EURASIP Journal on Applied Signal Processing 2005:2, 207–219
© 2005 Hindawi Publishing Corporation

Cross-Layer QoS Support for Multimedia Delivery over Wireless Internet

Qian Zhang
Microsoft Research Asia 3F, Beijing Sigma Center, 49 Zhichun Road, Haidian District, Beijing 100080, China
Email: qianz@microsoft.com

Fan Yang
Microsoft Research Asia 3F, Beijing Sigma Center, 49 Zhichun Road, Haidian District, Beijing 100080, China
Email: t-fanyan@microsoft.com

Wenwu Zhu
Microsoft Research Asia 3F, Beijing Sigma Center, 49 Zhichun Road, Haidian District, Beijing 100080, China
Email: wwzhu@microsoft.com

Received 17 August 2003; Revised 13 June 2004

Delivering multimedia over wireless Internet is a very challenging task. Multimedia delivery inherently has strict quality-of-service (QoS) requirement on bandwidth, delay, and delay jitter. However, the current Internet can only support best-effort service, which imposes varying network conditions during multimedia delivery. The advent of wireless networks further exacerbates the variance of network conditions and brings greater challenges for multimedia delivery. To improve perceived media quality by end users over wireless Internet, QoS supports can be addressed in different layers, including application layer, transport layer, link layer, and so forth. This paper presents a framework, which provides QoS support, for multimedia delivery over wireless Internet, across different layers. To provide efficient QoS support for different types of media over the best-effort networks, we first propose a cross-layer architecture, which combines the application-level, transport-layer, as well as link-layer controls, and then review recent advances in each individual component. Specifically, dynamic estimation of varying channel and network, adaptive and energy-efficient application and link-level error control, efficient congestion control, header compression, adaptive automatic repeat request (ARQ) and priority-based scheduling, as well as QoS-adaptive proxy caching technologies are explicitly reviewed in this paper.

Keywords and phrases: cross-layer design, video streaming, wireless Internet, congestion control, resource allocation.

1. INTRODUCTION

With the explosive growth of the Internet and dramatic increase in wireless access, there is a tremendous demand on multimedia delivery over wireless Internet. The third generation (3G) wireless networks, foreseen to be the enabling technology for multimedia services with up to 384 kbps outdoor and 2 Mbps indoor bandwidth, makes it feasible for multimedia communication over the wireless link [1]. Moreover, the proliferation of 802.11 systems, that can provide up to 100 Mbps bandwidth, has extended the role of traditional Internet to support media streaming services in the air [2]. However, multimedia over wireless Internet poses many challenges as follows.

Different QoS requirements for different types of media

In general, different types of media have different characteristics. Specifically, real-time media such as video and audio is delay-sensitive but capable of tolerating a certain degree of errors. Non-real-time media such as Web data is less delay-sensitive but requires reliable transmission. In addition, due to scalable media encoding technologies, different parts of real-time media are of different importance.

High packet loss rate and bit error rate

In wireline networks, packet losses are usually caused by congestion in intermediate routers. Meanwhile, wireless channels have higher bit error rate (BER) due to fading and multipath effects. The resulting packet losses and bit errors can have devastating effects on multimedia quality.

Bandwidth limitation and fluctuation

Network conditions and characteristics in the current Internet such as bandwidth, packet loss ratio, delay, and delay jitter vary from time to time. Meanwhile, the capacity of wireless network also fluctuates with the changing environment.

Low performance for traditional transport-layer protocol

Traditional transport-layer protocol assumes congestion to be the primary cause for packet losses and unusual delay in the network. It will decrease the transmitting rate in the case of packet losses. Unfortunately, in wireless networks, the packet may also be dropped due to channel errors, thereby resulting in unnecessary reduction in end-to-end throughput.

Limited battery life

Comparing with fixed nodes, there is a battery lifetime constraint in mobile devices. In general, maintaining good media quality and minimizing average power consumption, including processing power and transmission power at mobile devices, are in conflict with each other. From multimedia-coding point of view, achieving better media quality usually consumes more processing power in source coder. From the network point of view, multipath fading and multiple-access interference (MAI) in wireless network necessitate the use of high transmission power.

Heterogeneity among users and networks

Receivers in multimedia delivery systems are quite different in terms of latency requirements, visual quality requirements, processing capabilities, power limitations, and bandwidth constraints. Moreover, multimedia may traverse different type of networks such as wire-line networks, 3G, and wireless local area network (WLAN) systems, each of which has different characteristics such as reliability, delay, jitter, bandwidth, and medium access control (MAC) mechanisms.

To handle the above challenges, many studies had been performed from different aspects. More specifically, link-layer, transport-layer, and application-layer solutions are proposed, respectively.

Considering the limitation of bandwidth in wireless systems, in the link layer, the most important target is to increase link utilization. It is known that RTP/UDP/IP and TCP/IP have the problem of large header overhead on bandwidth-limited links. Header compression has been proven to be efficient for using those protocols. Unfortunately, existing header compression schemes [3, 4] do not work well on noisy links, especially the one with high BER and long round-trip time (RTT). Internet Engineering Task Force (IETF) had set up a working group (WG), called robust header compression (ROHC), to address the header compression issue.

To handle the severe bandwidth and delay fluctuation in wireless Internet, available network condition estimation and congestion control are key issues needed to be addressed in the transport layer. Throughput calculation, packet-pair, and packet-train bandwidth probing are several popular techniques for bandwidth measurement [5]. Other network information such as packet error rate, delay, and delay jitter is also quite useful for high-quality media delivery. Different congestion and rate control schemes can be performed so that multimedia such as video and audio can adapt to the estimated network information in a smooth way [6].

There are many studies in the application layer to improve media delivery quality. Error protection, power saving, and proxy management are several hot topics. To overcome the packet loss and residual bit error in wireless Internet, error control techniques such as forward error correction (FEC) and automatic repeat request (ARQ) are necessary to maintain high-quality media delivery. Unequal error control [7] can be adopted if further taking different importance of different types/parts of media into account. To compromise the power-quality dilemma, power control and joint source-channel coding (JSCC) are two effective approaches. Power control is conducted from the group point of view by controlling transmission power and spreading gain for a group of users so as to reduce interference [8]; while JSCC is conducted from the individual user's point of view to effectively combat the errors occurred during transmission by allocating bits between source and channel [9]. The heterogeneous networks and different requirements of receivers ask for an efficient proxy-caching mechanism to satisfy different characteristics of receivers. Traditional proxy servers were designed to serve web requests for noncontinuous media, such as textual and image objects. With the increasing advent of video and audio streaming applications, continuous-media caching has been studied in [10]. However, the varying wireless Internet condition and different media characteristics impose challenges on how to efficiently cache both continuous and noncontinuous media.

Figure 1 depicts a general architecture for multimedia delivery over wireless Internet. In this architecture, multimedia server, base station (BS) (gateway) with media proxy, and heterogeneous mobile clients are deployed. Different solutions in different layers, that are mentioned above, have been incorporated in this architecture. More specifically, application-layer, transport-layer, and link-layer control mechanisms are all taken into account in order to achieve good end-to-end quality of multimedia services. In the later sections, recent advances for those components are reviewed.

2. CROSS-LAYER ARCHITECTURE FOR MULTIMEDIA DELIVERY OVER WIRELESS INTERNET

As mentioned above, different layers have different impacts on the media delivery quality and meanwhile, different layers have different approaches to improving the media delivery quality. Figure 2 depicts the user plane protocol stack in Universal Mobile Telecommunications System (UMTS) highlighting the layers with significant impact on system performance. As shown in Figure 2, the application is transmitted via TCP or UDP in the Internet part based on the traffic characteristics. IP packets arriving in the downlink to the UMTS network are transported to the radio network controller (RNC). Necessary header compression techniques are applied to the packets in the packet data convergence protocol (PDCP) layer. Then the corresponding packets are

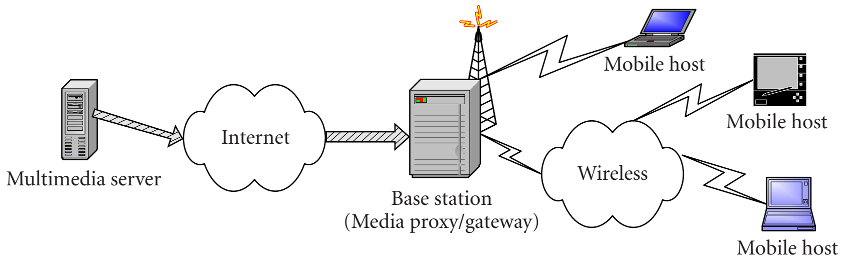


FIGURE 1: A general architecture for multimedia delivery over wireless Internet.

