Scalable Video Coding and Transport over Broad-Band Wireless Networks

DAPENG WU, STUDENT MEMBER, IEEE, YIWEI THOMAS HOU, MEMBER, IEEE, AND YA-QIN ZHANG, FELLOW, IEEE

Invited Paper

With the emergence of broadband wireless networks and increasing demand of multimedia information on the Internet, wireless multimedia services are foreseen to become widely deployed in the next decade. Real-time video transmission typically has requirements on quality of service (QoS). However, wireless channels are unreliable and the channel bandwidth varies with time, which may cause severe degradation to video quality. In addition, for video multicast, the heterogeneity of receivers makes it difficult to achieve efficiency and flexibility. To address these issues, three techniques, namely, scalable video coding, network-aware adaptation of end systems, and adaptive QoS support from networks, have been developed. This paper unifies the three techniques and presents an adaptive framework, which specifically addresses video transport over wireless networks. The adaptive framework consists of three basic components: 1) scalable video representations; 2) network-aware end systems; and 3) adaptive services. Under this framework, as wireless channel conditions change, mobile terminals and network elements can scale the video streams and transport the scaled video streams to receivers with a smooth change of perceptual quality. The key advantages of the adaptive framework are: 1) perceptual quality is changed gracefully during periods of QoS fluctuations and handoffs; and 2) the resources are shared in a fair manner.

Keywords*—*Adaptive framework, adaptive services, network-aware end systems, quality-of-service, real-time video, scalable video coding, wireless.

I. INTRODUCTION

Due to proliferation of multimedia on the World Wide Web and the emergence of broadband wireless networks, wireless video communication has received great interest from both industry and academia. Delivery of real-time video typically has quality of service (QoS) requirements, e.g., bandwidth, delay and error requirements. First, video transmission usually has minimum bandwidth requirements (e.g., 28 kb/s) to achieve acceptable presentation quality. Second, real-time video has strict delay constraints (e.g., 1 s). This is because real-time video must be played out continuously. If the video packet does not arrive in a timely manner, the playout process will pause, which is annoying to human eyes. Third, video applications typically impose upper limits on bit error rate (BER) (e.g., 1%) since too many bit errors would seriously degrade the video presentation quality. However, *unreliability* and *bandwidth fluctuations* of wireless channels can cause severe degradation to video quality. Furthermore, for video multicast, *heterogeneity* of receivers makes it difficult to achieve efficiency and flexibility. We discuss these issues in detail as follows.

Unreliability: Compared with wired links, wireless channels are typically much more noisy and have both small-scale (multipath) and large-scale (shadowing) fades [54], making the BER very high. The resulting bit errors can have a devastating effect on video presentation quality [62]. Therefore, it is crucial to develop robust transport mechanisms for video over wireless channels.

Bandwidth Fluctuations: The bandwidth fluctuates for several reasons. First, when a mobile terminal moves between different networks [e.g., from a wireless local area network (LAN) to a wireless wide area network (WAN)], the available bandwidth may vary drastically (e.g., from a few megabits per second to a few kilobits per second). Second, when a handoff happens, a base station may not have enough unused radio resource to meet the demand of a newly joined mobile host. Third, the throughput of a wireless channel may be reduced due to multipath fading, co-channel interference, and noise disturbances. Last but not least, the capacity of a wireless channel may fluctuate with the changing distance between the base station and the mobile host. Consequently, bandwidth fluctuations pose a serious problem for real-time video transmission over wireless networks.

Manuscript received February 26, 2000; revised September 5, 2000.

D. Wu is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA.

Y. T. Hou is with Fujitsu Laboratories of America, Sunnyvale, CA 94085 USA.

Y.-Q. Zhang is with Microsoft Research China, Haidian District, Beijing 100080, China.

Publisher Item Identifier S 0018-9219(01)00449-2.

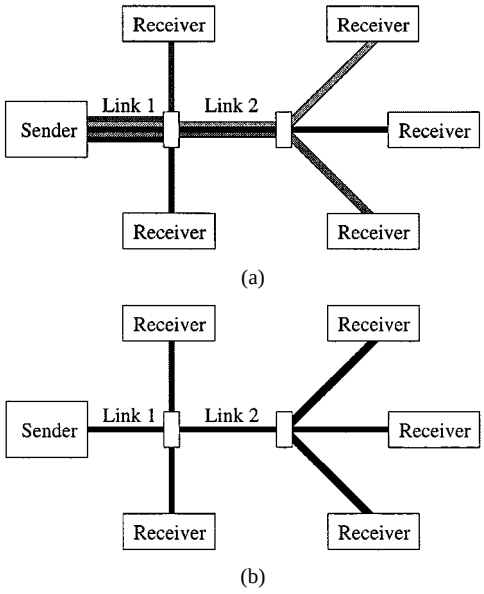

Fig. 1. (a) Unicast video distribution using multiple point-to-point connections. (b) Multicast video distribution using point-to-multipoint transmission.

Heterogeneity: To set the stage for our discussion of the heterogeneity problem, we first describe the pros and cons of unicast and multicast. Unicast delivery of real-time video uses point-to-point transmission, where only one sender and one receiver are involved. In contrast, multicast delivery of real-time video uses point-to-multipoint transmission,[1] where one sender and multiple receivers are involved. For applications such as video conferencing, delivery using multicast can achieve high-bandwidth efficiency, while unicast delivery of such applications is inefficient. An example is given in Fig. 1, where, for unicast, five copies of the same video content move across Link 1 and three copies move across Link 2 [Fig. 1(a)]. In contrast, multicast eliminates this replication. That is, there is only one copy of the video content traversing any link in the network [Fig. 1(b)], resulting in substantial bandwidth savings. However, the efficiency of multicast is achieved at the cost of losing the service flexibility of unicast (i.e., in unicast, each receiver can individually negotiate service parameters with the source). Such lack of flexibility in multicast can be problematic under a heterogeneous environment, where receivers may be different in terms of latency requirements, visual quality requirements, processing capabilities, power limitations (wireless versus wired), and bandwidth limitations. For example, the receivers in Fig. 1(b) may attempt to request different video quality with different bandwidth. But only one copy of the video content is sent out from the source. As a result, all the receivers have to receive the same video content with the same quality. It is thus a challenge to design a multicast mechanism that not only achieves efficiency in network bandwidth, but also meets the heterogeneous requirements of the receivers.

To address the above issues, three techniques have been studied in great depth individually. These techniques are scalable video coding, network-aware adaptation of end systems, and adaptive QoS support from networks, which are briefly described as follows.

Scalable Video Coding: In the example in Fig. 2 a raw video sequence is compressed into three layers: a base layer (i.e., Layer 0) and two enhancement layers (i.e., Layers 1 and 2). The base layer can be independently decoded and it provides basic video quality; the enhancement layers can only be decoded together with the base layer and they further refine the quality of the base layer. As shown in Fig. 2, the compressed video streams can adapt to three levels of bandwidth usage (i.e., 64 kb/s, 256 kb/s, and 1 Mb/s). In contrast, nonscalable video (say, a video stream with 1 Mb/s rate) is more susceptible to bandwidth fluctuations (e.g., a bandwidth change from 1 Mb/s to 100 kb/s) since it only has one representation. Thus, scalable video is more suitable than nonscalable video under a time-varying wireless environment. Second, scalable video representation is a good solution to the heterogeneity problem in the multicast case [33], [38]. In the example in Fig. 3, suppose that the wireless LAN can support at least 1 Mb/s; the path from the source to Receiver 2 can support 256 kb/s; the path from the source to Receiver 3 can support 64 kb/s. This makes each receiver have different bandwidth limitations. To accommodate this difference, the source uses scalable video and sends each video layer to a separate IP multicast group. At the receiver side, each receiver subscribes to a certain set of video layers by joining the corresponding IP multicast group. Specifically, Receiver 1 joins all three IP multicast groups. Accordingly, it consumes 1 Mb/s and receives all three layers. Receiver 2 joins the two IP multicast groups for Layers 0 and 1 with bandwidth usage of 256 kb/s. Receiver 3 only joins the IP multicast group for Layer 0 with bandwidth consumption of 64 kb/s. Hence, scalable video representations can effectively cope with the heterogeneity problem. Third, scalable video representations naturally fit unequal error protection, which can effectively combat bit errors induced by the wireless medium [71].

Network-Aware Adaptation of End Systems: Most current video applications are insensitive to changing network conditions. In a time-varying wireless environment, however, video applications must be robust and adaptive in the presence of QoS fluctuations (e.g., unreliability and bandwidth fluctuations) [9]. To address this issue, a new approach called *network-aware adaptation* was proposed [39], [47], [50], [64]. Network-aware adaptation, as the name implies, consists of two elements: network awareness and adaptation. Network awareness refers to having knowledge about the current status of underlying network resources (e.g., available bandwidth and bit error conditions) [9]. Adaptation is to adapt video streams based on network status. It has been shown that network-aware adaptation of end systems can significantly improve performance of the applications [48], [50].

Adaptive QoS Support from Networks: Adaptive QoS support (or *adaptive services*) is a technique to adapt video streams during periods of QoS fluctuations and handoffs. Adaptive services have been demonstrated to be able to effectively mitigate fluctuations of resource availability in wireless networks [4]. There have been many proposals on adaptive approaches and services in the literature, which include an "adaptive reserved

[1]Point-to-multipoint transmission can be regarded as a subset of multipoint-to-multipoint transmission.

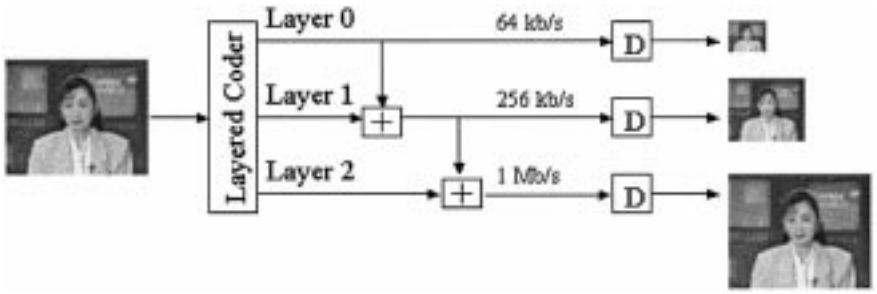

Fig. 2. Layered video encoding/decoding. D denotes the decoder.

