Scalable Video Coding and Transport over Broadband Wireless Networks

Dapeng Wu* Y. Thomas Hou† Ya-Qin Zhang‡

Abstract

Video application is becoming an important application under the next-generation wireless networks. However, supporting video communication over wireless networks poses many challenges due to fluctuations of wireless channel conditions. Scalable video coding and adaptive services have been shown to be capable of coping effectively under time-varying wireless environment. We present an adaptive framework to support quality video communication over wireless networks. The adaptive framework include: (1) scalable video representations, (2) network-aware video application, and (3) adaptive service. Under this framework, as wireless channel conditions change, the mobile terminal and network elements can scale the video streams and transport the scaled video streams to receivers with acceptable perceptual quality. The key advantages the adaptive framework are: (1) perceptual quality is degraded gracefully under severe channel conditions; (2) network resources are efficiently utilized; and (3) the resources are shared in a fair manner.

Key Words: Wireless, quality-of-service, adaptive framework, scalable video coding, network-aware application, application-aware network.

*Carnegie Mellon University, Dept. of Electrical & Computer Engineering, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA. Tel. (412) 268-7107, Fax (412) 268-3890, Email: dpwu@cs.cmu.edu.

†Fujitsu Laboratories of America, 595 Lawrence Expressway, Sunnyvale, CA 94086, USA. Tel. (408) 530-4529, Fax (408) 530-4515, Email: thou@fla.fujitsu.com.

‡Please direct all correspondence to Dr. Y.-Q. Zhang, Microsoft Research, China, 5F, Beijing Sigma Center, No. 49, Zhichun Road Haidian District, Beijing 100080, China. Tel. (86-10) 6261-7711 Ext. 5790, Fax (86-10) 8809-7305, Email: yzhang@microsoft.com.

1 Introduction

The proliferation of multimedia on the World Wide Web and the emergence of broadband wireless networks have brought great interest in wireless video communications. However, delivering quality video over wireless networks poses many challenges. This is primarily because of the following problems:

Bandwidth fluctuations: First, the throughput of a wireless channel may be reduced due to multipath fading, co-channel interference, and noise disturbances. Second, the capacity of a wireless channel may fluctuate with the changing distance between the base station and the mobile host. Third, when a mobile terminal moves between different networks (e.g., from wireless local area network to wireless wide area network), the available bandwidth may vary drastically (e.g., from a few megabits per second to a few kilobits per second). Finally, when a handoff happens, a base station may not have enough unused radio resource to meet the demand of a newly joined mobile host. Bandwidth fluctuations poses a challenging problem to meet the bandwidth requirement of video transmission.

High bit error rate: Compared with the wired links, wireless channels are typically much more noisy and have both small-scale (multipath) and large-scale (shadowing) fades [38], making the bit error rate (BER) very high. The resulting bit errors can have devastating effect on video presentation quality [44]. Therefore, there is a need for robust transmission of video over wireless channels.

Heterogeneity: In multicast scenario, receivers may be different in terms of latency requirements, visual quality requirements, processing capabilities, power limitations (wireless vs. wired) and bandwidth limitations. The heterogeneous requirements of receivers make it difficult to design an efficient multicast mechanism.

It has been shown that scalable video is capable of coping with the variability of bandwidth gracefully [3, 22, 27]. In contrast, non-scalable video is more susceptible to bandwidth fluctuations since it cannot adapt its video representation to bandwidth variations [27]. Thus, scalable video is more suitable than non-scalable video for use in a wireless environment to cope with the fluctuation of wireless channels. Furthermore, scalable video representation is a good solution to heterogeneity problem in multicast case [22, 27].

Recently, application-aware adaptive services have been demonstrated to be able to effectively mitigate fluctuations of resource availability in wireless networks [3]. Scalable video representation naturally fit unequal error protection, which can effectively combat bit errors induced by the wireless medium [53]. This motivates us to present an adaptive framework as a solution to support quality video communication over wireless networks.

For transporting video over wireless, there have been many proposals of adaptive approaches and services in the literature, which include an "adaptive reserved service" framework [20], an adaptive service based on QoS bounds and revenue [26], an adaptive framework targeted at end-to-end QoS provisioning [30], a utility-fair adaptive service [6], a framework for soft QoS control

[36], a teleservice model based on an adaptive QoS paradigm [14], an adaptive QoS management architecture [19], and an adaptive framework for scalable video over wireless [50].

In this paper we present an adaptive framework for future QoS-enabled broadband wireless networks. We envision that the future adaptive framework consists of three basic components: (1) scalable video representations, each of which has its own specified QoS requirement, (2) network-aware video application, and (3) adaptive service, which makes network elements support the QoS requirements of scalable video representations. Under such framework, as wireless channel conditions change, the mobile terminal and network elements can scale the video streams and transport the scaled video streams to receivers with acceptable perceptual quality.

The key features of our adaptive framework are:

1. Graceful quality degradation

   Different from non-scalable video, scalable video can adapt its video representation to bandwidth variations and the network can drop packets with awareness of the video representations. As a result, perceptual quality is gracefully degraded under severe channel conditions.

2. Efficiency

   When there is excess bandwidth (excluding reserved bandwidth), the excess bandwidth will be efficiently used in a way that maximizes the perceptual quality or revenue.

   If the objective is to maximize the perceptual quality, efficiency can be achieved by exploiting the interplay between video compression and transport. In other words, optimal bandwidth allocation is based on the relationship between the perceptual quality and the available bandwidth [6].

3. Fairness

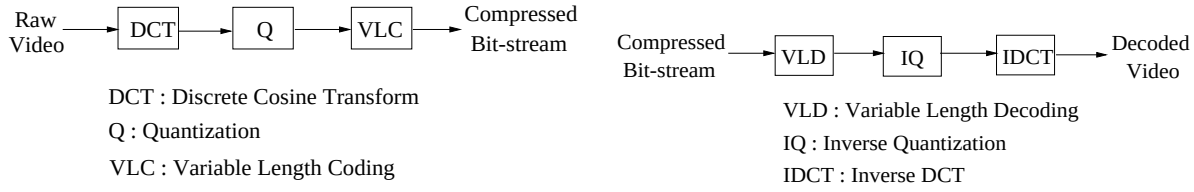
   The resources are shared in a fair manner. Specifically, the fairness could be either a utility-based fairness [6] or a max-min fairness [26].

