the implementation and evaluation of our simulation results.

Our analysis details those parametric interactions that significantly impact wireless performance. This analysis would then be used to inform the design, implementation, and evaluation of adaptive wireless networking algorithms. These algorithms will be able to dynamically adjust their behavior, thereby improving wireless performance in response to a changing radio frequency (RF) environment. Our work demonstrates that even conservative adaptation of the

physical, data link, and network layers will provide significant enhancement to the performance of the system, and that adaptation can strongly influence the ability of the system to meet quality of service (QoS) requirements.

2. BACKGROUND

Cognitive and software-defined radios (C/SDRs) are one of the most promising of the emerging technologies addressing spectrum use, wireless performance, reconfigurability, and interoperability. Many research groups are actively pursuing projects that exploit these opportunities [1–8]. Much of the research and development in software-defined radio has been characterized by the building of systems to solve specific problems. In the United States Military, efforts are focused on waveform portability and radio interoperability. Research done at the University of Kansas was centered on mobile and rapidly deployable disaster response communication. Vanu Bose's thesis focused on solving problems associated with moving traditionally analog or custom ASIC components and processes to a general-purpose processor. Most of the problems in software radio lie in the development of more capable hardware and interoperable software frameworks. Software radio research, although very active, has begun to give way to the recent popularity of cognitive radio.

In the radio space, Moore's law has also provided momentum in transitioning from special-purpose inflexible hardware and firmware to mutable virtual radios [9]. These virtual radios or SDRs make use of general-purpose processors and software to accomplish what used to be done with specialized hardware. One can envision an advanced radio network that dynamically allocates and reallocates spectrum or dynamically reconfigures itself in response to changes in policy and environmental conditions. This level of flexibility in a radio platform not only allows us to tackle the problem of spectrum utilization, but also serves as a highly capable platform from which one can exploit cross-layer interactions.

Much of the resistance in the cellular industry to emerging third generation technologies stems from the huge cost involved when removing and replacing existing infrastructure with technology that can support the new standard [10]. On the other hand, a C/SDR platform may be able to adapt to new standards by downloading and installing new software. The flexibility inherent in the C/SDR also allows it to adapt to changes in policy. The recent focus on homeland security, in light of the ineffective use of the emergency bands, has given impetus to many of the scenarios that illustrate the promise of C/SDR. One could imagine a government agency implementing a change to local spectrum policy in response to a disaster. Updates to policy, when acted upon by a C/SDR network, could affect reallocation of spectrum to support increased demand during the emergency. The C/SDR is also ideally suited for system-level cross-layer networking research. Additionally, a C/SDR could be used for research and experimentation with spatially aware applications, adaptive routing, cognitive media access control (MAC) layers, and mutable physical layers.

At the top end of the software radio taxonomy are radios that incorporate computational intelligence. These "cognitive" radios will be able to sense, learn, and act in response to changes in their environment [7]. The ultimate cognitive radio will be able to autonomously negotiate and propose entirely new optimized protocols for use in the networking environment. Although provocative, current radio technology is nowhere near mature enough to realize systems with these capabilities. At the lower end of the C/SDR spectrum are radios that are functionally equivalent to their analog predecessors, although the newer radio's functionality has been implemented on field programmable gate arrays (FPGA), digital signal processors (DSP), or general-purpose processors (GPP). By in large, the systems of today were built to solve a domain-specific problem, or are focused on tightly coupled manipulations of the lower layers of the protocol stack. Regardless of where one's research interests lie, there are a host of technical challenges to overcome.

The focus of this paper is on understanding how varying parameters at the physical, data link and network layers can affect the performance and reliability of a wireless system. Understanding these effects is a critical first step in the development and implementation of an algorithm for cognitive radio. In addition, once the performance implications of varying these parameters is understood, one must also consider a host of implications that arise when one alters such parameters. This includes decisions on when and how to change configurations, how these changes are communicated, and how much time can be spent calculating the next configuration. While these are all important questions, we focus on understanding the impact of varying a C/SR's settings on its performance.

3. RELATED WORK

Research in the area of cross-layer optimization for wireless systems has been an area of considerable focus in recent years. Others have also spent a considerable amount of time and effort investigating cognitive radios. However, the potential of improving the performance of a wireless system by combining cross-layer optimization with cognitive systems is just emerging as a research area.

Much of the work in the area of cross-layer optimization focuses on enhancing throughput, quality of service (QoS), and energy consumption [11–13]. These cross-layer optimizations tend to focus on two layers of the protocol stack with the goal of enhancing a specific performance measure. As such, they do not consider multifactor variation nor do they consider effects of this variation on inelastic applications, such as Voice over Internet Protocol (VoIP). Kawadia and Kumar present an interesting critique of cross-layer design in [14]. They warn that cross-layer optimization presents both advantages and dangers. The dangers they discuss include the potential for (1) spaghetti design, (2) proliferation problems and, (3) dependency issues. Such cautions (and others that we will identify) are easily overlooked in the hopes of gaining sometimes marginal performance improvements. Therefore, understanding the significance of the potential improvements is an important step to consider.

Given that the interactions among a set of parameters are determined, the next step is determining the significance of these interactions. In other words, those interactions provide the best response in a given situation. Vadde et al. have applied response surface methodology and design of experiments (DOE) techniques to determine the factors that impact the performance of mobile ad hoc networks (MANETs) [15–17]. Their research considers routing protocols, QoS architectures, media access control (MAC) protocols, mobility models, and offered load as input factors and throughput and latency as response factors. Their analysis demonstrates the usefulness of these techniques and shows where certain input factors can outperform others within a MANET.

Haykin provides a thorough overview of cognitive radios and describe the basic capabilities that a "smart" wireless device might offer [18]. Others describe techniques for applying C/SDRs to improving the coordinated use of spectrum [19, 20]. Sahai et al. describes some of the physical layer limits and limitations of cognitive radios, including the difficulties associated with determining whether or not a radio frequency band is occupied [21]. Nishra has implemented a test bed for evaluating the physical and data link layers of such networks [22]. Additionally, Thomas describes the basic concept of a C/SDR network and provides a case study to illustrate how such a network might operate [23]. It is also worth noting that the standards communities are focusing on cognitive radios. The IEEE 802.22 group is developing a wireless standard for the use of cognitive radios to utilize spectrum in geographically separated and vacant TV bands [24]. Also in the IEEE, the P.1900 workgroup is examining the general issue of spectrum management in next generation radio networks.

4. EXPERIMENTAL DESIGN

In this section, we describe the design and implementation of the simulation and experimental platforms. We also introduce design of experiments (DOE), a technique that we employ in the identification of those parameters that significantly impact performance.

4.1. Simulation tool

In order to determine the validity of our approach, we conducted our preliminary research on a simulation platform. Upon considering the potential complexity of a cognitive network composed of many nodes, we decided to begin by evaluating a simple network. The simulation itself consisted of two nodes communicating in the presence of an active noise source (e.g., a noncooperative node on a different network, or a radio frequency jammer). We used OPNET Modeler to simulate the effects of changing communication parameters in order to determine where best to employ cross-layer optimization [25]. The simulation suite provides a rich and readily extendable network modeling environment. While OPNET provides a wireless networking module for the data link layer, to obtain the flexibility that was required for interactions spanning protocol layers, we found it necessary to develop our own data link module. This module allows adaptation of the parameters affecting cross-layer interaction on a per-packet basis.

The simulation platform uses an additive white Gaussian noise (AWGN) model to simulate the effects of environmental noise. The jammer in our simulation emits RF energy in bursts of varying duration and interarrival using OPNET's 802.11 physical layer.

4.2. Platform

Much of the work in developing the platform for this research has already been completed. This research relies on the use of COTS products with C/SDR extensions. The extensions to the COTS platform, SoftMAC and MultiMAC, serve as the experimental framework for implementing and evaluating the results of the simulation. The following sections describe each component of the platform used in the research.

4.2.1. SoftMAC

This system was built to provide a flexible environment for experimenting with MAC protocols in the wireless domain. The ability to cheaply create, modify, and conduct system-level experimentation with hardware is often a goal of many research projects. However, many of these projects ultimately fail due to the cost, time, and effort involved in deploying a large-scale experimental platform. The SoftMAC platform fills this need. It uses a commodity 802.11b/g/a networking card with a chipset manufactured by the Atheros Corporation to build a software radio with predefined physical layers but a flexible MAC layer. Internally, the Atheros chipset provides considerable flexibility over the format of the transmitted packets, network drivers do not generally expose this flexibility. By reverse-engineering many of those controls, SoftMAC provides a driver that allows extensive control over the MAC layer while still allowing use of the waveforms defined by the underlying 802.11b/g/a physical layers.

4.2.2. MultiMAC

This system is intended to extend the basic SoftMAC environment to tackle problems in the areas of dynamic spectrum allocation and cognitive/software-defined radio. It builds upon the functionality in the SoftMAC platform with some specific features in mind. First, MultiMAC allows multiple MAC layers to coexist in the network stack with minimal switching impact. Second, it allows one to dynamically reconfigure the MAC and physical layers on a per-packet basis either from logic running as part of MultiMAC or from a user-level process. Finally, by leveraging these capabilities MultiMAC allows intelligent reconfiguration of the MAC and physical layers; thus achieving a cognitive MAC. The cognitive MAC layer couples efficient reconfiguration afforded by MultiMAC with computational intelligence. This combination allows the engine to make smart decisions about which MAC layer should be used and which physical layer properties should be set.

Table 1 lists parameters that might be available to a MultiMAC cognitive process running on the platform. This

Table 1: A potential set of mutable parameters.

Parameter	Datatype
Route	Enum
Frame size	Integer
Forward error correction	Enum
Automatic repeat request	Boolean
Encryption	Boolean
Media access protocol	Enum
Channel	Integer
Modulation	Enum
Bitrate	Enum
Antenna configuration	Integer
Transmit power	Float

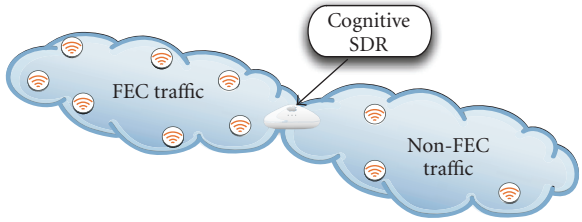


Figure 1: Dynamic MAC layer bridging using C/SDR.

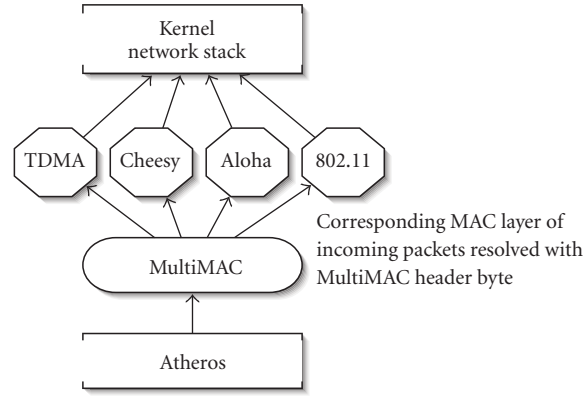


Figure 2: MultiMAC assigns received frames to the MAC layer that can decode them.

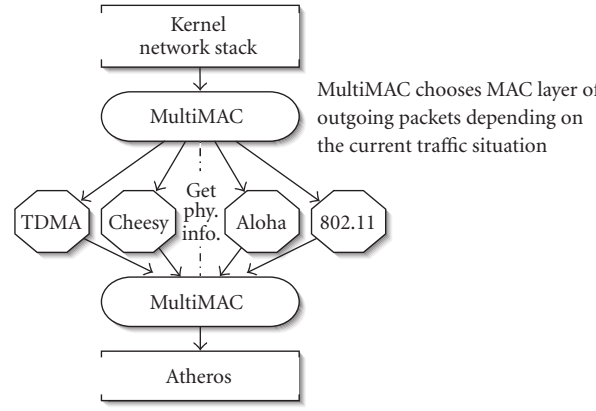


Figure 3: On the sending side, MultiMAC uses environmental stimuli to determine the path of a packet.

smart radio would use these parameters to effect changes in reliability due to its assessment of performance metrics and environmental inputs. The basic mechanism provided by MultiMAC can also be used to implement specific fixed MAC protocols. For example, one platform referenced in our work couples commodity 802.11 network cards with the Phocus phase array antenna. The directional phase array antenna will be able to use dynamic beam-forming to spatially create separate MAC zones (see Figure 1). One segment could use a forward error corrected (FEC) MAC layer while the other zone could use a MAC which does not have FEC enabled. The framework allows us to instantly transit from an FEC to a non-FEC MAC. Also, MultiMAC offers the ability to dynamically change properties while still decoding frames sent from previous configurations. Individual MAC variants will be used by MultiMAC for decoding their respective incoming frames as well as encoding outgoing frames with a MAC best suited for network conditions. This process is both completely transparent and highly adaptive in operation. MAC layers can be changed on the fly without interrupting radio service or dropping frames during the transition. When a decoded frame arrives, the appropriate MAC layer must "*claim*" and decode the frame. Once a packet is handed over to its corresponding MAC layer implementation, decoding happens the same way as in an unmodified network stack (see Figure 2). A mirrored procedure takes place on the encoding side; the process is more complex due to timing constraints imposed by the MAC layer protocol. (see Figure 3). The individual policies used to select an outgoing MAC protocol rely on cross-layer feedback from the physical and network layers. MultiMAC maintains a connection to the status and diagnosis API of the wireless chipset. By pulling out network status information such as the *time to transmit frames*, *queue lengths*, and *bit-errored frames*, MultiMAC policies can determine the appropriate MAC for the specified goal. (See [26] for details on SoftMAC and MultiMAC.)

4.3. Design of experiments

Design of experiments (DOE) is an approach for determining cause and effect relationships within an experiment [27]. Historically, this technique has been applied with great success in the process and materials industries. Later, we show that it also can be applied with success in the wireless domain. DOE provides a structured method for understanding the relationships among input and output variables. By systematically varying all the input factors, DOE allows researchers to identify the existence of interactions among these inputs and their impact on output factors. It allows researchers to determine what factors most influence an experiment and, moreover, determine the interaction among a group of input factors. In this paper, we rely on DOE to quantify the influence of single and multifactorial inputs illustrating how parameters across multiple layers might

Table 2: The set of mutable parameters.

Parameter	Settings	Layer
Automatic repeat request (ARQ)	Off/on	MAC
Frame size	2048, 9216, 18432 bits	MAC
Forward error correction	Off/on	MAC
Bitrate	1, 2, 5.5, 11 Mbs	Physical
Transmit power	5, 32, 100 mW	Physical
Selective queuing	Off/on	Network

improve (or degrade) the performance of a wireless system. This technique relies on the analysis of variance (ANOVA) statistical method to provide an assessment of the significance of the test results.

5. RESULTS AND DISCUSSION

The following section reports the results of our simulation work. The experimental trials were designed to cover a range of traffic sent between nodes in the presence of a noise source. FTP and VoIP traffic were selected due to their distinct tolerances for latency, jitter, throughput, and bit loss. The parameters that we examined included ARQ, frame size, bitrate, transmit power, FEC, and selective queueing (as described in Table 2). Stop-and-wait is the type of ARQ used, wherein the sending node will stop transmitting until it has received an acknowledgment from the receiver (or it times out; in which case the sender will retransmit the frame). When selective queuing is enabled, high priority frames (in our case, VoIP frames) are moved to the head of the transmit queue.

Each of the simulation trials was analyzed across the levels of the parameters and general trends were highlighted. For example, we looked at the average jitter, latency, bit loss, and throughput performance of FEC across all combinations of the other parameters. We then analyzed each of the traffic types using DOE techniques.

5.1. Simulation configuration

Figure 4 shows the physical layout of the two communicating nodes in relationship to the noise source. The uncooperative (or jamming) node is emitting noise in a Poisson distribution centered around an interarrival time of 0.05 seconds and a burst length of 1024 bits. The physical layout of the nodes and noise source as well as the power of the noise source is fixed across all of the trials; however, the duration and interarrival of the jamming bursts do vary. In our preliminary research, we examined a broad range of noise settings including different duration bursts, interarrival times, and power levels. We settled on a setting that provided appreciable interference without overwhelming the communicating nodes.

Each of the trials examines the performance of the experimental system at each of the potential parametric settings. Table 3 is a list of the metrics by which we evaluate each mutation of the settings. One can independently look at the performance of any of the parameters (alone or in combination with other parameters) against any one of the metrics used

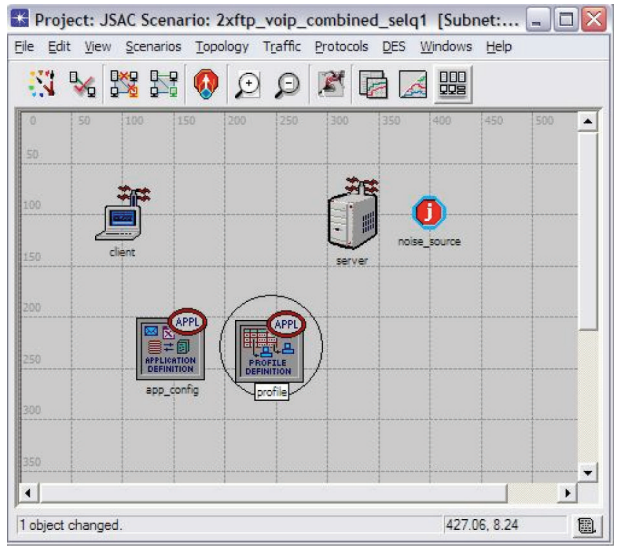


Figure 4: Experimental layout in OPNET.

Table 3: The set of metrics.

Metric	Units
Bit loss	Percent
Latency	Seconds
Jitter	Seconds
Throughput	bps

to evaluate the system. Next, we examine each of the trials in turn.

5.2. File transfer protocol (FTP)

5.2.1. Experimental setup

This first scenario was designed to isolate FTP traffic from the client to the server in the presence of a noise source. Here, FTP traffic is modeled using OPNET's client and server FTP traffic profiles. A 5 MB file is transfered from the client node to the server. In these experiments, we focused on optimizing throughput.

5.2.2. General trends

This analysis was done by fixing a parameter and reporting the average performance of that action across all permutations of the other parameters. For example, in order to investigate the general effect of ARQ on bit loss we started by first disabling ARQ and then running through all the permutations of the other parameters. This is followed by enabling ARQ and rerunning the simulation set. We then compare the average effect of enabling and disabling ARQ on bit loss. Figure 5 shows the performance of each of the parameter settings on throughput. One can see from the chart that increasing bitrate and frame size have a significant effect on throughput. This chart shows the average effect of changing a parameter; it does not show the overall best- or worst-case

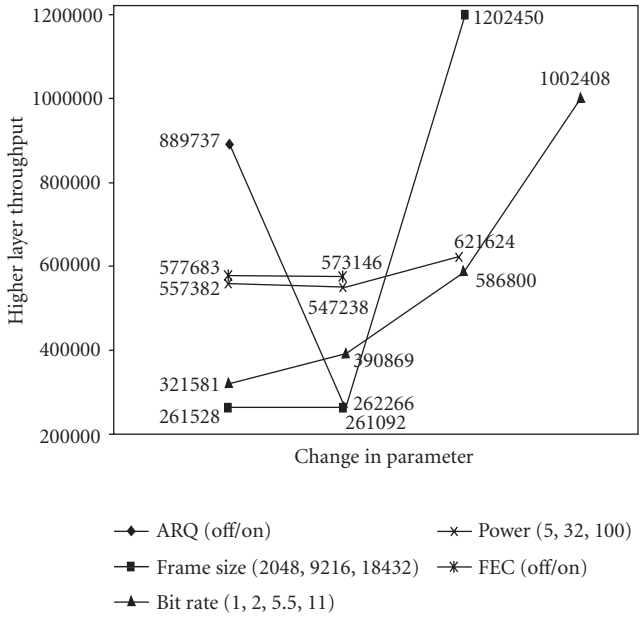


Figure 5: FTP throughput versus average effect of a parameter.

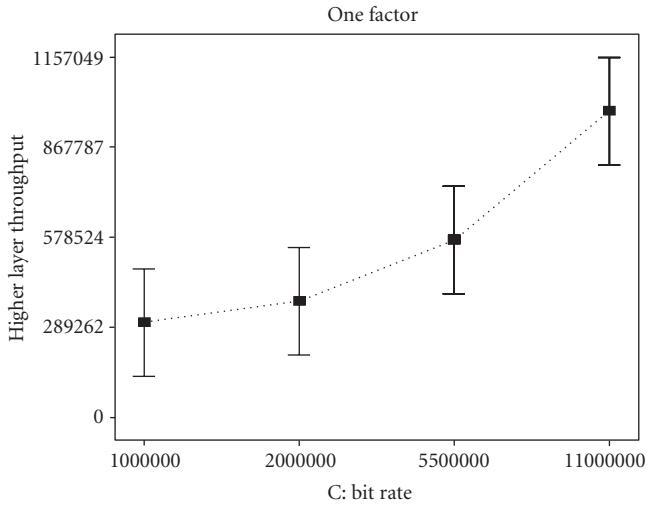


Figure 6: Analysis of bitrate's effect on FTP throughput.