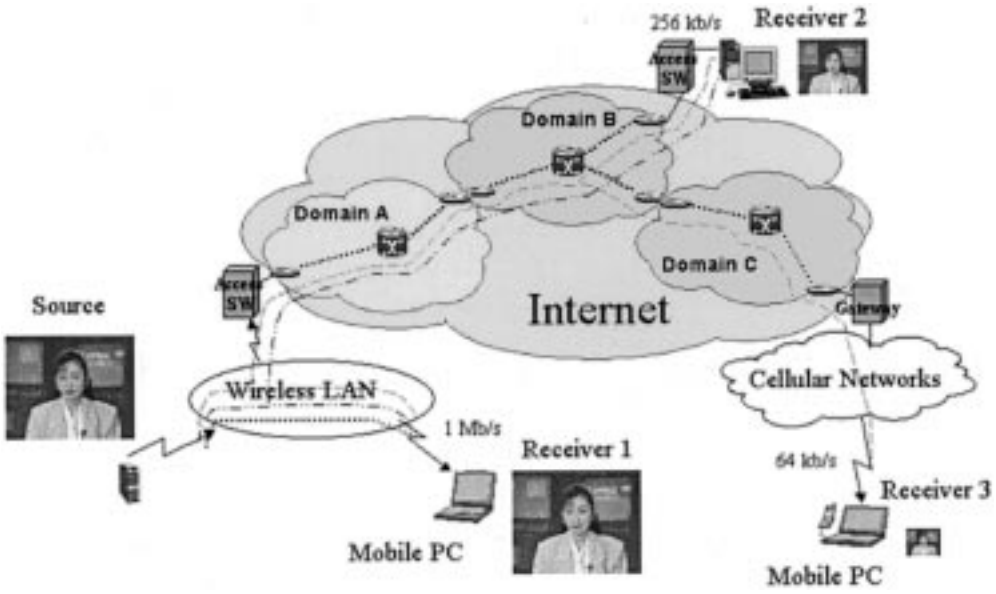

Fig. 3. IP multicast for layered video.

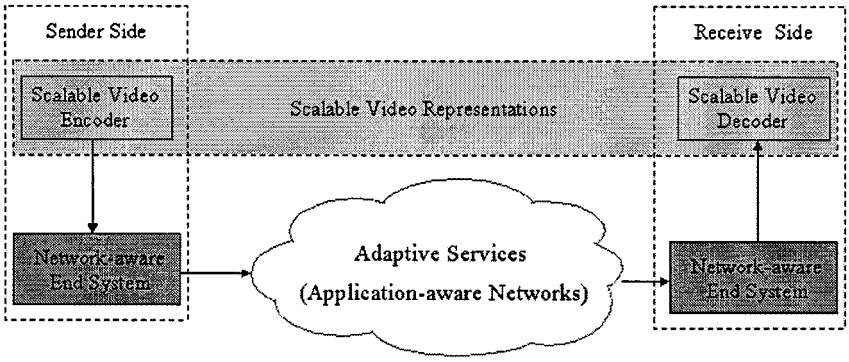

Fig. 4. Adaptive framework.

service" framework [27], a wireless adaptive mobile information system (WAMIS) [2], an adaptive service based on QoS bounds and revenue [37], an adaptive framework targeted at end-to-end QoS provisioning [43], a utility-fair adaptive service [7], a framework for soft QoS control [52], a teleservice model based on an adaptive QoS paradigm [21], an adaptive QoS framework called AQuaFWiN [59], and an adaptive QoS management architecture [26], among others.

This paper unifies the three techniques simultaneously and presents an adaptive framework, which specifically addresses scalable video transport over wireless networks. The adaptive framework consists of three basic components: 1) scalable video representations, each of which has its own specified QoS requirements; 2) network-aware end systems, which are aware of network status and can adapt the video streams accordingly; and 3) adaptive services, with which the networks support the adaptive QoS required by scalable video representations. Under this framework, as wireless channel conditions change, mobile terminals and network elements can scale the video streams and transport the scaled video streams to receivers with a smooth change of perceptual quality. Fig. 4 illustrates the adaptive framework.

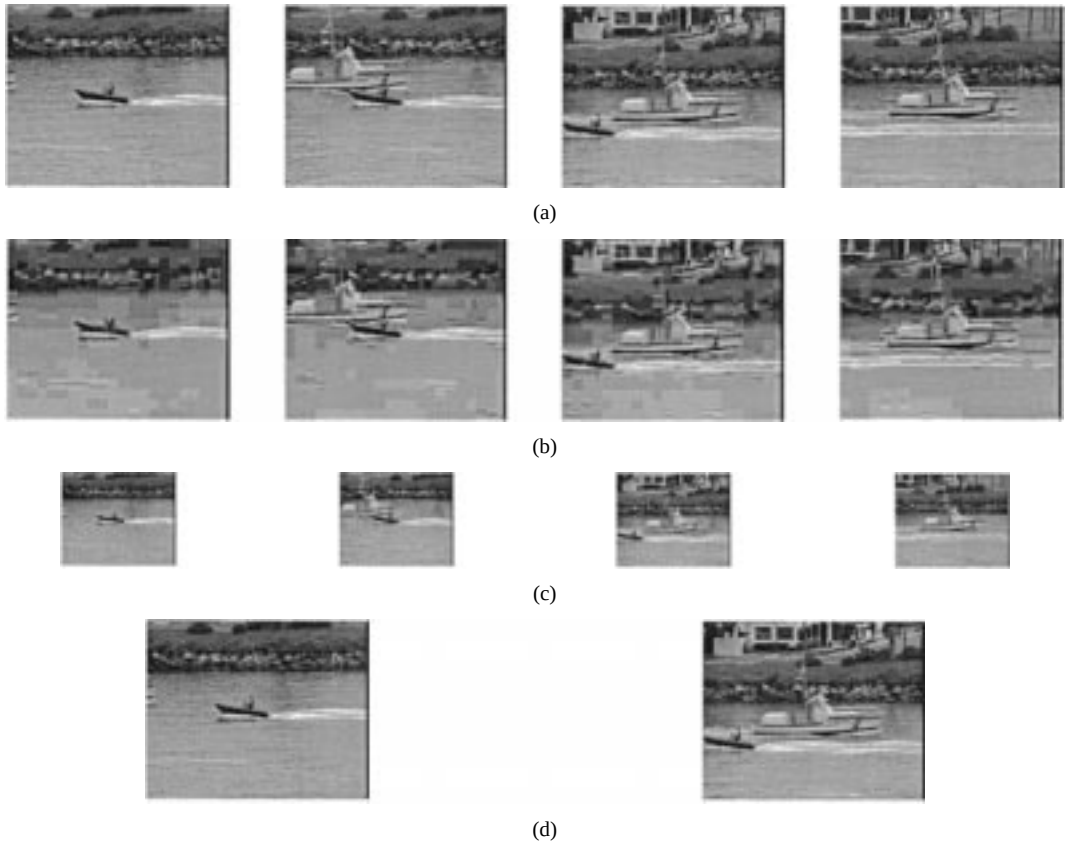

Fig. 5. Scalable video: (a) Video frames reconstructed from the complete bit stream. (b) Video frames with degraded quality. (c) Video frames with a smaller image size. (d) Video frames with a lower frame rate.

On the sender side, raw video is compressed by a scalable video encoder. Then the compressed video is sent to the networks by a network-aware end system, which monitors the network status and adapts the video streams accordingly. Inside the networks, the adaptive services provide adaptive QoS support to the scalable video. On the receiver side, a network-aware end system can sense the network status and coordinate with the networks in video transport. The received packets are decoded by a scalable video decoder.

The adaptive framework is a combination of network-aware end systems and *application-aware networks*. By application-aware networks, we mean network elements are capable of processing application-specific information such as video formats. With network-aware end systems and application-aware networks, the adaptive framework is able to achieve the following advantages.

Graceful Quality Degradation: Scalable video can adapt its video representations to bandwidth variations. If the available bandwidth becomes smaller than the sending rate of the scalable video, application-aware network elements can perform scaling to groom the video streams rather than drop packets indiscriminately. In other words, the network elements understand the format of the scalable video representations so that they can drop packets in a way that gracefully degrades the stream's quality instead of corrupting the flow outright.

Fairness: When there is excess bandwidth (excluding reserved bandwidth), the competing video streams can share the excess bandwidth in a fair manner. Specifically, the fairness could be either a utility-based fairness [7] or a max–min fairness [20], [37].

The remainder of this paper is organized as follows. In Sections II–IV, we describe each component in the adaptive framework. Specifically, Section II presents various scalable video coding mechanisms, Section III discusses network-aware end systems, and Section IV describes the adaptive services. In Section V, we summarize the paper and point out future research directions.

II. Scalable Video Coding

A scalable video coding scheme is to compress a raw video sequence into multiple substreams. One of the compressed substreams is a base substream, which can be independently decoded and provides coarse visual quality; other compressed substreams are enhancement substreams, which can only be decoded together with the base substream and provide better visual quality; the complete bit stream (i.e., combination of all the substreams) provides the highest quality. Specifically, compared with decoding the complete bit stream Fig. 5(a), decoding the base substream or multiple substreams produces pictures with either degraded quality [Fig. 5(b)], or a smaller image size [Fig. 5(c)], or a lower frame rate [Fig. 5(d)].

Scalable video coding schemes have found a number of applications. For video applications over the Internet, scal-

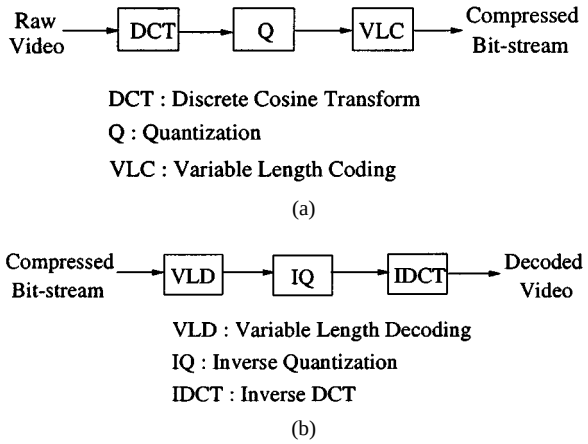

Fig. 6. (a) Nonscalable video encoder. (b) Nonscalable video decoder.