The adaptive framework is a combination of network-aware applications and application-aware networks. That is, network-aware applications are aware of network status (e.g., available bandwidth) while application-aware networks are capable of processing application-specific information such as video format.

The remainder of this paper is organized as follows. Section 2 presents various scalable video coding mechanisms. In Section 3, we describe the adaptive framework for transporting scalable video over wireless networks. Section 4 summarizes this paper and points out future research directions.

2 Scalable Video Coding

A scalable video coding scheme is to produce a compressed bit-stream, parts of which are decodable. Compared with decoding the complete bit-stream, decoding part of the compressed bit-stream



(a) (b)

Figure 1: (a) A non-scalable video encoder; (b) a non-scalable video decoder.

produces pictures with degraded quality, or smaller image size, or smaller frame rate.

Scalable video coding schemes have found a number of applications. For video applications over the Internet, scalable coding can assist rate control during network congestions; for web browsing of video library, scalable coding can generate a low-resolution video preview without decoding a full resolution picture; for multicasting applications, scalable coding can provide a range of picture qualities suited to heterogeneous requirements of receivers.

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 Section 2.1 to 2.3, respectively.

To give a clear picture about scalable coding mechanisms, we first briefly describe a non-scalable encoder/decoder shown in Fig. 1. At the non-scalable 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 non-scalable decoder, the received compressed video stream is first decoded by variable length decoding (VLD), then inversely quantized, and inversely DCT-transformed.

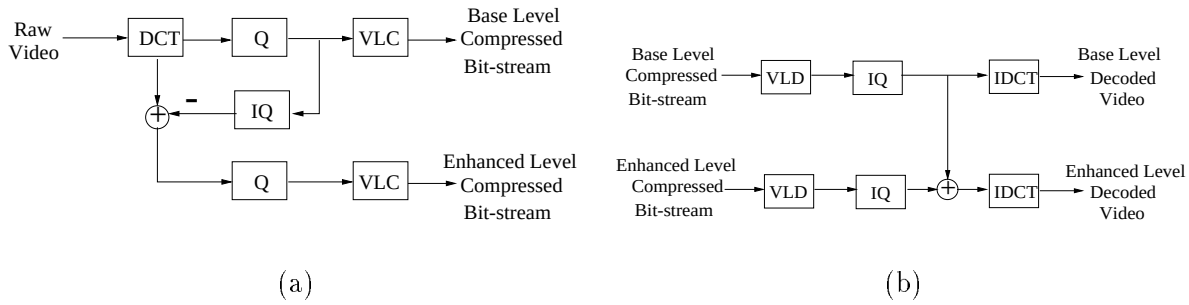
For purpose of simplicity, we only showed intra mode and only used DCT as an example in the above codec.[1] Similarly, Section 2.1 to 2.3 only describe intra mode for scalable video coding mechanisms and only use DCT. For wavelet-based scalable video coding, please refer to Refs. [9, 15, 25, 39, 40, 42] and references therein.

2.1 SNR Scalability

SNR scalability is defined as representing the same video in different SNR or perceptual quality. 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.

An SNR-scalable encoder with two-level scalability is depicted in Fig. 2(a). For the base level,

[1] Intra mode means that no motion compensation is involved in encoding/decoding process.



(a) (b)

Figure 2: (a) An SNR-scalable encoder; (b) an SNR-scalable decoder.

the SNR-scalable encoder operates in the same manner as that of the non-scalable video encoder. For the enhanced level, the operations are as follows:

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 coefficient.
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 enhanced level uses smaller quantization parameter, it achieves better quality than the base level.

An SNR-scalable decoder with two-level scalability is depicted in Fig. 2(b). For the base level, the SNR-scalable decoder operates exactly as the non-scalable video encoder. For the enhanced level, both levels must be received, decoded by VLD, and inversely quantized. Then the base-level DCT coefficient values are added to the enhanced-level DCT coefficient refinements. After this stage, the summed DCT coefficients are inversely DCT-transformed, resulting in enhanced-level decoded video.

2.2 Spatial Scalability

Spatial scalability is defined as representing the same video in different spatial resolutions or sizes (see Fig. 3(a) and 3(b)). Typically, spatially scalable video is encoded in such an efficient way: making use of spatially up-sampled pictures from a lower layer as a prediction in a higher layer. Figure 4(a) shows a block diagram of a two-layer spatially scalable encoder. For the base layer, the raw video is first spatially down-sampled,[2] then DCT-transformed, quantized and VLC-coded. For the enhanced layer, the operations are as follows:

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

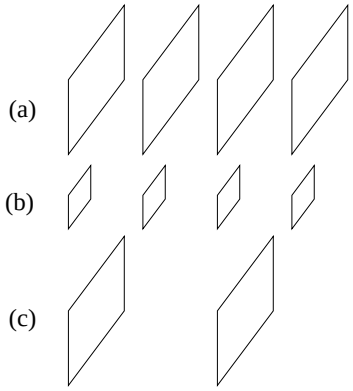
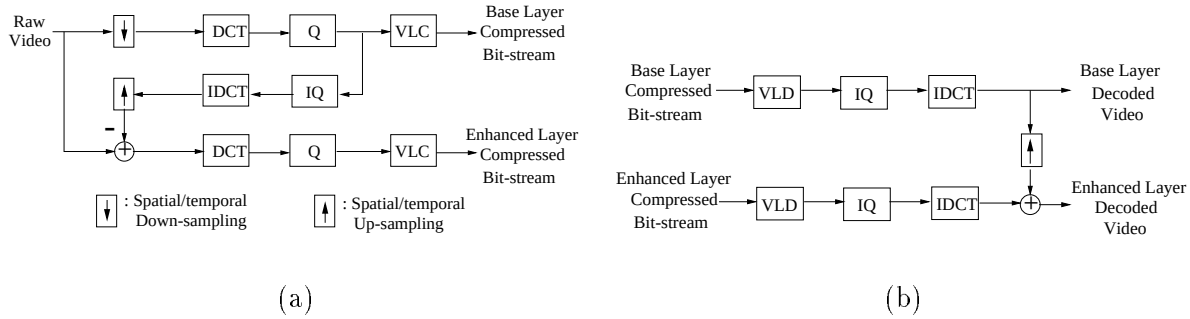


Figure 3: Spatial/temporal scaling of a video stream: (a) original video frames; (b) frames scaled to 1/4 original size; (c) temporally scaled frames.