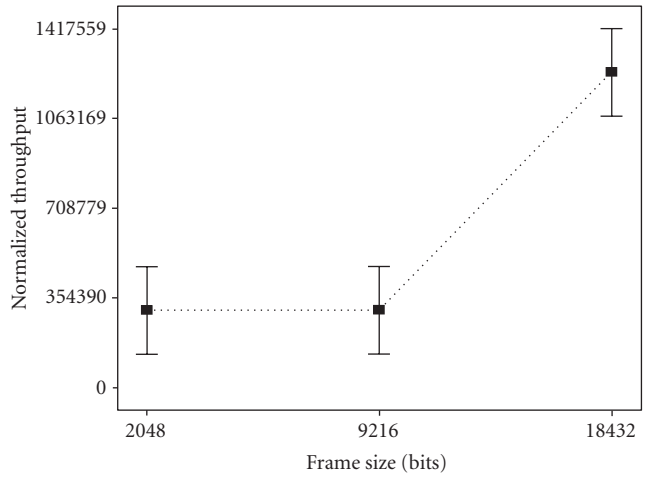


Figure 7: Analysis of frame size's effect on FTP throughput.

parameter settings. For FTP, the worst selection of parameters yielded a throughput of 10 Kbps and the best slightly above 9.5 Mbs.

5.2.3. Design of experiments (DOE)

It is appropriate to begin this section with some supporting information on the use of DOE. DOE is ideally suited to help us identify those cross-layer interactions that are statistically significant. DOE provides a design methodology and set of statistical tools for setting up and running experimental trials in a manner that allows one to identify those factors that significantly impact what you are measuring. In our case, DOE serves to identify both intra- and cross-layer interactions that effect the response of interest. The core statistical process at work in DOE is the calcuation of the *F-test*. This test compares the variance among the treatment means versus the variance of the individuals within the specific treatments. Another way of looking at F is as a ratio of signal to noise.

DOE analysis of the *main factor effects* of the parameteric change on throughput yeilded some interesting results. We first considered the main effects on FTP traffic (*main effects* can be defined as the change in response caused by altering a single factor). We found bitrate and frame size (as shown in Figures 6 and 7) to most improve throughput and ARQ to have a detrimental influence on throughput. Note that the DOE Y axis provides a normalized scale and therefore we do not discuss the quantitative results of these experiments; rather we focus on the trends and the interactions among the parameters. As expected, large frames and/or high bitrate improve throughput. Additionally, we found power and forward error correction to have little or no effect. Power does have a significant impact when we increase power and or frequency at the noise source; however, our general intention is to minimize the use of power because of the detrimental impact on neighboring nodes.

Next, we examined the two-factor effects on FTP throughput. As shown in Figure 8, frame size and data rate show a strong synergistic effect on improving throughput. Note that the top two lines both contain *least significant difference* bars that do not overlap, which indicates that the result is significant. Sending large packets at a high rate should improve throughput. The significance here is the magnitude of the improvement. Again, given the noise level, power provided little improvement in the achieved throughput.

5.3. Voice over IP (VoIP)

5.3.1. Experimental setup

This next scenario was designed to isolate VoIP traffic between the client and the server in the presence of a noise source. Here, VoIP traffic is modeled using OPNET's IP Telephony Model. We designed this experiment to demonstrate

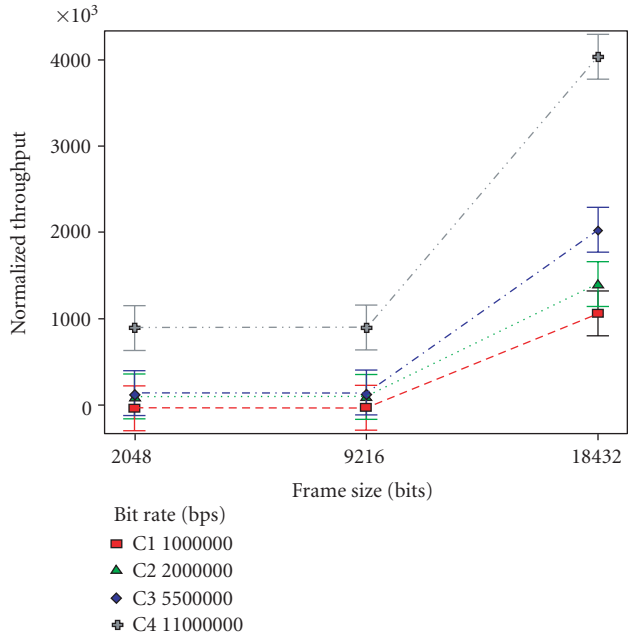


Figure 8: Analysis of frame size and data rate's effect on FTP throughput.

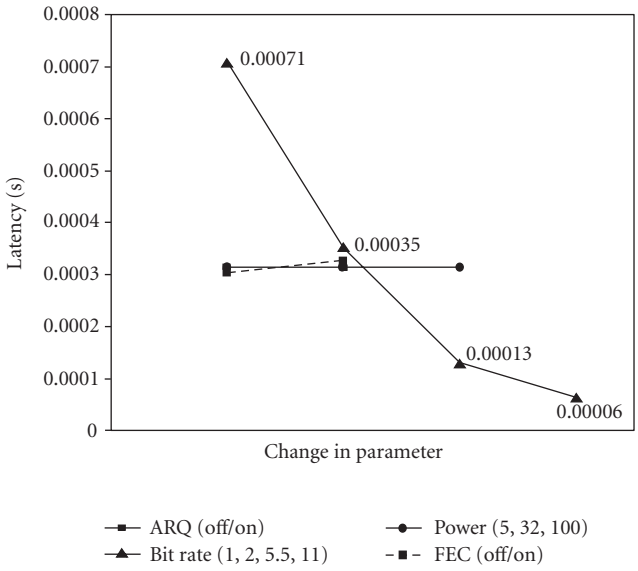


Figure 9: VoIP latency versus average affect of a parameter.

the behavior of VoIP on a lightly loaded network (later we look at VoIP performance on a heavily loaded network).

5.3.2. General trends

Figure 9 shows the performance of each of the parameter settings on latency (we also examined *jitter* and later indicated where it negatively and positively impacted VoIP traffic). One can see that increasing bitrate has the most significant effect on latency. Again, this chart shows the average effect of a parameter setting; it does not show the best- or worst-case

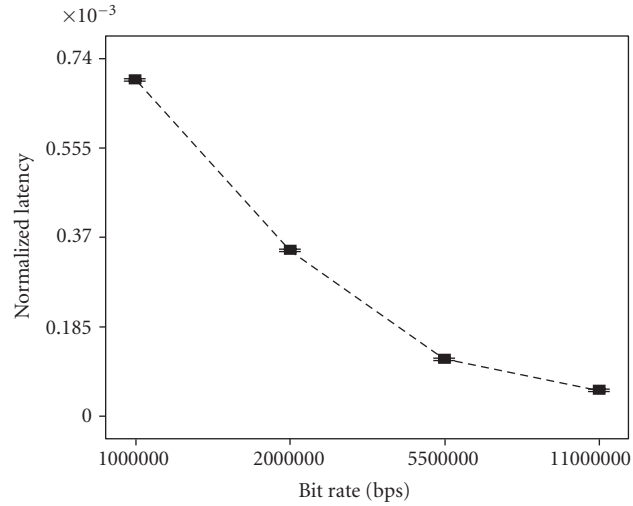


Figure 10: Analysis of bitrate's effect on VoIP latency.

configuration. The worst-case configuration yielded an average latency of 0.0073 seconds and the best 0.00006 seconds. Since this trial was conducted on a lightly loaded network, the other parameters (namely, ARQ, frame size, FEC, and power) had little impact on latency in the average case.

5.3.3. Design of experiments (DOE)

As shown in Figure 10, the DOE analysis confirmed the impact of bitrate on VoIP latency. The DOE analysis also confirmed that none of the other parameters had a significant main or multifactor effect on latency.

One can see from the chart that increasing bitrate has the most significant effect on latency. DOE multifactor analysis did not yield any statistically significant results.

5.4. FTP/VOIP combined

5.4.1. Experimental setup

Our goal in combining the two traffic types was to see how the parameter changes effect our metrics when the traffic profiles differ and overlap. This trial models the sending of a file from the client to the server during a VoIP call. Both of the traffic sources in this simulation are modeled with OPNET's constant bitrate sources. For this trial, we developed a selective queuing mechanism above our MAC layer. Selective queuing gives priority to VoIP frames by moving them to the front of the transmit queue.

5.4.2. General trends

We found a number of interesting results in this set of experiments, as shown in Figures 11 and 12. By selectively queueing VoIP frames, we were able to drastically improve latency while minimally impacting FTP throughput. Furthermore, both bitrate and power had a beneficial impact on both VoIP and FTP traffics. One can see from the chart that turning selective queuing on has a significant effect on latency

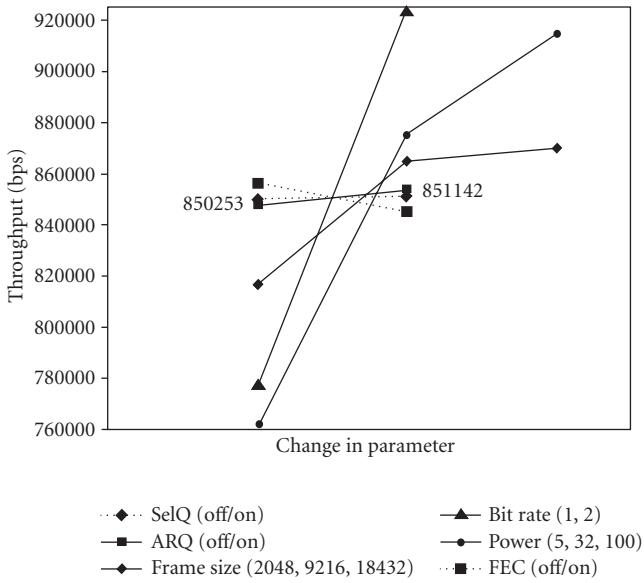


Figure 11: FTP throughput versus average affect of a parameter.

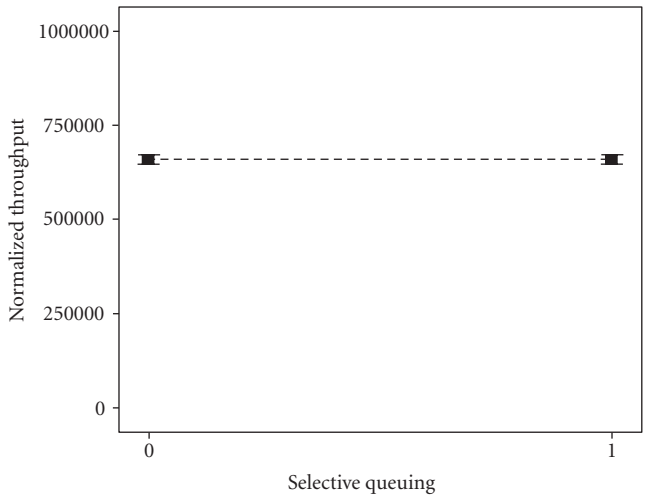


Figure 13: Analysis of selective queuing's impact on FTP throughput.

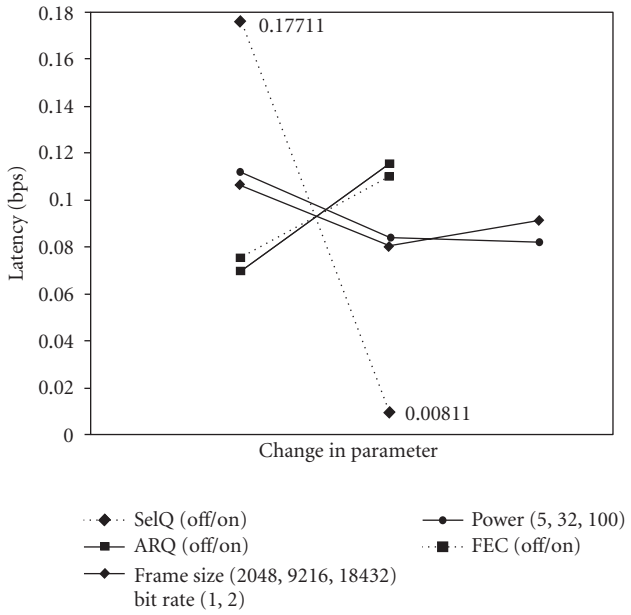


Figure 12: VoIP latency versus average effect of a parameter.

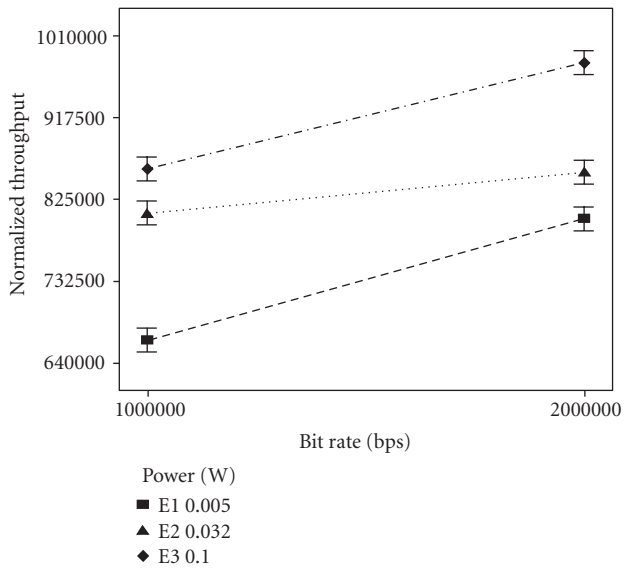


Figure 14: Analysis of power and data rate's impact on FTP throughput.

(the worst case being an average latency of 1.36 seconds and the best being 0.00937 seconds). This offers an impressive exploitation of cross-layer information yielding several orders of magnitude improvement in latency with minimal impact on FTP throughput (see Figure 11).

5.4.3. Design of experiments (DOE)

On the single-factor analysis for FTP, frame size, data rate, and power all had a positive impact, while ARQ and FEC were detrimental. More significantly, selective queueing greatly improved VoIP delay while not adversely impacting FTP throughput (as shown in Figure 13). Within the two factor analysis, we found that both ARQ and FEC had a negative effect with all parameter interactions. As shown in Figure 14, data rate demonstrated a strong synergistic effect with power.

On the single factor analysis for VoIP, Figure 15 shows that the selective queueing had the strongest impact. Likewise, data rate also demonstrated a positive effect on latency. Figure 16 shows that selective queueing on a heavily loaded channel significantly improves latency; conversely, selective queueing has no impact on a lightly loaded channel.

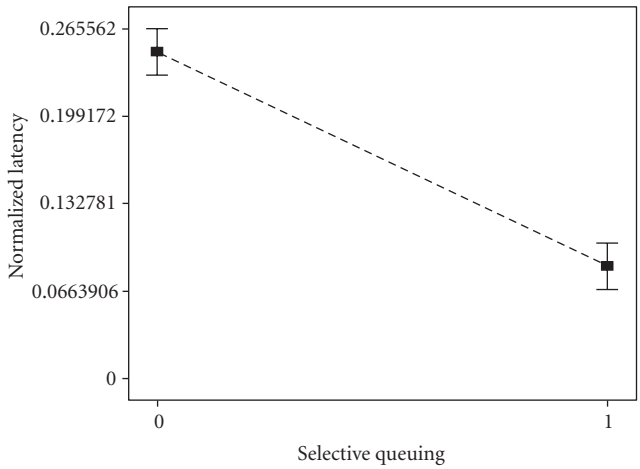


Figure 15: Analysis of selective queuing's impact on VoIP latency.

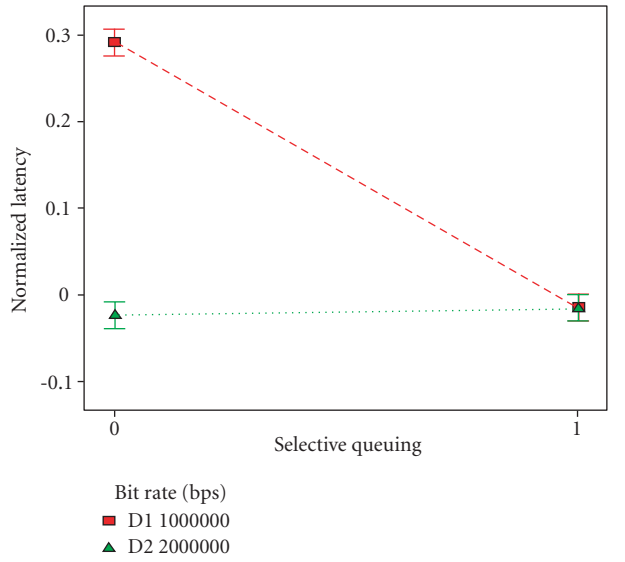


Figure 16: Analysis of selective queuing and data rate's impact on VOIP latency.

Power and selective queueing also showed a synergistic effect on improving latency. Additionally, data rate had a beneficial effect on latency regardless of power and FEC settings.

5.5. Summary

In these experiments, we were able to demonstrate a number of promising interactions that developed across parameters within the physical, data link, and network layers. Of particular significance was the ability for the platform to simultaneously support the needs of QoS-diverse applications by adapting the parameters of the system. Furthermore, this was done without increasing transmit power (while this maybe useful in improving link performance in noisy environment, it also negatively impacts neighboring nodes). In nearly all experiments, we consistently found a set of cross-layer settings that are able to support the needs of QoS-diverse applications. What makes this useful is that it shows the feasibility and promise of developing an algorithm that can adapt to the time-varying nature of environmental conditions and an active jammer.

6. AN ILLUSTRATIVE EXAMPLE

Looking forward, it is clear that there are huge performance gains to be obtained by exploiting cross-layer interactions in wireless networking. A goal of this research is the development of an operational system incorporating the full functionality of a cognitive radio. This system would autonomously select and optimize parameters according to input from the user and environment, capitalizing on cross-layer interactions as reported in this article.

To demonstrate this capability, Jeff Fifield developed an adaptive Reed-Solomon MAC that utilizes Reed-Solomon (RS) forward error correction to detect and fix bit errors in the MAC data payload. RS codes are a well-known method of encoding data for protection against transmission errors. In the RS MAC, the common $(255, 223)$ encoding scheme is used. Using this scheme, data is broken up into 223-byte blocks and each block is encoded separately, resulting in 255 bytes of encoded data. Because of the additional space and computational overhead associated with RS encoding, the MAC is adaptive, only using FEC if bit errors occur. This MAC was implemented as a click [28] application using the SoftMAC click elements and a standard RS software package. The CSMA/CA mechanism provided by SoftMAC was used for channel access. In RS MAC, all outgoing packets are either RS encoded or not. Since an endpoint cannot determine whether or not a packet it transmitted was received without error, it must rely on feedback from its peer to make transmission decisions. A simple algorithm with three configurable parameters governs the sending of these feedback packets. The parmeters are the sample period s, the error threshold e, and the no-error threshold c. Packets are observed over a sample period of s packets. If an endpoint is receiving unencoded packets and e or more packets with errors are received during a sample period, a packet is sent indicating that RS encoding should be used. Similarly, if an endpoint is currently receiving RS-encoded packets and c or more packets are received without errors during the sample period, the MAC sends a message telling its peer to stop encoding packets. In unencoded packets, errors are detected using a CRC32 checksum. In RS encoded packets, errors are detected during the RS-decoding process.

To test the functionality and performance of the adaptive Reed-Solomon MAC, we performed an experiment wherein two nodes try to send 1000-byte packets to each other at a rate of 100 packets per second. To decrease the probability of errors occurring in control frames relative to probability of errors occurring in data frames, a data rate of 1 Mbps was used for control information, while data was sent at a rate of 54 Mbps. Nodes were placed far enough apart to induce significant error when using the 54 Mbps waveform. The result of 10 trials are shown in Figure 17. For each test, 2000 packets were sent by each node for a total of 4000 packets.

Reed-Solomon MAC for $s = 10$, $e = 2$, $c = 10$				
	Recv	Valid Recv	RS Recv	Corrections
RS	3859	3660	2971	23013
No RS	3845	1850	0	0

Figure 17: Packets received, Packets correctly received, Reed-Solomon packets received, and number of Reed-Solomon corrected bytes. Averages for 10 trials of 4000 packets each.

The results show that the adaptive RS encoding scheme reduces the transmission error rate. On average, about 75 percent of packets were RS encoded, reducing the number of packets dropped due to errors from greater than 50 percent to less than 10 percent. The results also suggest that most errors occur in the 54 Mbps payload portion of the packet and not in the 1 Mbps and 2 Mbps PLCP headers. Errors were observed in more than half of the packets received, the wireless header (which cannot be disabled in SoftMAC) accounts for about 15 percent of the transmission time of a large packet for high data rates. If we assume that errors occur in half of all transmissions, that errors are equally distributed within a transmission, and that each transmission has only a single error (although the RS results suggest more than 7 errors per corrupted packet), we would expect errors in at least 7.5 percent of all headers, or in about 300 of 4000 packets. If an error occurs in the header, the frame is dropped by the hardware and the device driver never sees it. Thus, the number of packets with errors in the header is just the number of packets sent minus the number of packets received. Since the observed error rate is roughly half the predicted error rate, it must be the case that errors occur less frequently in the header than in the payload. This also validates our assumption that sending control data at 1 Mbps decreases the probability of error occuring in those frames. This simple implementation only hints at the promise of DOE as a technique for identifying beneficial cross-layer interactions (in this test, we achieved a reduction in error rate of 40 percent).