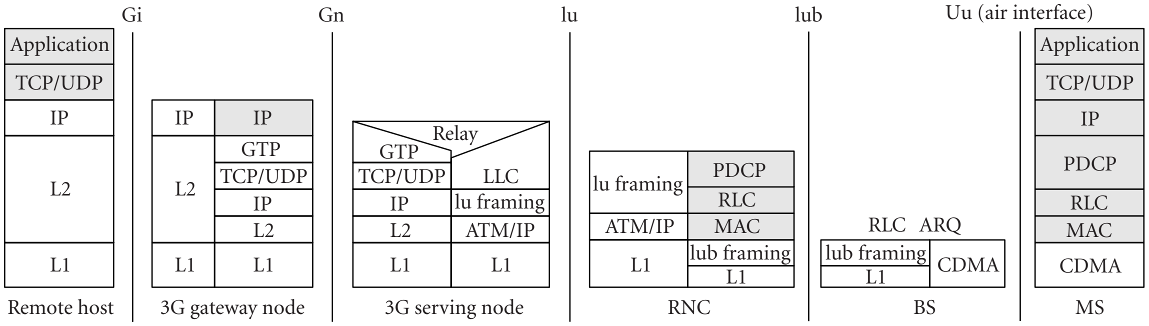


FIGURE 2: The user plane protocol stack of the end-to-end packet delivery in UMTS. (GTP: GPRS tunneling protocol; LLC: logical link control; BS: base station; MS: mobile station.)

transferred to the radio link layer (RLC), where they are segmented into smaller RLC protocol data units (PDUs). Diverse RLC ARQ schemes can be used to achieve the required reliability. The RLC PDU queues of a particular IP connection are served by the MAC layer. In deterministic transmission time intervals (TTIs), the MAC layer entities ask the corresponding RLC layer entities for a certain number of RLC PDUs, which are then transferred through the radio interface in MAC frames.

Considering the three key components in wireless Internet architecture, that is, multimedia server, BS (gateway), and mobile hosts, the cross-layer architecture should fulfill the functionalities as shown in Figure 3.

Dynamic wireless Internet condition estimation

Network estimation in different layers on the server, BS, and mobile-host side works together to track the varying wireless Internet conditions.

Network condition adaptation

Adaptively, adjust the amount of wireless Internet resources (i.e., the bandwidth) according to the varying network condition. It is fulfilled in the congestion control module in the multimedia server and BS.

Network-aware media adaptation

In response to the changing network conditions, media encoding mechanisms and different parts of media can be adaptively adjusted or tailored to maximize the system efficiency and perceived end-to-end quality.

Power efficiency and error robustness

Application and link-layer error control schemes can be used together for error robustness. Meanwhile, the overall power consumption in the mobile station (MS) should be minimized.

Efficient network utilization

To improve the network utilization, especially in wireless channels, header compression is performed in both BS and mobile hosts.

Multiservices supporting

Priority-based scheduling is an efficient way to support multiservices.

Network and clients heterogeneity

It can be supported by QoS-adaptive proxy caching.

In the following sections, we will review various QoS-support technologies for media streaming over wireless networks from application-layer, transport-layer, and link-layer aspects, respectively.

3. APPLICATION-LAYER MULTIMEDIA TRANSMISSION CONTROL

There are many topics that need to be addressed in application level to improve media delivery quality. More specifically, recent progress on adaptive media codec, error protection schemes, power saving approaches, and resource allocation solutions are reviewed in this section.

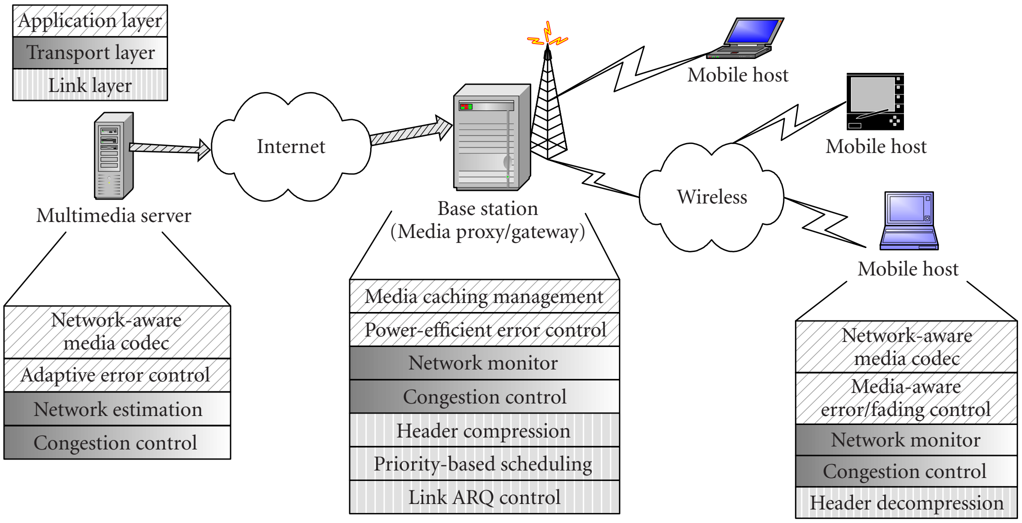


Figure 3: The cross-layer architecture for multimedia delivery over wireless Internet.

3.1. Network-aware media codec

In general, media codec has the ability to dynamically change the coding rate and other coding parameters according to the varying network conditions (bandwidth, loss, delay, etc.). Scalable coding techniques are introduced to realize this type of media adaptation. The major technique to achieve scalability is layered coding technology, which divides multimedia information into several layers, and the incremental reception of layers also increases the media fidelity. In video coding techniques which utilize DCT-based transform such as H.263 and MPEG-4, layered coding techniques can be categorized into three classes: temporal, spatial, and signal-to-noise ratio (SNR) scalability [11]. Video codecs with temporal (spatial) scalability encode a video sequence into one base layer and multiple enhancement layers. The base layer is encoded by itself with the lowest temporal (spatial) resolution, and the enhancement layers are coded based upon the temporal (spatial) prediction to lower layers. SNR scalable codecs encode a video sequence into several layers with the same temporal and spatial fidelity. Base layer is coded by itself to provide the basic quality, and enhancement layers are coded to enhance the video quality when added back to the base layer [12]. However, traditional layered codec requires the enhancement layers to be fully received before decoding. Otherwise, they will not provide any quality improvement at all. The fine granularity scalability (FGS), which adopts a method called bit-plane coding, can overcome this problem, that is, the enhancement layers can be truncated anywhere while maintaining some degree of quality improvement [11]. Further enhancement based upon FGS such as PFGS [13], improves the coding efficiency by using as many predictions as possible from the same and immediate lower layers in the current and previous frames. Another kind of video codec can divide a video sequence into multiple spatial-temporal subbands via 3D wavelet transform. Each of these subbands could be coded into multiple bit planes which can be further divided into several coding passes. Thus each layer of the coded video consists of some specific coding passes among all the subbands to achieve the scalability [14].