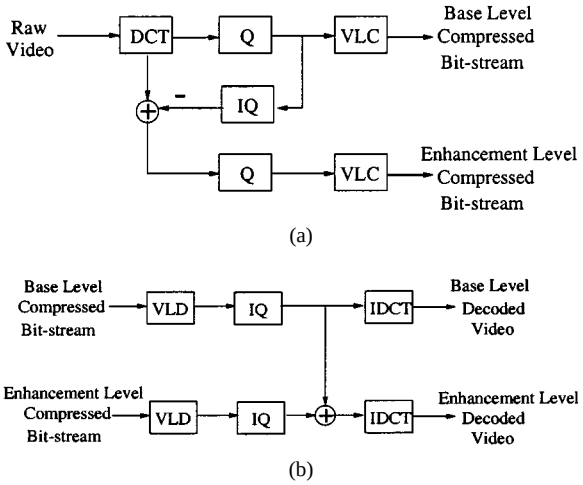

Fig. 7. (a) SNR-scalable encoder. (b) SNR-scalable decoder.

able coding can assist rate control during network congestion [66]; for web browsing of a video library, scalable coding can generate a low-resolution video preview without decoding a full-resolution picture [29]; for multicast applications, scalable coding can provide a range of picture quality suited to heterogeneous requirements of receivers (as shown in Fig. 3) [38].

As we mentioned before, scalable video can withstand bandwidth variations. This is due to its bandwidth scalability. Basically, the bandwidth scalability of video consists of SNR scalability, spatial scalability, and temporal scalability, which will be presented in Sections II-A to II-C, respectively.

To depict a clear picture about scalable coding mechanisms, we first briefly describe a nonscalable encoder/decoder as shown in Fig. 6. At the nonscalable encoder, the raw video is transformed by discrete cosine transform (DCT), quantized, and coded by variable-length coding (VLC). Then the compressed video stream is transmitted to the decoder through the networks. At the nonscalable decoder, the received compressed video stream is first decoded by variable-length decoding (VLD), then inversely quantized, and finally inversely DCT transformed.

For simplicity, we only show intramode[2] and only use DCT as an example in the above codec. Similarly, Sections II-A–II-C only describe intramode for scalable video coding mechanisms and only use DCT. For wavelet-based scalable video coding, please refer to [14], [22], [36], [55], [56], [58], and references therein.

A. SNR Scalability

SNR scalability is defined as representing the same video in different SNR or perceptual quality [see Figs. 5(a) and (b)]. To be specific, SNR-scalable coding quantizes the DCT coefficients to different levels of accuracy by using different quantization parameters. The resulting streams have different SNR levels or quality levels. In other words, the smaller the quantization parameter is, the better quality the video stream can achieve.

[2]Intramode coding refers to coding a video unit without any reference to previously coded data.

An SNR-scalable encoder with two-level scalability is depicted in Fig. 7(a). For the base level, the SNR-scalable encoder operates in the same manner as the nonscalable video encoder. For the enhancement level, the operations are performed in the following order:

1) The raw video is DCT transformed and quantized at the base level.
2) The base-level DCT coefficients are reconstructed by inverse quantization.
3) Subtract the base-level DCT coefficients from the original DCT coefficients.
4) The residual is quantized by a quantization parameter, which is smaller than that of the base level.
5) The quantized bits are coded by VLC.

Since the enhancement level uses a smaller quantization parameter, it achieves better quality than the base level.

An SNR-scalable decoder with two-level scalability is depicted in Fig. 7(b). For the base level, the SNR-scalable decoder operates exactly the same as the nonscalable video encoder. For the enhancement level, both levels must be received, decoded by VLD, and inversely quantized. Then the base-level DCT coefficient values are added to the enhancement-level DCT coefficient refinements. After this stage, the summed DCT coefficients are inversely DCT transformed, resulting in enhancement-level decoded video.

B. Spatial Scalability

Spatial scalability is defined as representing the same video in different spatial resolutions or sizes [see Fig. 5(a) and (c)]. Typically, spatially scalable video is efficiently encoded by making use of spatially up-sampled pictures from a lower layer as a prediction in a higher layer. Fig. 8(a) shows a block diagram of a two-layer spatially scalable encoder. For the base layer, the raw video is first spatially down-sampled,[3] then DCT transformed, quantized, and

[3]For example, spatially down-sampling with ratio 4 : 1 is to select one pixel from four pixels and discard the nonselected pixels.

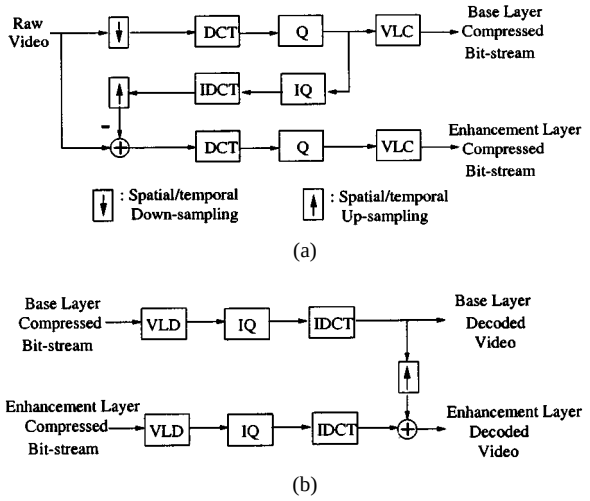

Fig. 8. (a) Spatially/temporally scalable encoder. (b) Spatially/temporally scalable decoder.

VLC coded. For the enhancement layer, the operations are performed in the following order:

1) The raw video is spatially down-sampled, DCT transformed, and quantized at the base layer.
2) The base-layer image is reconstructed by inverse quantization and inverse DCT.
3) The base-layer image is spatially up-sampled.[4]
4) Subtract the up-sampled base-layer image from the original image.
5) The residual is DCT transformed, and quantized by a quantization parameter, which is smaller than that of the base layer.
6) The quantized bits are coded by VLC.

Since the enhancement layer uses a smaller quantization parameter, it achieves finer quality than the base layer.

A spatially scalable decoder with two-layer scalability is depicted in Fig. 8(b). For the base layer, the spatially scalable decoder operates exactly the same as the nonscalable video encoder. For the enhancement layer, both layers must be received, decoded by VLD, inversely quantized, and inversely DCT transformed. Then the base-layer image is spatially up-sampled. The up-sampled base-layer image is combined with the enhancement-layer refinements to form enhanced video.

C. Temporal Scalability

Temporal scalability is defined as representing the same video in different temporal resolutions or frame rates [see Fig. 5(a) and (d)]. Typically, temporally scalable video is encoded by making use of temporally up-sampled pictures from a lower layer as a prediction in a higher layer. The block diagram of temporally scalable codec is the same as that of spatially scalable codec (see Fig. 8). The only difference is that the spatially scalable codec uses spatial down-sampling and spatial up-sampling while the temporally scalable codec uses temporal down-sampling and temporal up-sampling. Temporal down-sampling uses frame skipping. For example, a temporal down-sampling with ratio 2 : 1 is to discard one frame from every two frames [see Fig. 5(d)]. Temporal up-sampling uses frame copying. For example, a temporal up-sampling with ratio 1 : 2 is to make a copy for each frame and transmit the two frames to the next stage.

So far, we have discussed SNR, spatial, and temporal scalability, which provide multiple video representations in different SNR/spatial/temporal resolutions, respectively. Each video representation has different significance and bandwidth requirement. The base layer is more important, while an enhancement layer is less important. The base layer needs less transmission bandwidth due to its coarser quality; an enhancement layer requires more transmission bandwidth due to its finer quality. As a result, SNR/spatial/temporal scalability achieves bandwidth scalability. That is, the same video content can be transported at different rates (i.e., in different representations).

The different video layers can be transmitted in different bit streams called substreams. On the other hand, they can also be transmitted in the same bit stream, which is called an embedded bit stream. As shown in Fig. 9, an embedded bit stream is formed by interleaving the base layer with the enhancement layer(s). An embedded bit stream is also bandwidth scalable since application-aware networks can select a certain layer(s) from an embedded bit stream and discard it (them) to match the available bandwidth.

We would like to point out that we have described only basic scalable mechanisms, that is, SNR, spatial, and temporal scalability. There can be combinations of the basic mechanisms, such as spatiotemporal scalability [15]. Other scalability mechanisms include frequency scalability for MPEG-1/2 [42], object-based scalability for MPEG-4 [61], and fine-granular scalability [30]–[32], [49], [60].

In the above section, we have discussed the technique of scalable video coding. The primary goal of using bandwidth-scalable video coding is to obtain smooth change of perceptual quality in the presence of bandwidth fluctuations in wireless channels. However, without appropriate transport mechanisms, this goal may not be accomplished. So we ask the following question: What transport mechanisms are needed to achieve this goal? Sections III and IV will answer this question and present network-aware end systems and the adaptive services for scalable video over wireless networks.

III. Network-Aware End Systems

Network-aware adaptation of end systems is an effective technique for scalable video over wireless networks [4], [13]. The use of network-aware end systems is motivated by the following facts: 1) the BER is very high when channel status is poor; and 2) packet loss is unavoidable if the available bandwidth is less than that required. If a sender attempts to transmit each layer with no awareness of channel status, all layers may get corrupted with equal probability, resulting in very poor picture quality. To address this problem, network-aware adaptation was proposed to preemptively discard enhancement layers at the sender in an intelligent manner by considering network status [4], [13].

[4]For example, spatially up-sampling with ratio 1 : 4 is to make three copies for each pixel and transmit the four pixels to the next stage.

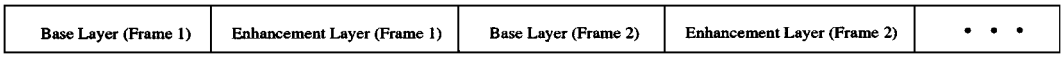

Fig. 9. Embedded bit stream.