(a) (b)

Figure 4: (a) A spatially/temporally scalable encoder; (b) a spatially/temporally scalable decoder.

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.[3]
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 enhanced layer uses smaller quantization parameter, it achieves finer quality than the base layer.

A spatially scalable decoder with two-layer scalability is depicted in Fig. 4(b). For the base layer, the spatially scalable decoder operates exactly as the non-scalable video encoder. For the

[3]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.

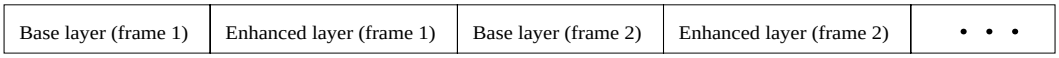


Figure 5: Embedded bit-stream.

enhanced 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 enhanced-layer refinements to form enhanced-layer decoded video.

2.3 Temporal Scalability

Temporal scalability is defined as representing the same video in different temporal resolutions or frame rates (see Fig. 3(a) and 3(c)). Typically, temporally scalable video is encoded in such an efficient way: 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. 4). 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. 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.

In sum, SNR/spatial/temporal scalability provides multiple video representations in different SNRs/spatial/temporal resolutions. Each video representation has different significance and bandwidth requirement. The base layer is more important while the enhanced layer is less important. The base layer needs less transmission bandwidth due to its coarser quality; the enhanced 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 rate (i.e., in different representation).

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. 5, an embedded bit-stream is formed by interleaving the base layer with the enhanced layer(s). An embedded bit-stream is also bandwidth-scalable since application-aware networks can select 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 only described basic scalable mechanisms, that is, SNR, spatial and temporal scalability. There can be combinations of the basic mechanisms, such as spatiotemporal scalability [10]. Other scalabilities include frequency scalability for MPEG [29] and object-based scalability for MPEG-4 [43].

We have presented three basic scalable video coding mechanisms for transmission over wireless networks. The primary goal of using bandwidth-scalable video coding is to achieve acceptable perceptual quality in presence of bandwidth fluctuations in wireless channels. However, without

appropriate network support, this goal cannot be achieved. So what kind of network support is needed to achieve this goal? Section 3 is to answer this question and present an adaptive framework as network support for transporting scalable video over wireless networks.

3 Adaptive Framework

In this section, we discuss the concept of adaptive framework for transporting scalable video over wireless network. The adaptive framework consists of (1) scalable video representations, each of which has its own specified QoS requirement, (2) network-aware video application, and (3) adaptive service, which makes network elements support the QoS requirements of scalable video representations. Under this framework, as wireless channel conditions change, the video sender and network elements can scale the video streams and transport the scaled video streams to receivers with acceptable perceptual quality.

The rest of the section is organized as follows. We first discuss the concept of network-aware application in Section 3.1 and then proceed to present the adaptive service in Section 3.2. Finally, we compare the adaptive service with other well-known services in Section 3.3.

3.1 Network-aware Application

The use of network-aware application is motivated by the fact: (1) bit error rate is very high when channel status is bad; and (2) packet loss is unavoidable if the available bandwidth is less than 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, a network-aware application was proposed to preemptively discard enhanced layers at the sender in an intelligent manner by considering network status [50].

For the purpose of illustration, we present an architecture including a network-aware sender, an application-aware base station, and a receiver in Fig. 6. The architecture in Fig. 6 is applicable to both live and stored video. In Fig. 6, 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 scales them (i.e., select suitable video representation) and then retransmits them to the destination through the network.

Note that a scaler can distinguish the video layers and drop layers according to importance. 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 enhanced layer(s); (2) transmit what is received, i.e., do not scale the received video representation.

Under our architecture, a bandwidth manager is maintained in the base station. One function of the bandwidth manager is to notify the sender about the available bandwidth of the wireless channel through signaling channel [31]. Upon receiving this information, the rate control module at the

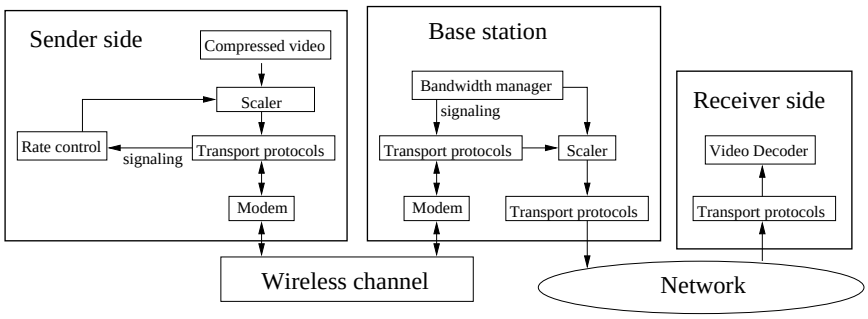


Figure 6: An architecture for transporting scalable video from a mobile terminal to a wired terminal.

sender 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 base station notifies the sender about the channel quality (i.e., BER) [4]. Upon receiving this information, the rate control module at the sender commands the scaler to perform as follows (suppose that the video is compressed into two layers): (1) if the BER is above a threshold, discard the enhanced layer so that the bandwidth allocated for the enhanced layer can be utilized by forward error correction (FEC) to protect the base layer; (2) otherwise transmit both layers. Here, for representations of multiple layers (more than 2), an open problem is:

Given a fixed bit budget, how many less important layers (higher layers) should be discarded in favor of heavily FEC-shielded more important layers (lower layers)?