Speech coding is totally different from that for video, in that speech signals can be characterized by some particular model. There are generally three types of speech codec, that is, waveform, model-based, and hybrid codec. Waveform codec quantizes the amplitude of the signals at each point and improves the coding efficiency via an adaptive prediction filter which captures the correlation among the signals. A well-known example of this codec is adaptive differential pulse-code modulation (ADPCM). Model-based codec encodes speech signals based on a specific speech model. The parameters of the model, instead of the original signal samples, are quantized and transmitted. The majority of the modern speech codec is based upon the hybrid method which combines the waveform and model-based codec. Coded-excited linear predictive (CELP) coding is the most widely used codec of hybrid methods [15]. ITU G.729 and GSM-EFR are two standards based on CELP. All the above codecs encode speech signals into a voice stream with a fixed rate. The prevalent speech adaptation technique is adaptive multirate codec (AMR) [16], which is a mandatory standard for 3G systems. AMR allows a dynamic rate adaptation which is controlled by an in-band signaling procedure. In contrast to speech, audio codec cannot benefit from general models of the audio signal. The adaptation of audio codec is achieved similar to scalable video codec, which also has a layered structure [17, 18].

Error resilience and concealment techniques are employed in video/audio coding to prevent from or minimize the quality impairment in the case of packet losses/errors during media transmission. Common error resilience and concealment approaches include data partition, insertion of synchronization marks to prevent from drifting errors, concealing lost information from temporally/spatially adjacent region, and so forth. A good review of robustness techniques on streaming audio can be found in [19]. The introduction of the most updated error concealment/resilient tools in video coding can be found in [20], and the overview of existing error resilient and error concealment techniques in video coding can be found in [21, 22], respectively.

3.2. Network-adaptive and energy-efficient error protection

3.2.1. Network-adaptive error control

Besides the varying network conditions, there are also packet losses and bit errors in wireless Internet. Thus, efficient error protection scheme is essential for improving end-to-end media quality. As mentioned above, ARQ and FEC are two basic error correction mechanisms. FEC is a channel coding technique protecting the source data at the expense of adding redundant data during transmission. FEC has been commonly suggested for applications with strict delay requirements such as voice communications [23]. In the case of media transmission, where delay requirements are not that strict or round-trip delay is small (e.g., video/audio delivery over a single wireless channel), ARQ is applicable and usually plays a role as a complement to FEC. In [24], hybrid FEC/ARQ method was adopted and the delay bound can be achieved by limiting the number of retransmissions. A similar technology is also used in [25, 26, 27, 28]. Targeting at video streaming over WLAN, in [29], hybrid ARQ/FEC was adopted in unicast and FEC was used in multicast. A similar work in WLAN can also be found in [30], where unequal error protection (UEP) and ARQ are applied, taking the characteristics of WLAN into account. To reduce the bursty effect of packet losses, packet interleaving can also be adopted in conjunction with FEC and ARQ [31]. As mentioned in Section 3.1, layered scalable media codec usually divides media into a base layer and multiple enhancement layers. Since the correct decoding of enhancement layers depends on the errorless receipt of base layer, base layer is more important than enhancement layers. Therefore, it is natural to adopt UEP for layered scalable media. Specifically, stronger FEC protection can be applied to the base and lower layer data while weaker channel coding protection level is applied to the higher layer parts.

To keep the residual error/loss rate under the targeted level, it is efficient to adjust the FEC protection level in response to the underlying changing network condition. For example, Global System for Mobile Communications (GSM) systems can dynamically distribute the voice data and channel coding among the overall bandwidth to the possible best voice quality. Moreover, studying how to add FEC codes to layered scalable media is of a great interest recently [7, 26, 27, 31]. In [7, 31], the protection level of every layer in video bitstream is dynamically adjusted according to the changing network conditions. In [26], a channel-adaptive application level error control scheme utilizing UEP and delay constrained ARQ, has been proposed for scalable video streaming. Current and estimated RTTs are used to maximize retransmissions times while meeting the delay requirements. Similar approaches are also applicable to the layered audio stream [27].

3.2.2. Bit allocation between source and channel coding

Considering the limited resource in the media delivery system, an important question raised is how to decide the distribution between source codes and channel codes; and specifically for layered codec, to what extend a specific layer should be protected. This is generally known as the bit allocation problem. To answer the above questions, an analytic model describing the relation between media quality and source/channel parameters should be developed. The most common metrics to evaluate media quality is the expected end-to-end distortion D_T, where D_T consists of source distortion D_S and channel distortion D_C. Source distortion is caused during the media source encoding. Variable encoding methods such as quantization, motion estimation in video coding, linear prediction in voice coding, and rate control can be the cause of source distortion. Channel distortion occurs when fragments of media stream are lost due to network congestion, or incorrectly received due to wireless channel noise. Therefore, the bit allocation problem can be formulated as the optimization problem

$$\min D_T(D_S, D_C) \quad \text{s.t. } R_S + R_C \le R_T, \tag{1}$$

where R_T is the total available bandwidth, and R_S and R_C are the rates for source coding and channel coding, respectively. Based upon the analytical model above, optimal bit allocation can be resolved by numeric methods such as dynamic programming, penalty function, or Lagrange multiplier. Several bit allocation schemes have been developed according to the above model taking different kinds of scalable media codec and channel models into account [26, 27, 31].

In wireless Internet scenario, the packet losses consist of those due to network congestion and those caused by wireless transmission errors, which in turn may have different loss patterns. Since different loss patterns lead to different perceived QoS at the application level [32], Yang et al. [33] proposed a loss differentiated rate-distortion based bit allocation scheme which minimizes the end-to-end video distortion taking the above different loss patterns into account.

From the above rate-distortion analytical model, we can see that both source coding and channel coding parameters can affect the final media quality. JSCC schemes are thus proposed to achieve the optimal end-to-end quality by adjusting the source and channel coding parameters, simultaneously. From the source-coding point of view, adjusting quantization parameters or entropy coder to control source rate [34], selecting inter- or intracoding mode to trade off the coding efficiency and the error resilient ability [35], are several key issues that can be jointly considered with channel coding. In [28], an integrated JSCC scheme has been proposed to study the performance of FEC/ARQ, meanwhile the joint effect of FEC/ARQ and error-resilient source coding is considered. Because of the large amount of source and network parameters that could be jointly adjusted, the computational complexity of searching for the optimal solution is extraordinary high. Some works have to limit the dependencies among the parameters so that it can be solvable to dynamic programming [36]. Another method is to use the local optimal solution instead of the global one [37].

3.2.2.1 Energy-efficient bit allocation

In addition to optimizing the quality of media streaming, mobile users in wireless Internet also need to considering the constraints imposed by limited battery power. How to achieve the good user's perceived QoS while minimizing the power consumption is yet another challenge. As mentioned in Section 1, there is a tradeoff between maintaining good media quality and minimizing power consumption. In order to maintain a certain transmission quality, larger transmission rate in wireless channels inherently needs more power, and more power also allows adopting more complicated media encoding algorithms with higher complexity and thus can achieve better coding efficiency. Therefore, an efficient way to obtain optimal media quality is jointly considering source-channel coding and power consumption issues. In end systems, the power consumption of media streaming over wireless mainly consists of transmission power and processing power.

Traditional joint source coding and power control schemes are mainly targeted at minimizing the power consumed for a single user. For example, a low-power communication system for image transmission has been investigated in [38]. In [39], quantization and mode selection have been discussed together with transmission power consumption. Moreover, rate adaptation could be further taken into account [40]. In addition to transmission power, processing power, which further consists of power consumption for source coding and channel coding, has been considered in [41]. In [41], the analytical model characterizing the relation among power consumption, source and channel coding can be denoted as

$$\min P_T(P_S, P_C, P_t) \quad \text{s.t. } R_S + R_C \le R_T, \ \text{s.t. } D_T \le D_0, \tag{2}$$

or

$$\min D(D_S, D_C) \quad \text{s.t. } R_S + R_C \le R_T, \ \text{s.t. } P_T \le P_0, \tag{3}$$

where P_T is the total power consumption, P_S, P_C, and P_t are the power required by source coding, channel coding, and transmission, respectively, and D_0 and P_0 are system- or user-specified distortion and power thresholds.