Table 1
Taxonomy of Network Monitoring

Criteria	Type of monitoring	
Method of monitoring	Active	Passive
Monitoring frequency	On demand	Continuous
Replication of information	Centralized	Distributed

Network-aware adaptation consists of two elements: network awareness and adaptation. The process of network awareness, or *network monitoring*, is to collect the information about the current status of underlying network resources (e.g., available bandwidth and bit error conditions) [9]. Adaptation is to adapt video streams based on network status. Hence, network-aware end systems are able to monitor relevant QoS fluctuations in wireless networks and react accordingly to achieve graceful change in perceptual quality. We describe mechanisms for network monitoring and adaptation in Sections III-A and III-B, respectively.

A. Network Monitoring

Classical networking technology has effectively separated communications issues from end-user applications through the abstractions of the open systems interconnection (OSI) framework and other reference models [9]. These models have successfully defined protocols, by which developers could focus their work at a level appropriate to their development needs, for example, physical, data link, network, transport, and application layers. This architectural concept has been enormously successful and implemented almost ubiquitously. However, in a time-varying wireless environment, applications with classical networking technology (e.g., the OSI model) may experience very poor performance due to lack of awareness of network status [9]. To address this problem, network awareness or network monitoring was proposed [9]. Network-aware applications, executing in wireless environments, have the ability to react in response to changes in the status of the network, with the ultimate goal of minimizing the impact of these changes on the application's performance.

Network monitoring aims to collect information about network status. Existing network-aware systems monitor such parameters as available bandwidth and BER [47]. Here we use the term network monitor to refer to an entity in charge of the network-sensing tasks in wireless networks. The network-monitoring process can be classified according to the criteria in Table 1 [9]. The first classification is based on the method of monitoring: in passive monitoring, network monitors infer status information on existing messages, whereas in active monitoring, network measurements are done by sending additional control messages [9]. The second classification is based on whether it is performed on demand or continuously. On-demand monitoring occurs when applications ask the monitor to collect status information about a certain resource in an online fashion. In continuous monitoring, on the other hand, the monitor notifies the application when the status of a previously requested resource changes in a certain way (e.g., falls below a predefined threshold). The latter scheme requires mechanisms for applications to register their resource interests with the monitor, either synchronously or asynchronously [47]. The third classification is based on how status information is replicated. Under this classification, network monitoring can be either centralized or distributed. In the centralized case, status information from the entire network is maintained at a central host and shared by all other hosts (most commonly, this information is duplicated at several central hosts). In the distributed case, monitors collect only local network status information and obtain nonlocal status information on demand from other network monitors. The centralized scheme is not scalable, since the network monitors would maintain virtually the same status information, leading to a large amount of wasted storage. In the distributed scheme, collaboration between monitors is necessary if applications need status information about resources outside the vicinity of the local network monitor.

B. Adaptation

With the status information collected by network monitors, end systems can adapt video streams so that perceptual quality is changed gracefully during periods of QoS fluctuations and handoffs.

To illustrate the adaptation process, in Fig. 10 we present an architecture including a network-aware mobile sender, a base station, and a receiver. The architecture in Fig. 10 is applicable to both live and stored video. In Fig. 10, at the sender side, the compressed video bit stream is first filtered by the scaler, the operation of which is to select certain video layers to transmit. Then the selected video representation is passed through transport protocols. Before being transmitted to the base station, the bit stream has to be modulated by a modem (i.e., modulator/demodulator). Upon receipt of the video packets, the base station transmits them to the destination through the networks (e.g., the Internet).

In the above example, the adaptation is performed by a scaler, which can distinguish video layers and drop layers according to their significance. The dropping order is from the highest enhancement layer down to the base layer. A scaler only performs two operations: 1) scale down the received video representation, that is, drop the enhancement layer(s); and 2) transmit what is received, i.e., do not scale the received video representation.

Under our architecture, a network monitor is maintained in the base station. One function of the network monitor is to notify the sender about the available bandwidth of the wireless channel through a signaling channel [44]. Upon receiving this information, the rate control module at the sender

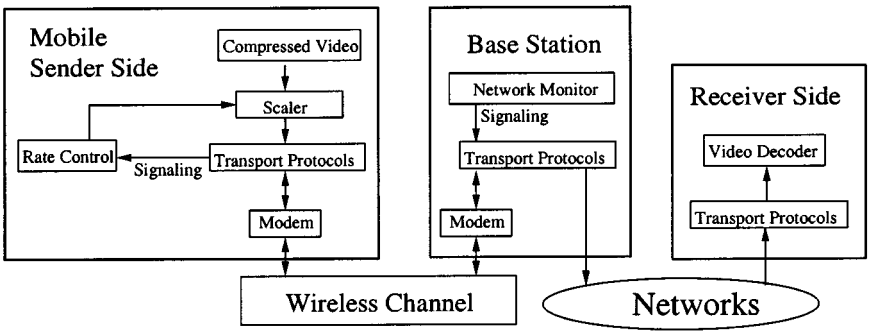

Fig. 10. Architecture for transporting scalable video from a mobile terminal to a wired terminal.

conveys the bandwidth parameter to the scaler. Then, the scaler regulates the output rate of the video stream so that the transmission rate is less than or equal to the available bandwidth.

Another scenario is that the network monitor notifies the sender about the channel quality (i.e., BER) [5]. Upon receiving this information, the rate control module at the sender commands the scaler to perform the following operations (suppose that the video is compressed into two layers): 1) if the BER is above a threshold, discard the enhancement layer so that the bandwidth allocated for the enhancement layer can be utilized by forward error correction (FEC) to protect the base layer; 2) otherwise, transmit both layers. For representations with multiple layers (more than two), an open problem is: Given a fixed bit budget, how many less important layers (higher layers) should be discarded so that more important layers (lower layers) can be protected by FEC?

With network monitoring conveying the available bandwidth or the channel quality, network-aware end systems achieve two advantages. First, by taking the available bandwidth into account, the sender can make the best use of network resources by selectively discarding enhancement layers in order to minimize the likelihood of more significant layers being corrupted, thereby increasing the perceptual quality of the video. Second, by considering the channel error status, the sender can discard the enhancement layers and then FEC can utilize the bandwidth allocated for the enhancement layer to protect the base layer, thereby maximizing the possibility of the base layer being correctly received.

Note that adaptive techniques at the physical/link layer are required to support network-aware end systems. Such adaptive techniques include software radio [19], [40], a combination of variable spreading, coding, and code aggregation in code division multiple access (CDMA) systems, adaptive coding and modulation in time division multiple access (TDMA) systems, channel quality estimation, and a measurement feedback mechanism [44]. In addition, the feedback interval is typically constrained on the order of tens to hundreds of milliseconds [44].

IV. Adaptive Services

Adaptive services are designed for scalable video transport over wireless networks. The objective of adaptive services is to achieve smooth change of perceptual quality in the presence of bandwidth fluctuations in wireless channels. As we discussed in Section II, a scalable video encoder can generate multiple layers or substreams to the network. In support of scalable video transport, the adaptive services provide scaling of the substreams based on the resource availability conditions in the wired and wireless networks. Specifically, the adaptive services include the following functions:

- Reserve a minimum bandwidth to meet the demand of the base layer. As a result, the perceptual quality can always be achieved at an acceptable level.
- Adapt the enhancement layers based on the available bandwidth and the fairness policy. In other words, it scales the video streams based on resource availability and the fairness policy.

In addition, using scaling inside the network has the following advantages.

a) Improved video quality: For example, when an upstream link with larger bandwidth feeds a downstream link with smaller bandwidth, use of a scaler at the connection point (between the upstream link and the downstream link) could help improve the video quality. This is because the scaler understands the structure of the video streams and can selectively drop substreams instead of random dropping, which could corrupt the video streams outright.

b) Low latency and low complexity: Scalable video representations make the operation at a scaler very simple, i.e., only discarding enhancement layers. Thus, the processing is fast, compared with processing on nonscalable video.

c) Lower call blocking and handoff dropping probability: The adaptability of scalable video at base stations can translate into lower call blocking and handoff dropping probability. For example, a request from a nonscalable video sender may be rejected since the required bandwidth (say, 256 kb/s) is larger than the available bandwidth (e.g., 100 kb/s). In contrast, a request from a scalable video sender can be accepted since it can transmit only the base layer (e.g., 64 kb/s) instead of both layers with larger bandwidth usage (e.g., 256 kb/s). Hence, call blocking probability is reduced. Similarly, handoff dropping probability is also reduced.

The adaptive service can be deployed in the whole network (i.e., end-to-end provisioning) or only at base stations (i.e., local provisioning). Since local provisioning of the adaptive

service is just a subset of end-to-end provisioning, we will focus on end-to-end provisioning.

The required components of the end-to-end adaptive services include [43]: 1) service contract; 2) call admission control and resource reservation; 3) mobile multicast mechanism; 4) substream scaling; 5) substream scheduling; and 6) link-layer error control.

The rest of the section is organized as follows. Section IV-A–F describes each component of the end-to-end adaptive services, respectively. Finally, we compare the adaptive services with other well-known services in Section IV-G.

A. Service Contract

The service contract between the application and the network could consist of multiple subcontracts, each of which corresponds to one or more substreams with similar QoS guarantees [43]. Each subcontract has to specify traffic characteristics and QoS requirements of the corresponding substream(s). A typical scenario is that a subcontract for the base layer specifies the reserved bandwidth, while a subcontract for the enhancement layers does not specify any QoS guarantee. For simplicity, we will use this scenario for two-layered video example in the rest of the paper.

At a video source, substreams must be generated according to subcontracts used by the application and shaped at the network access point [26]. In addition, a substream is assigned a priority according to its significance. For example, the base layer is assigned the highest priority. The priority can be used by routing, scheduling, scaling, and error control components of the adaptive network.

B. Call Admission Control and Resource Reservation

Call admission control and resource reservation are two major components in end-to-end QoS provisioning [8], [12], [53], [68].

The objective of call admission control (CAC) is to provide a QoS guarantee for individual connections while efficiently utilizing network resources. This is achieved by preventing the admission of an excessive number of calls to the network. Specifically, a CAC has to make a decision on the following question: Given a call arriving, requesting a connection with specified QoS (e.g., packet loss, delay, and bandwidth), should it be admitted? To answer this, the CAC algorithm has to check whether admitting the connection would reduce the service quality of existing connections, and whether the incoming connection's QoS requirements can be met. The admission decision is based on the availability of resources as well as the information provided by the users (e.g., traffic characteristics and QoS requirements).