7. CONCLUSION

In this paper, we have described how parameters at the physical, data link, and network layers interact with respect to a variety of performance metrics. First, through simulation and then through experimental design techniques, we describe how parameters including power, bitrate, forward error correction, automatic repeat request, frame size, and selective queueing interact to influence throughput, bit error rate, delay and jitter. We show how such optimization can be used to improve the performance of applications by matching the settings of the lower protocol layers to the demands of the application. We then illustrate this optimization by showing how forward error correction can be used to decrease error on a noisy link by 40 percent. It is our intention to capitalize on DOE as a technique for identifying beneficial cross-layer interactions. However, the adaptive and dynamic nature of cognitive wireless systems leads to other interesting questions. It will be important to quantify the amount of time that a cognitive process can devote to computing an adaptive radio configuration, thus allowing one to characterize the types of processing that can be done without negatively affecting communication. This line of research should also provide insight into what processing should be done in real time, offline, or in the background. We plan to investigate each of these questions in future work.

ACKNOWLEDGMENTS

The authors would like to thank Eric Anderson, Gary Yee, Christian Doerr, Jeff Fifield, Mike Neufeld, and Mike Buetner for their assistance in this research.

REFERENCES

[1] J. O. Neel, J. H. Reed, and R. P. Gilles, "Game models for cognitive radio algorithm analysis," in *Software Define Radio Forum Technical Conference (SDR '04)*, pp. 27–32, Phoenix, Ariz, USA, November 2004.

[2] J. Neel, R. Buehrer, B. Reed, and R. P. Gilles, "Game theoretic analysis of a network of cognitive radios," in *Proceedings of the 45th Midwest Symposium on Circuits and Systems (MWSCAS '02)*, vol. 3, pp. 409–412, Tulsa, Okla, USA, August 2002.

[3] J. Neel, J. H. Reed, and R. P. Gilles, "Convergence of cognitive radio networks," in *IEEE Wireless Communications and Networking Conference (WCNC '04)*, vol. 4, pp. 2250–2255, Atlanta, Ga, USA, March 2004.

[4] C. Rieser, *Biologically inspired cognitive radio engine model utilizing distributed genetic algorithms for secure and robust wireless communications and networking*, Ph.D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Va, USA, 2004.

[5] R. I. Lackey and D. W. Upmal, "Speakeasy: the military software radio," *IEEE Communications Magazine*, vol. 33, no. 5, pp. 56–61, 1995.

[6] R. J. Sánchez, J. B. Evans, G. J. Minden, V. S. Frost, and K. S. Shanmugan, "Rdrn: a prototype for a rapidly deployable radio network," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 2, no. 2, pp. 15–22, 1998.

[7] J. Mitola III, *Cognitive radio an integrated agent architecture for software defined radio*, Ph.D. dissertation, Royal Institute of Technology (KTH), Stockholm, Sweden, 2000.

[8] Website, "Defense advanced research projects agency (DARPA) next generation (XG) communications," http://www.darpa.mil/sto/smallunitops/xg.html.

[9] V. Bose, M. Ismert, M. Welborn, and J. Guttag, "Virtual radios," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 4, pp. 591–602, 1999, special issue on software radios.

[10] V. Bose, *Design and implementation of software radios using a general purpose processor*, Ph.D. dissertation, Massachusetts Institute of Technology (MIT), Cambridge, Mass, USA, 1999.

[11] A. Goldsmith and S. Wicker, "Design challenges for energy-constrained ad hoc wireless networks," *IEEE Wireless Communications*, vol. 9, no. 4, pp. 8–27, 2002.

[12] C. Barrett, A. Marathe, M. V. Marathe, and M. Drozda, "Characterizing the interaction between routing and MAC protocols in ad-hoc networks," in *Proceedings of the 3rd ACM*

International Symposium on Mobile Ad Hoc Networking & Computing (MobiHoc '02), pp. 92–103, ACM Press, Lausanne, Switzerland, June 2002.

[13] R. Jiang, V. Gupta, and C. Ravishankar, "Interactions between TCP and the IEEE 802.11 MAC protocol," in *Proceedings of DARPA Information Survivability Conference and Exposition (DISCEX '03)*, vol. 1, pp. 273–282, Washington, DC, USA, April 2003.

[14] V. Kawadia and P. R. Kumar, "A cautionary perspective on cross-layer design," *IEEE Wireless Communications*, vol. 12, no. 1, pp. 3–11, 2005.

[15] K. K. Vadde and V. R. Syrotiuk, "Factor interaction on service delivery in mobile ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 7, pp. 1335–1346, 2004.

[16] K. K. Vadde and V. R. Syrotiuk, "Quantifying factors affecting quality of service in mobile ad hoc networks," *Simulation*, vol. 81, no. 8, pp. 547–560, 2005.

[17] K. K. Vadde, V. R. Syrotiuk, and D. C. Montgomery, "Optimizing protocol interaction using response surface methodology," *IEEE Transactions on Mobile Computing*, vol. 5, no. 6, pp. 627–639, 2006.

[18] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, 2005.

[19] L. Berlemann, S. Mangold, and B. H. Walke, "Policy-based reasoning for spectrum sharing in cognitive radio networks," in *Proceedings of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, pp. 1–10, Baltimore, Md, USA, November 2005.

[20] M. Buddhikot, P. Kolodzy, S. Miller, K. Ryan, and J. Evans, "DIMSUMnet: new directions in wireless networking using coordinated dynamic spectrum access," in *Proceedings of the 6th IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM '05)*, pp. 78–85, Taormina, Italy, June 2005.

[21] A. Sahai, N. Hoven, and R. Tandra, "Some fundamental limits in cognitive radio," in *Allerton Conference on Communication, Control and Computing*, Monticello, Va, USA, October 2004.

[22] S. Mishra, D. Cabric, C. Chang, et al., "A real time cognitive radio testbed for physical and link layer experiments," in *Proceedings of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, pp. 562–567, Baltimore, Md, USA, November 2005.

[23] R. W. Thomas, L. A. DaSilva, and A. B. MacKenzie, "Cognitive networks," in *Proceedings of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, pp. 352–360, Baltimore, Md, USA, November 2005.

[24] C. Cordeiro, K. Challapali, D. Birru, and N. Shankar, "IEEE 802.22: the first worldwide wireless standard based on cognitive radios," in *Proceedings of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, pp. 328–337, Baltimore, Md, USA, November 2005.

[25] Website, http://www.opnet.com/.

[26] C. Doerr, M. Neufeld, J. Fifield, T. Weingart, D. C. Sicker, and D. Grunwald, "MultiMAC—an adaptive MAC framework for dynamic radio networking," in *Proceedings of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, pp. 548–555, Baltimore, Md, USA, November 2005.

[27] M. J. Anderson and P. J. Whitcomb, *Doe Simplified: Practical Tools for Effective Experimentation*, Productivity Press, Portland, Ore, USA, 2000.

[28] E. Kohler, R. Morris, B. Chen, J. Jannotti, and M. F. Kaashoek, "The click modular router," *ACM Transactions on Computer Systems*, vol. 18, no. 3, pp. 263–297, 2000.

Hindawi Publishing Corporation
Advances in Multimedia
Volume 2007, Article ID 94040, 11 pages
doi:10.1155/2007/94040

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 Wu[1]

[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.

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

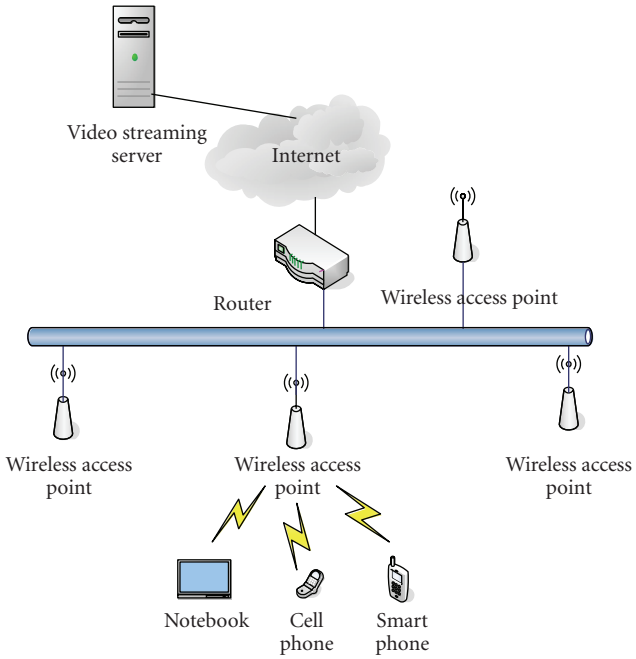


Figure 1: An example of video streaming relayed by access points.

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

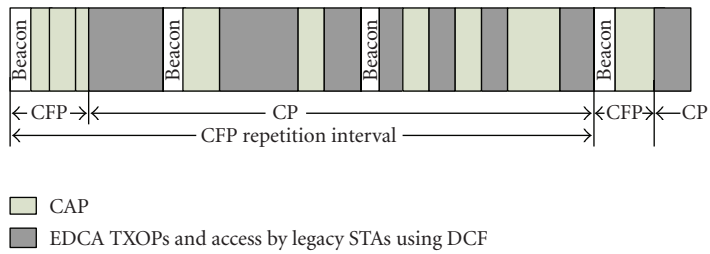


Figure 2: The CAP/CFP/CP periods.

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 (SImax)*: 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* ($A_{\min}^{\text{PHY}}$): the minimum physical bitrate 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 nonstationary. 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

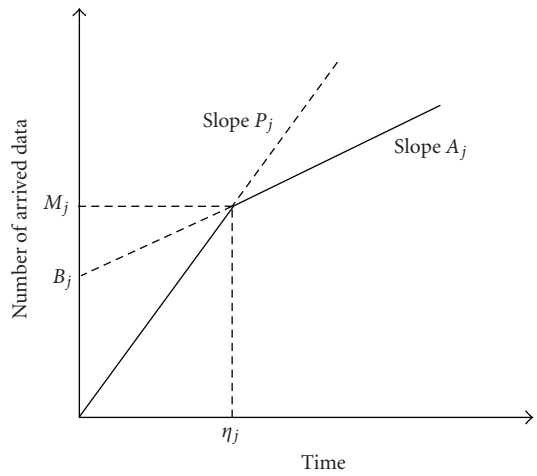


Figure 3: The traffic envelop for a traffic flow with the DTB shaper.

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 pregenerated.

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

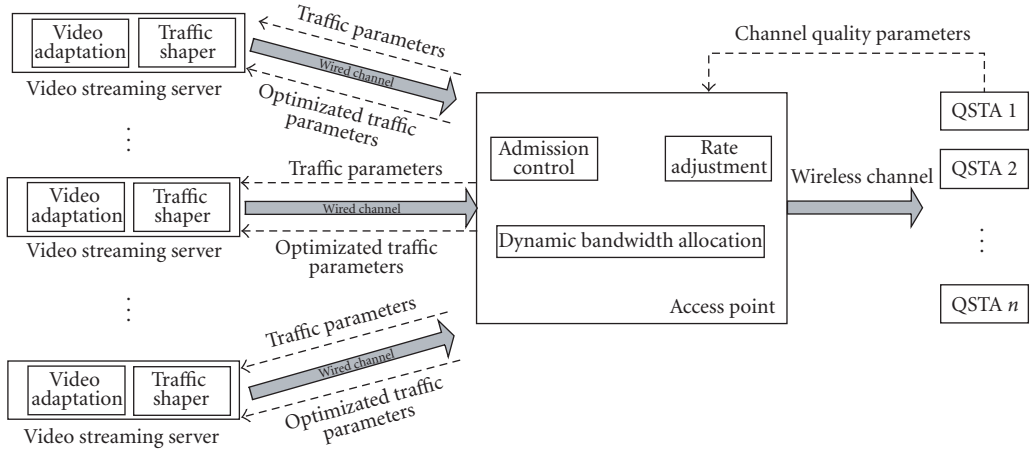


Figure 4: The cross-layer architecture for video streaming over HCCA.

(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 decison 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 torelable 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 queusing delay is deterimined 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 typcially 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 ith 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, $(i = 1,\ldots,N)$ denote the minimum bandwidth needed to guarantee the QoS requirements, $P\{d_i > T_i^D\} \le \epsilon_i$, for each class. Clearly, C_i^g depends on the aggregated traffic rate statistics of the ith 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 ith 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 ith class as

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

with

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

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 jth flow of the ith class, respectively. If a flow j is stationary and its arrival $a_{ij}[t, t+\tau]$ is upper bounded, that is, $a_{ij}[t, t+\tau] \le q_{ij}(\tau)$ for all $t, \tau > 0$, according to [27], its rate-variance envelop should be upper bounded as

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

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

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

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

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

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

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

$$\phi_{ij} = A_{ij}^{\min} r_{ij}. \tag{8}$$

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

$$RV_{ij}(\tau) = \begin{cases} A_{ij}^{\min}\left(P_{ij}^{\min} - A_{ij}^{\min}\right)\left(r_{ij}\right)^2, & 0 \le \tau \le \eta_{ij}, \\ \dfrac{A_{ij}^{\min} B_{ij}\left(r_{ij}\right)^2}{\tau}, & \eta_{ij} \le \tau \le \beta_i, \end{cases} \tag{9}$$

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]

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

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 $q_{ij}(\tau) \le (B_{ij} + A_{ij}^{\min}\tau)r_{ij}$ and we obtain

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

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 ith 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 $P\{d_i > T_i^D\} \le \epsilon_i$, we can also derive the minimum bandwidth needed for the ith 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 $C^{\mathrm{PHY}}\delta$, where C^{PHY} is the physical bandwidth of a WLAN and δ is the percentage of polling-based transmission specified in HCF. If $C^g \le C^{\mathrm{PHY}}\delta$, 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,

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

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,

$$C^{\mathrm{rm}} = C^{\mathrm{PHY}}\delta - C^{g}. \tag{13}$$

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,

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

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 $A_{ij}^{\max}$ 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

$$\omega_i = \frac{C_i^g}{\sum_{i=1}^{N} C_i^g} \tag{15}$$

and the total bandwidth C_i for the ith class becomes

$$C_i = \omega_i C^{\mathrm{rm}} + C_i^g. \tag{16}$$

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

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

Solving the equation above, we obtain

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

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

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

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 ith 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

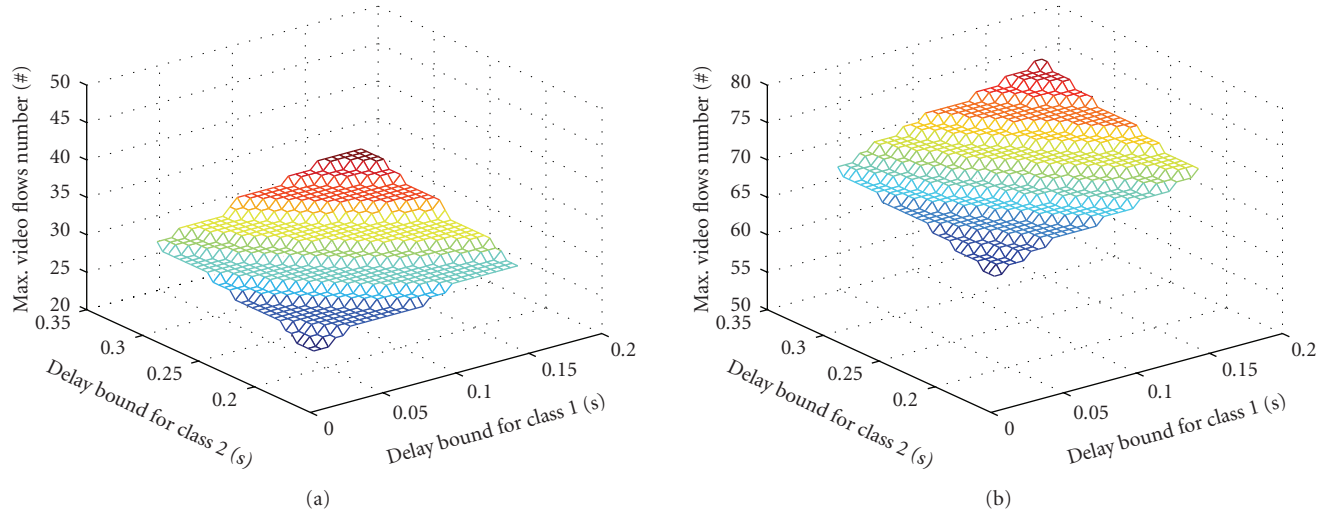


FIGURE 5: The number of admitted flows using different admission control schemes. (a): using the rate guarantee-based admission control. (b): our proposed rate-variance-based admission control.

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

$$\Omega_{ij} = \begin{cases} \Omega_{ij}^{\max}, & j = 1, 2, \ldots, k-1, \\ \Omega_{ij}^{\min}, & j = k, \ldots, K_i, \end{cases} \tag{20}$$

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.

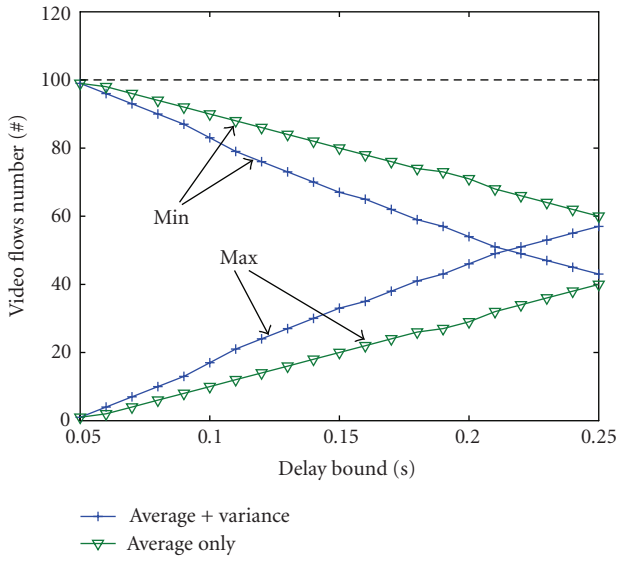


FIGURE 6: The number of flows admitted at $\Omega^{\max}$ and $\Omega^{\min}$ using different resource allocation schemes.

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 nth 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)

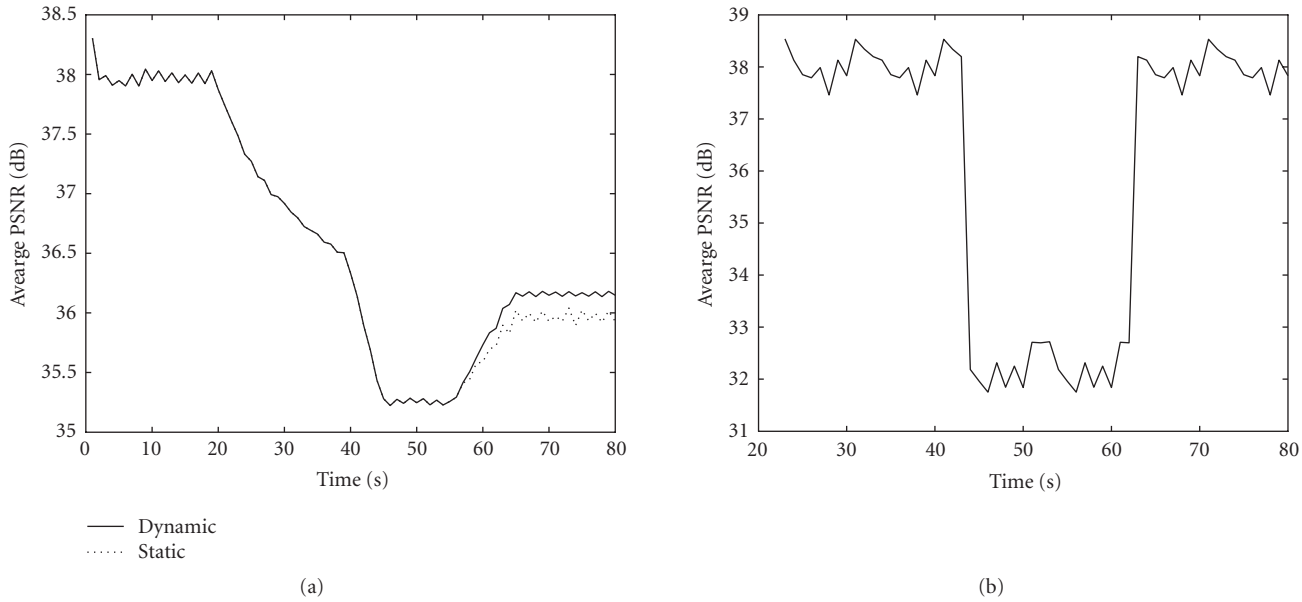


Figure 7: (a): the average PSNR. (b): the PSNR result of the 46th video flow.

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. Guéroui, "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, Febrauary 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 802.1 IE based networks," in *IEE 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.