It is worth noting that power-quality optimization for a single user would potentially increase the interference to other mobile users in the interference range, which results in QoS reduction in those users who have been in the optimized state. In order to rereach the optimal status, those interfered users may adjust their transmission powers, which will also introduce interference to all the other users. Therefore, the global power optimization should be studied from the group point of view [39, 41].

3.3. QoS-adaptive proxy caching for multimedia delivery over wireless Internet

Multimedia applications usually have stronger QoS requirements than that of best-effort services, which bring great challenges to the Internet and unreliable wireless networks. Moreover, the heterogeneity among different devices in different networks implies that their demands are different in terms of delay, bandwidth, and visual quality. Deploying multimedia proxies at the edge of Internet connecting both remote servers and end clients is an efficient way to satisfy the heterogeneous requirements of end users. On the proxy-server side, the backbone network between proxy and server is a best-effort network, that is, the network conditions such as bandwidth, packet loss ratio, delay, and jitter vary from time to time. On the proxy-client side, two types of clients, namely Internet clients and wireless clients, access the proxy via different networks. By caching the popular media content and treating end users from different networks differently, multimedia proxies can also alleviate network congestion and reduce the latency and workload on multimedia servers.

In order to provide efficient streaming service for both Internet and wireless clients, media proxy should deal with the following issues:

(1) providing high quality video streaming service for both Internet clients and various wireless clients;
(2) managing limited cache resource in proxy so as to provide optimal performance for heterogeneous users;
(3) evaluating and selecting multimedia replicas from the servers in the Internet to relay streaming for end clients.

Traditional proxy servers are designed to serve web requests for noncontinuous media such as texts and images. In contrast to these objects, continuous media has very different characteristics such as high delay and bandwidth sensibility and tolerance of moderate data loss. Moreover, within a scalable media stream, different part of data is usually of different impacts on media quality. And caching the real-time traffic should also take the varying conditions in the Internet and wireless networks into account. All these call for a different design for multimedia proxies.

Cache replacement policy is one of the key components in the proxy design. Traditional caching replacement schemes designed for web data can be roughly classified as recency-based and frequency-based. Recency-based schemes such as least recently used (LRU) [42] exploit temporal locality among cache objects or recency of reference, that is, objects which have been referenced recently are more likely to be rereferenced in the near future. Frequency-based policies, for example, least frequently used (LFU) [43] make cache replacement decision according to the popularity of the content in the cache. Some other solutions are proposed to balance the frequency and recency-based algorithms such as LRU-k and LRFU [44, 45]. To support continuous media, Rejaie et al. introduced a replacement policy for layered media [46]. Tewari et al. proposed a resource-based caching

(RBC) policy for both continuous and noncontinuous media, balancing the usage of disk space and I/O [47]. All the above mentioned works use hit rate as the performance metric. In [48], Yu et al. proposed a QoS-adaptive replacement policy for mixed media. In this scheme, different priorities among and within media, along with the network conditions, are considered and the goal of cache management is to maximize the hit rate of noncontinuous media and the perceived QoS for the real-time continuous media. Q. Zhang et al. further proposed a unified cost metrics to measure the cache performance balancing the issues of network, latency, and media distortion [49].

Prefetching between proxies and servers is another effective technique in proxy design, if the user access pattern can be accurately estimated [10, 50, 51]. Since continuous media is more likely to be accessed sequentially, Sen et al. proposed a proxy prefetching scheme for multimedia stream [10]. In [48, 52], the QoS-adaptive prefetching schemes took the network condition into account. Other caching techniques such as batching and merging were incorporated in [52]. In order to handle heterogeneous users with different QoS requirement, multimedia proxies are also assigned tasks such as transcoding [53], rate control [54], or any network-adaptive techniques mentioned in Section 3.2. If there are multiple servers between proxies and servers, selecting a proper server to achieve load balance, and meanwhile maximizing media quality, is another issue [49].

4. TRANSPORT-LAYER MULTIMEDIA TRANSMISSION CONTROL

To efficiently deliver multimedia over wireless Internet, it is important to estimate the status of underlying networks so that multimedia applications can adapt accordingly. IETF has developed several standards such as real-time transport control protocol (RTP/RTCP), real-time streaming protocol (RTSP), session initiation protocol (SIP), session description protocol (SDP), and streaming control transport protocol (SCTP) to monitor and control the media streaming process. However, how well these protocols can work to achieve a desirable media streaming quality relies on the accuracy of the estimation of the network conditions.

One of the most important issues in the estimation of network conditions is to detect current available bandwidth and perform efficient congestion control. A proper congestion control scheme should maximize the bandwidth utilization and at the same time should avoid overusing network resource which may cause network collapse. Since TCP is the dominantly used transport protocol in the Internet, it is very important for multimedia streaming applications to be TCP-friendly, which means the long-term throughput of a multimedia stream is close to that of a TCP flow under similar network conditions [55]. Generally there are two kinds of TCP-friendly streaming control protocol: window-based and model-based. Window-based congestion control scheme performs additive increase and multiplicative decrease (AIMD) rate adjustment which is similar to TCP [56]. The rate-adaptive protocol (RAP) [57] mimics the behavior of the TCP congestion window and acknowledgement is triggered by every incoming packet on the receiver side to measure packet loss and RTT. TCP emulation at receivers (TEAR) [58] maintains a "virtual" congestion window at receivers and tries to derive from the incoming packets whether the congestion window should increase or decrease, note that the window adjustment is also in AIMD manner. Model-based congestion control algorithms model TCP throughput by packet loss rate, RTT, and retransmission timeout (RTO) [59]. They use the derived model and current measured network parameters to determine current available bandwidth for streaming protocols [60]. TFRC is a well-known streaming protocol based upon this kind of model [61].

4.1. Packet loss differentiation and estimation

To design a streaming protocol for wireless Internet, several issues need to be considered. We will discuss these issues in the remaining sections. The most important one is congestion-loss estimation. In wireless environment, packet losses can be caused by either network congestion or transmission errors in wireless channels. TCP and TCP-friendly streaming protocols treat any packet loss as a signal of network congestion and consequently reduce the transmission rate. However, this rate reduction is not necessary when the loss is due to wireless errors, which in turn underutilizes the network resource.

Generally, there are two types of solutions to discriminate the losses, which are split connection and end-to-end method, respectively [62]. In the former case, a proxy locates at the edge of the wired and wireless network to handle the two types of network separately [6, 63, 64]. However, this type of solution introduces the deployment and cost issues for network operators. In the end-to-end method, one solution is to incorporate explicit congestion notification (ECN) to detect whether the network is in congestion status [65] and ignore the signal of packet losses. This method requires ECN scheme to be enabled at any intermediate router. Another type of end-to-end method is to use heuristic methods to differentiate the congestive packet loss from the erroneous loss. In [66], Biaz and Vaidya used packet interarrival time to differentiate the cause of losses. While in [67], Cen et al. extended their idea and further incorporated relative one-way trip time (ROTT) as an additional metrics to discriminate the losses. In [68], packet pair, that is, two packets sent back by back, was used in conjunction with hidden Markov model to achieve loss differentiation. Tsaoussidis and C. Zhang further proposed a technique called packet "wave" which is a series of back-by-back packets to detect the cause of losses [69]. All the heuristic methods expect a packet to exhibit a certain behavior under network congestion or erroneous losses. However, a specific behavior of a packet in the wireless Internet reflects the joint effect of several factors. Moreover, the traffic pattern in the Internet itself is a complicated research topic, and using a simple pattern to predict the behaviors of the packets is risky. In [33], Yang et al. proposed an end-to-end loss differential method, which utilizes the link information in wireless channel.

4.2. Available bandwidth estimation

Another way to avoid the packet loss ambiguity is to bypass this problem, that is, use metrics other than packet loss as a signal of network congestion.

Packet pair is one of the effective methods to measure bandwidth. TCP-Westwood measures the ACK pair to derive the short-term bandwidth and counts the amount of acknowledged data during a time period to get the relative long-term bandwidth estimation [70]. In [71], Wu et al. also used packet pair to probe the current bottleneck of the network and interpacket arrival time is adopted to measure the available bandwidth. By feeding these two values into an exponentially weighted moving average (EMWA) filter, the bandwidth estimation can be stable when possible, and agile when necessary.