The network-aware application has two advantages. Firstly, by taking the available bandwidth into account, the sender can make the best use of network resources by selectively discarding enhanced layers in order to minimize the likelihood of more significant layers being corrupted, thereby increasing the perceptual quality of the video delivered. Secondly, by considering the channel error status, the sender can discard the enhanced layers and FEC can utilize the bandwidth allocated for the enhanced layer to protect the base layer, thereby maximizing the possibility of the base layer being correctly received.

Note that adaptive techniques at physical/link layer are required to support network-aware applications. Such adaptive techniques include 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 measurement feedback channel [31]. In addition, the feedback interval is typically constrained to be of the order of tens to hundreds of milliseconds [31].

3.2 Adaptive Service

As we discussed earlier, a scalable video encoder can generate multiple layers or substreams to the network. The adaptive service is to provide scaling of the substreams based on the resource

availability conditions in the fixed and wireless network. Specifically, the adaptive service includes the following functions [50]:

- 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 enhance layers based on the available bandwidth and the fair policy. In other words, it scales the video streams based on resource availability.

Advantages of using scaling inside the network include:

(1) Adaptiveness to network heterogeneity. For example, when an upstream link with larger bandwidth feeds a downstream link with smaller bandwidth, use of a scaler at the connection point could help improve the video quality. This is because it selectively drops substreams instead of randomly dropping.

(2) Low latency and low complexity. Scalable video representations make the operation at a scaler very simple, i.e., only discarding enhanced layers. Thus, the processing is fast, compared with processing on non-scalable video.

(3) 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.

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 only address end-to-end provisioning in this paper.

The required components of the end-to-end adaptive service include [30, 50]:

1. service contract
2. call admission control and resource reservation
3. mobile multicast mechanism
4. substream scaling
5. substream scheduling
6. link-layer error control

Next, we describe the above components in Section 3.2.1 to 3.2.6, respectively.

3.2.1 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. 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 reserved bandwidth while a subcontract for the enhanced layers does not specify any QoS guarantee. As an example, we will use this typical scenario for two-layered video 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 [19]. 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.

3.2.2 Call admission control and resource reservation

Call admission control and resource reservation are two of the major components in end-to-end QoS provisioning [46, 47].

The objective of call admission control (CAC) is to provide a QoS guarantee for individual connections while efficiently utilizing network resources by preventing admission to an excessive number of calls or sources to the network. Specifically speaking, a CAC has to make such a decision: given a call arriving, requiring a connection with specified QoS (e.g., packet loss and 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).

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 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 at short time-scale, some duplication must be done in the transport of the connection to neighboring base stations of a connection so that in the event of a handoff, an outage in the link can be avoided. The resource reservation is done during connection admission and can be renewed by re-negotiation during lifetime of the connection.

The scalable video representation (i.e., substream) 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 base-layer substream needs QoS guarantee. 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 enhanced-layer substream(s) of one connection could share the leftover bandwidth with the enhanced-layer substreams of other connections. The enhanced-layer substreams are subject to scaling under bandwidth shortage and/or severe error conditions, which

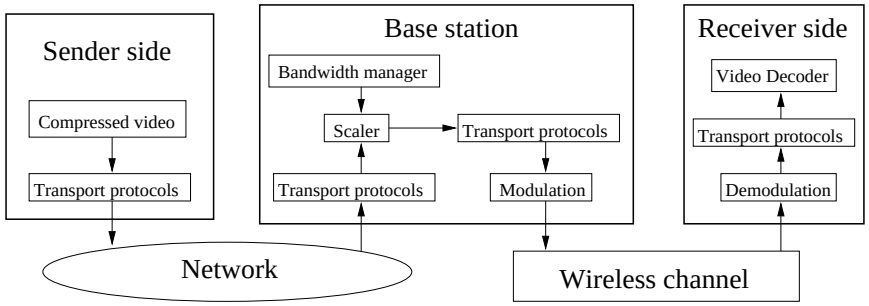


Figure 7: An architecture for transporting scalable video from a wired terminal to a mobile terminal.

will be discussed next.

For interested readers, more information about radio resource management can be found in Ref. [51].

3.2.3 Mobile multicast mechanism

CAC and resource reservation can provide connection-lever QoS guarantee. To seamless guarantee QoS at packet level, mobile multicast mechanism has to be used. That is, while being transported along its current path, the base-layer stream is also multicasted to its neighboring base stations. In this way, the QoS in short time-scale can be seamlessly achieved.

To support seamless QoS, 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 which 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 added and old paths are deleted [30].

3.2.4 Substream scaling

Scaling is employed during bandwidth fluctuations and/or under bad channel conditions. As the available bandwidth on a path reduces due to mobility or fading, lower-priority substreams are dropped by the scaler(s) on the path and substreams with higher priority are transmitted. As more bandwidth becomes available, lower-priority substreams are passed through the scaler, and the perceptual quality at the receivers increases. Figure 6 showed an architecture for transporting scalable video from a mobile terminal to a wired terminal. Figure 7 depicts an architecture for transporting scalable video from a wired terminal to a mobile terminal. We do not show the case of transporting scalable video from a mobile terminal to a mobile terminal since it is a combination of Fig. 6 and Fig. 7.

Figure 8: Utility functions, where BL, EL and E2 are the bandwidths for the base layer, the first enhanced layer and the second enhanced layer, respectively.