Hindawi Publishing Corporation
Advances in Multimedia
Volume 2007, Article ID 56592, 12 pages
doi:10.1155/2007/56592

Research Article

A Cross-Layer Optimization Approach for Energy Efficient Wireless Sensor Networks: Coalition-Aided Data Aggregation, Cooperative Communication, and Energy Balancing

Qinghai Gao,[1] Junshan Zhang,[1] Xuemin (Sherman) Shen,[2] and Bryan Larish[3]

[1] *Electrical Engineering Department, Arizona State University, Tempe, AZ 85287, USA*
[2] *Electrical and Computer Engineering Department, University of Waterloo, Waterloo, ON, Canada N2L 3G1*
[3] *Space and Naval Warfare Systems Center, 53560 Hull Street, San Diego, CA 92152, USA*

Received 29 December 2006; Revised 16 March 2007; Accepted 17 March 2007

Recommended by Jianwei Huang

We take a cross-layer optimization approach to study energy efficient data transport in coalition-based wireless sensor networks, where neighboring nodes are organized into groups to form coalitions and sensor nodes within one coalition carry out cooperative communications. In particular, we investigate two network models: (1) many-to-one sensor networks where data from one coalition are transmitted to the sink directly, and (2) multihop sensor networks where data are transported by intermediate nodes to reach the sink. For the many-to-one network model, we propose three schemes for data transmission from a coalition to the sink. In scheme 1, one node in the coalition is selected randomly to transmit the data; in scheme 2, the node with the best channel condition in the coalition transmits the data; and in scheme 3, all the nodes in the coalition transmit in a cooperative manner. Next, we investigate energy balancing with cooperative data transport in multihop sensor networks. Built on the above coalition-aided data transmission schemes, the optimal coalition planning is then carried out in multihop networks, in the sense that unequal coalition sizes are applied to minimize the difference of energy consumption among sensor nodes. Numerical analysis reveals that energy efficiency can be improved significantly by the coalition-aided transmission schemes, and that energy balancing across the sensor nodes can be achieved with the proposed coalition structures.

1. INTRODUCTION

Wireless sensor networks have received much attention in recent years because of their great potential in many application domains, including environmental monitoring, target tracking, security, or system control (see [1] and the references therein). Depending on the specific applications, different wireless sensor networks have different traffic patterns. For example, sensors deployed for intrusion detection may only need to send very basic signal to the control center, while sensors monitoring enemy movements may need to send multimedia signals. With the increase of storage space and computing power of sensors, wireless multimedia sensor networks emerge as a very promising technology. One important example of multimedia sensor networks is a surveillance system with video cameras, which has great potential for environmental monitoring, patient care, or security. On the other hand, the multimedia data generated in such settings have a variety of different quality of service (QoS) requirements such as stringent delay constraints for high data rate video services. Supporting multimedia applications and services puts forth great challenges on the design of wireless sensor networks to meet these QoS demands.

In wireless sensor networks, sensor nodes are often powered by batteries with limited energy. It is difficult, if not impossible, to replace or recharge the batteries in many practical scenarios. As a result, improving energy efficiency is of great importance for the design of wireless sensor networks. For sensor networks supporting multimedia applications, the energy issue becomes even more critical because of possibly larger traffic demand. Thus motivated, in this paper we study two fundamental aspects impacting the network lifetime: *energy saving for data transport* and *energy balancing across sensor nodes*. Simply put, energy saving is concerned with the total energy consumption for transporting data to the sink, and energy balancing is concerned with the difference of energy consumption among sensor nodes.

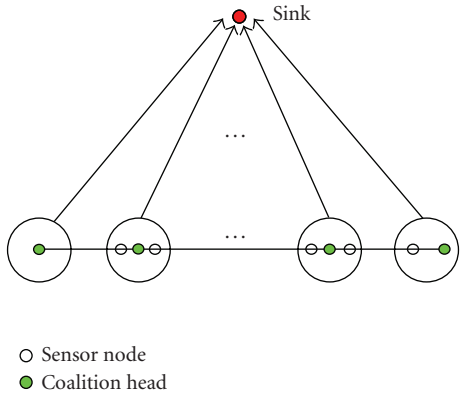


Figure 1: A many-to-one sensor network.

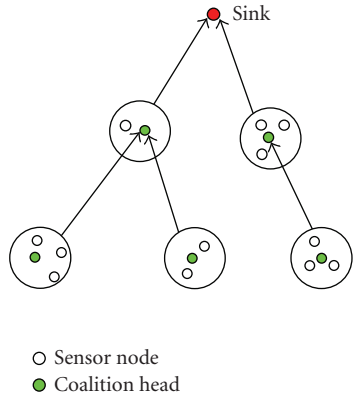


Figure 2: A multihop sensor network.

The clustering approach has proved to be one of the most effective mechanisms to improve energy efficiency in wireless sensor networks (see, e.g., [2–7]). In a cluster-based sensor network, sensor nodes are organized into groups, each with a cluster head. Traditionally, sensor nodes in a cluster send their data to the corresponding cluster head, and the cluster head forwards the data to the neighboring cluster along the route or to the sink directly. Building on the cluster-based model, we propose a coalition-aided network structure, where sensor nodes within one coalition can carry out cooperative data transmissions. This structure is motivated by the two key features of wireless sensor networks: *node cooperation* and *data correlation*, which differentiate wireless sensor networks from conventional wireless networks. In particular, the coalition head (CH) carries out data aggregation and coordinates the sensor nodes within the coalition, but not necessarily transmits the data itself. We use the term *coalition* instead of cluster to emphasize the cooperation among sensor nodes in a coalition, whereas in a traditional cluster the cluster head performs the bulk of the communication tasks.

We consider two network models in our work, that is, a many-to-one network model and a multihop network model. In a many-to-one network, data from one coalition are transmitted to the sink in one hop (see Figure 1). We propose three schemes for data transmission from each coalition to the sink. In scheme 1, one node in the coalition is selected randomly by the CH to transmit the data, implying that the energy consumption is balanced across the sensor nodes within the coalition. In scheme 2, the sensor node with the best channel condition transmits the data, yielding multiuser diversity gain. In scheme 3, all the sensor nodes within the coalition transmit as a virtual antenna array, so that cooperative diversity gain could be achieved. For the sake of fair comparison, we also take into account the energy consumption for intracoalition communications and channel contention. Our results show that significant energy saving can be achieved by the coalition-aided transmission schemes, and as expected, scheme 3 achieves the best performance.

In some practical scenarios, some sensor nodes may not be capable of communicating directly with the sink (e.g., due to limited power), and the data need to be relayed by intermediate nodes to reach the sink. Building on the studies for the single-hop networks, we investigate coalition-based multihop networks, where one coalition sends the data to another along the route until the sink is reached, as illustrated in Figure 2. Besides using coalition-aided data transmission schemes to balance the energy consumption across the sensors within one coalition, we investigate optimal coalition planning, using unequal coalition sizes, to balance the energy consumption across different coalitions. Based on the energy consumption model for intracoalition and intercoalition communications, we treat energy balancing as an optimization problem that is targeted at minimizing the difference of energy consumption among the sensor nodes. To the best of our knowledge, our study is the first work addressing both intracoalition and intercoalition energy balancing issues in wireless sensor networks. In particular, we investigate two types of multihop network models with different traffic patterns. In a Type I network, only part of the sensor nodes have data to transmit and others serve as relays; and in a Type II network, all sensor nodes have data to transmit. Numerical examples and simulations show that energy balancing across the sensor nodes can be achieved with the proposed coalition structures.

Many methods have been developed to improve energy efficiency at individual protocol layers (see, e.g., [3, 8–11] and the references therein). Since energy consumption takes place in all layers, the methods considering layers separately leave much room for improving energy efficiency further from a cross-layer point of view. In particular, we explore the interplay between physical layer, MAC layer, and coalition planning at the routing layer. The formation of coalitions facilitates data aggregation and mitigates channel contention; and the MAC layer transmissions exploit the physical layer channel conditions. For instance, in the scheme with multiuser diversity, the channel state information is used to choose the node with the best channel condition within one coalition for data transmission. In a multihop network, the

data transmission schemes serve as the basis for the optimal coalition planning, which helps to achieve energy balancing across the sensor nodes.

The remainder of this paper is organized as follows. In Section 2 we give a brief review of the related work. Section 3 analyzes the energy efficiency of the coalition-aided data transmission schemes in many-to-one sensor networks. In Section 4, the optimal coalition planning with unequal coalition sizes is investigated for multihop sensor networks. Finally, Section 5 concludes this paper.

2. RELATED WORK

Energy efficiency of wireless sensor networks has received much attention in recent years. In particular, hierarchical protocols, in which sensor nodes are organized into clusters, have been studied extensively (see, e.g., [2–7]). In [3], the authors proposed the LEACH (low energy adaptive clustering hierarchy) protocol, in which a cluster head aggregates data from sensor nodes within the cluster and send the aggregated data directly to the sink. Furthermore, cluster head rotation scheme was proposed such that the role of cluster head is dynamically rotated among sensor nodes. It is shown that LEACH can improve the energy efficiency, at the cost of extra overhead due to dynamic clustering. As an enhancement of LEACH, Younis and Fahmy [7] proposed HEED (hybrid energy-efficient distributed clustering), where the cluster head selection is carried out periodically according to a hybrid of the node residual energy and a secondary parameter such as node proximity to its neighbors or node degree, with the assumption that multiple power levels are available at sensor nodes. It is shown that HEED prolongs network lifetime and achieves a well-distributed set of clusters. Note that in both of the protocols mentioned above, an energy consumption model is assumed such that a fixed amount of energy is needed to transmit one information bit, given the transmit distance. This model does not take into account the time varying channel condition, which can be exploited to adapt the data rate. For instance, given transmission power, if the channel condition is better, the data rate could be larger, and more information bits could be transmitted with certain energy.

Cooperative communication has also been studied in recent years (see, e.g., [12–14]). A survey about cooperative communication can be found in [14], where three cooperative methods, namely detect and forward, amplify and forward, and coded cooperation, are presented. Simply put, distributed sensor nodes can "share" their antennas in a cooperative manner to form a virtual antenna array. In this way, some benefits of MIMO (multiple input multiple output) systems can be achieved. In [12], the authors proposed cooperative MIMO in sensor networks. In their scheme, each node fist broadcasts its data to other local nodes and then the nodes encode the transmission sequence according to the Alamouti diversity scheme [15]. They assume that each node has its own data to transmit and the data correlation property of sensor networks is not exploited. It is shown that both energy saving and delay reduction can be achieved.

There have been a number of studies on energy balancing in wireless sensor networks (see, e.g., [3, 16–24]). In [19], the authors proposed and analyzed four strategies that are used to balance the energy consumption of the nodes, including distance variation, balanced data compression, routing, and equalization of the end-to-end reliability. For cluster-based sensor networks, most of the existing studies focus on energy balancing across CHs, assuming that CHs take the full responsibility to forward the data. In [22], the authors proposed the routing-aware optimal cluster planning to achieve the balanced power consumption. The difference of energy consumption among cluster heads is minimized with respect to the clustering profile. Their analytical solutions and simulation results show that energy balancing across the CHs can be improved. In [21], the authors proposed a clustering scheme which takes into account the distances between the sensor nodes and the sink. Accordingly, the clusters close to the base station have smaller sizes than those farther away from the base station. In [23], the authors considered a heterogeneous network where some powerful nodes take on the cluster head role to control network operation, and an unequal clustering approach was proposed to balance the energy consumption of CHs in multihop wireless sensor networks.

3. MANY-TO-ONE SENSOR NETWORKS: COALITION-AIDED DATA TRANSPORT

3.1. System model

Following [25, 26], we consider a one-dimensional network model which consists of N sensor nodes and one sink, and the N sensors are randomly placed on a line of length L (see Figure 1). Based on the positions of sensor nodes, local neighboring nodes form coalitions. Let M be the number of coalitions and n_i the number of sensor nodes in the ith coalition. Then we have $\sum_{i=1}^{M} n_i = N$.

As is standard in [3], we assume that the distances between the sensor nodes and the sink are much larger than those among sensor nodes, and accordingly, we treat the distances between the sensor nodes and the sink as more or less the same (denoted as d). We assume that all the intracoalition communication channels can be modeled as AWGN (additive white Gaussian noise) channels, and this is applicable to scenarios where there exists strong line of sight (LOS) between neighboring sensor nodes in a densely deployed wireless sensor network. In contrast, we assume the communications between the sensor nodes and the sink undergo Rayleigh fading. We assume that the sink does the network training by broadcasting pilot signals periodically, so that the sensor nodes can estimate the corresponding fading channel gain. We also assume that the channels from different sensor nodes to the sink are independent.

We assume a homogeneous random field, and denote by H_0 the information entropy of each sensor node. In the ith coalition, we define the joint entropy of the n_i sensor nodes as $\mathbf{H_i}$. Note that the number of information bits from different coalitions may be different.

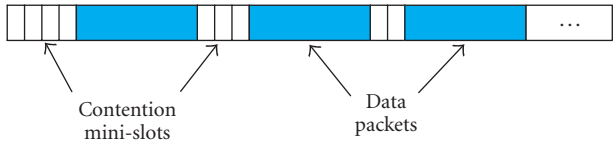


Figure 3: CSMA-based random access.

3.2. Coalition-aided data transmission

We assume TDMA (time division multiple access) scheduling for the intracoalition communications, which take place as follows:

(i) the sensor nodes send their data to the CH in their time slots;
(ii) the CH carries out data aggregation;
(iii) the CH broadcasts the aggregated data back to the sensor nodes.

We note that some overhead may be incurred by broadcasting the data back to sensor nodes. However, since this broadcast occurs only over a short distance within a coalition, the overhead is negligible. We elaborate further on this in Section 3.3 where the energy consumption is analyzed.

We assume that the intracoalition communications of different coalitions do not interfere each other. For instance, as proposed in [3], if a unique CDMA (code division multiple access) code is used by sensor nodes within each coalition, then the neighboring coalitions' radio signals would be filtered out and not corrupt the communication in the coalition.

For the data transmissions from the coalitions to the sink, the CHs compete for the channel on behalf of the coalitions. We assume that CSMA (carrier sensing multiple access) based random access scheme is used to reserve the channel, as illustrated in Figure 3. Let the CHs probe the channel in a minislot with probability p. If the pilot packet from one CH is transmitted successfully, then an ACK signal is sent out by the sink and the channel is reserved for the coalition. We assume that the ACK signal can be received by all the sensors within the coalition, so that the data transmission can be triggered in the subsequent time slot. If collisions occur, the CHs would probe the channel again in the next contention minislot.

We study three schemes for data transmissions from the coalition with reservation to the sink. In scheme 1, one of the sensor nodes is selected randomly by the CH to transmit the data. In this scheme, the CH does not have the channel status information between the sensor nodes and the sink, but just aims to balance the energy consumption across the sensors within the coalition. In scheme 2, the sensor node with the best channel condition in the coalition transmits the data. To achieve the multiuser diversity gain, the sensor nodes need to send their channel status information (between the sensor nodes and the sink) to the CH in their time slots, (which can be simply inserted in the header of the data packets,) so that CH can choose the one with the best channel gain to send the data. In scheme 3, all the sensor nodes within the coalition transmit in a cooperative manner to form a virtual antenna array. In this scheme, the CH also needs the channel conditions between sensor nodes and the sink to apply the transmitter beamforming across the sensor nodes [27, 28]. We illustrate these three schemes by the following example.

Example 1. As illustrated in Figure 4, there are three sensor nodes in the coalition. The channel gains in a given time slot are assumed to be $g_1 < g_2 < g_3$. The solid line indicates the data transmission from the corresponding sensor node. In scheme 1, node A is chosen "unfortunately" although it has the worst channel condition. In scheme 2, node B is chosen because it has the best channel gain g_3. All three sensor nodes transmit the data to the sink in scheme 3.

We observe that there are benefits from at least three perspectives in a coalition-based sensor network. First, data aggregation can be carried out for the data from sensor nodes within one coalition since the data collected by neighboring nodes are typically correlated. That is, the amount of total information to be transmitted to the sink is less than that in the noncoalition case. Second, after the formation of a coalition, the coalition behaves as one metanode to communicate with the sink, and as a result, the channel contention is reduced significantly. Third, the sensor nodes within one coalition could transmit the data to the sink in a cooperative manner such that cooperative diversity gain can be achieved [14].

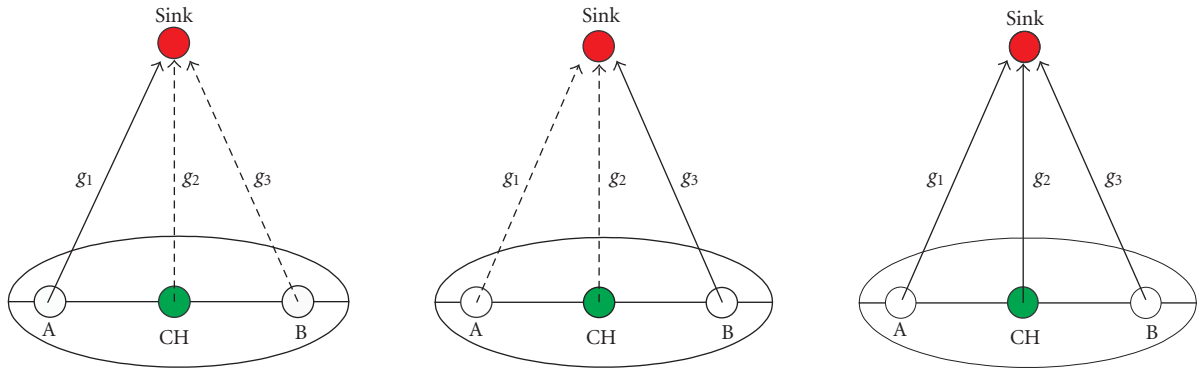
Needless to say, intracoalition communications are needed to carry out coalition-aided data transmissions. Specifically, channel conditions of nodes within one coalition are needed for the multiuser diversity scheme and the cooperative diversity scheme, which would incur additional message passing. Then, a natural question to ask is how much net gain the coalition-aided data transmission schemes yield, and that is the main subject of this section. In the following, we analyze the energy consumption of the proposed data transmission schemes, and compare them with the noncoalition case and the traditional cluster scheme where the CHs take the responsibility to transmit the data to the sink.

3.3. Energy consumption analysis

In what follows, we analyze the energy consumption corresponding to three parts, namely the intracoalition communications, the channel contention, and the data transmissions from coalitions to the sink.

3.3.1. Intracoalition communications

We first examine the cost of intracoalition communications. Recall that we assume an AWGN channel model and the TDMA mechanism for intracoalition communications. Each node transmits with fixed power P_t. In the ith coalition, the $n_i - 1$ sensor nodes send their information of H_0 bits to the CH and the CH broadcasts back the $\mathbf{H_i}$ bits to the sensor nodes. As a result, the intracoalition communications involve totally n_i transmissions. Let R_0 be the data rate between sensor nodes and the CH.



(a) Random selection (b) Multiuser diversity (c) Cooperative diversity

Figure 4: Data transmission schemes from a coalition to the sink.

The energy consumption for the intra-coalition communications is the sum of those from all the coalitions:

$$\mathcal{E}_{\text{intra}} = P_t \sum_{i=1}^{M} \frac{(n_i - 1)H_0 + \mathbf{H_i}}{R_0}. \tag{1}$$

Note that in traditional cluster-based sensor network, the CHs transmit the data to the sink, without sending data back to the sensors. So the intracluster energy consumption for traditional cluster-based sensor network is given by

$$\mathcal{E}_{\text{trad\_intra}} = P_t \sum_{i=1}^{M} \frac{(n_i - 1)H_0}{R_0}. \tag{2}$$

3.3.2. Channel contention among coalitions

For the data transmissions from coalitions to the sink, CSMA is used to reserve the channel. In the coalition case, the CHs contend for the channel on behalf of the coalitions, whereas in the noncoalition case all sensor nodes with data contend for the channel. We assume that each node knows the number of contending nodes (m) and contends the channel with the optimal probability $p = 1/m$. The probability that one contending node wins the channel is $p_{\text{succ}} = (1 - 1/m)^{m-1}$. Since the number of slots needed until the successful reservation is a geometric random variable, the average total number of contending slots for the coalition case (M coalitions) is given by

$$S = \sum_{i=1}^{M} \frac{1}{(1 - 1/(M - i + 1))^{M-i}}. \tag{3}$$

Assuming that the time length of the contention minislot is τ, we have the energy consumption for channel contention in the coalition case:

$$\mathcal{E}_{\text{cont}} = \tau P_t \sum_{i=1}^{M} \frac{1}{(1 - 1/(M - i + 1))^{M-i}}. \tag{4}$$

Similarly, the energy consumption for channel contention in the noncoalition case is given by

$$\mathcal{E}'_{\text{cont}} = \tau P_t \sum_{i=1}^{N} \frac{1}{(1 - 1/(N - i + 1))^{N-i}}. \tag{5}$$

3.3.3. Data transmissions from one coalition to the sink: the fixed transmission power case

In this subsection, we examine the energy consumption for data transmissions from a coalition with reservation to the sink. We consider two popular transmission power allocation schemes, namely the fixed transmission power scheme and the channel inversion scheme. For the sake of fair comparison, we assume that, in the fixed transmission power case, the total transmission power from the coalition is fixed for all the data transmission schemes; and that the total received power from the coalition is a constant in the channel inversion case.

With fixed transmission power, the transmission data rate changes with the channel gain. We use Shannon capacity to approximate the transmission data rate. Let P_t be the transmission power and ρ the average received SNR in a corresponding SISO (single input single output) fading link [28]. We assume Rayleigh fading with unit average channel gain, that is, $E[g] = 1$ where g is exponentially distributed. In scheme 1, one node in a coalition is chosen randomly to transmit the data. The average transmission data rate from the sensor node to the sink is given by

$$E[W \log(1 + \rho g)] = \int_0^{\infty} W \log(1 + \rho g) e^{-g} dg = -\frac{W}{\ln 2} e^{1/\rho} E_i\left(-\frac{1}{\rho}\right), \tag{6}$$

where $E_i(\cdot)$ is the exponential integral function defined as $E_i(x) = -\int_{-x}^{\infty} (e^{-t}/t)dt$ [29]. Then the energy consumption for data transmission from the coalition to the sink is given by

$$\mathcal{E}_{\text{tosink}} = P_t \sum_{i=1}^{M} \frac{\mathbf{H_i}}{E[W \log(1 + \rho g)]}. \tag{7}$$

Note that for traditional cluster scheme, the energy consumption for data transmission from a CH to the sink is the same as above.