Another method to estimate the available bandwidth is based on delay. TCP-Vegas is the most famous solution based upon this idea [72]. TCP-Vegas maintains a minimum RTT as the "base RTT," and compares the current measured RTT to base RTT. If difference between them is small, TCP-Vegas deduces that network is not in the congestion status and increases the transmission rate. Otherwise, it reduces the rate to avoid congestion. Based upon this idea, TCP-Veno proposed a packet loss differentiation method which claims that only packet losses during the period when RTT varies greatly are congestive losses [73]. A severe problem of delay-based available bandwidth measurement is how to achieve fairness with TCP, which is not clear.

4.3. Delay variation and estimation

In order to alleviate the packet losses due to transmission errors in wireless networks, ARQ is often adopted in modern wireless systems. This will introduce large delay variation in data transmission, which may cause inaccurate estimation of RTT and RTO. This in turn will result in performance degradation in window-based congestion control solution [74]. Chan and Ramjee [75] proposed to use receiver's window field to convey the current wireless channel conditions to the TCP sender and an ACK regulator to manage (postpone) the release of ACKs to the sender to absorb the delay variation. For model-based streaming protocols, they adjust sending rate based on the estimated packet loss ratio and RTT. To reduce reverse path traffic, many streaming protocols send only a single acknowledgement back to measure the RTT during a predefined period of time. However, in wireless environment, the aforementioned delay fluctuation causes a dramatic variation of RTT value. That is to say, the rate estimation counted on RTT may be inaccurate and fluctuate greatly, which is not favorable to delay-sensitive real-time multimedia streaming applications. Yang et al. [33] proposed a streaming protocol which can collect as many RTT observations as possible during a period of time without increasing additional reverse traffic. It turns out that this method can smooth the rate variation of the congestion control behavior.

5. LINK-LAYER MULTIMEDIA TRANSMISSION CONTROL

5.1. Channel estimation

To date, large amount of wireless multimedia studies have focused on robust media delivery over wireless channels [76, 77, 78]. Most of the works assumes they have perfect knowledge of the fluctuating wireless channel conditions such as BER, delay, bandwidth, and so forth. However, it is very complicated to convert the physical channel QoS parameters into the desired QoS requirements, which can be understood by multimedia applications. For example, the raw physical-layer data rate is not equal to that obtained in link layer, considering the header and modulation overheads, and the channel decoding efficiency. Moreover, it is not obvious how to achieve end-to-end optimality for multimedia delivery although a single layer (e.g., physical or link layer) can reach the optimum.

The physical wireless channel can be characterized by the large-scale loss and the small-scaling fading models [79]. Large-scale loss models can predict the physical channel variation caused by user location and background interference level, while small-scale fading models statistically describe the radio signal strength fluctuations in very short time durations or over very short traverse distance [79]. The physical channel state can be characterized by a finite-state Markov chain, which classifies the physical status into several states in terms of different BERs or data rates [80, 81]. A series of work from Zorzi et al. [81, 82] shows that Markov model is a good approximation on block transmission over fading wireless channels.

Due to the different QoS metrics used in different layers addressed above, researchers propose to move the physical channel models up to link layer, that is, converting physical QoS parameters into application-understandable QoS metrics. Effective capacity (EC) theory [83] is proposed to model a wireless link by two EC functions: the probability of nonempty buffer and the QoS exponent of a connection, which characterize the queuing behavior in the link layer. Therefore, EC model is a powerful tool and can easily be used to provide the multimedia QoS metrics such as delay bound and available bandwidth [84]. According to the results of [82], Q. Zhang et al. adopted a first-order Markov process to model the RLC layer frames in 3G wireless channels [26]. Based upon the characteristics of 3G wireless channels [82, 85] and considering the interaction between UDP and RLC protocol, they derived the available UDP throughput by physical channel parameters.

5.2. Header compression

As stated above, IETF had set up ROHC WG to address header compression issues. The goal of the ROHC WG is to develop header compression schemes that perform well over links with high error rates and long link RTT. In the ROHC framework, relevant information from past packets is maintained in a context. The context information is used to compress (decompress) subsequent packets. The compressor and decompressor update their contexts upon certain events.

It is known that, impairment events may lead to inconsistencies between the contexts of the compressor and the decompressor, which in turn may cause incorrect decompression. Thus, ROHC scheme needs some mechanisms for avoiding context inconsistencies and also mechanisms for making the contexts consistent when they are not.

Due to the limited packet loss robustness of existing realtime traffic compression scheme, CRTP, and the demands of the cellular industry for an efficient way of transporting voice over IP over wireless, ROHC has designed an ROHC scheme for IP/UDP/RTP headers [86], which are generous in size, especially compared to the payloads often carried by packets with such headers. This scheme had been accepted as IETF RFC 3095. ROHC-RTP has become a very efficient, robust and capable compression scheme, able to compress the headers down to a total size of one octet only. Also, transparency is guaranteed to an extremely great extent even when residual bit errors are present in compressed headers delivered to the decompressor.

The work on TCP header compression, called ROHC-TCP, has recently started in ROHC group. Compared with previous works such as CTCP (RFC 1144) or IPHC (RFC 2507), ROHC-TCP focused on ROHC over lossy links. In addition, ROHC tries to improve compression efficiency taking the optional fields (e.g., timestamp, SACK) into account [87]. TCP-aware robust header compression (TAROC) scheme [88] can significantly improve the compression efficiency in unidirectional link by using congestion window tracking mechanisms and window-based least significant bit (LSB) encoding technique.

5.3. Application-adaptive ARQ and priority-based scheduling

Error control can be performed both at application/transport layer and link-layer. In general, the link-layer ARQ is more efficient than that in the application/transport layer. This is because, firstly, ARQ across a single link has a shorter control loop than that in upper layer, thus can recover the lost data more quickly. Secondly, link-layer ARQ can operate on frames which are much smaller than IP datagram, and therefore become more efficient in terms of error/loss recovery. Thirdly, a link-layer ARQ procedure is able to use local knowledge which is not available to end hosts [89]. However, optimal performance can hardly be achieved based upon the link-layer ARQ. ARQ running both at the link layer and in end-to-end method could lead to undesirable competition on data retransmission. Such kind of “contention” of data retransmission in different layers will result in severe performance degradation in transport protocols [64], and may potentially have two or more copies of one packet residing in intermediate routes at the same time.

Link ARQ schemes, according to their willingness to retransmit lost frames to ensure reliable data delivery, can be classified into perfectly-persistent, high-persistent, and low-persistent ARQ. Those schemes are differed in delay and reliability, which inspires people to adopt an upper-layer-aware link ARQ for applications which may have different QoS requirements. The idea is that the applications signal their QoS requirements to each link along the path on a per-packet basis. Link-layer ARQ can therefore adaptively adjust its behavior in accordance to different QoS requirements [26]. The effects of the adaptive ARQ are implicitly fed back to applications through packet drops or delay.

In addition to the adaptive ARQ, it is well known that priority-based packet scheduling can also support differential QoS services. In priority-based schedulers, packets are grouped into several classes with different priority according to their QoS requirements. Packets in the class with higher priority are more likely to be transmitted first. And packets in the same class are served in a FIFO manner. Based upon the priority scheduling mechanism, each QoS class will have some sort of statistical QoS guarantees. Traditional priority packet-scheduling algorithms based upon generalized processor-sharing (GPS) fluid model such as weight fair queueing (WFQ) [90] inherently couple delay bound and bandwidth requirement, which lack flexible QoS provision. Liao and Zhu [91] proposed a priority packet-scheduling algorithm by relaxing the packet service order. In [84], the authors employed the simplest strict (nonpreemptive) prioritized scheduling policy and derived the rate constraints for different video substreams with different QoS requirements according to the EC theory [83]. The EC theory can also be applied to the QoS-provision scheme which exploits multiuser diversity [92]. The advantage of this scheme is that it can achieve capacity gain under strict QoS requirements where traditional multiuser diversity scheduling cannot be applied directly.