The scaling decision is made by a bandwidth manager. When there is no excess bandwidth (excluding reserved bandwidth), the bandwidth manager instructs the scaler to drop the enhanced layer. If there is excess bandwidth, an issue is how to fairly allocate the excess resources among contending adaptive flows when the excess resources cannot meet all the bandwidth demands of adaptive flows. Lu, Lee, and Bharghavan presented one solution which maximizes network revenue and achieves max-min fair allocation among the adaptive flows [26]. Another solution is based on a utility function [6], which captures the adaptive nature over which an application can successfully adapt to available bandwidth in terms of a utility curve that represents the range of observed quality to bandwidth. The observed quality index refers to the level of quality perceived by an application, as illustrated in Fig. 8. By using the utility function, Bianchi, Campbell and Liao proposed a utility-fair bandwidth allocation scheme that supports the dynamic bandwidth needs of adaptive flows [6]. Therefore, a good design of bandwidth manager has two features:

- Efficiency

  When there is excess bandwidth (excluding reserved bandwidth), the excess bandwidth will be efficiently used in a way that maximizes the perceptual quality or revenue.

  If the objective is to maximize the perceptual quality, efficiency can be achieved by exploiting the interplay between video compression and transport. In other words, optimal bandwidth allocation is based on a utility function, which characterizes the relationship between the perceptual quality and the available bandwidth [6].

- Fairness

  The resources are shared in a fair manner. The fairness can be defined as either a utility-based fairness [6] or a max-min fairness [26].

Note that rate adaptive techniques at physical/link layer [31] are required to support scaling the traffic, which will be transported over the wireless link.

3.2.5 Substream scheduling

The substream scheduler is used in mobile terminals as well as base stations. Its function is to schedule the transmission of packets on 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. Thus, a scheduler provides a scaling function as well; however, its scaling function is a result of its scheduling function. 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 regarding wireless channel conditions to determine the state of the wireless channel and also to predict its near-term state. For determining the transmission time of any packet in a specific substream (or its position in the transmission queue), the scheduler takes the following factors into account:

- The QoS parameters of the substream
- The relative importance of the substream compared to other substreams
- Wireless channel conditions (as well as past and future predicted conditions)

To achieve both QoS (e.g., bounded delay and reserved bandwidth) and fairness, algorithms like packet fair queueing have to be employed [5]. While the 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 in which a session receives significantly less service than it is supposed to, while another receives more. This results in large discrepancies between the sessions' virtual times, making it difficult to provide both delay-guarantees and fairness simultaneously.

To apply packet fair queueing algorithms, Ng, Stoica, and Zhang [32] identified a set of properties, called Channel-condition Independent Fair (CIF), that a packet fair queueing algorithm should have in a wireless environment: (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. Then 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 [11], called Channel-condition Independent packet Fair Queueing (CIF-Q), that achieves all the above properties [32].

Consider two-layered video as an example. Suppose that a subcontract for the base layer specifies reserved bandwidth while a subcontract for the enhanced layer does not specify any QoS guarantee, which is a typical case. We design an architecture for substream scheduling shown in Fig. 9 [50].

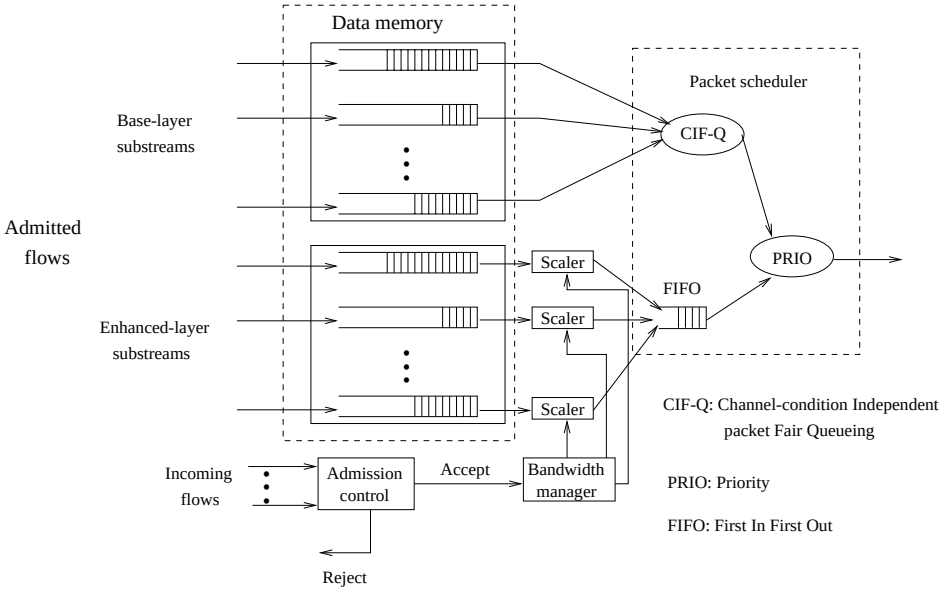


Figure 9: An architecture for substream scheduling at a base station.

Under our architecture, we partition the buffer pool (i.e., data memory in Fig. 9) into two parts: one for base-layer substreams, and one for enhanced-layer substreams. Within the same buffer partition for base or enhance 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 requested QoS and fairness. The first part of our architecture is CAC and bandwidth allocation. Video connections are admitted by CAC based on their base-layer QoS requirements. And bandwidth reservations for the admitted base-layer substreams are made accordingly. For admitted enhanced-layer substreams, their bandwidths are dynamically allocated by a bandwidth manager, which has been addressed in Section 3.2.4. The scaled enhanced-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. Shown in Fig. 9 is 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 enhanced-layer substreams. Service priority is first given to the CIF-Q scheduler and then to the FIFO scheduler.

3.2.6 Link-layer error control

To provide quality video over wireless, link-layer error control is required. Basically, there are two kinds of error control mechanisms, namely, FEC and automatic repeat request (ARQ).

The principle of FEC is to add redundant information so that original message can be recovered

in presence of bit errors. The use of FEC is primarily because throughput can be kept constant and delay can be bounded under FEC. However, the redundancy ratio (the ratio of redundant bit number to total bit number) should be made large enough to guarantee target QoS requirements under the worst channel conditions. In addition, FEC is not adaptive to varying wireless channel condition and it works best only when BER is stable. 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 will cause unnecessary overhead and waste bandwidth.