In scheme 2, the node with the best channel condition within the coalition is chosen to transmit the data. Denote by $g_{mi} = \max\{g_{i1}, g_{i2}, \ldots, g_{in_i}\}$ the best channel gain in the ith coalition, where g_{ij} is the channel gain of the jth node in the ith coalition. The expected value of data rate is given by [30]

$$E[W\log(1+\rho g_{mi})] = \int_0^\infty W\log(1+\rho g_{mi}) n_i e^{-g_{mi}}\left[1-e^{-g_{mi}}\right]^{n_i-1} dg_{mi}. \tag{8}$$

This integration can be evaluated by numerical methods. The average energy consumption of scheme 2 is given by

$$\mathcal{E}_{\text{tosink}} = P_t \sum_{i=1}^{M} \frac{\mathbf{H_i}}{E[W\log(1+\rho g_{mi})]}. \tag{9}$$

The cooperative transmission technique is employed in scheme 3. More specifically, all the sensor nodes within a coalition transmit in a cooperative manner to form a virtual antenna array, that is, transmitter beamforming is applied across the sensor nodes such that the signal received at the sink can be combined coherently [27, 28]. Let $g_{ci} = \sum_{j=1}^{n_i} g_{ij}$ denote the sum of channel gains in the ith coalition. Then, the average data rate for cooperative diversity techniques can be derived as [30]

$$E[W\log(1+\rho g_{ci})] = \int_0^\infty W\log(1+\rho g_{ci}) \frac{1}{(n_i-1)!} {g_{ci}}^{n_i-1} e^{-g_{ci}} dg_{ci} \tag{10}$$

and it can be evaluated numerically. We obtain the average energy consumption of scheme 3 as

$$\mathcal{E}_{\text{tosink}} = P_t \sum_{i=1}^{M} \frac{\mathbf{H_i}}{E[W\log(1+\rho g_{ci})]}. \tag{11}$$

Combining the energy consumption for intra-coalition communications, the channel contention, and the data transmissions from coalitions to the sink, we have the total energy consumption of the three schemes:

$$\mathcal{E} = \mathcal{E}_{\text{intra}} + \mathcal{E}_{\text{cont}} + \mathcal{E}_{\text{tosink}}, \tag{12}$$

where $\mathcal{E}_{\text{tosink}}$ for scheme 1 to 3 is given by (7), (9), and (11), respectively.

For comparison, we also derive the performance of traditional cluster scheme and the non-coalition case with fixed transmission power. For the traditional cluster scheme, the energy consumption is given by

$$\mathcal{E}_{\text{trad}} = \mathcal{E}_{\text{trad\_intra}} + \mathcal{E}_{\text{trad\_cont}} + \mathcal{E}_{\text{trad\_tosink}}, \tag{13}$$

where $\mathcal{E}_{\text{trad\_intra}}$, $\mathcal{E}_{\text{trad\_cont}}$, and $\mathcal{E}_{\text{trad\_tosink}}$ are given by (2), (4), and (7), respectively.

For the non-coalition case, the energy consumption comes from the channel contention and the data transmission from the sensors to the sink. Then the average total energy consumption of the non-coalition case is given by

$$\mathcal{E}' = \mathcal{E}'_{\text{cont}} + P_t \frac{NH_0}{E[W\log(1+\rho g)]}, \tag{14}$$

where $\mathcal{E}'_{\text{cont}}$ is the energy consumption for channel contention given by (5).

3.3.4. Data transmissions from one coalition to the sink: the channel inversion case

With channel inversion, the transmitter adjusts the transmission power with the channel gain, that is, $P_t = P/g$, such that the received power at the sink is a constant (P). Accordingly, the data rate R is also a constant and the time needed for data transmission is the same for all the coalition-aided data transmission schemes. We consider the energy consumption for the three data transmission schemes in the following.

In scheme 1, one node in the coalition is selected randomly to transmit the data. We assume that the sensor node does not transmit if the channel gain is below a threshold g_{th}. Then, the average energy consumption is given by

$$\mathcal{E}_{\text{tosink}} = E\left[\frac{P}{g}\right] \sum_{i=1}^{M} \frac{\mathbf{H_i}}{R}, \tag{15}$$

where the average transmission power $E[P/g]$ is given by [29]

$$E\left[\frac{P}{g}\right] = \int_{g_{th}}^\infty \frac{P}{g} e^{-g} dg = -E_i(-g_{th})P. \tag{16}$$

In scheme 2, the average energy consumption with multiuser diversity is given by

$$\mathcal{E}_{\text{tosink}} = \sum_{i=1}^{M} E\left[\frac{P}{g_{mi}}\right] \frac{\mathbf{H_i}}{R}, \tag{17}$$

where the average transmission power in the ith coalition is given by [29]

$$E\left[\frac{P}{g_{mi}}\right] = \int_0^\infty \frac{P}{g_{mi}} n_i e^{-g_{mi}}\left[1-e^{-g_{mi}}\right]^{n_i-1} dg_{mi} = Pn_i(-1)^{n_i} \sum_{k=0}^{n_i-1} (-1)^k \binom{n_i-1}{k} \ln(n_i-k). \tag{18}$$

In scheme 3, the average energy consumption with cooperative diversity is given by

$$\mathcal{E}_{\text{tosink}} = \sum_{i=1}^{M} E\left[\frac{P}{g_{ci}}\right] \frac{\mathbf{H_i}}{R}, \tag{19}$$

where the average transmission power in the ith coalition is given by [29]

$$E\left[\frac{P}{g_{ci}}\right] = \int_0^\infty \frac{P}{g_{ci}} \frac{1}{(n_i-1)!} {g_{ci}}^{n_i-1} e^{-g_{ci}} dg_{ci} = \frac{P}{n_i-1}. \tag{20}$$

Table 1: Numerical parameters.

P_t	1 mw	Transmission power of sensor nodes
d	100–200 m	Distance between sensor nodes and the sink
r	10 m	Distance between sensor nodes and CH
N	10	Number of sensors
n	2	Number of sensors within a coalition
H_0	2 k	Information bits of each sensor node
H_p	200	Number of bits in a pilot packet
W	1 MHz	Frequency bandwidth (Hz)
α	4	Path loss factor between sensors and sink
α_0	2	Path loss factor between sensors and CH
G_t	0 dB	Transmit antenna gain
G_r	0 dB	Receive antenna gain
f_c	2.4 GHz	Carrier frequency

Then we can get the total energy consumption with channel inversion by combining the energy consumption for intra-coalition communications, the channel contention and the data transmissions from coalitions to the sink.

We also consider the performance of the tradition cluster scheme and the non-coalition case for comparison. For the traditional cluster scheme with channel inversion, the energy consumption is the same as in (13) except that $\mathcal{E}_{\text{trad\_tosink}}$ is given by (15) here. For the non-coalition case, the average energy consumption of the non-coalition case is given by

$$\mathcal{E}' = E\left[\frac{P}{g}\right]\frac{NH_0}{R} + \mathcal{E}'_{\text{cont}}, \tag{21}$$

where $E[P/g]$ is shown in (16) and $\mathcal{E}'_{\text{cont}}$ is given by (5).

3.4. Numerical examples

In this section, we illustrate our findings via examples. We consider a one-dimensional network where sensors are uniformly deployed. Each coalition comprises of two sensor nodes. For a transmitter-receiver separation d, the average received power is given by $P_r(d) = P_r(d_0)(d_0/d)^{\alpha}$, where α is the path loss factor and $P_r(d_0) = (P_t G_t G_r \lambda^2)/(4\pi)^2 {d_0}^2$ is the received power at the close-in distance d_0, with d_0 normalized to 1 meter [30]. The Shannon capacity is used to approximate the data rate. The parameters for our numerical examples are summarized in Table 1.

A simple empirical model is used to model the joint entropy of two sources as a function of their distance r: $H'(r) = H_0 + [1 - 1/(r/c + 1)]H_0$, where c is a constant that characterizes the extent of spatial correlation in the data [31]. Assuming the correlation constant $c = r$, we have the joint entropy of a coalition $\mathbf{H_i} = 1.5H_0$.

We define the energy saving gain as the ratio of saved energy for each transmission scheme against the energy consumption of the non-coalition case: $\eta(E) \triangleq (\mathcal{E}' - \mathcal{E})/\mathcal{E}'$. The energy consumption with fixed transmission power are shown in Figure 5. (For the channel inversion scheme, numerical examples can be achieved similarly.) From the results it can be seen that the coalition-aided data transmission schemes have better performance than the non-coalition case. Comparing with scheme 1 (random selection), the traditional cluster scheme has almost the same performance, because of the dominating energy consumption for data transmissions from coalitions to the sink. We can also see that scheme 2 and scheme 3 outperform scheme 1, and scheme 3 has the best performance. For example, when the distance between sensor nodes and the sink is 100 meters, the energy saving gain for the three data transmission schemes are 29.74%, 49.47%, and 58.84%, respectively. Note that some overhead is incurred by schemes 2 and 3 since channel status of each node should be maintained and updated from time to time. Moreover, since all the sensor nodes within a coalition transmit in scheme 3, the overhead of circuit energy consumption may become an issue which we ignore in this study.

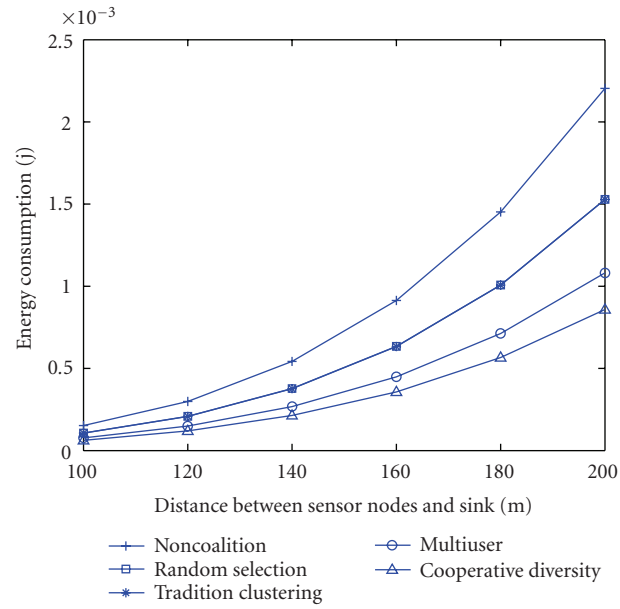


Figure 5: Energy consumption over fading channels.

4. MULTIHOP SENSOR NETWORKS: OPTIMAL COALITION PLANNING AND ENERGY BALANCING

In this section, we focus on energy balancing in multihop sensor networks. In a multihop network, the sensor nodes close to the sink are called in the "hot-spot," in the sense that more traffic is forwarded by these nodes to the sink. Sensor nodes in the hot-spot may deplete their energy faster than other sensors. As a result, the network may not function properly after some nodes die, because of either network partition or insufficient field covering. Motivated by this observation, we investigate the optimal coalition planning to balance the energy consumption among the sensor nodes in the network.

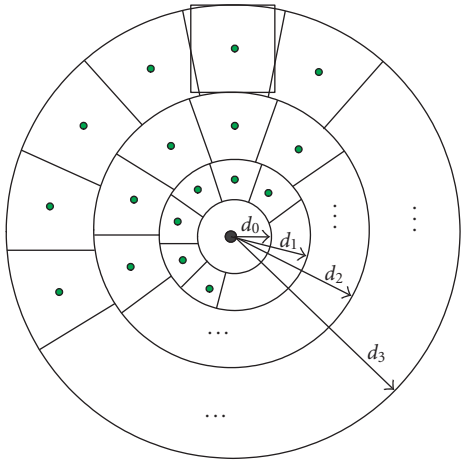


Figure 6: A coalition-based multihop network.

4.1. System model

4.1.1. Network model

Following [22], we consider a homogeneous circular network where the sink is located at the center and the sensors are uniformly deployed in the area $\mathcal{A}$: $\{(x, y) \mid {d_0}^2 \le x^2 + y^2 \le D^2\}$ with node density δ, as illustrated in Figure 6. In light of the symmetric property of this network, we assume that the sensor nodes are divided into K rings, and the ith ring denotes the sensors in the area $\{(x, y) \mid {d_{i-1}}^2 < x^2 + y^2 \le {d_i}^2\}$, $i = 1, \ldots, K$, where d_i is the distance between the outer boundary of the ith ring and the sink and $d_K = D$. The sensor nodes of each ring are grouped into multiple coalitions and the area covered by a coalition is represented by a sector within the ring. In [22], a cluster is approximated by a small circle to facilitate analysis. In this study, based on the shape of a coalition, we approximate it as a square with side length $l_i = d_i - d_{i-1}$, which we believe is more accurate than the circle approximation. (Indeed, as indicated by the simulation results in Section 4.3, the approximation has a negligible impact on network performance.) We also assume that each CH is located at the center of its coalition.

We assume AWGN channels for intracoalition communications and Rayleigh fading channels for intercoalition communications, respectively. The intra-coalition communications take place the same way as proposed in Section 3.2, except that in the Type I model the sensor nodes do not transmit data to the CH. For the intercoalition communications, data from a coalition in the ith ring is sent to the closest CH in the $(i-1)$th ring until the sink is reached. Coalition-aided data transmission schemes can be employed for each hop.

4.1.2. Traffic model

Depending on the specific applications, different wireless sensor networks have different traffic patterns. Roughly speaking, sensor networks can be classified into four categories [32, 33]: continuous, event-driven, query-driven, and hybrid. In the continuous delivery model, each sensor sends out data periodically. In the event-driven data delivery model, the sensors report information only if an event of interest occurs. In the query-based model, the sensors only report their results in response to an explicit request from the sink. Some networks apply a hybrid model using a combination of these models.

The aforementioned traffic patterns can be categorized into two types of traffic models:

(i) Type I: only part of the sensor nodes have data to transmit and other sensors serve as relays;
(ii) Type II: all sensor nodes in the area have data to transmit.

These two models represent different traffic patterns. Type I provides a good model for the event-driven (e.g., for intrusion detection) and query-based wireless sensor networks; and Type II corresponds to the periodical data transmission model (e.g., for field monitoring).

4.1.3. Energy consumption model

In this subsection we examine the transmit energy required for reliable data transmission. We consider intra-coalition communications first. Let e_{i_intra} denote the transmit energy per bit and x_i the communication distance in a coalition of the ith ring. Then, the received energy per bit is given by $e_{i\_\text{intra}}/x_i^{\alpha}$, where α is the path loss factor for intra-coalition communications. To ensure reliable reception, the received energy per bit should be no less than the threshold γ_{intra}. So the required transmit energy per bit for intra-coalition communications is given by

$$e_{i\_\text{intra}} = \gamma_{\text{intra}} x_i^{\alpha}. \tag{22}$$

Next we turn to model the energy consumption for inter-coalition communications. Let e_{i_inter} denote the transmit energy per bit and y_i the communication distance. The received energy per bit is given by $e_{i\_\text{inter}}\xi/y_i^{\beta}$, where ξ is the Rayleigh fading gain seen by the sink and β is the path loss factor for inter-coalition communications. For reliable reception, it is assumed that the expected value of received energy per bit should be no less than a predefined threshold γ_{inter}, that is, $E[e_{i\_\text{inter}}\xi/y_i^{\beta}] \ge \gamma_{\text{inter}}$, where the expectation is taken with respect to the channel variation seen by the receiver (γ_{inter} and γ_{intra} could but not necessarily be the same). Let n_i denote the number of sensor nodes in the coalition. For scheme 1, one sensor node is selected randomly to transmit the data. Assuming normalized channel fading, that is, $E[\xi^{(1)}] = 1$, we have

$$e_{i\_\text{inter}}^{(1)} = \gamma_{\text{inter}} y_i^{\beta}. \tag{23}$$

For scheme 2, since the node with the best channel gain is chosen, the average channel gain is $E[\xi^{(2)}] = \sum_{j=1}^{n_i} 1/j$ [30]. Thus the required transmit energy per bit is given by

$$e_{i\_\text{inter}}^{(2)} = \frac{\gamma_{\text{inter}} y_i^{\beta}}{\sum_{j=1}^{n_i} 1/j}. \tag{24}$$

For scheme 3, the received signal can be added coherently, so the average channel gain is given by $E[\xi^{(3)}] = n_i$ [30]. Then the required transmit energy per bit is given by

$$e_{i\_\text{inter}}^{(3)} = \frac{\gamma_{\text{inter}} y_i^{\beta}}{n_i}. \tag{25}$$

4.2. Optimal coalition planning

4.2.1. The Type I network model

In the Type I model, the data (H) need to be forwarded to the sink through K rings of coalitions. Due to the symmetry of the rings and the uniform distribution of sensors, the H information bits are evenly distributed throughout all the coalitions in the ith ring. Since the number of coalitions in the ith ring is

$$N_i \approx \frac{\pi(d_i + d_{i-1})}{d_i - d_{i-1}}, \tag{26}$$

the average amount of information bits received by a CH in the ith ring is given by

$$H_i = \frac{H(d_i - d_{i-1})}{\pi(d_i + d_{i-1})}. \tag{27}$$

After the CH receives the data, it broadcasts within the coalition. Approximating the transmission distance as $x_i = (d_i - d_{i-1})/2$, we have that the required transmit energy per bit is

$$e_{i\_\text{intra}} = \gamma_{\text{intra}} \left(\frac{d_i - d_{i-1}}{2} \right)^{\alpha}. \tag{28}$$

So the total energy consumption for the intra-coalition communications is given by

$$\mathcal{E}_{i\_\text{intra}} = H_i \cdot e_{i\_\text{intra}}. \tag{29}$$

Next, we consider the energy consumption for inter-coalition communications. Denote the coordinates of the CH as $(0, (d_i + d_{i-1})/2)$ and the sink as (0,0). The sensor nodes within the coalition are uniformly deployed in the area $\{(x, y) : x \in (-l_i/2, l_i/2],\ y \in (d_{i-1}, d_i]\}$, where $l_i = d_i - d_{i-1}$. We approximate the position of the next-hop CH as $(0, (d_{i-1} + d_{i-2})/2)$ (which actually leads the lower bound of the distance). Note that for $i = 1$, the next hop reaches the sink. Then, the average path loss for the inter-coalition communications is given by

$$y_i^{\beta} = \begin{cases} \displaystyle\int_{-l_i/2}^{l_i/2} \int_{d_{i-1}}^{d_i} \frac{1}{l_i^2} \left[x^2 + y^2\right]^{\beta/2} dx\, dy, & \text{for } i = 1, \\ \displaystyle\int_{-l_i/2}^{l_i/2} \int_{d_{i-1}}^{d_i} \frac{1}{l_i^2} \left[x^2 + \left(y - \frac{d_{i-1} + d_{i-2}}{2} \right)^2 \right]^{\beta/2} dx\, dy, & \text{for } i = 2, \ldots, K. \end{cases} \tag{30}$$

Substituting y_i into (23), (24), and (25), we get the required transmit energy per bit of the three proposed schemes for inter-coalition communications. Then, the energy consumption for the inter-coalition communications is given by

$$\mathcal{E}_{i\_\text{inter}} = H_i \cdot e_{i\_\text{inter}}. \tag{31}$$

The total energy consumption of a coalition in the ith ring is given by

$$\mathcal{E}_i = \mathcal{E}_{i\_\text{intra}} + \mathcal{E}_{i\_\text{inter}} = \frac{H(d_i - d_{i-1})}{\pi(d_i + d_{i-1})} (e_{i\_\text{intra}} + e_{i\_\text{inter}}). \tag{32}$$

Since each sensor node has the same probability to transmit the data, and the average number of sensor nodes in a coalition in the ith ring is $n_i = \delta(d_i - d_{i-1})^2$, the average energy consumption of one sensor node in the ith ring is given by

$$\frac{\mathcal{E}_i}{n_i} = \frac{H}{\delta\pi(d_i^2 - d_{i-1}^2)} (e_{i\_\text{intra}} + e_{i\_\text{inter}}). \tag{33}$$

Then energy balancing boils down to the following optimization problem:

$$\mathbf{P1}: \min_{\{d_1, \ldots, d_K\}} \ \max_i \left\{ \frac{\mathcal{E}_i}{n_i} \right\} - \min_i \left\{ \frac{\mathcal{E}_i}{n_i} \right\} \quad \text{s.t.} \quad d_0 \le d_1 \le \cdots \le d_K. \tag{34}$$

By introducing auxiliary variables $t \ge \mathcal{E}_i/n_i$ and $s \le \mathcal{E}_i/n_i$, the optimization problem (34) can be transformed into the following equivalent form:

$$\begin{aligned} \mathbf{P2}: \min_{\{d_1, \ldots, d_K\}} \ & t - s \\ \text{s.t.} \ & \frac{\mathcal{E}_i}{n_i} \le t, \quad i = 1, 2, \ldots, K, \\ & \frac{\mathcal{E}_i}{n_i} \ge s, \quad i = 1, 2, \ldots, K, \\ & d_0 \le d_1 \le \cdots \le d_K. \end{aligned} \tag{35}$$

Clearly, this problem in general involves nonlinear optimization. In light of this, we turn to numerical methods to find the optimal solution. In particular, we use the nonlinear optimization solver KNITRO [34] which implements algorithms of both the interior (or barrier) type and the active-set type, and using trust regions to promote convergence [35]. We will elaborate further on this in Section 4.3.