If a connection request is accepted, resources need to be reserved for this connection. Under the adaptive framework containing wireless links, resource reservation is more complex than that in wired networks. Specifically, the reserved bandwidth may not be rigidly guaranteed in wireless networks. This is because the available bandwidth may be less than the reserved bandwidth due to mobility and fading. Typically, there are two parts of resource reservation. First of all, in order to maintain the specified QoS in the long time scale, the network must reserve some resource along the current path of a mobile connection. Second, in order to seamlessly achieve the QoS on the short time scale, bandwidth must be reserved on the paths from the current base stations to the neighboring base stations so that in the event of a handoff, a termination of the connection can be avoided (i.e., the reserved bandwidth can be used to transport the traffic of the connection to neighboring base stations during a handoff). The resource reservation is done during connection admission and can be renewed by renegotiation during the lifetime of the connection.

The scalable video representation (i.e., substreams) concept provides a very flexible and efficient solution to the problem of CAC and resource reservation. First, there is no need to reserve bandwidth for the complete stream since typically only the base-layer substream needs QoS guarantees. As a result, CAC is only based on the requirement of the base layer and resource is reserved only for the base-layer substream. Second, the enhancement-layer substream(s) of one connection could share the leftover bandwidth with the enhancement-layer substreams of other connections. The enhancement-layer substreams are subject to scaling under bandwidth shortage and/or severe error conditions (see Section IV-D).

For interested readers, more information about radio resource management can be found in [67].

C. Mobile Multicast Mechanism

To seamlessly guarantee QoS during a handoff, a mobile multicast mechanism has to be used. That is, while being transported along its current path, the base-layer stream is also sent through multicast to its neighboring base stations so that in the event of a handoff, the base-layer stream can still reach the receiver in a timely manner.

To support seamless QoS during a handoff, the mobile routing protocol needs to be proactive and anticipatory in order to match the delay, loss, and jitter constraints of a substream. According to the requirements of a substream, multicast paths might need to be established. The multicast paths terminate at base stations that are potential access-point candidates of a mobile terminal. The coverage of such a multicast path depends on the QoS requirements and the mobility, as well as handoff characteristics of a mobile receiver. As a mobile station hands off from a base station to another, new paths are set up and old paths are torn down [43].

D. Substream Scaling

Scaling is employed during bandwidth fluctuations and/or under poor channel conditions. As the available bandwidth on a path decreases due to mobility or fading, lower-priority substreams are dropped by the scaler(s) on the path and substreams with higher priorities are transmitted. As more bandwidth becomes available, lower-priority substreams are passed through the scaler, and the perceptual quality at the receivers increases. Fig. 10 shows an architecture for transporting scalable video from a mobile terminal to a wired terminal. Fig. 11 depicts an architecture for transporting scal-

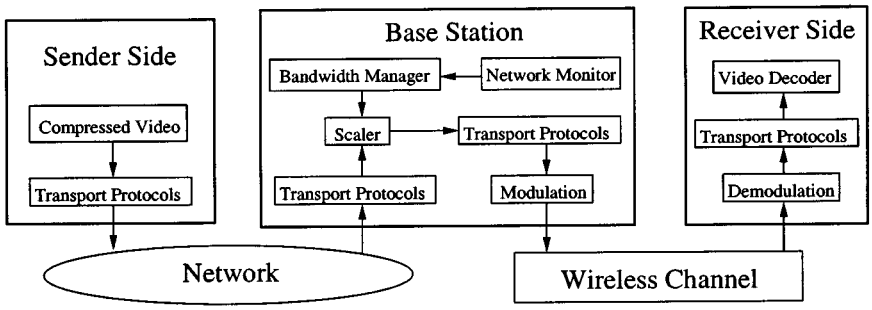

Fig. 11. Architecture for transporting scalable video from a wired terminal to a mobile terminal.

able video from a wired terminal to a mobile terminal. The case of transporting scalable video from a mobile terminal to a mobile terminal would be a combination of Figs. 10 and 11.

The scaling decision is made by a bandwidth manager, which obtains the available bandwidth from a network monitor. When there is no excess bandwidth (excluding reserved bandwidth), the bandwidth manager instructs the scaler to drop the enhancement layer. When there is excess bandwidth and the excess bandwidth cannot meet all the demands of adaptive flows, it is desirable to "fairly" allocate the excess bandwidth among contending adaptive flows. To address this issue, several solutions were proposed [7], [37]. One solution [37] is to maximize network revenue and achieve max–min fair allocation among the adaptive flows. Another solution [7] is based on a utility function, which represents the relationship between observed quality (i.e., utility) and bandwidth. Fig. 12 illustrates several kinds of utility functions, where the utility index refers to the level of quality perceived by an application. As shown in Fig. 12, a utility function captures the adaptive characteristic of an application: an application could be linearly adaptive, discretely adaptive, weakly adaptive or strongly adaptive. By using the utility function, Bianchi *et al.* [7] proposed a utility-fair bandwidth allocation scheme that supports the dynamic bandwidth needs of adaptive flows.

It can be seen that a good design of bandwidth manager should achieve fairness. That is, when there is excess bandwidth (excluding reserved bandwidth), the competing video streams can share the excess bandwidth in a fair manner. The fairness could be either a utility-based fairness [7] or a max–min fairness [37].

Note that rate-adaptive techniques [44] at the physical/link layer are required to support scaling the traffic, which will be transported over the wireless link.

E. Substream Scheduling

The substream scheduler is used in mobile terminals as well as base stations. Its function is to schedule the transmission of packets over the wireless medium according to their substream QoS specifications and priorities.

When a short fading period is observed, a mobile terminal tries to prioritize the transmission of its substreams in order to achieve a minimum QoS. Here, depending on channel conditions, a substream might be dropped for a period of time in order to accommodate higher-priority substreams. To determine the transmission time of any packet in a specific substream (or its position in the transmission queue), the scheduler takes two factors into account: 1) the relative importance of the substream compared to other substreams; and 2) wireless channel conditions. It is important to note that the scheduler reacts to the fluctuations in the wireless channel due to error and fading conditions, and requires feedback from the wireless transmitter and receiver to infer the condition of the wireless channel and also to predict its near-term condition [43].

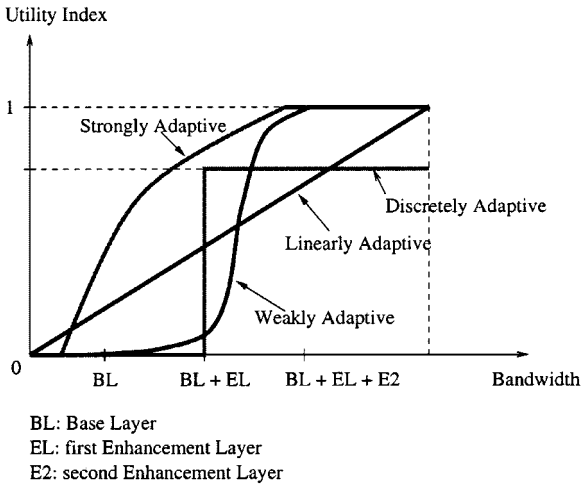

Fig. 12. Utility functions.

To achieve both QoS (e.g., bounded delay and reserved bandwidth) and fairness, algorithms like packet fair queueing may be employed [6]. While existing packet fair queueing algorithms provide both bounded delay and fairness in wired networks, they cannot be applied directly to wireless networks. The key difficulty is that in wireless networks, sessions can experience location-dependent channel errors. This may lead to situations where a session receives significantly less service time than it is supposed to receive, while another receives more. This results in large discrepancies between the sessions' virtual times,[5] making it difficult to provide both delay guarantees and fairness simultaneously.

To apply packet fair queueing algorithms, Ng *et al.* [45] identified a set of properties, called channel-condition independent fair (CIF), that a packet fair queueing

[5]Virtual times are used in packet fair queueing algorithms to determine the transmission order.

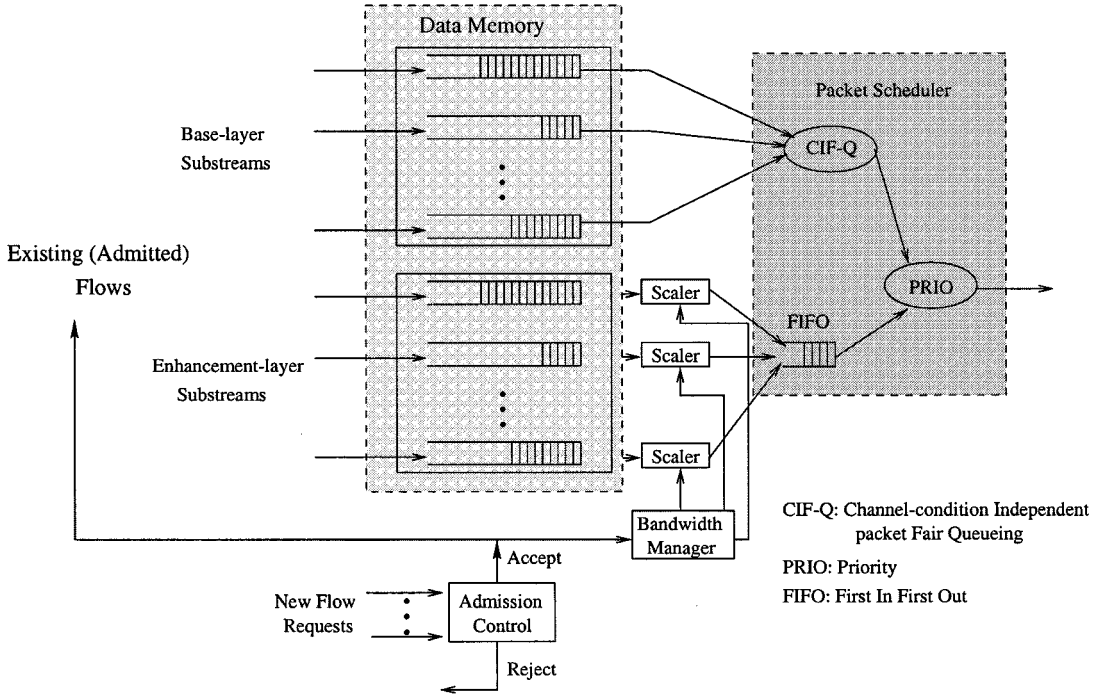

Fig. 13. Architecture for substream scheduling at a base station.

algorithm should have in a wireless environment. The CIF properties include: 1) delay and throughput guarantees for error-free sessions; 2) long-term fairness for error sessions; 3) short-term fairness for error-free sessions; and 4) graceful degradation for sessions that have received excess service time. Further, they presented a methodology for adapting packet fair queueing algorithms for wireless networks and applied the methodology to derive an algorithm based on the start-time fair queueing [16], called channel-condition independent packet fair queueing (CIF-Q), that achieves all the above properties [45].