Different from FEC, ARQ is adaptive to varying wireless channel condition. When the channel is in good state, no retransmissions are required and no bandwidth is wasted. Only when the channel condition becomes poor, the retransmissions will be used to recover the errors. However, adaptiveness and efficiency of ARQ come with the cost of unbounded delay. That is, 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 [24, 52]. Different from conventional type-II hybrid ARQ [12, 17, 23, 45], the truncated type-II hybrid ARQ has the restriction of maximum number of transmissions for a packet. Due to the maximum number of transmissions, 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 transmissions N_m is assumed to be fixed and known *a priori* [24, 52], which may not reflect the time-varying nature of delay. If N_m is set too large, retransmitted packets may arrive too late for play-out and thereby be discarded, resulting in waste of bandwidth; if N_m 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 [49]. In our delay-constrained hybrid ARQ, the receiver makes retransmission request in an intelligent way: when errors in the received packet is detected, the receiver decides whether to send a retransmission request according to the delay bound of the packet.

Next, we briefly describe how to achieve bounded delay in our delay-constrained hybrid ARQ. In our scheme, a receiver-based control is employed to minimize the request of retransmissions that will not arrive timely for display. Under the receiver-based control, the receiver executes the following algorithm.

When the receiver detects the loss of packet N:
    if $(T_c + RTT + D_s < T_d(N))$
        send the request for retransmission of packet 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). It can be seen that if $T_c + RTT + D_s < T_d(N)$, the retransmitted packet is expected to arrive timely for display. The timing diagram for receiver-based control is shown in

Figure 10, where D_s is only the receiver's decoding delay.

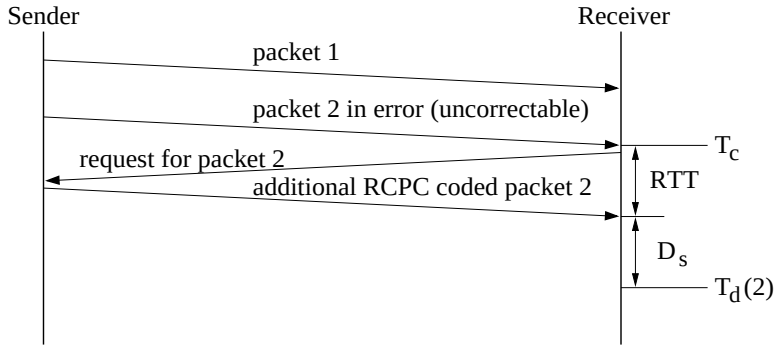


Figure 10: Timing diagram for delay-constrained retransmission (RCPC: Rate-Compatible Punctured Convolution).

Through simulations, our delay-constrained hybrid ARQ is shown to be capable of achieving bounded delay, adaptiveness, and efficiency [49]. It is also suitable for scalable video over wireless. Specifically, our delay-constrained hybrid ARQ can be employed in this way: based on the loss and delay requirements of a substream, an appropriate combination of FEC and retransmission can be selected by the error control module at the base and mobile stations.

On the other hand, unequal error protection [13] naturally fit the hierarchical structure of scalable video. Specifically, the base layer can be better protected against transmission errors than the enhanced layers. This form of unequal error protection is much more desirable than having to protect all the substreams. An open issue is how to combine unequal error protection with our delay-constrained hybrid ARQ.

3.3 Service Comparison

To give readers a clearer picture of the adaptive service, we would like to compare the adaptive service with other well-known services, i.e., the guaranteed service [37] and the best-effort service.

The guaranteed service guarantees that packets will arrive within the guaranteed delivery time, and will not get lost, provided that the flow's traffic conforms to its specified traffic parameters [37]. This service is intended for applications which need a hard guarantee that a packet will arrive no later than a certain time after it was transmitted by its sender. Examples that have hard real-time requirements and require guaranteed service include distant nuclear plant control, distant weapon control, and distant surgery control.

The best-effort service class offers the same type of service as that provided by the current Internet. Under best-effort service, the network makes 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., TCP) to make sure that all packets are delivered correctly. These include most popular applications like FTP, email, web browsing, and so on. All of these applications can work without guarantees of timely delivery of data.

Table 1: Comparison of Different Network Services

Services	Mode of transfer	Traffic charater-ization	End-to-end QoS guarantee	Explicit network feedback	Reserved resources	Isolation from other traffic	Emphasis	Target appli-cations
Guaranteed service	Connection oriented	Yes	Yes	No	Yes	Complete	Throughput, delay/loss	Nonadaptive CBR/VBR
Adaptive Service B	Connection oriented	Yes	If needed	No	Yes	Partial	Throughput delay/loss	Adaptive CBR/VBR
Adaptive Service E		No	No	If needed	No	Partial	Efficiency fairness	
Best effort	Connection-less	No	No	No	No	No	Fairness	Bursty data

A thorough comparison among the three service classes is given in Table 1, where adaptive service B and adaptive service E are provided for the base layer and the enhanced layer, respectively. Regarding target applications, both the guaranteed service and the adaptive service can support constant bit-rate (CBR) and variable bit-rate (VBR) applications.

In selecting a specific type of service for video transport, a trade-off must be made between two conflicting requirements: QoS guarantees (reflecting cost) and network utilization. The cost of the guaranteed service is too high for non-time-critical video applications. As a result, the guaranteed service is typically not chosen to transport video. The current best-effort service is not acceptable in many cases due to its poor QoS. The adaptive service provides users with another option. It achieves acceptable perceptual quality with medium cost. Specifically, adaptive service B provides basic perceptual quality at the cost of reservation; at almost no cost, adaptive service E takes advantage of statistical multiplexing gain to achieve better perceptual quality if possible. Therefore, the adaptive service can achieve better quality than the best-effort service while it costs less than the guaranteed service.

4 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. We presented an adaptive framework to support quality video communication over wireless networks. The adaptive framework is a combination of network-aware application and application-aware network.