For the sake of comparison, we also study the case that considers the energy balancing across CHs only. In this case, because the CHs always transmit the data, there is no energy consumption for intra-coalition communications. Then the energy consumption of a CH in the ith ring is given by

$$\mathcal{E}'_i = \frac{H(d_i - d_{i-1})}{\pi(d_i + d_{i-1})} (\gamma_{\text{inter}} y_i^{\beta}). \tag{36}$$

Accordingly, the energy balancing problem can be formulated as following:

$$\mathbf{P3}: \min_{\{d_1, \ldots, d_K\}} \ \max_i \{\mathcal{E}'_i\} - \min_i \{\mathcal{E}'_i\} \quad \text{s.t.} \quad d_0 \le d_1 \le \cdots \le d_K. \tag{37}$$

4.2.2. The Type II network model

In the Type II model, all the sensor nodes in the area generate information of H_0 bits. In each coalition, the CH receives the data from the coalition members and from outside rings. The CH carries out the aggregation for data from its own coalition and combine them with the relaying traffic. Let η denote the compression ratio. Then, the compressed data from its own coalition is given by

$$H_{i\_\text{own}} = \eta\delta(d_i - d_{i-1})^2 H_0, \tag{38}$$

and the relaying data received by a CH in the ith ring is given by

$$H_{i\_\text{relay}} = \frac{\eta\delta\pi(D^2 - d_i^2)(d_i - d_{i-1})H_0}{\pi(d_i + d_{i-1})}. \tag{39}$$

Thus the total information bits to be sent by a coalition in the ith ring is given by

$$H_i = \frac{\eta\delta\pi(D^2 - d_{i-1}^2)(d_i - d_{i-1})H_0}{\pi(d_i + d_{i-1})}. \tag{40}$$

Accordingly, the intra-coalition energy consumption is given by

$$\mathcal{E}_{i\_\text{intra}} = (n_i H_0 + H_i)e_{i\_\text{intra}}, \tag{41}$$

where e_{i_intra} is given by (28), and the inter-coalition energy consumption is given by

$$\mathcal{E}_{i\_\text{inter}} = H_i \cdot e_{i\_\text{inter}}, \tag{42}$$

where e_{i_inter} is given by (23), (24), and (25) for the three coalition-aided data transmission schemes, respectively. The total energy consumption of a coalition in the ith ring is given by

$$\mathcal{E}_i = n_i H_0 \cdot e_{i\_\text{intra}} + H_i(e_{i\_\text{intra}} + e_{i\_\text{inter}}), \tag{43}$$

and the average energy consumption of one sensor node in the ith ring is given by

$$\frac{\mathcal{E}_i}{n_i} = H_0 \cdot e_{i\_\text{intra}} + \frac{\eta\delta H_0(D^2 - d_{i-1}^2)}{d_i^2 - d_{i-1}^2}(e_{i\_\text{intra}} + e_{i\_\text{inter}}). \tag{44}$$

Then, the energy balancing problem can be formulated the same as $P1$.

We also present the problem which considers the energy balancing across CHs only for the sake of comparison. The energy consumption of a CH in the ith ring is given by

$$\mathcal{E}'_i = \frac{\eta\delta H_0\pi(D^2 - d_i^2)(d_i - d_{i-1})}{\pi(d_i + d_{i-1})}\gamma_{\text{inter}}\gamma_i^{\beta}, \tag{45}$$

and the energy balancing problem across CHs can be formulated the same as $P3$.

Table 2: Numerical parameters.

K	5	Number of rings
d_0	10	$X(0) = d0$
D	200	$X(K) = D$
H_0	2000	Information bits at each node
H	5 M	Total information bits for Type I network
α	2	Path loss factor for intracoalition communications
β	4	Path loss factor for intercoalition communications
γ	10^{-15}	Received energy threshold
η	0.5	Data compress ratio
δ	0.02	Sensor node density

4.3. Numerical examples

In this section, we illustrate by numerical examples the performance of the proposed schemes, and compare them with the one considering energy balancing across CHs only. We characterize the solutions to the nonlinear optimization problems in Section 4.2. To solve the nonlinear optimization problems, we use the solver KNITRO [34] with the AMPL [36] interface. KNITRO is a powerful solver for nonlinear optimization problems, by implementing novel and state-of-the-art algorithms of both the interior (or barrier) type and the active-set type, and using trust regions to promote convergence [35]. AMPL is a comprehensive and powerful algebraic modeling language for linear and nonlinear optimization problems, in discrete or continuous variables. We convert our problems into the AMPL format and get the numerical results from the KNITRO solver. The parameters of our problem are summarized in Table 2.

First, we examine the coalition size profile in the network. Using the analytical solution, we show in Figure 7 the coalition sizes of different rings of the three transmission schemes for the Type I network model, as well as the one considering the energy balancing across CHs only (numerical studies can be carried out for the Type II network model similarly). It can be seen that the coalition size profiles of these schemes are very different. In particular, for the scheme considering CHs only and the random selection scheme, the coalition size becomes larger for coalitions farther away from the sink, while for the schemes with multiuser diversity or cooperative diversity, the middle coalitions have larger coalition sizes. This is because that the communication distance is the dominant factor for the scheme considering CH only and the random selection scheme, whereas the number of sensors becomes an important factor affecting energy consumption for the schemes with multiuser diversity or cooperative diversity. In summary, the optimal coalition structure depends on the specific data transmission scheme and therefore should be designed carefully to achieve energy balancing across nodes.

Then, we examine the energy consumption among all these schemes. The analytical results for energy consumption are shown in Table 3. From Table 3 it can seen that the coalition-based schemes reduce the burden of the CHs a lot and hence help to prolong the life time significantly. Note that for each of the coalition aided transmission schemes, the

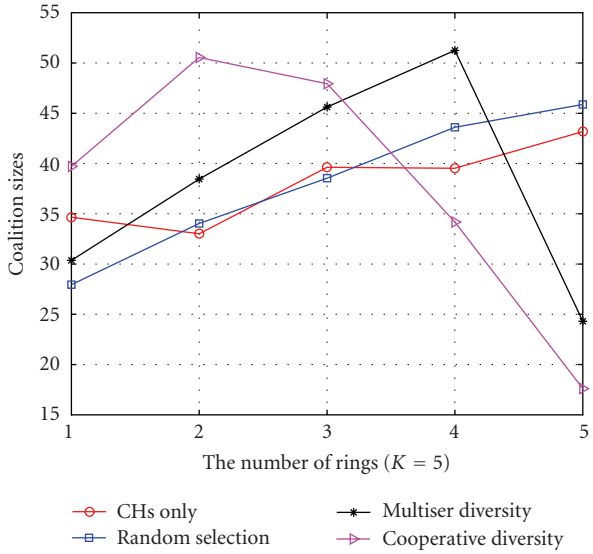


Figure 7: The coalition sizes.

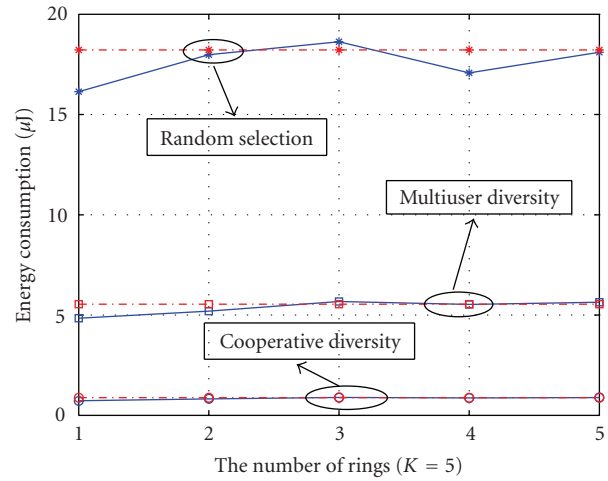


Figure 8: Energy balancing across coalitions.

Table 3: The energy consumption of all the schemes.

CHs only	Random selection	Multiuser	Cooperative
521.08 μJ	18.22 μJ	5.54 μJ	0.87 μJ

result in Table 3 denotes the energy consumption of one sensor node, while for the scheme considering CHs only it denotes the energy consumption of a CH.

The analysis above is based on certain simplified assumptions (e.g., square coalitions, lower-bounded next-hop distance, etc.). To corroborate our analytical studies, we conduct simulations in a more "realistic" setting, where the sensor nodes are randomly placed in the area $\mathcal{A}$. The distances between the sink and the CHs of different rings are based on the analytical results obtained. Each sensor node joins the closest CH according to its location. The average energy consumption of one sensor node in different rings are shown in Figure 8. It can be seen that energy balancing across the sensor nodes can be achieved for all the coalition-aided data transmission schemes, and that scheme 3 has the best energy saving performance among the three schemes.

5. CONCLUSIONS

We take a cross-layer optimization approach to study energy efficient data transport in coalition-based wireless sensor networks, where neighboring nodes are organized into groups to form coalitions and data aggregation and cooperative communications can be carried out within one coalition. The interplay among data aggregation, medium access control, cooperative communication, and coalition planning are exploited. In particular, we investigate two network models, that is, many-to-one sensor networks and multihop sensor networks. In a many-to-one sensor network, data from one coalition are transmitted to the sink directly. We propose three schemes for data transmission from a coalition to the sink. In scheme 1, one node in the coalition is selected randomly by the CH to transmit the data, so that each node within the coalition consumes energy in the same pace. In scheme 2, the sensor node with the best channel condition transmits the data, yielding multiuser diversity gain. In scheme 3, all the sensors within the coalition transmit as a virtual antenna array, so the cooperative diversity gain could be achieved.

Building on the coalition-aided data transmission schemes for one hop, we study energy balancing across sensor nodes in multihop networks, where data are relayed by intermediate coalitions to reach the sink. Optimal coalition planning is carried out, in the sense that unequal coalition sizes are applied to minimize the difference of energy consumption among sensor nodes. In particular, we investigate multihop networks with two different traffic patterns. In a Type I network, only part of the sensor nodes have data to transmit and others serve as relays; and in a Type II network, all sensor nodes have data to transmit. Numerical analysis shows that energy efficiency can be improved significantly by the coalition-aided transmission schemes, and that energy balancing across the sensor nodes can be achieved with the proposed coalition structures.

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, 2002.

[2] S. Bandyopadhyay and E. J. Coyle, "An energy efficient hierarchical clustering algorithm for wireless sensor networks," in *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '03)*, vol. 3, pp. 1713–1723, San Francisco, Calif, USA, March-April 2003.

[3] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences (HICSS '00)*, vol. 2, pp. 3005–3014, Maui, Hawaii, USA, January 2000.

[4] R. Krishnan and D. Starobinski, "Efficient clustering algorithms for self-organizing wireless sensor networks," *Ad Hoc Networks*, vol. 4, no. 1, pp. 36–59, 2006.

[5] S. Lindsey and C. S. Raghavendra, "PEGASIS: power efficient gathering in sensor information systems," in *Proceedings of IEEE Aerospace Conference*, vol. 3, pp. 1125–1130, Big Sky, Mont, USA, March 2002.

[6] A. Manjeshwar and D. P. Agrawal, "TEEN: a routing protocol for enhanced efficiency in wireless sensor networks," in *Proceedings of the 15th International Parallel and Distributed Processing Symposium (IPDPS '01)*, pp. 2009–2015, San Francisco, Calif, USA, April 2001.

[7] O. Younis and S. Fahmy, "HEED: a hybrid, energy efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Transactions on Mobile Computing*, vol. 3, no. 4, pp. 366–379, 2004.

[8] B. Chen, K. Jamieson, H. Balakrishnan, and R. Morris, "Span: an energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks," *Wireless Networks*, vol. 8, no. 5, pp. 481–494, 2002.

[9] C. Schurgers and M. B. Srivastava, "Energy efficient routing in wireless sensor networks," in *Proceedings of IEEE Military Communications Conference on Communications for Network-Centric Operations: Creating the Information Force (MILCOM '01)*, vol. 1, pp. 357–361, McLean, Va, USA, October 2001.

[10] Y. Xu, J. Heidemann, and D. Estrin, "Geography-informed energy conservation for ad hoc routing," in *Proceedings of the 7th Annual International Conference on Mobile Computing and Networking, (MOBICOM '01)*, pp. 70–84, Rome, Italy, July 2001.

[11] W. Ye, J. Heidemann, and D. Estrin, "An energy-efficient MAC protocol for wireless sensor networks," in *Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '01)*, vol. 3, pp. 1567–1576, New York, NY, USA, June 2002.

[12] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 6, pp. 1089–1098, 2004.

[13] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.

[14] A. Nosratinia, T. E. Hunter, and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Communications Magazine*, vol. 42, no. 10, pp. 74–80, 2004.

[15] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451–1458, 1998.

[16] J. Ai, D. Turgut, and L. Boloni, "A cluster-based energy balancing scheme in heterogeneous wireless sensor networks," in *Proceedings of the 4th International Conference on Networking (ICN '05)*, pp. 467–474, Reunion, France, April 2005.

[17] C. Efthymiou, S. Nikoletseas, and J. Rolim, "Energy balanced data propagation in wireless sensor networks," in *Proceedings of the 18th International Parallel and Distributed Processing Symposium (IPDPS '04)*, p. 225a, Santa Fe, NM, USA, April 2004.

[18] Q. Gao, J. Zhang, B. Larish, and S. Shen, "Coalition-aided data transmissions in wireless sensor networks," in *Proceedings of IEEE International Conference on Communications (ICC '06)*, pp. 3426–3431, Istanbul, Turkey, June 2006.

[19] M. Haenggi, "Energy-balancing strategies for wireless sensor networks," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '03)*, vol. 4, pp. 828–831, Bangkok, Thailand, May 2003.

[20] I. Howitt and J. Wang, "Energy balanced chain in distributed sensor networks," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '04)*, vol. 3, pp. 1721–1726, Atlanta, Ga, USA, March 2004.

[21] C. Li, M. Ye, G. Chen, and J. Wu, "An energy-efficient unequal clustering mechanism for wireless sensor networks," in *Proceedings of the 2nd IEEE International Conference on Mobile Ad-Hoc and Sensor Systems (MASS '05)*, pp. 8 pages, Washington, DC, USA, November 2005.

[22] T. Shu, M. Krunz, and S. Vrudhula, "Power balanced coverage-time optimization for clustered wireless sensor networks," in *Proceedings of the 6th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC '05)*, pp. 111–120, Urbana-Champaign, Ill, USA, May 2005.

[23] S. Soro and W. Heinzelman, "Prolonging the lifetime of wireless sensor networks via unequal clustering," in *Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS '05)*, pp. 8 pages, Denver, Colo, USA, April 2005.

[24] Y. Yu and V. K. Prasanna, "Energy-balanced multi-hop packet transmission in wireless sensor networks," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '03)*, vol. 1, pp. 480–486, San Francisco, Calif, USA, December 2003.

[25] M. Dong, L. Tong, and B. M. Sadler, "Effect of MAC design on source estimation in dense sensor networks," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 3, pp. 853–856, Montreal, Quebec, Canada, May 2004.

[26] D. Marco, E. J. Duarte-Melo, M. Liu, and D. L. Neuhoff, "On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data," in *Proceedings of the 2nd International Symposium on Information Processing in Sensor Networks (IPSN '03)*, pp. 1–16, Palo Alto, Calif, USA, April 2003.

[27] T. K. Y. Lo, "Maximum ratio transmission," *IEEE Transactions on Communications*, vol. 47, no. 10, pp. 1458–1461, 1999.

[28] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*, Cambridge University Press, Cambridge, UK, 2003.

[29] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, San Diego, Calif, USA, 2002.

[30] T. S. Rappaport, *Wireless Communications*, Prentice-Hall, Upper Saddle River, NJ, USA, 2nd edition, 2002.

[31] S. Pattem, B. Krishnamachari, and R. Govindan, "The impact of spatial correlation on routing with compression in wireless sensor networks," in *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks (IPSN '04)*, pp. 28–35, Berkerley, Calif, USA, April 2004.

[32] S. Tilak, N. B. Abu-Ghazaleh, and W. Heinzelman, "A taxonomy of wireless micro-sensor network models," *ACM Mobile Computing and Communications Review*, vol. 6, no. 2, pp. 28–36, 2002.

[33] K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks," *Ad Hoc Networks*, vol. 3, no. 3, pp. 325–349, 2005.

[34] KNITRO, http://www.ziena.com/knitro.htm.

[35] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, NY, USA, 1999.

[36] AMPL, http://www.ampl.com/.

Hindawi Publishing Corporation
Advances in Multimedia
Volume 2007, Article ID 74591, 9 pages
doi:10.1155/2007/74591

Research Article

Location-Aware Cross-Layer Design Using Overlay Watermarks

Xianbin Wang,[1] Paul Ho,[2] and Yiyan Wu[1]

[1] *Communications Research Centre Canada, 3701 Carling Avenue, P.O. Box 11490, Station H, Ottawa, ON, Canada K2H 8S2*
[2] *School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6*

Received 28 December 2006; Accepted 7 March 2007

Recommended by Jianwei Huang

A new orthogonal frequency division multiplexing (OFDM) system embedded with overlay watermarks for location-aware cross-layer design is proposed in this paper. One major advantage of the proposed system is the multiple functionalities the overlay watermark provides, which includes a cross-layer signaling interface, a transceiver identification for position-aware routing, as well as its basic role as a training sequence for channel estimation. Wireless terminals are typically battery powered and have limited wireless communication bandwidth. Therefore, efficient collaborative signal processing algorithms that consume less energy for computation and less bandwidth for communication are needed. Transceiver aware of its location can also improve the routing efficiency by selective flooding or selective forwarding data only in the desired direction, since in most cases the location of a wireless host is unknown. In the proposed OFDM system, location information of a mobile for efficient routing can be easily derived when a unique watermark is associated with each individual transceiver. In addition, cross-layer signaling and other interlayer interactive information can be exchanged with a new data pipe created by modulating the overlay watermarks. We also study the channel estimation and watermark removal techniques at the physical layer for the proposed overlay OFDM. Our channel estimator iteratively estimates the channel impulse response and the combined signal vector from the overlay OFDM signal. Cross-layer design that leads to low-power consumption and more efficient routing is investigated.

1. INTRODUCTION

The growth of wireless packet data applications (e.g., wireless Web access, interactive mobile multimedia applications, and interactive gaming) drives the rapid evolution of next-generation wireless networks. One of the key challenges for next-generation broadband wireless networks is to devise end-to-end protocol solutions across wired and wireless links through cross-layer design. Traditional network protocol design is based on a layered approach in which each layer in the protocol stack is designed and operated independently with interfaces between layers that are static and independent of the individual network constraints and applications. With this approach, information regarding the changing wireless channel condition is often hidden from the higher network layers. In the meantime, the ability for mobile stations to determine their position through automatic means is recognized as an essential feature, since location information is particularly important for network optimization, including energy conservation and location-aware routing. Mobile hosts are typically battery powered and have limited wireless communication bandwidth. Therefore, the transmission power should just be at the right level and this can be achieved if the mobile is aware of its location. As a result, the unique characteristics of wireless networks such as user mobility, fast channel variation, limited link capacity, and limited battery and computational resources in mobile devices, along with the diverse quality of service (QoS) requirements for wireless applications, pose significant challenges in codesigning different layers of network protocols for high-speed mobile communications. Various positioning approaches have been proposed, of which some were even constructed and deployed on a large scale, for example, Global Positioning System (GPS) [1]. GPS is effective and accurate outdoors, but it works very poorly, if at all, indoors and in urban canyon environments [2]. As a result, reliable position location solution is needed for wireless communication devices, particularly for indoor applications. Cellular telephone networks can be used to provide location services, where the mobile receivers are located by measuring the strength of signals traveling to and from a set of fixed cellular base stations. However, owing to the narrow bandwidth and variation of signal strength, position systems based on cellular networks can only achieve very limited

accuracy with locationing error often large than few hundred meters [3, 4]. Other positioning alternatives based on ultra wide band (UWB) devices and wireless local area networks (WLAN) can only provide very limited coverage [5].

In this paper, a new OFDM system with overlay watermark for location-aware cross-layer design is proposed and investigated. Orthogonal Frequency Division Multiplexing (OFDM) has been widely accepted as the major transmission technology for next generation wireless communication systems due to its high spectral efficiency, robustness to multipath distortion and simple frequency domain equalization [6]. Accurate channel estimation is indispensable for an OFDM system to achieve coherent demodulation and consequently higher data rate. For OFDM systems operating in a mobile wireless environment, estimation of the time-frequency varying channel requires closely-spaced pilot subcarriers in both the time and frequency domains, resulting in a significant loss in bandwidth efficiency. As an alternative to improve the bandwidth efficiency, pilot symbols can be superimposed upon the data symbols to enable channel estimation without sacrificing the data rate. This idea was first proposed for analog communication in [7] and was later extended to digital single carrier systems in [8]. Recently, the idea of superimposed training has received renewed attention in OFDM systems [9–11]. However, superimposed pilots in the frequency domain will deteriorate the peak to average power ratio (PAPR) problem of the OFDM signals. The high PAPR associated with a frequency-domain overlay pilot signal and the need of cross-layer interface inspire us to consider a time-domain overlay sequence with constant amplitude as a cross-layer signaling and transmitter identification, which can be used to determine the location of the transmitter.