6. CONCLUSION

This paper presents a framework with cross-layer architecture for multimedia delivery over wireless Internet. We review various media delivery techniques at the application layer, transport layer, and link layer to achieve good user’s perceived quality of multimedia data. More specifically, network-aware adaptive media source coding, dynamical estimation of the varying channel, adaptive and energy-efficient application and link-level error control, efficient congestion control, header compression, adaptive ARQ and priority-based scheduling, as well as the QoS-adaptive proxy caching are explicitly reviewed in the architecture.

Cross-layer design for multimedia delivery over heterogeneous wireless Internet presents many challenges and opportunities. There are still a lot of issues needed to be further investigated. Moreover, this paper mainly focuses on QoS support in unicast media streaming. Efficient work on QoS provisions for multicast media streaming is an area that requires lots of efforts [29]. Mobility also has significant impact on perceived QoS during multimedia streaming. How to maintain an acceptable media quality when handoff happens is another research direction [93]. Enabling media streaming over ad hoc network is more challenging than over traditional wireless networks, where mobile hosts communicate with BS. In wireless ad hoc networks, dynamic changing topology and interference result in even greater QoS fluctuation. Recently, multipath media streaming and QoS-aware MAC design are two promising cross-layer approaches to providing QoS support for ad hoc networks [94, 95].

REFERENCES

[1] E. Dahlman, P. Beming, J. Knutsson, F. Ovesjo, M. Persson, and C. Roobol, "WCDMA—the radio interface for future mobile multimedia communications," *IEEE Trans. Veh. Technol.*, vol. 47, no. 4, pp. 1105–1118, 1998.

[2] B. O'Hara and A. Petrick, *IEEE 802.11 Handbook: A Designer's Companion*, IEEE Press, New York, NY, USA, 1999.

[3] M. Degermark, B. Nordgren, and S. Pink, "IP header compression," IETF RFC 2507, February 1999.

[4] S. Casner and V. Jacobson, "Compressing IP/UDP/RTP headers for low-speed serial links," IETF RFC 2508, February 1999.

[5] K. Lai and M. Baker, "Measuring bandwidth," in *Proc. Conference on Computer Communications (INFOCOM '99)*, vol. 1, pp. 235–245, New York, NY, USA, March 1999.

[6] F. Yang, Q. Zhang, W. Zhu, and Y.-Q. Zhang, "An efficient tranport scheme for multimedia transport over wireless Internet," in *Proc. IEEE International Conference on 3G Wireless and Beyond (3Gwireless '01)*, San Francisco, Calif, USA, June 2001.

[7] T. Zhang and Y. Xu, "Unequal packet loss protection for layered video transmission," *IEEE Trans. Broadcast.*, vol. 45, no. 2, pp. 243–252, 1999.

[8] A. Sampath, P. S. Kumar, and J. M. Holtzman, "Power control and resource management for a multimedia CDMA wireless system," in *Proc. 6th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '95)*, vol. 1, pp. 21–25, Toronto, Ontario, Canada, September 1995.

[9] L. Qian, D. L. Jones, K. Ramchandran, and S. Appadwedula, "A general joint source-channel matching method for wireless video transmission," in *Proc. IEEE Data Compression Conference (DCC '99)*, pp. 414–423, Snowbird, Utah, USA, March 1999.

[10] S. Sen, J. Rexford, and D. Towsley, "Proxy prefix caching for multimedia streams," in *Proc. 18th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '99)*, vol. 3, pp. 1310–1319, New York, NY, USA, March 1999.

[11] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 301–317, 2001.

[12] M. Ghanbari, *Video Coding: An Introduction to Standard Codecs*, IEE Telecommunications Series, INSPEC/IEE Publishing, London, UK, 1999.

[13] F. Wu, S. Li, and Y.-Q. Zhang, "A framework for efficient progressive fine granularity scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 332–344, 2001.

[14] J. Hua, Z. Xiong, and X. Wu, "High-performance 3-D embedded wavelet video (EWV) coding," in *IEEE 4th Workshop on Multimedia Signal Processing*, pp. 569–574, Cannes, France, October 2001.

[15] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '85)*, vol. 10, pp. 937–940, Tampa, Fla, USA, April 1985.

[16] 3GPP, "Adaptive multi-rate (AMR) speech transcoding," Ts 26.090 v. 4.0.0, March 2001.

[17] D. Pan, "Digital audio compression," *Digital Technical Journal*, vol. 5, no. 2, pp. 28–40, 1993.

[18] D. Pan, "A tutorial on MPEG/audio compression," *IEEE Multimedia*, vol. 2, no. 2, pp. 60–74, 1995.

[19] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 40–48, 1998.

[20] T. Wiegand, G. Sullivan, G. Bjntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, 2003.

[21] G. Cote, F. Kossentini, and S. Wenger, "Error resilience coding," in *Compressed Video over Networks*, M. T. Sun and A. Reibman, Eds., pp. 309–341, Marcel Dekker, New York, NY, USA, 2000.

[22] Y. Wang and Q.-F. Zhu, "Error control and concealment for video communication: a review," *Proceedings of IEEE*, vol. 86, no. 5, pp. 974–997, 1998.

[23] C. Boutremans and J.-Y. Le Boudec, "Adaptive joint playout buffer and FEC adjustment for Internet telephony," in *Proc. 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '03)*, vol. 1, pp. 652–662, San Francisco, Calif, USA, March–April 2003.

[24] Q. Zhang and S. Kassam, "Hybrid ARQ with selective combining for fading channels," *IEEE J. Select. Areas Commun.*, vol. 17, no. 5, pp. 867–880, 1999.

[25] G. Wang, Q. Zhang, W. Zhu, and Y.-Q. Zhang, "Channel-adaptive error control for scalable video over wireless channel," in *Proc. 7th International Workshop on Mobile Multimedia Communications (MoMuC '00)*, Tokyo, Japan, October 2000.

[26] Q. Zhang, W. Zhu, and Y.-Q. Zhang, "Channel-adaptive resource allocation for scalable video transmission over 3G wireless network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 8, pp. 1049–1063, 2004.

[27] Q. Zhang, G. Wang, Z. Xiong, J. Zhou, and W. Zhu, "Error robust scalable audio streaming over wireless IP networks," *IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 897–909, 2004.

[28] F. Zhai, Y. Eisenberg, T. N. Pappas, R. Berry, and A. K. Katsaggelos, "Rate-distortion optimized hybrid error control for real-time packetized video transmission," in *Proc. IEEE International Conference on Communications (ICC '04)*, vol. 3, pp. 1318–1322, Paris, France, June 2004.

[29] A. Majumda, D. Sachs, I. Kozintsev, K. Ramchandran, and M. Yeung, "Multicast and unicast real-time video streaming over wireless LANs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 524–534, 2002.

[30] Q. Li and M. Schaar, "Providing adaptive QoS to layered video over wireless local area networks through real-time retry limit adaptation," *IEEE Trans. Multimedia*, vol. 6, no. 2, pp. 278–290, 2004.

[31] Q. Zhang, G. Wang, W. Zhu, and Y.-Q. Zhang, "Robust scalable video streaming over Internet with network-adaptive congestion control and unequal loss protection," in *Proc. Packet Video Workshop*, Kyongju, Korea, April 2001.

[32] W. Jiang and H. Schulzrinne, "Modeling of packet loss and delay and their effect on real-time multimedia service quality," in *Proc. 10th International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV '00)*, Chapel Hill, NC, USA, June 2000.

[33] 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 J. Select. Areas Commun.*, vol. 22, no. 4, pp. 777–790, 2004.

[34] B. Tao, H. A. Peterson, and B. W. Dickinson, "A rate-quantization model for MPEG encoders," in *Proc. IEEE*

International Conference on Image Processing (ICIP '97), vol. 1, pp. 338–341, Santa Barbara, Calif, USA, October 1997.