We envision that the future adaptive framework consists of (1) scalable video representations, (2) network-aware video application, and (3) adaptive service. Under this framework, the mobile terminal and network elements can adapt the video streams to the channel conditions and transport the adapted video streams to receivers with acceptable perceptual quality. The advantages of deploying such an adaptive framework are that it can achieve suitable QoS for video over wireless,

bandwidth efficiency, and fairness in sharing resources.

As this paper outlines a high level framework, for implementation, some details need to be carefully considered.

- We have to consider the particular multiple access control protocol (e.g., CDMA or TDMA), modulation, channel allocation and mobile terminal being used [1, 18, 21, 28].
- We also need to take into account how to adapt the rate at link and physical layers [31]. In addition, channel quality feedback mechanisms have been defined in link/physical layer standards to carry out rate adaptation. As of the emerging broadband wireless networks, we might also need to design new rate adaptation techniques.
- A software platform might be necessary to support adaptive applications (e.g., [34]).
- A scalable video coding scheme needs to be carefully designed so that it is robust to multiple time-scale QoS fluctuations in the wireless/wireline network [8]. A scalable video coding scheme should achieve high efficiency with less complexity. It 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 design of efficient CAC schemes and in resource reservation [16].

Note that the above details can be implemented transparently to the adaptive framework (e.g., in a programmable way like that in Mobiware [2]).

As a final note, we would like to stress that each service (e.g., the adaptive service, the best-effort service, or the guaranteed service) has a trade-off 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 service could be provisioned at a single base station or provisioned in the whole network. In the real interconnected wireless networks, even though we cannot require each router deploy the adaptive service, a partial deployment of the adaptive service can still have clear benefits. For example, a service provider can deploy the adaptive service in its own network and its customers can enjoy the quality offered by the adaptive service in this network. Furthermore, it is entirely feasible to fully deploy the adaptive service 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 Mag.,* pp. 39–47, July 1999.

[2] O. Angin, A. T. Campbell, M. E. Kounavis, and R. Liao, "The Mobiware toolkit: programmable support for adaptive mobile networking," *IEEE Personal Commun. Mag.,* pp. 32–43, Aug. 1998.

[3] A. Balachandran, A. T. Campbell, M. E. Kounavis, "Active filters: delivering scalable media to mobile devices," *Proc. Seventh International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV'97),* May 1997.

[4] K. Balachandran, S. Kadaba, and S. Nanda, "Rate adaptation over mobile radio channels using channel quality information," *Proc. IEEE GLOBECOM'98,* Nov. 1998.

[5] V. Bharghavan, S. Lu, and T. Nandagopal, "Fair queuing in wireless networks: issues and approaches," *IEEE Personal Commun. Mag.,* pp. 44–53, Feb. 1999.

[6] G. Bianchi, A. T. Campbell, R. Liao, "On utility-fair adaptive services in wireless networks," *6th International Workshop on Quality of Service (IWQOS'98),* Napa Valley, CA, May 1998.

[7] M. C. Chan and T. Y. C. Woo, "Next-generation wireless data services: architecture and experience," *IEEE Personal Commun. Mag.,* pp. 20–33, Feb. 1999.

[8] Y.-C. Chang, and D. G. Messerschmitt, "Adaptive layered video coding for multi-time scale bandwidth fluctuations," submitted to *IEEE J. on Selected Areas in Communications.*

[9] T. Ebrahimi and M. Kunt, "Visual data compression for multimedia applications," *Proceedings of the IEEE,* vol. 86, no. 6, pp. 1109–1125, June 1998.

[10] B. Girod, U. Horn, B. Belzer, "Scalable video coding with multiscale motion compensation and unequal error protection," *Proc. Symp. on Multimedia Communications and Video Coding,* New York: Plenum Press, pp. 475–482, Oct. 1995.

[11] P. Goyal, H. M. Vin, and H. Chen, "Start-time fair queuing: a scheduling algorithm for integrated service access," *Proc. ACM SIGCOMM'96,* Aug. 1996.

[12] J. Hagenauer, "Rate-compatible punctured convolutional codes (RCPC codes) and their applications," *IEEE Trans. Commun.,* vol. 36, pp. 389–400, April 1988.

[13] J. Hagenauer and T. Stockhammer, "Channel coding and transmission aspects for wireless multimedia," *Proceedings of the IEEE,* vol. 87, no. 10, pp. 1764–1777, Oct. 1999.

[14] A. Iera, A. Molinaro, and S. Marano, "Wireless broadband applications: the teleservice model and adaptive QoS provisioning," *IEEE Communications Magazine,* pp. 71–75, Oct. 1999.

[15] K. Illgner and F. Mueller, "Spatially scalable video compression employing resolution pyramids," *IEEE J. Select. Areas Commun.,* vol. 15, no. 9, pp. 1688–1703, Dec. 1997.

[16] B. Jabbari, "Teletraffic aspects of evolving and next-generation wireless communication networks," *IEEE Personal Commun. Mag.,* pp. 4–9, Dec. 1996.

[17] 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.

[18] 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.,* pp. 47–51, Nov. 1999.

[19] O. Lataoui, T. Rachidi, L. G. Samuel, S. Gruhl, and R.-H., Yan, "A QoS management architecture for packet switched 3rd generation mobile systems," *Proc. Networld+Interop 2000 Engineers Conference,* May 2000.

[20] K. Lee, "Adaptive network support for mobile multimedia," *Proc. ACM Mobicom'95,* Nov. 1995.

[21] P. Lettieri and M. B. Srivastava, "Advances in wireless terminals," *IEEE Personal Commun. Mag.,* pp. 6–19, Feb. 1999.

[22] X. Li, S. Paul, M. H. Ammar, "Layered video multicast with retransmissions (LVMR): evaluation of hierarchical rate control," *Proc. IEEE INFOCOM'98,* March 1998.

[23] S. Lin and D. Costello, "Error control coding: fundamentals and applications," Englewood Cliffs, NJ: Prentice-Hall, 1983.