In the proposed overlay OFDM system, time-domain orthogonal Kasami sequences [12–14] are used as overlay watermarks for cross-layer signaling and channel estimation training sequence. We propose to modulate the watermark so that a new, low-rate, parallel data pipe is created for the purpose of transporting cross-layer signaling and control information without interruptions to the physical link. Note that there is no redundancy introduced since the overlay watermark will also be used as training sequence for channel estimation. Preambles or training sequences are always required either in frequency or time domain in traditional communications system for channel estimation purpose. For instance, normally more than 10 percent of total bandwidth is used as in-band pilots for channel estimation purpose in conventional OFDM system. In this paper, the in-band pilots of OFDM system is converted as an overlay watermark for channel estimation purposes. It will not introduce extra redundancy since channel estimation preambles are always needed. As an added advantage, the overlay watermark provides an independent data pipe for cross-layer signalling transmission. The use of Kasami watermarks provide the following advantages (i) The availability of a large set of orthogonal Kasami sequences ensures that a unique watermark can be assigned to each individual OFDM transceiver, which may be used for transceiver identification and position location. As a result, position-aware routing algorithms can be used to improve the network efficiency. (ii) A parallel data link can be created by modulating the watermarks for data link controlling purposes. Information related to the adaptive modulation and coding schemes employed can be transmitted over this extra data link. (iii) Simple channel estimation and watermark removal algorithm can be readily employed. The organization of the paper is as follows. The transceiver structure of the proposed OFDM is illustrated in Section 2. To eliminate the impact of the watermark on OFDM signal detection, we also propose an iterative channel estimation and data detection algorithm. Initial channel estimation is obtained from the overlay watermark with the OFDM signal acting as interference. Decision for the transmitted OFDM data is then made based on the tentative channel estimate. The accuracy of the channel estimates is then progressively improved by reestimating the channel by using a new composite channel estimation sequence consisting of the watermark and a tentative OFDM signal derived from the data detection results. Location-aware cross-layer design is investigated in Section 3 with the proposed Kasami watermarks. The design and detection of cross-layer signaling through the modulation of the overlay watermarks are analyzed. A position location technique based on the overlay watermark is investigated. Numerical results are presented in the next section and the paper is summarized in Section 5.

Notations

$()^H$ and $()^T$ represents the conjugate transpose and transpose; N and L indicate the number of OFDM subcarriers and length of the channel impulse response, respectively; Tr{} denotes the trace of a matrix; $\mathbf{X}$, a vector of size N, representing the OFDM data in the frequency domain; $\mathbf{x}$ is the corresponding time domain OFDM signal vector; $\mathbf{y}$ is the received time-domain signal vector; $\mathbf{h}$ and $\mathbf{H}$ are channel vectors in the time and frequency domains with size of L and N, respectively; $\mathbf{n}$ is an additive white Gaussian noise (AWGN) vector. Unless otherwise stated, all vectors in the paper are column vectors.

2. TRANSCEIVER STRUCTURE FOR OVERLAY OFDM SYSTEMS

2.1. Transmitter for overlay OFDM system

The transceiver block diagram of the proposed overlay OFDM system is depicted in Figure 1. Each time-domain OFDM symbol $\mathbf{x}$ in the transmitter side of proposed overlay system is represented by an N-point complex sequence through an inverse discrete Fourier transformation (IDFT) of the subchannels data $\mathbf{X}$ as follows:

$$x(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X(k) e^{j2\pi(nk/N)}, \quad n = 0, 1, 2, \ldots, N-1. \tag{1}$$

The signal in (1) consists of N complex sinusoids modulated by the complex data symbols $X(1), X(2), \ldots, X(N-1)$.

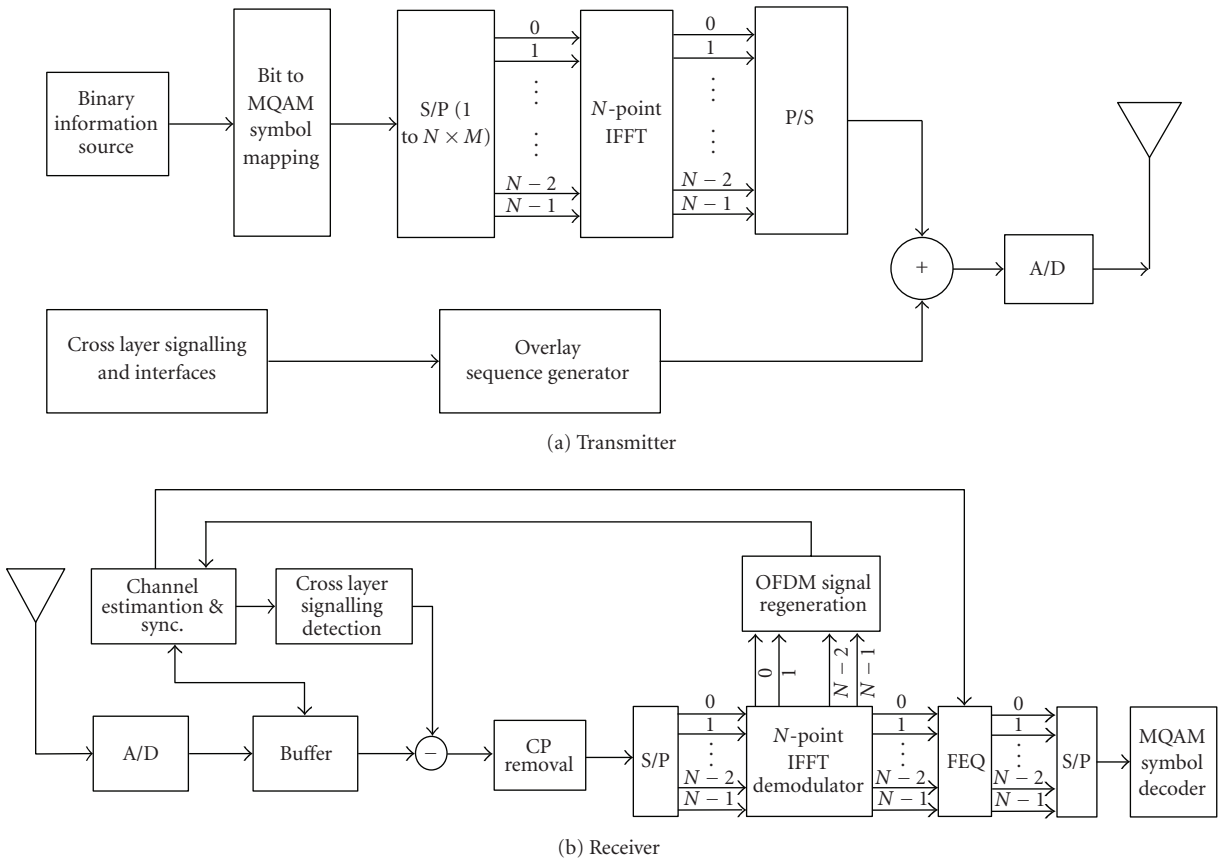


Figure 1: The transceiver block diagram for the OFDM system with overlay watermark.

The watermark signal is superimposed on the OFDM symbol before the cyclic prefix is added. Since the addition and elimination of the cyclic prefix has no impact on the statistics of the signal as well as the subsequent analysis, these steps are omitted from the discussion throughout the paper. The superimposed watermark vector $\mathbf{p}$ is added to the OFDM signal according to

$$\mathbf{s} = \mathbf{x} + \mathbf{p}, \tag{2}$$

where $\mathbf{p} = [p(0), p(1), \ldots, p(N-1)]^T = \alpha \mathbf{p}_{\text{kasami}}$, α is a gain control parameter that determines the power of the superimposed watermark, and $\mathbf{p}_{\text{kasami}}$ represents a complex vector whose real and imaginary parts are two orthogonal Kasami sequences. The duration of the Kasami sequences and hence the watermark is identical to one OFDM symbol.

2.2. Receiver for overlay OFDM system

A slow varying multipath channel model is adopted in this paper. The received signal can be expressed as

$$y(n) = \sum_{l=0}^{L-1} h_l s(n-l) + n(n). \tag{3}$$

After the removal of cyclic prefix, the received signal vector $\mathbf{y} = [y(0), y(1), \ldots, y(N-1)]^T$ can be written as

$$\mathbf{y} = \mathbf{X}_M \mathbf{h} + \mathbf{P}\mathbf{h} + \mathbf{n}, \tag{4}$$

where $\mathbf{h} = [h(0), h(1), \ldots, h(L)]^T$ is the channel vector, $\mathbf{n} = [n(0), n(1), \ldots, n(N-1)]^T$ is an AWGN vector with variance σ_n^2,

$$\mathbf{X}_M = \begin{bmatrix} x(0) & x(N-1) & \cdots & x(N-L+1) \\ x(1) & x(0) & \cdots & x(N-L+2) \\ \vdots & \vdots & & \vdots \\ x(N-1) & x(N-2) & \cdots & x(N-L) \end{bmatrix} \tag{5}$$

is the data matrix derived from $\mathbf{x}$, and

$$\mathbf{P} = \begin{bmatrix} p(0) & p(N-1) & \cdots & p(N-L+1) \\ p(1) & p(0) & \cdots & p(N-L+2) \\ \vdots & \vdots & & \vdots \\ p(N-1) & p(N-2) & \cdots & p(N-L) \end{bmatrix} \tag{6}$$

is the watermark matrix obtained from $\mathbf{p}$. Here we assume the watermark vector $\mathbf{p}$ (and hence $\mathbf{P}$) is known to the receiver. Note that the polarity of the imaginary part of the watermark has to be determined when the watermark is modulated for data transmission; see Section 3. Given the transmitted signal vector $\mathbf{s}$ and the channel $\mathbf{h}$, the conditional

likelihood function of the received signal can be expressed as

$$\Lambda(\mathbf{y} \mid \mathbf{x}, \mathbf{h}) = \frac{1}{\left(\pi\sigma_{\mathbf{n}}^2\right)^N} \exp\left\{ -\frac{1}{\sigma_{\mathbf{n}}^2}[\mathbf{y}-\mathbf{X}_M\mathbf{h}-\mathbf{P}\mathbf{h}]^{\mathbf{H}}[\mathbf{y}-\mathbf{X}_M\mathbf{h}-\mathbf{P}\mathbf{h}]\right\}. \tag{7}$$

The goal of the receiver is to find the data $\mathbf{x}$ and the channel $\mathbf{h}$ that maximizes this conditional likelihood function. With a brute force implementation, the complexity associated with this joint optimization is huge. Here we propose a much simpler iterative algorithm, as shown in Figure 1(b). Below is a description of this iterative procedure.

First, consider (4). This equation can be rewritten as

$$\mathbf{y} = \mathbf{A}\mathbf{h} + \mathbf{n}, \tag{8}$$

where the $N \times L$ matrix $\mathbf{A}$ is derived from the composite signal $\mathbf{s}$ in (2) as follows:

$$\mathbf{A} = \begin{bmatrix} s(0) & s(N-1) & \cdots & s(N-L+1) \\ s(1) & s(0) & \cdots & s(N-L+2) \\ \vdots & \vdots & & \vdots \\ s(N-1) & s(N-2) & \cdots & s(N-L) \end{bmatrix}. \tag{9}$$

When the OFDM signal $\mathbf{x}$ is known to the receiver, then the above matrix is also known and hence can be treated as a training sequence for channel estimation purpose. In this case, the conditional likelihood function of the received signal becomes

$$\Lambda(\mathbf{y} \mid \mathbf{h}) = \frac{1}{\left(\pi\sigma_{\mathbf{n}}^2\right)^N} \exp\left\{-\frac{1}{\sigma_{\mathbf{n}}^2}[\mathbf{y}-\mathbf{A}\mathbf{h}]^{\mathbf{H}}[\mathbf{y}-\mathbf{A}\mathbf{h}]\right\}. \tag{10}$$

The maximum likelihood (ML) channel estimate, $\widetilde{\mathbf{h}}$, is the value of $\mathbf{h}$ that maximizes the argument of the above exponential function, that is,

$$\widetilde{\mathbf{h}} = \min_{\mathbf{h}} \{[\mathbf{y}-\mathbf{A}\mathbf{h}]^{\mathbf{H}}[\mathbf{y}-\mathbf{A}\mathbf{h}]\}. \tag{11}$$

Since the right-hand side of the above equation, denoted by $\Lambda_L(\mathbf{y} \mid \mathbf{h}) = [\mathbf{y}-\mathbf{A}\mathbf{h}]^{\mathbf{H}}[\mathbf{y}-\mathbf{A}\mathbf{h}]$, is a convex function over $\mathbf{h}$, the ML channel estimate satisfies

$$\left.\frac{\partial \Lambda_L(\mathbf{y} \mid \mathbf{h})}{\partial \mathbf{h}}\right|_{\mathbf{h}=\widetilde{\mathbf{h}}} = 0. \tag{12}$$

This implies that the ML channel estimate $\widetilde{\mathbf{h}}$ is, in principle, given by

$$\widetilde{\mathbf{h}} = \left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{y}. \tag{13}$$

Unfortunately, the matrix $\mathbf{A}$ in (9) is not known initially to the receiver, since each element of the matrix is the superimposition of the unknown OFDM signal sequence and the watermark. To circumvent this problem, the receiver resorts to obtaining an approximation of $\mathbf{A}$ based on the tentative data estimates derived from the rather crude initial channel estimates provided by the superimposed watermark. The iterative receiver then progressively provides more reliable information on matrix $\mathbf{A}$. As a result, the accuracy of the channel and data estimates will also be improved accordingly. We list below the procedure of this iterative channel estimator.

Initial channel estimates will be derived solely from the embedded watermark signal. That is, the OFDM signal will be treated as noise. This is because OFDM signal at any time instant is the summation of N independent subcarriers. When the number of the subcarriers is large enough, the OFDM signal can be approximated as Gaussian distributed random variable. Due to the Gaussian nature of the OFDM signal, the combined effect of the channel AWGN and the OFDM signal, $\mathbf{w} = \mathbf{X}_M\mathbf{h} + \mathbf{n}$, will still be Gaussian. The received signal is now expressed as

$$\mathbf{y} = \mathbf{P}\mathbf{h} + \mathbf{w}. \tag{14}$$

It is straightforward to verify that the variance of the effective noise $\mathbf{w}$ is

$$\sigma_w^2 = \sigma_n^2 + \sigma_x^2, \tag{15}$$

where σ_x^2 is the variance of the OFDM signal. Similar to (13), the initial channel estimate is given by

$$\widetilde{\mathbf{h}} = \left(\mathbf{P}^H\mathbf{P}\right)^{-1}\mathbf{P}^H\mathbf{y}. \tag{16}$$

One of the key ideas of the proposed iterative channel estimator is that, instead of using only the embedded watermark as the training sequence, the tentative estimated OFDM sequence in the time domain will also be used for that purpose. With this approach, the performance of the estimator is expected to be significantly improved, since now, the power of the training sequence is increased. To improve the channel estimate, the iterative receiver subtracts $\mathbf{P}\widetilde{\mathbf{h}}$ from the received vector $\mathbf{y}$ to obtain the new observation $\mathbf{y}' = \mathbf{y} - \mathbf{P}\widetilde{\mathbf{h}}$. After converting $\mathbf{y}'$ to $\mathbf{Y}' = \mathrm{DFT}(\mathbf{y}')$ via DFT, individual subchannels are gain/phase compensated by dividing the components of $\mathbf{Y}'$ by the corresponding components in frequency domain channel estimates $\widetilde{\mathbf{H}}$. Decisions on the data in individual channels are then made. These data estimates $\widetilde{\mathbf{X}}$ are then used to generate the estimated OFDM signal in the time domain $\widetilde{\mathbf{x}}$ using IDFT. At this point, the channel estimator employs $\widetilde{\mathbf{s}} = \widetilde{\mathbf{x}} + \mathbf{p}$ as the effective training sequence to update the channel estimates. A similar matrix $\widetilde{\mathbf{A}}$ will be constructed using $\widetilde{\mathbf{s}}$ to get the improved channel estimation according to $\widetilde{\mathbf{h}} = (\widetilde{\mathbf{A}}^H\widetilde{\mathbf{A}})^{-1}\widetilde{\mathbf{A}}^H\mathbf{y}$. This process will be iterated for the improved performance of the receiver.

3. CROSS-LAYER DESIGN WITH THE OVERLAY WATERMARKS

In this section, cross-layer signaling and interface design using the embedded watermarks are investigated. In addition to the basic function of the overlay watermark as a training sequence for channel estimation, channel quality information, adaptive rate control information, and network timing

information can be transmitted by modulating the watermarks. This "free" physical signaling pipe, which is in parallel to the OFDM signal, provides interfaces between different network layers directly. Interruption to the physical layers can be reduced since the cross-layer interactive information can be transmitted with the new link. Note here that only the bottom three layers of the OSI model under investigation here are depicted. The physical layer defines all the electrical and physical specifications for the communications devices, and is responsible for OFDM data transmission. The data link layer responds to service requests from the network layer and issues service requests to the physical layer. The network layer performs network routing, flow control, segmentation/de-segmentation, and error control functions.

3.1. Cross-layer signaling detection

We propose in this subsection a technique for transmitting medium access layer (MAC), layer controlling information as well as other protocol information via the superimposed watermark. Specifically, we propose to modulate the imaginary part of watermark with the incoming control data. Assuming that antipodal signalling is employed in this low data-rate digital pipe, then the received time-domain signal in (4) becomes

$$\mathbf{y} = \mathbf{X}_M\mathbf{h} + \mathbf{P}_r\mathbf{h} + jD\mathbf{P}_i\mathbf{h} + \mathbf{n}, \tag{17}$$

where D is the data bit (-1 or $+1$) containing cross-layer signaling, and $\mathbf{P}_r$ and $\mathbf{P}_i$ represents the real and imaginary parts of the matrix $\mathbf{P}$ in (6). Assuming perfect timing and frequency synchronization are achieved, a simple demodulator for the control data bit D is

$$\widetilde{D} = \text{sign}\left(\text{Re}\left\langle \mathbf{y} \cdot \mathbf{P}_i\widetilde{\mathbf{h}}\right\rangle\right), \tag{18}$$

where $\mathbf{P}_i\widetilde{\mathbf{h}}$ is the locally generated watermark, sign$(\cdot)$ is the sign operator, Re$(\cdot)$ is the real operator, and $\langle\cdot\rangle$ denotes inner product. The average signal-to-interference and noise ratio (SINR) in the above decision variable can be shown equal to

$$\text{SINR} = 10\log_{10}\left(\frac{N\alpha^2}{\sigma_x^2 + \sigma_n^2 + \alpha^2\Delta\widetilde{h}_{\text{MSE}}}\right), \tag{19}$$

where $\Delta\widetilde{h}_{\text{MSE}} = \sum_{l=0}^{L-1}|\widetilde{h}_l - h_l|^2$ is the channel estimation error. Here we assume $\sum_{l=0}^{L-1}|h_l|^2 = 1$. The SINR in (19) provides a rough idea on the performance of the new data pipe by modulating the watermarks. Specifically, for a given channel response $\mathbf{h}$, the bit error rate (BER) is related to SINR according to

$$P_b = \frac{1}{2}\,\text{erfc}\left(\sqrt{\text{SINR}}\right), \tag{20}$$

where $\text{erfc}(x) = (2/\sqrt{\pi})\int_x^\infty e^{-t^2}dt$.

3.2. Position location for mobile receivers

Due to the mobility of the wireless transceivers, the ability for mobile station to determine their position through automatic means is recognized as an essential feature, since location information is particularly important for network optimization, including energy conservation and location-aware routing [15]. Mobile hosts are typically battery powered and have limited wireless communication bandwidth. Therefore, the transmission power should just be at the right level and this can be achieved if the mobile is aware of its location. By assigning different orthogonal Kasami sequences to different transmitters, the source of a received signal can be easily identified. Location awareness can also improve the routing efficiency by selectively forwarding data in the desired direction [16, 17].

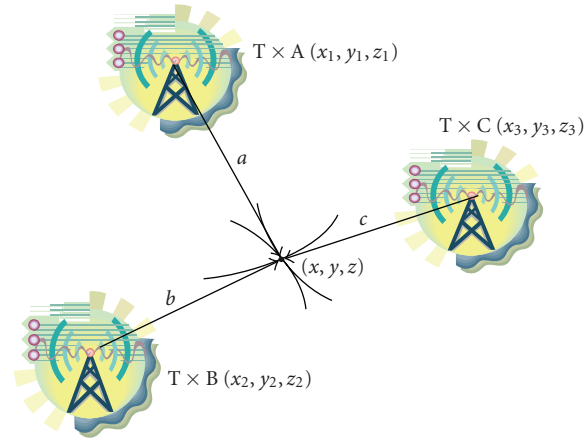


Figure 2: Position location using three transmitters

There are several different approaches to determine the location of receiving devices in a wireless network, ranging from direction-of-arrival detection to determination of received signal strength. The technique considered herein is based on triangulation. This method derives its name from the availability of at least three distance measurements between known points. When the total number of known transmitters is less than three, position location can be achieved by direction-based techniques, aided by the strength of the received signal. Since direction-based techniques require the availability of an antenna array, they involve more complicated signal processing and the accuracy of the position information is also lower.