[35] Z. He, J. Cai, and C. Chen, "Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 511–523, 2002.

[36] A. Ortego and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 23–50, 1998.

[37] P. A. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media," Tech. Rep. MSR-TR2001-35, Microsoft Research, Redmond, Wash, USA, February 2001.

[38] M. Goel, S. Appadwedula, N. R. Shambhag, K. Ramchandran, and D. L. Jones, "A low-power multimedia communication system for indoor wireless applications," in *Proc. IEEE Workshop on Signal Processing Systems (SiPS '99)*, pp. 473–482, Taipei, Taiwan, October 1999.

[39] Y. Eisenberg, C. E. Luna, T. N. Pappas, R. Berry, and A. K. Katsaggelos, "Joint source coding and transmission power management for energy efficient wireless video communications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 411–424, 2002.

[40] 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 J. Select. Areas Commun.*, vol. 21, no. 10, pp. 1710–1720, 2003.

[41] Q. Zhang, Z. Ji, W. Zhu, and Y.-Q. Zhang, "Power-minimized bit allocation for video communication over wireless channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 398–410, 2002.

[42] A. Dan and D. Towsley, "An approximate analysis of the LRU and FIFO buffer replacement schemes," in *Proc. ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pp. 143–152, ACM Press, Boulder, Colo, USA, May 1990.

[43] H. Chou and D. DeWitt, "An evaluation of buffer management strategies for relational database systems," in *Proc. 11th International Conference on Very Large Data Bases (VLDB '85)*, pp. 127–141, Stockholm, Sweden, August 1985.

[44] E. O'Neil, P. O'Neil, and G. Weikum, "The LRU-K page replacement algorithm for database disk buffering," in *Proc. ACM SIGMOD International Conference on Management of Data*, pp. 297–306, ACM Press, Washington, DC, USA, May 1993.

[45] T. Kelly, Y. Chan, S. Jamin, and J. MacKie-Mason, "Biased replacement policies for web caches: differential quality-of-service and aggregate user value," in *Proc. 4th International Web Caching Workshop*, San Diego, Calif, USA, March 1999.

[46] R. Rejaie, M. Handley, H. Yu, and D. Estrin, "Proxy caching mechanism for multimedia playback streams in the Internet," in *Proc. 4th International Web Caching Workshop*, San Diego, Calif, USA, March 1999.

[47] R. Tewari, H. M. Vin, A. Dan, and D. Sitaram, "Resource-based caching for web servers," in *Proc. SPIE/ACM Multimedia Computing and Networking (MMCN '98)*, vol. 3310 of *Proceedings of SPIE*, pp. 191–204, San Jose, Calif, USA, January 1998.

[48] F. Yu, Q. Zhang, W. Zhu, and Y.-Q. Zhang, "QoS-adaptive proxy caching for multimedia streaming over the Internet," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 3, pp. 257–269, 2003.

[49] Q. Zhang, Z. Xiang, W. Zhu, and L. Gao, "Cost-based cache replacement and server selection for multimedia proxy across wireless Internet," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 587–598, 2004.

[50] E. Cohen and H. Kaplan, "Prefetching the means for document transfer: a new approach for reducing web latency," in *Proc. 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '00)*, vol. 2, pp. 854–863, Tel Aviv, Israel, March 2000.

[51] L. Fan, Q. Jacobson, P. Cao, and W. Lin, "Web prefetching between low-bandwidth clients and proxies: potential and performance," in *Proc. ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 178–187, ACM Press, Atlanta, Ga, USA, May 1999.

[52] B. Wang, S. Sen, M. Adler, and D. Towsley, "Optimal proxy cache allocation for efficient streaming media distribution," *IEEE Trans. Multimedia*, vol. 6, no. 2, pp. 366–374, 2004.

[53] B. Shen, S. Lee, and S. Basu, "Caching strategies in transcoding-enabled proxy systems for streaming media distribution networks," *IEEE Trans. Multimedia*, vol. 6, no. 2, pp. 375–386, 2004.

[54] D. Nam and S. Park, "Adaptive multimedia stream presentation in mobile computing environment," in *Proc. IEEE Region 10 Conference (TENCON '99)*, vol. 2, pp. 966–969, Cheju Island, South Korea, September 1999.

[55] S. Floyd and K. Fall, "Promoting the use of end-to-end congestion control in the Internet," *IEEE/ACM Trans. Networking*, vol. 7, no. 4, pp. 458–472, 1999.

[56] M. Allman, V. Paxson, and W. Stevens, "TCP congestion control," IETF RFC 2581, April 1999.

[57] R. Rejaie, M. Handley, and D. Estrin, "Rap: an end-to-end rate-based congestion control mechanism for realtime streams in the Internet," in *Proc. 18th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '99)*, vol. 3, pp. 1337–1345, New York, NY, USA, March 1999.

[58] I. Rhee, V. Ozdemir, and Y. Yi, "TEAR: TCP emulation at receivers-flow control for multimedia streaming," Tech. Rep., Department of Computer Science, North Carolina State University, Raleigh, NC, USA, April 2000.

[59] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, "Modeling TCP Reno performance: a simple model and its empirical validation," *IEEE/ACM Trans. Networking*, vol. 8, no. 2, pp. 133–145, 2000.

[60] Q. Zhang, W. Zhu, and Y.-Q. Zhang, "Resource allocation for multimedia streaming over the Internet," *IEEE Trans. Multimedia*, vol. 3, no. 3, pp. 339–355, 2001.

[61] S. Floyd, M. Handley, J. Padhye, and J. Widmer, "Equation based congestion control for unicast applications," in *Proc. ACM SIGCOMM 2000*, pp. 43–56, Stockholm, Sweden, August 2000, http://www.aciri.org/tfrc.

[62] G. Montenegro, S. Dawkins, M. Kojo, V. Magret, and N. Vaidya, "Long thin networks," IETF RFC 2757, January 2000.

[63] A. Bakre and B. Badrinath, "I-TCP: Indirect TCP for mobile hosts," in *Proc. IEEE 15th International Conference on Distributed Computing Systems (ICDCS '95)*, pp. 136–143, Vancouver, British Columbia, Canada, May–June 1995.

[64] H. Balakrishnan, V. Padmanabhan, S. Seshan, and R. Katz, "A comparison of mechanisms for improving TCP performance over wireless links," *IEEE/ACM Trans. Networking*, vol. 5, no. 6, pp. 756–769, 1997.

[65] S. Floyd, "TCP and explicit congestion notification," *ACM SIGCOMM Computer Communication Review*, vol. 24, no. 5, pp. 8–23, 1994.

[66] S. Biaz and N. Vaidya, "Discriminating congestion losses from wireless losses using inter-arrival times at the receiver," in *Proc. IEEE Symposium on Application-Specific Systems and Software Engineering and Technology (ASSET '99)*, pp. 10–17, Richardson, Tex, USA, March 1999.

[67] S. Cen, P. Cosman, and G. Voelker, "End-to-end differentiation of congestion and wireless losses," in *Proc. SPIE Multimedia Computing and Networking (MMCN '02)*, M. G. Kienzle and P. J. Shenoy, Eds., vol. 4673 of *Proceedings of SPIE*, pp. 1–15, San Jose, Calif, USA, January 2002.

[68] D. Barman and I. Matta, "Effectiveness of loss labeling in improving TCP performance in wired/wireless networks," in *Proc. 10th IEEE International Conference on Network Protocols (ICNP '02)*, pp. 2–11, Paris, France, November 2002.

[69] V. Tsaoussidis and C. Zhang, "TCP-Real: receiver-oriented congestion control," *Journal of Computer Networks*, vol. 40, no. 4, pp. 477–497, 2002.

[70] M. Gerla, B. K. F. Ng, M. Y. Sanadidi, M. Valla, and R. Wang, "TCP westwood with adaptive bandwidth estimation to improve efficiency/friendliness tradeoffs," *Computer Communications*, vol. 27, no. 1, pp. 41–58, 2004.