[24] H. Liu and M. El Zarki, "Performance of H.263 video transmission over wireless channels using hybrid ARQ," *IEEE J. on Selected Areas in Communications,* vol. 15, no. 9, pp. 1775–1786, Dec. 1997.

[25] Y.-J. Liu and Y.-Q. Zhang, "Wavelet-coded image transmission over land mobile radio channels," *IEEE GLOBECOM'92,* Orlando, FL, USA, Dec. 1992.

[26] S. Lu, K.-W. Lee and V. Bharghavan, "Adaptive service in mobile computing environments," *5th International Workshop on Quality of Service (IWQOS'97),* May 1997.

[27] S. McCanne, V. Jacobson, and M. Vetterli, "Receiver-driven layered multicast," *Proc. ACM SIGCOMM'96,* pp. 117–130, Aug. 1996.

[28] 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.

[29] J. Moura, R. S. Jasinschi, H. Shiojiri, and J.-C. Lin, "Video over wireless," *IEEE Personal Communications Magazine,* pp. 44–54, Feb. 1996.

[30] M. Naghshineh and M. Willebeek-LeMair, "End-to-end QoS provisioning in multimedia wireless/mobile networks using an adaptive framework," *IEEE Communications Magazine,* pp. 72–81, Nov. 1997.

[31] S. Nanda, K. Balachandran, and S. Kumar, "Adaptation techniques in wireless packet data services," *IEEE Communications Magazine,* pp. 54–64, Jan. 2000.

[32] T. S. E. Ng, I. Stoica and H. Zhang, "Packet fair queueing algorithms for wireless networks with location-dependent errors," *Proc. IEEE INFOCOM'98,* pp. 1103–1111, March 1998.

[33] M. Nishio, N. Shinagawa, and T. Kobayashi, "A lossless handover method for video transmission in mobile ATM networks and its subjective quality assessment," *IEEE Communications Magazine,* pp. 38–44, Nov. 1999.

[34] B. Noble, "System support for mobile, adaptive applications," *IEEE Personal Communications Magazine,* pp. 44–49, Feb. 2000.

[35] D. Reininger, D. Raychaudhuri, and M. Ott, "A dynamic quality of service framework for video in broadband networks," *IEEE Network Mag.,* Nov. 1998.

[36] D. Reininger, R. Izmailov, B. Rajagopalan, M. Ott, and D. Raychaudhuri, "Soft QoS control in the WATMnet broadband wireless system," *IEEE Personal Communications Magazine,* pp. 34–43, Feb. 1999.

[37] S. Shenker, C. Partridge and R. Guerin, "Specification of guaranteed quality of service," *RFC 2212,* Internet Engineering Task Force, Sept. 1997.

[38] B. Sklar, "Rayleigh fading channels in mobile digital communication systems Part I: characterization," *IEEE Commun. Mag.,* vol. 35, pp. 90–100, July 1997.

[39] I. Sodagar, H.-J. Lee, P. Hatrack, and Y.-Q. Zhang, "Scalable wavelet coding for synthetic/natural hybrid images," *IEEE Trans. on Circuits and Systems for Video Technology,* vol. 9, no. 2, pp. 244–254, March 1999.

[40] D. Taubman and A. Zakhor, "A common framework for rate and distortion based scaling of highly scalable compressed video," *IEEE Trans. on Circuits and Systems for Video Technology,* vol. 6, no. 4, pp. 329–354, Aug. 1996.

[41] L. Taylor, R. Titmuss, and C. Lebre, "The challenges of seamless handover in future mobile multimedia networks," *IEEE Personal Commun. Mag.,* pp. 32–37, April 1999.

[42] 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, no. 1, pp. 12–27, Jan. 1998.

[43] A. Vetro, H. Sun, and Y. Wang, "Object-based transcoding for scalable quality of service," *Proc. IEEE ISCAS'2000,* Geneva, Switzerland, May 28–31, 2000.

[44] J. Villasenor, Y.-Q. Zhang and J. Wen, "Robust video coding algorithms and systems," *Proceedings of the IEEE,* vol. 87, no. 10, pp. 1724–1733, Oct. 1999.

[45] 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.

[46] D. Wu and H. J. Chao, "On efficient bandwidth allocation and call admission control for VBR service using UPC parameters," to appear in *International Journal of Communication Systems.*

[47] D. Wu, Y. T. Hou, Z.-L. Zhang, H. J. Chao, "A framework of architecture and traffic management algorithms for achieving QoS provisioning in integrated services networks," to appear in *International Journal of Parallel and Distributed Systems and Networks, 1st Quarter 2000.*

[48] D. Wu, Y. T. Hou, W. Zhu, H.-J. Lee, T. Chiang, Y.-Q. Zhang, H. J. Chao, "On end-to-end architecture for transporting MPEG-4 video over the Internet," to appear in *IEEE Trans. on Circuits and Systems for Video Technology.*

[49] D. Wu, Y. T. Hou, Y.-Q. Zhang, W. Zhu, H. J. Chao, "Adaptive QoS control for MPEG-4 video communication over wireless channels," *Proc. IEEE ISCAS'2000,* Geneva, Switzerland, May 28–31, 2000.

[50] D. Wu, T. Hou, J. Yao, T. Chujo, "An adaptive approach for video transport over wireless networks," submitted to *IEEE GLOBECOM'2000,* San Francisco, CA, USA, Nov. 27–Dec. 1, 2000.

[51] J. Zander, "Radio resource management in future wireless networks: requirements and limitations," *IEEE Commun. Mag.,* vol. 35, pp. 30–36, Aug. 1997.

[52] Q. Zhang and S. A. Kassam, "Hybrid ARQ with selective combining for fading channels," *IEEE J. on Selected Areas in Communications,* vol. 17, no. 5, pp. 867–880, May 1999.

[53] Y.-Q. Zhang, Y.-J. Liu and R. Pickholtz, "Layered image transmission over cellular radio channels," *IEEE Trans. Vehicular Technology,* vol. 43, no. 3, Aug. 1994.