If one can measure the precise time a signal is transmitted and the precise time the signal arrives at a receiver, the distance between the transmitter and receiver can then be determined. The extra signaling link obtained via modulating the embedded watermarks is an excellent candidate for the distribution of this network timing information. Consider the three base station (backbone node) transmitters and the positioning receiver shown in Figure 2. The coordinates of the three transmitters are (x_1, y_1, z_1), (x_2, y_2, z_2), and (x_3, y_3, z_3), respectively. For base station transmitters, these coordinates are known a priori to the positioning receiver. Denoting the propagation time from the ith transmitter to the receiver as t_i, then in the absence of any measurement error, the co-ordinate of the receiver, (x, y, z), is the solution to

the following equations: [18–20]

$$t_1 c = \sqrt{(x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2},$$
$$t_2 c = \sqrt{(x - x_2)^2 + (y - y_2)^2 + (z - z_2)^2}, \quad (21)$$
$$t_3 c = \sqrt{(x - x_3)^2 + (y - y_3)^2 + (z - z_3)^2},$$

where c is the speed of light.

The first step in solving the above equations is to identify the operating transmitters. To identify the existence of the lth transmitter, a cross-correlation between the received signal from the lth transmitter, $y_l(n)$, and the locally generated watermark, $p_{r,l}(n)$, has to be performed. Mathematically, this correlation is

$$\begin{aligned} R_{y_l p_{r,l}}(m) &= \sum_{n=0}^{N-1} y_l(n) p_{r,l}(n-m) \\ &= \sum_{n=0}^{N-1} \{[x_l(n) + p_l(n)] \otimes h_l + n_l(n)\} \cdot p_{r,l}(n-m) \\ &= \alpha R_{p_{r,l} p_{r,l}} \otimes h_l + \left\{ \sum_{n=0}^{N-1} [x_l(n) + j p_{i,l}(n)] p_{r,l}(n-m) \right\} \\ &\quad \otimes h_l + \sum_{n=0}^{N-1} n_l(n)\, p_{r,l}(n-m), \end{aligned} \quad (22)$$

where N is the length of the transmitter identification watermark. The first term on the last line of (22), that is, the autocorrelation function $R_{p_{r,l} p_{r,l}}$, exists only when the watermark signal $\alpha p_{r,l}(n)$ is found in the received signal. The existence of the lth transmitter can then be determined by the correlation peak in (22), because the watermark signal $\alpha p_{r,l}(n)$ is uniquely associated with the lth transmitter. Equation (22) also indicates that the correlation peak in the first term on the last line undergoes the same attenuation and channel distortion as the OFDM signal described by the second term. Due to the orthogonal property of the Kasami sequences, $R_{p_{r,l} p_{r,l}}$ can be approximated as a delta function. The second term in (22) is only a noise-like sequence resulting from the in-band data signal from the same transmitter. Therefore, the channel response h_l from the lth transmitter can be approximated by $R_{y_l p_{r,l}}$, that is,

$$R_{y_l p_{r,l}} = A h_l + \text{noise}, \quad (23)$$

where A is a constant determined by $R_{p_{r,l} p_{r,l}}$ and the gain coefficient α. The earliest correlation peak that exceeds a particular threshold is chosen to be the direct propagation path from the lth transmitter to the position location receiver. The threshold for each transmitting station is decided by the station's transmission power, the approximate distance between the station and the receiver (as determined by the propagation delay in the main path), as well as the maximum expected excess path loss due to building penetration [21, 22]. The arrival time of the earliest correlation peak can then be converted to a relative propagation time in terms of second. However, the strength of the first arrived signal sometimes is very weak and it is difficult to discriminate multipath echoes from interference. In such circumstances, the interference in (22) or (23) from the OFDM data signal can be cancelled to improve the precision of position location after the OFDM signal is demodulated. Another approach to reduce the interference is through time-domain averaging of the correlation functions from different OFDM symbols. In this case, the main path can always be used as a timing reference for averaging a number of adjacent transmitter identification results. Simple averaging of the transmitter identification results in the time domain would reduce the impact of the interference by $10 \log_{10} V$, where V is the number of averaging.

Regarding the implementation complexity, the proposed position location algorithm can be divided into two separate steps, that is, transmitter identification and position location. Computation complexity associated with the transmitter identification, which is the major part of the position location algorithm, is proportional to the total number of the transmitters used in this process. The total number of the multiplications for identification of each transmitter can be approximately estimated as $TN\Delta M$, where T is the total number of the transmitters in the network and ΔM is the correlation range in (22) for transmitter identification. The position of the mobile receiver can then be determined by (21). When the number of the available transmitters is more than needed, a nonlinear optimization process can be invoked to finalize the location. The complexity associated position location and optimization process is minimal compared to the transmitter identification process. Therefore, the overall implementation complexity of the proposed algorithm is approximately proportional to the total number of the transmitter used for position location.

3.3. Position aware routing algorithms

Mobile hosts in a wireless network are dynamically located and continuously changing their locations. The mobility in wireless networks makes it difficult to predetermine "optimal" routes between mobile hosts. It therefore becomes important to design efficient and reliable routing protocols to maintain, discover, and organize the routes based on the most recent locations of the mobile hosts. Assuming that each node can obtain its position through the proposed position location technique in Section 3.2 and update the location information using the new signaling link proposed in Section 3.1, then various efficient position-based routing algorithms can then be readily applied [23]. Position-based routing algorithms eliminate some of the limitations of topology-based routing by using additional location information. The routing decision by a node is primarily based on the position of a packet's destination and the position of the node's immediate one-hop neighbors.

Before a packet can be sent, it is essential to determine the position of the destination host. Typically, a location service is required for this task. Different techniques, for example, grid and quorum-based location service, are available.

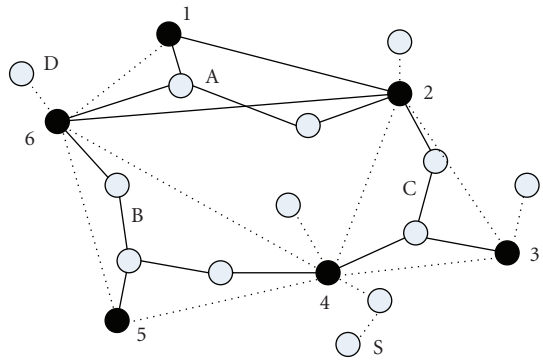


FIGURE 3: Example of position-based routing for wireless communication networks [23].

Example of a quorum-based location service is shown in Figure 3 [23]. After a mobile node determines its position using the technique in Section 3.2, it then sends position update messages to the nearest backbone node, which then chooses a quorum of backbone nodes to host the position information. Thus, node D sends its updates to node 6, which might then select quorum A with the nodes 1, 2, and 6 to host the information. When the node S wants to obtain the position information, it sends a query to its nearest backbone node, which in turn contacts the nodes of (a usually different) quorum.1. Since by definition, the intersection of two quorums is nonempty, the querying node is guaranteed to obtain at least one response with the desired position information. It is also important to time-stamp position updates, since some nodes in the queried quorum might have been in the quorum of previous updates and would then report outdated position information. If several responses are received, the one representing the most current position update is chosen. Once the position of the destination host is obtained, three forwarding strategies for position-based routing could be used: greedy forwarding, restricted directional flooding, and hierarchical approaches. The watermark signal in (17) can be used to indicate the selected forwarding route for the chosen forwarding strategy.

4. NUMERICAL RESULTS

Numerical simulations have been conducted to quantify the performance of the proposed overlay OFDM system and the corresponding cross-layer design. The demonstration system considered has the FFT size of 512 with cyclic prefix of length 1/8 of the symbol duration. Choice for the modulation format in the demonstration system is QPSK. Note that the transmission power of the overlay OFDM signal is normalized to that of a conventional OFDM signal. Unless otherwise stated, the parameter α is set to 0.5774 in all the figures. As for the channel model, we consider the channel

$$\mathbf{h} = [0.0855, 0, 0.8334, 0, 0, -0.3419, 0, 0, 0, 0.1282, 0, 0, 0, 0, 0, -0.4060]^T. \quad (24)$$

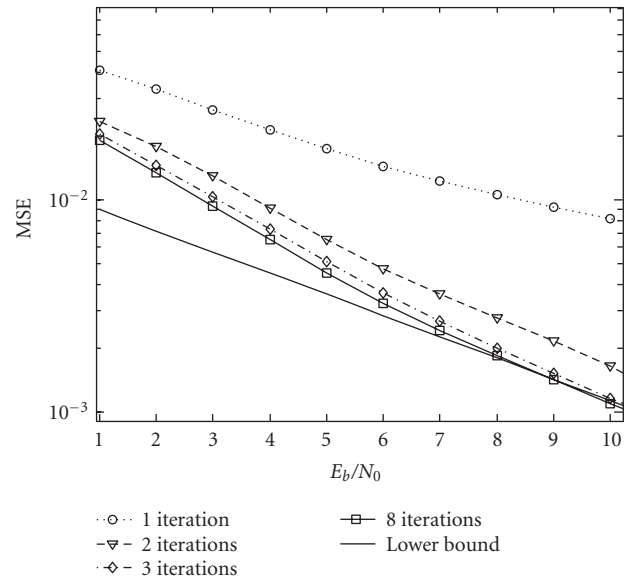


FIGURE 4: Mean square error of the iterative receiver for the overlay OFDM system with QPSK modulations.

We would like to point out that the emphasis of the investigation is to demonstrate the workability of overlay OFDM and its functionalities in future communication systems. Consequently, the "exact" channel model and parameter selections are only of secondary concern.

The MSE of the proposed iterative ML channel estimator in Section 2 was simulated. The results are plotted in Figure 4, with different number of iterations as a parameter. Here E_b/N_0 is defined as the signal-to-noise ratio (SNR) per bit. The results in Figure 4 indicate that for QPSK modulation, only three iterations are needed to approach the lower bound. Similar observations can be found in the symbol error rate (SER) simulation results for the QPSK in Figure 5. The results in Figure 5 show that good SER performance can be achieved with only three iterations for QPSK. More iterations may be needed for higher-modulation schemes like 16 QAM. However, we would like to point out that the number of iterations in practical systems could be significantly reduced when error correction coding is used, since the desired bit error rate after decoding could be easily achieved when the SER before decoding is less than 10^{-2}. In addition, the complexity of the iterative channel estimation can be further reduced when the channel estimate from the previous OFDM symbol is used as the initial input for the first round of the iteration.

The signal-to-interference and noise ratio (SINR) in the signaling link created from watermark modulation is also simulated and shown in Figure 6. We assume here the multipath channel is known to the receiver. It is found that at an OFDM data signal-to-noise ratio (SNR) of 10 dB and beyond, the SINR in the signaling link is insensitive to the SNR and attain fairly high values. At the SINR values plotted in Figure 6, very robust transmission of the cross-layer signaling can be achieved even without the assistance of error

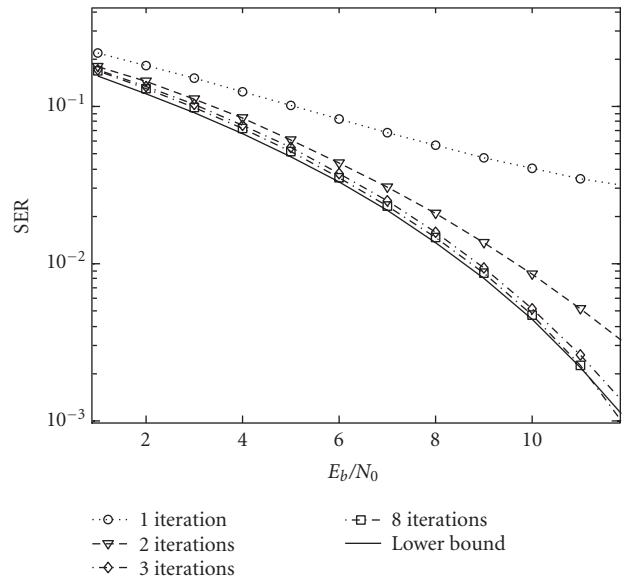


Figure 5: Symbol error rate of the overlay OFDM system with QPSK modulations.

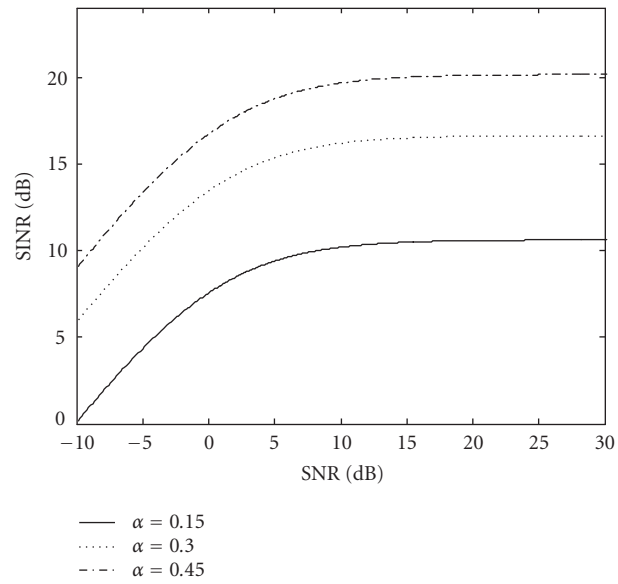


Figure 6: Signal-to-interference and noise ratio (SINR) for the cross-layer data link pipe based on watermark polarity modulation at difference signal-to-noise ratio (SNR).

correction doing. Note that a 0 dB of SNR for the OFDM data link is an extremely low SNR in typical wireless communication systems. The curves in Figure 6 also show that the SINRs are mainly decided by the amplitude of the overlay watermark and the length of the OFDM symbol. When a large α is used, higher-order amplitude modulation schemes could be used to increase the capacity of the signalling link. The capacity of the proposed watermark transmission technique can easily reach a few thousand bits per second for any broadband OFDM system. It is therefore more than sufficient to provide media access control information and adaptive rate control purposes. In order to design an overlay OFDM signaling link with the desired system performance, the target bit error rate P_b in (20) has to be selected first. For instance, 10^{-6} could be used for an uncoded system. With the desired bit error rate performance, the required SINR can be determined through table lookup approach based on (20). The watermark injection level is then determined with the OFDM symbol size given in (19).

Two base stations with known locations in a 2D Cartesian coordinate are used to test the proposed position location algorithm. The coordinates for the two stations, and the mobile receiver are (0, 0), (2000, 0), and (1000, 1000) meters. The channel model in the previous MSE and SER simulations is used as the propagation models for the signals from the two stations. The location results from the simulation were shown in Figure 7, where each star represents one round of location processing. The accuracy of the position location process can be evaluated by the distance between the location results and the true location of the receiver (origin of the coordinates). As independent random noise is added for each transmitter in position location simulation, ambiguity will be inevitably introduced to the positions obtained

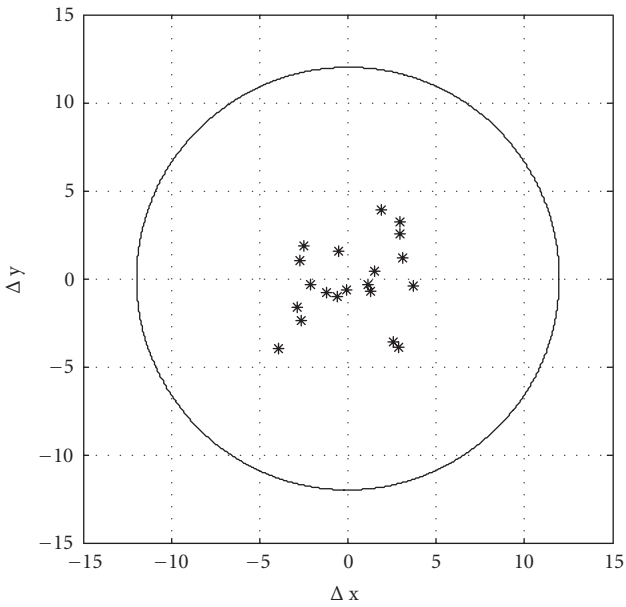


Figure 7: Numerical results for the proposed location position system based on watermark signal.

through the proposed position location algorithm in (22). Each star in Figure 7, which represents one simulation, will be driven away randomly from its true position at the original of the coordinate. The simulation results indicate that the accuracy of the proposed location system is within a few meters. Position-based routing algorithms and transmission power control can be effectively implemented at this precision level.

5. CONCLUSIONS

An OFDM system with overlay watermark for cross-layer design is proposed in this paper. The multiple roles played by the overlay watermark are investigated. It is demonstrated that an extra cross-layer signaling pipe can be created by modulating the overlay watermarks. New interfaces for cross-layer design can be established on top of this new supplementary data link. The major benefit of the proposed system is the improved network and bandwidth efficiency when compared to the conventional in-band pilot approach. Interruption to the physical link due to the cross-layer interaction can be significantly reduced with the introduction of the supplementary data link. When unique orthogonal Kasami sequences are assigned to individual transceivers as identifications, the location of the transmitter can be easily identified for position-based routing algorithms. An iterative channel estimation and data detection algorithm is investigated for the overlay system. Our analysis and simulations show that the impact from the overlay watermark to OFDM data detection is minimal.

REFERENCES

[1] G. Sun, J. Chen, W. Guo, and K. J. R. Liu, "Signal processing techniques in network-aided positioning: a survey of state-of-the-art positioning designs," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 12–23, 2005.

[2] X. Wang, Y. Wu, B. Caron, and J.-Y. Chouinard, "A new position location system using ATSC TxID signals," in *Proceedings of 61st. IEEE Vehicular Technology Conference*, pp. 2815–2819, Stockholm, Sweden, May 2005.

[3] J. Caffery Jr. and G. L. Stuber, "Subscriber location in CDMA cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 47, no. 2, pp. 406–416, 1998.

[4] J. Caffery Jr. and G. L. Stuber, "Overview of radiolocation in CDMA cellular systems," *IEEE Communications Magazine*, vol. 36, no. 4, pp. 38–45, 1998.

[5] P. Prasithsangaree, P. Krishnamurthy, and P. K. Chrysanthis, "On indoor position location with wireless LANs," in *Proceedings of the 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications Conference (PIMRC '02)*, vol. 2, pp. 720–724, Lisboa, Portugal, September 2002.

[6] Y. Wu and W. Zou, "COFDM: an overview," *IEEE Transactions on Broadcasting*, vol. 41, no. 1, pp. 1–8, 1995.

[7] C. Kastenholz and W. Birkemeier, "A simultaneous information transfer and channel-sounding modulation technique for wide-band channels," *IEEE Transactions on Communications*, vol. 13, no. 2, pp. 162–165, 1965.

[8] B. Farhang-Boroujeny, "Pilot-based channel identification: proposal for semi-blind identification of communication channels," *Electronics Letters*, vol. 31, no. 13, pp. 1044–1046, 1995.

[9] C. K. Ho, B. Farhang-Boroujeny, and F. Chin, "Added pilot semi-blind channel estimation scheme for OFDM in fading channels," in *Proceedings of Conference IEEE Global Telecommunicatins Conference (GLOBECOM '01)*, vol. 5, pp. 3075–3079, San Antonio, Tex, USA, November 2001.

[10] N. Chen and G. T. Zhou, "A superimposed periodic pilot scheme for semi-blind channel estimation of OFDM systems," in *Proceedings of the 10th IEEE Digital Signal Processing Workshop & the 2nd Signal Processing Education Workshop*, pp. 362–365, Pine Mountain, Ga, USA, October 2002.

[11] M. Ghogho, D. McLernon, E. Alameda-Hernandez, and A. Swami, "Channel estimation and symbol detection for block transmission using data-dependent superimposed training," *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 226–229, 2005.

[12] D. V. Sarwate and M. B. Pursley, "Cross correlation properties of pseudorandom and related sequences," *Proceedings of the IEEE*, vol. 68, no. 5, pp. 593–619, 1980.

[13] X. Wang, Y. Wu, and B. Caron, "Transmitter identification using embedded pseudo random sequences," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 244–252, 2004.

[14] X. Wang, Y. Wu, and J.-Y. Chouinard, "A new position location system using DTV transmitter identification watermark signals," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 42737, 11 pages, 2006.

[15] K. K. Chintalapudi, A. Dhariwal, R. Govindan, and G. Sukhatme, "Ad-hoc localization using ranging and sectoring," in *Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '04)*, vol. 4, pp. 2662–2672, Hong Kong, March 2004.

[16] Z. Ye and Y. Hua, "Stability of wireless relays in mobile ad hoc networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 3, pp. 505–508, Philadelphia, Pa, USA, March 2005.

[17] J. Liu, F. Zhao, and D. Petrovic, "Information-directed routing in ad hoc sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 851–861, 2005.

[18] E. D. Kaplan, *Understanding GPS: Principles and Applications*, Artech House Publishers, London, UK, 1996.

[19] J. A. Farrell and M. Barth, *The Global Positioning System & Inertial Navigation*, McGraw-Hill, London, UK, 1999.

[20] M. S. Grewal, L. R. Weill, and A. P. Andrews, *Global Positioning Systems, Inertial Navigation, and Integration*, John Wiley & Sons, Hoboken, NJ, USA, 2001.

[21] H. Hashemi, "The indoor radio propagation channel," *Proceedings of the IEEE*, vol. 81, no. 7, pp. 943–968, 1993.

[22] D. Molkdar, "Review on radio propagation into and within buildings," *IEE Proceedings H: Microwaves, Antennas and Propagation*, vol. 138, no. 1, pp. 61–73, 1991.

[23] M. Mauve, J. Widmer, and H. Hartenstein, "A survey on position-based routing in mobile ad hoc networks," *IEEE Network*, vol. 15, no. 6, pp. 30–39, 2001.