As an example, we consider two-layer video. Suppose that a subcontract for the base layer specifies the reserved bandwidth while a subcontract for the enhancement layer does not specify any QoS guarantee, which is a typical case. We design an architecture for substream scheduling as shown in Fig. 13. Under this architecture, we partition the buffer pool (i.e., data memory in Fig. 13) into two parts: one for base-layer substreams, and one for enhancement-layer substreams. Within the same buffer partition for the base or the enhancement layer, we employ per-flow queueing for each substream. Furthermore, substreams within the same buffer partition share the buffer pool of that partition while there is no buffer sharing across partitions. We believe this approach offers an excellent balance between traffic isolation and buffer sharing.

Under the above buffering architecture, we design our per-flow-based traffic management algorithms with the aim of achieving QoS requirements and fairness. The first part of our architecture is CAC and bandwidth allocation. Video connections are admitted by CAC based only on their base-layer QoS requirements. For those admitted base-layer substreams, bandwidth reservations are made accordingly. For admitted enhancement-layer substreams, their bandwidths are dynamically allocated by a bandwidth manager (see Section IV-D). The scaled enhancement-layer substreams enter a shared buffer and are scheduled by a first-in–first-out (FIFO) scheduler. The second part of our architecture is packet scheduling. In Fig. 13, we use a hierarchical packet-scheduling architecture where a priority link scheduler is shared among a CIF-Q scheduler for base-layer substreams, and an FIFO scheduler for enhancement-layer substreams. Service priority is first given to the CIF-Q scheduler and then to the FIFO scheduler.

F. Link-Layer Error Control

In wireless environments, bit errors are unavoidable, which consequently degrades the video quality. To compensate for these errors, link-layer error control is employed. Basically, there are two kinds of link-layer error control mechanisms, namely, FEC and automatic repeat request (ARQ).

FEC is used to add redundant information, like a kit of spare parts, so that the original message can be reconstructed in the event of bit errors. The advantages of FEC are: 1) the throughput can be kept constant; and 2) delay can be bounded. However, the redundancy ratio (the ratio of the redundant bit number to the total bit number) should be made large enough to guarantee recovery of corrupted bits under the worst channel conditions. In addition, FEC is not adaptive to varying wireless channel conditions and it works best only when the BER is stable. Specifically, if the number of bit errors exceeds the FEC code's recovery capability, the FEC code cannot recover any portion of the original data. In other words, FEC is useless when the short-term BER exceeds the recovery capability of the FEC code. On the other hand,

when the wireless channel is in good state (i.e., the BER is very small), using FEC with large redundancy ratio will cause unnecessary overhead and waste bandwidth.

Different from FEC, ARQ is adaptive to varying wireless channel conditions. That is, with ARQ, the receiver notifies the source only when packets are corrupted and need to be retransmitted. In other words, when the channel is in good state, no retransmission is required and no bandwidth is wasted. However, adaptiveness and efficiency of ARQ come with the cost of unbounded delay, e.g., in the worst case, a packet may be retransmitted in unlimited times to recover bit errors.

To deal with the problems associated with FEC and ARQ, truncated type-II hybrid ARQ schemes have been proposed [35], [69]. Different from conventional type-II hybrid ARQ [17], [24], [34], [63], the truncated type-II hybrid ARQ has a constraint on the maximum number of retransmissions for a packet. Consequently, delay can be bounded. The truncated type-II hybrid ARQ combines the good features of FEC and ARQ: bounded delay and adaptiveness. However, the maximum number of retransmissions N_r is assumed to be fixed and known *a priori* [35], [69], which may not reflect the time-varying nature of delay. If N_r is set too large, retransmitted packets may arrive too late for play-out and thereby be discarded, resulting in wastage of bandwidth; if N_r is set too small, the perceptual quality will be reduced due to unrecoverable errors that could have been corrected with more retransmissions. We address this problem by introducing delay-constrained hybrid ARQ [65]. Under this scheme, the receiver makes retransmission requests in such a way: when errors are detected in the received packet, the receiver decides whether to send a retransmission request according to the delay bound of the packet. The following pseudocode describes the delay-constrained hybrid ARQ:

```
When the receiver detects the loss of packet N:
  if (T_c + RTT + D_s < T_d(N))
    send the request for retransmission of
    package N to the sender;
```

where T_c is the current time, RTT is an estimated round trip time, D_s is a slack term, and $T_d(N)$ is the time when packet N is scheduled for display. The slack term D_s could include tolerance of error in estimating RTT, the sender's response time to a request, and/or the receiver's processing delay (e.g., decoding). The scheme is aimed at minimizing the request of retransmissions that will not arrive in a timely manner for display. It is clear that if $T_c + RTT + D_s < T_d(N)$, the retransmitted packet is expected to arrive in a timely manner for display. The timing diagram for receiver-based control is shown in Fig. 14, where D_s is the receiver's decoding delay.

The delay-constrained hybrid ARQ is capable of achieving bounded delay, adaptiveness, and efficiency [65]. It is also suitable for scalable video over wireless [70]. In addition, unequal error protection [18] naturally fits the hierarchical structure of scalable video. With unequal error protection, the base layer is given more protection than the enhancement layers. This form of unequal error protection achieves better quality than protecting all the substreams equally [70].

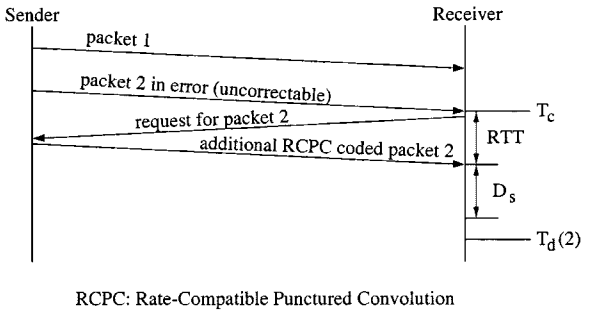

Fig. 14. Timing diagram for delay-constrained retransmission.

G. Service Comparison

To give the reader a clear picture of the adaptive services, we compare the adaptive services with other well-known services, i.e., the guaranteed service [53] and the best-effort service.

The guaranteed service assures that packets will arrive within the required delivery time, and will not get lost, provided that the flow's traffic conforms to its specified traffic parameters [53]. This service is intended for applications that require a stringent delay, e.g., distant nuclear plant control, distant weapon control, and distant surgery control, all of which are mission critical.

The best-effort service class offers the same type of service as that provided by the current Internet. Under the best-effort service, the network makes every effort to deliver data packets but makes no guarantees. This works well for non-real-time applications which can use reliable transport protocol [e.g., transmission control protocol (TCP)] to make sure that all packets are delivered correctly. These applications include file transfer protocol (FTP), e-mail, and web browsing, all of which can work without stringent delay requirements.

A comparison of the three service classes is summarized in Table 2. The guaranteed service and the adaptive services need to set up a path for an admitted connection. In contrast, the best-effort service does not require path setup. Regarding target applications, both the guaranteed service and the adaptive services can support constant bit rate (CBR) and variable bit rate (VBR) applications.

In selecting a specific type of service for video transport, a tradeoff must be made between two conflicting requirements: QoS guarantees (reflecting cost) and network utilization. The cost of the guaranteed service is high for nontime-critical video applications. As a result, the guaranteed service is usually not chosen for video transport. The current best-effort service is not acceptable in many cases due to its poor QoS support. The adaptive services provide users with a viable option. They achieve acceptable perceptual quality at a medium cost. Specifically, the adaptive service for the base layer provides basic perceptual quality at the cost of resource reservation; at almost no cost, the adaptive service for the enhancement layer takes advantage of statistical multiplexing gain to achieve better perceptual quality if possible. Therefore, the adaptive services can achieve better quality than the best-effort service while they cost less than the guaranteed service.

Table 2
Comparison of Different Network Services

Services		Path set-up	Traffic charaterization	End-to-end QoS guarantee	Network feedback	Resource reservation	QoS	Target applications
Guaranteed service		Yes	Yes	Yes	No	Yes	Bounded delay, zero loss	Non-adaptive CBR/VBR
Adaptive services	Base layer	Yes	Yes	If needed	No	Yes	Small delay, low loss	Adaptive CBR/VBR
	Enhancement layer	Yes	No	No	If needed	No	Better than best-effort	
Best-effort service		No	No	No	No	No	None	Non-real-time data

V. Summary

Recent years have witnessed a rapid growth of research and development to provide mobile users with video communication through wireless media. In this paper, we examined the challenges in QoS provisioning for wireless video transport. To address the challenges, three techniques (i.e., scalable video coding, network-aware adaptation of end systems, and adaptive QoS support from network) have been studied in great depth individually. This paper aims to unify the three techniques simultaneously and presents an adaptive framework, which specifically addresses scalable video transport over wireless networks. The adaptive framework consists of: 1) scalable video representations; 2) network-aware end systems; and 3) adaptive services. Under this framework, mobile terminals and network elements can adapt the video streams according to the channel conditions and transport the video streams to receivers with a smooth change of perceptual quality. The advantage of deploying such an adaptive framework is that it can provide suitable QoS for video over wireless while achieving fairness in sharing resources.

As this paper only sketches a high-level framework, for the purpose of implementation, some details remain to be addressed. We list some of them as follows.