[71] H. Wu, Q. Zhang, and W. Zhu, "Design study for multimedia transport protocol in heterogeneous networks," in *Proc. IEEE International Conference on Communications (ICC '03)*, vol. 1, pp. 567–571, Anchorage, Alaska, USA, May 2003.

[72] L. S. Brakmo, S. W. O'Malley, and L. L. Peterson, "TCP Vegas: new techniques for congestion detection and avoidance," in *Proc. Conference on Communications Architectures, Protocols and Applications*, pp. 24–35, ACM Press, London, UK, August 1994.

[73] C. P. Fu and S. C. Liew, "TCP Veno: TCP enhancement for transmission over wireless access networks," *IEEE J. Select. Areas Commun.*, vol. 21, no. 2, pp. 216–228, 2003.

[74] R. Ludwig and R. Katz, "The Eifel algorithm: making TCP robust against spurious retransmissions," *ACM SIGCOMM Computer Communication Review*, vol. 30, no. 1, pp. 30–36, 2000.

[75] M. Chan and R. Ramjee, "Improving TCP/IP performance over third generation wireless networks," in *Proc. Conf. on Computer Communications (INFOCOM '04)*, Hong Kong, China, March 2004.

[76] J. Hagenauer, T. Stockhammer, C. Weiss, and A. Donner, "Progressive source coding combined with regressive channel coding on varying channels," in *Proc. 3rd ITG Conference Source and Channel Coding*, vol. 159, pp. 123–130, VDE Publishing House, Munich, Germany, January 2000.

[77] D. Wu, Y. Hou, Y.-Q. Zhang, W. Zhu, and H. J. Chao, "Adaptive QoS control for MPEG-4 video communication over wireless channels," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS '00)*, vol. 1, pp. 48–51, Geneva, Switzerland, May 2000.

[78] T. Stockhammer, H. Jenkac, and C. Weiss, "Error control for wireless progressive video transmission," in *Proc. International Conference on Image Processing (ICIP '02)*, vol. 1, pp. 545–548, Rochester, NY, USA, September 2002.

[79] T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice Hall, New Jersey, NJ, USA, 1996.

[80] Q. Zhang and S. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Communications*, vol. 47, no. 11, pp. 1688–1692, 1999.

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

[82] M. Zorzi, R. R. Rao, and L. B. Milstein, "On the accuracy of a first-order Markov model for data transmission on fading channels," in *Proc. 4th IEEE International Conference on Universal Personal Communications Record (ICUPC '95)*, pp. 211–215, Tokyo, Japan, November 1995.

[83] 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.

[84] W. Kumwilaisak, Y. Hou, Q. Zhang, W. Zhu, C.-C. Kuo, and Y.-Q. Zhang, "A cross-layer quality-of-service mapping architecture for video delivery in wireless networks," *IEEE J. Select. Areas Commun.*, vol. 21, no. 10, pp. 1685–1698, 2003.

[85] A. Chockalingam and G. Bao, "Performance of TCP/RLP protocol stack on correlated fading DS-CDMA wireless links," *IEEE Trans. Vehicular Technology*, vol. 49, no. 1, pp. 28–33, 2000.

[86] C. Bormann, C. Burmeister, M. Degermark, et al., "Robust header compression (ROHC)," IETF RFC 3095, July 2001.

[87] G. Pelletier, L.-E. Jonsson, M. West, R. Price, and K. Sandlund, "RObust header compression (ROHC): a profile for TCP/IP (ROHC-TCP)," IETF Internet Draft, July 2004.

[88] H. Liao, Q. Zhang, W. Zhu, and Y.-Q. Zhang, "TCP-Aware robust header compression (TAROC)," IETF Internet Draft, November 2001.

[89] G. Fairhurst and L. Wood, "Link ARQ issues for IP traffic," IETF Internet Draft, November 2000.

[90] K. Parekh and R. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, pp. 344–357, 1993.

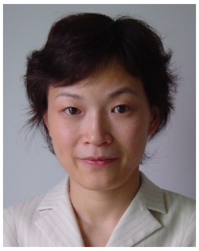
[91] H. Liao and W. Zhu, "Enable fair queueing with decoupled bandwidth-delay guarantees by relaxing the packet service order," in *Proc. IEEE International Conference on Communications (ICC '03)*, vol. 1, pp. 147–151, Anchorage, Alaska, USA, May 2003.

[92] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," in *Proc. IEEE International Conference on Communications (ICC '03)*, vol. 3, pp. 2202–2207, Anchorage, Alaska, USA, May 2003.

[93] Y. Pan, M. Lee, J. Kim, and T. Suda, "An end-to-end multipath smooth handoff scheme for stream media," *IEEE J. Select. Areas Commun.*, vol. 22, no. 4, pp. 653–663, 2004.

[94] S. Mao, S. Lin, S. S. Panwar, Y. Wang, and E. Celebi, "Video transport over ad hoc networks: multistream coding with multipath transport," *IEEE J. Select. Areas Commun.*, vol. 21, no. 10, pp. 1721–1737, 2003.

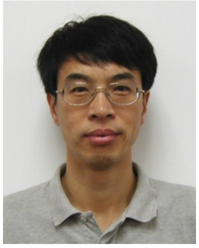
[95] S. Kumar, V. S. Raghavan, and J. Deng, "Medium access control protocols for ad-hoc wireless networks: a survey," to appear in *Elsevier Ad-Hoc Network Journal*.



Qian Zhang received the B.S., M.S., and Ph.D. degrees from Wuhan University, China, in 1994, 1996, and 1999, respectively, all in computer science. She joined Microsoft Research Asia in July 1999 as an Associate Researcher in the Internet Media Group, and now she is a Researcher and Project Leader of the Wireless and Networking Group. She has published about 80 refereed papers in international leading journals and key conferences. She is the inventor of about 20 pending patents. Her current research interests include seamless roaming across wireless networks, multimedia delivery over wireless Internet, next-generation wireless networks, and P2P network/ad hoc network. She also participated in many activities in the IETF ROHC WG for TCP/IP header compression. Dr. Zhang is a Member of the Visual Signal Processing and Communication Technical Committee, and the Multimedia System and Application Technical Committee, IEEE Circuits and Systems Society. She is also a Member and Chair of QoSIG of the Multimedia Communication Technical Committee, IEEE Communications Society.

Dr. Zhang is now serving as a Guest Editor for a special issue on wireless video in IEEE Wireless Communication Magazine. She is named one of the top 100 young creative individuals by MIT Technology Review (TR100) in 2004.

Fan Yang received the B.S., M.S., and Ph.D. degrees in computer science from Nanjing University, Nanjing, China, in 1997, 2000, and 2003, respectively. He joined Microsoft Research Asia, Beijing, China, in March 2004 as an Associate Researcher in the Wireless and Network Group. His research interests include multimedia delivery and resource allocation optimization for wireless communication.



Wenwu Zhu received the B.E. and M.E. degrees from the National University of Science and Technology, China, in 1985 and 1988, respectively, the M.S. degree from Illinois Institute of Technology, Chicago, and the Ph.D. degree from Polytechnic University, Brooklyn, NY, in 1993 and 1996, respectively, all in electrical engineering. From 1999 till now, he is with Microsoft Research Asia and now he is a Research Manager of Wireless and Networking Group. Prior to his current post, he was with Bell Labs, Lucent Technologies, NJ, as a member of the technical staff from 1996 to 1999. He has published over 180 refereed journal and conference papers, and patents. His current research interests include multimedia communication and networking, and wireless communication and networking. Dr. Zhu has been on the Editorial Board of 8 IEEE journals such as AE for IEEE Transactions on Mobile Computing, IEEE Transactions on Multimedia, and IEEE Transactions on Circuits and Systems for Video Technology. He received the Best Paper Award in IEEE Transactions on Circuits and Systems for Video Technology in 2001. Dr. Zhu is now the Chairman of IEEE Circuits and System Society Beijing Chapter. He serves as the Secretary of Visual Signal Processing and Communication Technical Committee.