- We have to consider the particular multiple access control protocol (e.g., CDMA or TDMA), modulation, channel allocation, and mobile terminals being used [1], [25], [28], [41].
- We also need to take into account how to adapt the rate at the link and physical layers [44]. In addition, channel quality feedback mechanisms have been defined in link/physical layer standards to carry out rate adaptation. For the emerging broadband wireless networks, we might also need to design new rate adaptation techniques.
- A software platform like Odyssey [48] may be necessary to support adaptive applications. Such a software platform can provide mechanisms enabling adaptation, leaving applications free to set adaptive policies.
- A scalable video coding scheme needs to be carefully designed so that it is robust to multiple time-scale QoS fluctuations in the wireless/wired network [11]. A scalable video coding scheme should achieve high efficiency with less complexity and should try to optimally decompose video into multiple substreams without loss of compression efficiency.
- It is necessary to characterize scalable video streams (i.e., traffic modeling) and use the characterization in the design of efficient CAC and resource reservation schemes [23].

Note that the above details can be implemented transparently to the adaptive framework (e.g., in a programmable way as that in Mobiware [3]).

There are many promising and interesting research directions under the adaptive framework. One topic is the design of mechanisms to achieve seamless QoS for the base layer of scalable video. One such mechanism is a lossless handover method for mobile asynchronous transfer mode (ATM) communication networks [46], which helps to prevent cell loss and suppress cell delay variation. More investigations need to be done for handoffs between networks using different network technologies (e.g., from wireless LAN to wireless WAN), and between network domains [57]. Another direction is the seamless integration of wireless networks and wired networks. Since wireless segments and wired segments have different QoS provision mechanisms [10], [51], for the adaptive services, how to provide seamless integration of wireless networks and wired networks needs further study.

As a final note, we stress that each service (e.g., the adaptive services, the best-effort service, or the guaranteed service) has a tradeoff between cost/complexity and performance. The adaptive framework is targeted at quality video transport over near-term QoS-enabled broadband wireless networks. In addition, the adaptive services could be provisioned at a single base station or provisioned in the entire network. In the real interconnected wireless networks, even though we cannot require each router to provide the adaptive services, a partial deployment of the adaptive services can still have clear benefits. Furthermore, it is entirely feasible to fully deploy the adaptive services within a single administrative domain (e.g., intranet) and achieve high statistical multiplexing gain and acceptable QoS.

References

[1] I. F. Akyildiz, J. McNair, L. C. Martorell, R. Puigjaner, and Y. Yesha, "Medium access control protocols for multimedia traffic in wireless networks," *IEEE Network*, vol. 13, pp. 39–47, July 1999.

[2] A. Alwan, R. Bagrodia, N. Bambos, M. Gerla, L. Kleinrock, J. Short, and J. Villasenor, "Adaptive mobile multimedia networks," *IEEE Pers. Commun.*, vol. 3, pp. 34–51, Apr. 1996.

[3] O. Angin, A. T. Campbell, M. E. Kounavis, and R. Liao, "The Mobiware toolkit: Programmable support for adaptive mobile networking," *IEEE Pers. Commun.*, vol. 5, pp. 32–43, Aug. 1998.

[4] A. Balachandran, A. T. Campbell, and M. E. Kounavis, "Active filters: Delivering scalable media to mobile devices," presented at the 7th Int. Workshop Network and Operating System Support for Digital Audio and Video (NOSSDAV'97), St. Louis, MO, May 1997.

[5] K. Balachandran, S. Kadaba, and S. Nanda, "Rate adaptation over mobile radio channels using channel quality information," presented at the IEEE GLOBECOM'98, Sydney, Australia, Nov 8–12, 1998.
[6] V. Bharghavan, S. Lu, and T. Nandagopal, "Fair queuing in wireless networks: Issues and approaches," *IEEE Pers. Commun.*, vol. 6, pp. 44–53, Feb. 1999.
[7] G. Bianchi, A. T. Campbell, and R. Liao, "On utility-fair adaptive services in wireless networks," presented at the 6th Int. Workshop Quality of Service (IWQOS'98), Napa Valley, CA, May 1998.
[8] R. Braden, D. Clark, and S. Shenker, "Integrated services in the internet architecture: An overview," Internet Engineering Task Force, RFC 1633, July 1994.
[9] W. Caripe, G. Cybenko, K. Moizumi, and R. Gray, "Network awareness and mobile agent systems," *IEEE Commun. Mag.*, vol. 36, pp. 44–49, July 1998.
[10] M. C. Chan and T. Y. C. Woo, "Next-generation wireless data services: Architecture and experience," *IEEE Pers. Commun.*, vol. 6, pp. 20–33, Feb. 1999.
[11] *Adaptive layered video coding for multi-time scale bandwidth fluctuations*, submitted for publication.
[12] D. Clark, S. Shenker, and L. Zhang, "Supporting real-time applications in an integrated services packet network: Architecture and mechanisms," presented at the ACM SIGCOMM'92, Baltimore, MD, Aug. 1992.
[13] N. Davies, J. Finney, A. Friday, and A. Scott, "Supporting adaptive video applications in mobile environments," *IEEE Commun. Mag.*, vol. 36, pp. 138–143, June 1998.
[14] T. Ebrahimi and M. Kunt, "Visual data compression for multimedia applications," *Proc. IEEE*, vol. 86, pp. 1109–1125, June 1998.
[15] B. Girod, U. Horn, and B. Belzer, "Scalable video coding with multiscale motion compensation and unequal error protection," in *Proc. Symp. Multimedia Communications and Video Coding*. New York: Plenum, Oct. 1995, pp. 475–482.
[16] P. Goyal, H. M. Vin, and H. Chen, "Start-time fair queuing: A scheduling algorithm for integrated service access," presented at the ACM SIGCOMM'96, Stanford, CA, Aug. 1996.
[17] J. Hagenauer, "Rate-compatible punctured convolutional codes (RCPC codes) and their applications," *IEEE Trans. Commun.*, vol. 36, pp. 389–400, Apr. 1988.
[18] J. Hagenauer and T. Stockhammer, "Channel coding and transmission aspects for wireless multimedia," *Proc. IEEE*, vol. 87, pp. 1764–1777, Oct. 1999.
[19] T. Hentschel, M. Henker, and G. Fettweis, "The digital front-end of software radio terminals," *IEEE Pers. Commun.*, vol. 6, pp. 40–46, Aug. 1999.
[20] Y. T. Hou, S. S. Panwar, Z.-L. Zhang, H. Tzeng, and Y.-Q. Zhang, "On network bandwidth sharing for transporting rate-adaptive packet video using feedback," *Int. J. Commun. Syst.*, vol. 13, no. 2, pp. 117–143, Mar. 2000.
[21] A. Iera, A. Molinaro, and S. Marano, "Wireless broadband applications: The teleservice model and adaptive QoS provisioning," *IEEE Commun. Mag.*, vol. 37, pp. 71–75, Oct. 1999.
[22] K. Illgner and F. Mueller, "Spatially scalable video compression employing resolution pyramids," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1688–1703, Dec. 1997.
[23] B. Jabbari, "Teletraffic aspects of evolving and next-generation wireless communication networks," *IEEE Pers. Commun.*, vol. 3, pp. 4–9, Dec. 1996.
[24] S. Kallel and D. Haccoun, "Generalized type-II hybrid ARQ scheme using punctured convolutional coding," *IEEE Trans. Commun.*, vol. 38, pp. 1938–1946, Nov. 1990.
[25] Y. C. Kim, D. E. Lee, B. J. Lee, Y. S. Kim, and B. Mukherjee, "Dynamic channel reservation based on mobility in wireless ATM networks," *IEEE Commun. Mag.*, vol. 37, pp. 47–51, Nov. 1999.
[26] O. Lataoui, T. Rachidi, L. G. Samuel, S. Gruhl, and R.-H. Yan, "A QoS management architecture for packet switched 3rd generation mobile systems," presented at the Networld+Interop 2000 Engineers Conf., Las Vegas, NV, May 10–11, 2000.
[27] K. Lee, "Adaptive network support for mobile multimedia," presented at the ACM Mobicom'95, Berkeley, CA, Nov. 13–15, 1995.
[28] P. Lettieri and M. B. Srivastava, "Advances in wireless terminals," *IEEE Pers. Commun.*, vol. 6, pp. 6–19, Feb. 1999.
[29] J. Li, "Visual progressive coding," presented at the Visual Communications and Image Processing (VCIP'99), Jan. 1999.
[30] S. Li, F. Wu, and Y.-Q. Zhang, "Study of a new approach to improve FGS video coding efficiency," ISO/IEC JTC1/SC29/WG11, MPEG99/M5583, Dec. 1999.
[31] W. Li, "Bit-plane coding of DCT coefficients for fine granularity scalability," ISO/IEC JTC1/SC29/WG11, MPEG98/M3989, Oct. 1998.
[32] ——, "Syntax of fine granularity scalability for video coding," ISO/IEC JTC1/SC29/WG11, MPEG99/M4792, July 1999.
[33] X. Li, S. Paul, and M. H. Ammar, "Layered video multicast with retransmissions (LVMR): Evaluation of hierarchical rate control," presented at the IEEE INFOCOM'98, San Francisco, CA, Mar. 1998.
[34] S. Lin and D. Costello, *Error Control Coding: Fundamentals and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
[35] H. Liu and M. El Zarki, "Performance of H.263 video transmission over wireless channels using hybrid ARQ," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1775–1786, Dec. 1997.
[36] Y.-J. Liu and Y.-Q. Zhang, "Wavelet-coded image transmission over land mobile radio channels," presented at the IEEE GLOBECOM'92, Orlando, FL, Dec. 1992.
[37] S. Lu, K.-W. Lee, and V. Bharghavan, "Adaptive service in mobile computing environments," presented at the 5th Int. Workshop Quality of Service (IWQOS'97), Columbia University, New York, May 21–23, 1997.
[38] S. McCanne, V. Jacobson, and M. Vetterli, "Receiver-driven layered multicast," in *Proc. ACM SIGCOMM'96*, Aug. 1996, pp. 117–130.
[39] N. Miller and P. Steenkiste, "Collecting network status information for network-aware applications," presented at the IEEE INFOCOM'2000, Tel Aviv, Israel, Mar. 2000.
[40] J. Mitola, "The software radio architecture," *IEEE Commun. Mag.*, vol. 33, pp. 26–38, May 1995.
[41] N. Morinaga, M. Nakagawa, and R. Kohno, "New concepts and technologies for achieving highly reliable and high-capacity multimedia wireless communications systems," *IEEE Commun. Mag.*, vol. 35, pp. 90–100, Jan. 1997.
[42] J. Moura, R. S. Jasinschi, H. Shiojiri, and J.-C. Lin, "Video over wireless," *IEEE Pers. Commun.*, vol. 3, pp. 44–54, Feb. 1996.
[43] M. Naghshineh and M. Willebeek-LeMair, "End-to-end QoS provisioning in multimedia wireless/mobile networks using an adaptive framework," *IEEE Commun. Mag.*, vol. 35, pp. 72–81, Nov. 1997.
[44] S. Nanda, K. Balachandran, and S. Kumar, "Adaptation techniques in wireless packet data services," *IEEE Commun. Mag.*, vol. 38, pp. 54–64, Jan. 2000.
[45] T. S. E. Ng, I. Stoica, and H. Zhang, "Packet fair queueing algorithms for wireless networks with location-dependent errors," presented at the Proc. IEEE INFOCOM'98, San Francisco, CA, Mar. 1998, pp. 1103–1111.
[46] M. Nishio, N. Shinagawa, and T. Kobayashi, "A lossless handover method for video transmission in mobile ATM networks and its subjective quality assessment," *IEEE Commun. Mag.*, vol. 37, pp. 38–44, Nov. 1999.
[47] B. Noble, M. Satyanarayanan, D. Narayanan, J. E. Tilton, J. Flinn, and K. Walker, "Agile application-aware adaption for mobility," presented at the 16th ACM Symp. Operating System Principles, St. Malo, France, Oct. 1997.
[48] B. Noble, "System support for mobile, adaptive applications," *IEEE Pers. Commun.*, vol. 7, pp. 44–49, Feb. 2000.
[49] H. Radha and Y. Chen, "Fine-granular-scalable video for packet networks," presented at the Packet Video'99, Columbia University, New York, Apr. 1999.
[50] M. Ranganathan, A. Acharya, S. Sharma, and J. Saltz, "Network-aware mobile programs," presented at the USENIX Annual Technical Conf., Anaheim, CA, Jan. 1997.
[51] D. Reininger, D. Raychaudhuri, and M. Ott, "A dynamic quality of service framework for video in broadband networks," *IEEE Network*, vol. 12, pp. 22–34, Nov. 1998.
[52] D. Reininger, R. Izmailov, B. Rajagopalan, M. Ott, and D. Raychaudhuri, "Soft QoS control in the WATMnet broadband wireless system," *IEEE Pers. Commun.*, vol. 6, pp. 34–43, Feb. 1999.
[53] S. Shenker, C. Partridge, and R. Guerin, "Specification of guaranteed quality of service," Internet Engineering Task Force, RFC 2212, Sept. 1997.
[54] B. Sklar, "Rayleigh fading channels in mobile digital communication systems—Part I: Characterization," *IEEE Commun. Mag.*, vol. 35, pp. 90–100, July 1997.
[55] I. Sodagar, H.-J. Lee, P. Hatrack, and Y.-Q. Zhang, "Scalable wavelet coding for synthetic/natural hybrid images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 244–254, Mar. 1999.
[56] D. Taubman and A. Zakhor, "A common framework for rate and distortion based scaling of highly scalable compressed video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 329–354, Aug. 1996.

[57] L. Taylor, R. Titmuss, and C. Lebre, "The challenges of seamless handover in future mobile multimedia networks," *IEEE Pers. Commun.*, vol. 6, pp. 32–37, Apr. 1999.
[58] J. Y. Tham, S. Ranganath, and A. A. Kassim, "Highly scalable wavelet-based video codec for very low bit-rate environment," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 12–27, Jan. 1998.
[59] B. Vandalore, R. Jain, S. Fahmy, and S. Dixit, "AQuaFWiN: Adaptive QoS framework for multimedia in wireless networks and its comparison with other QoS frameworks," presented at the IEEE Conf. Local Computer Networks (LCN'99), Boston, MA, Oct. 17–20, 1999.
[60] M. van der Schaar, H. Radha, and Y. Chen, "An all FGS solution for hybrid temporal-SNR scalability," ISO/IEC JTC1/SC29/WG11, MPEG99/M5552, Dec. 1999.
[61] A. Vetro, H. Sun, and Y. Wang, "Object-based transcoding for scalable quality of service," presented at the IEEE Int. Symp. Circuits and Systems (ISCAS'2000), Geneva, Switzerland, May 28–31, 2000.
[62] J. Villasenor, Y.-Q. Zhang, and J. Wen, "Robust video coding algorithms and systems," *Proc. IEEE*, vol. 87, pp. 1724–1733, Oct. 1999.
[63] Y. Wang and S. Lin, "A modified selective-repeat type-II hybrid ARQ system and its performance analysis," *IEEE Trans. Commun.*, vol. 31, pp. 593–608, May 1983.
[64] G. Welling and B. R. Badrinath, "An architecture for exporting environment awareness to mobile computing applications," *IEEE Trans. Software Eng.*, vol. 24, pp. 391–400, May 1998.
[65] D. Wu, Y. T. Hou, Y.-Q. Zhang, W. Zhu, and H. J. Chao, "Adaptive QoS control for MPEG-4 video communication over wireless channels," presented at the IEEE Int. Symp. Circuits and Systems (ISCAS'2000), Geneva, Switzerland, May 28–31, 2000.
[66] D. Wu, Y. T. Hou, and Y.-Q. Zhang, "Transporting real-time video over the internet: Challenges and approaches," *Proc. IEEE*, vol. 88, Dec. 2000.
[67] J. Zander, "Radio resource management in future wireless networks: Requirements and limitations," *IEEE Commun. Mag.*, vol. 35, pp. 30–36, Aug. 1997.
[68] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala, "RSVP: A new resource Reservation protocol," *IEEE Network*, vol. 7, pp. 8–18, Sept. 1993.
[69] Q. Zhang and S. A. Kassam, "Hybrid ARQ with selective combining for fading channels," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 867–880, May 1999.
[70] Q. Zhang, W. Zhu, G. Wang, and Y.-Q. Zhang, "Resource allocation with adaptive QoS for multimedia transmission over W-CDMA channels," presented at the IEEE Wireless Communications and Networking Conf. (WCNC'2000), Chicago, IL, Sept. 23–28, 2000.
[71] Y.-Q. Zhang, Y.-J. Liu, and R. Pickholtz, "Layered image transmission over cellular radio channels," *IEEE Trans. Veh. Technol.*, vol. 43, pp. 786–794, Aug. 1994.

Dapeng Wu (Student Member, IEEE) received the B.E. degree from Huazhong University of Science and Technology, Wuhan, China, and the M.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 1990 and 1997, respectively, both in electrical engineering. From July 1997 to December 1999, he conducted graduate research at Polytechnic University, Brooklyn, NY. Since January 2000, he has been working toward the Ph.D. degree in electrical and computer engineering at Carnegie Mellon University, Pittsburgh, PA.

During the summers of 1998, 1999, and 2000, he conducted research at Fujitsu Laboratories of America, Sunnyvale, CA, on architectures and traffic management algorithms on the Internet and wireless networks for multimedia applications. His current interests are in the areas of rate control and error control for video communications over the Internet and wireless networks, and next-generation Internet architecture, protocols, and implementations for integrated and differentiated services.

Mr. Wu is a Student Member of the Association for Computing Machinery.

Yiwei Thomas Hou (Member, IEEE) received the B.E. degree *(summa cum laude)* from the City College of New York in 1991, the M.S. degree from Columbia University, New York, in 1993, and the Ph.D. degree from Polytechnic University, Brooklyn, NY, in 1997, all in electrical engineering. He was awarded a National Science Foundation Graduate Research Traineeship for pursuing the Ph.D. degree in high-speed networking, and was recipient of the Alexander Hessel award for outstanding Ph.D. dissertation (1997–1998 academic year) from Polytechnic University.

While a graduate student, he worked at AT&T Bell Labs, Murray Hill, NJ, during the summers of 1994 and 1995, on internetworking of IP and ATM networks; he conducted research at Bell Labs, Lucent Technologies, Holmdel, NJ, during the summer of 1996, on fundamental problems on network traffic management. Since September 1997, he has been a Research Scientist at Fujitsu Laboratories of America, Sunnyvale, CA. He has received several awards from Fujitsu Laboratories of America for intellectual property contributions. His current research interests are in the areas of scalable architecture, protocols, and implementations for differentiated services Internet; terabit switching; quality of service (QoS) support for multimedia over wired and wireless Internet; and emerging service overlay infrastructure. He has authored or coauthored over 50 refereed papers in the above areas, including over 20 papers in major international archival journals.

Dr. Hou is a Member of the Association for Computing Machinery, Sigma Xi, and the New York Academy of Sciences.

Ya-Qin Zhang (Fellow, IEEE) is currently the Managing Director of Microsoft Research China. He was previously the Director of the Multimedia Technology Laboratory at the Sarnoff Corporation in Princeton, NJ (formerly the David Sarnoff Research Center and RCA Laboratories). His laboratory is a world leader in MPEG2/DTV, MPEG4/VLBR, and multimedia information technologies. He was with GTE Laboratories Inc. in Waltham, MA, and Contel Technology Center in Chantilly, VA, from 1989 to 1994. He has authored or coauthored over 200 refereed papers and 40 U.S. patents granted or pending in digital video, Internet multimedia, and wireless and satellite communications. Many of the technologies that he and his team developed have become the basis for start-up ventures, commercial products, and international standards.

Dr. Zhang was Editor-in-Chief for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from July 1997 to July 1999. He was a Guest Editor for the special issue on "Advances in Image and Video Compression" for the PROCEEDINGS OF THE IEEE (February 1995). He serves on the editorial boards of seven other professional journals and over a dozen conference committees. He has been an active contributor to the ISO/MPEG and ITU standardization efforts in digital video and multimedia. He has received numerous awards, including several industry technical achievement awards and IEEE awards. He was named the "Research Engineer of the Year" in 1998 by the Central Jersey Engineering Council for his "leadership and invention in communications technology, which has enabled dramatic advances in digital video compression and manipulation for broadcast and interactive television and networking applications."