Scheduling Scalable Multimedia Streams for 3G Cellular Broadcast and Multicast Services

Kyungtae Kang, *Member, IEEE*, Yongwoo Cho, *Student Member, IEEE*, Jinsung Cho, *Member, IEEE*, and Heonshik Shin, *Member, IEEE*

***Abstract*—The cdma2000 1×evolution-data-only (EV-DO) mobile communication system provides broadcast and multicast services (BCMCS) to meet an increasing demand for multimedia data services. Currently, broadcast and multicast streams are scheduled using a slot-based static algorithm which cannot support dynamic environments where broadcast content is added or removed online. We propose a dynamic packet-scheduling algorithm that works with a retransmission scheme to enable a scalable and adaptive service across the cdma2000 1×EV-DO BCMCS environments. Since it is integrated with earliest deadline first real-time scheduling, the proposed algorithm not only adapts efficiently to dynamic contexts but also satisfies the real-time requirements of broadcast streams. Furthermore, by exploiting the fine granular scalability features of the MPEG-4 Part 2 standard, our scheme can avoid abrupt playback quality degradation by protecting the base layer using a retransmission technique and can also adapt more flexibly to an environment in which the resource requirements of video streams change dynamically. Simulation results show a significant improvement in average playback quality while quantitatively validating the efficiency of our approach.**

***Index Terms*—Automatic repeat request, broadcasting, code division multiaccess, multimedia communication, scheduling.**

I. Introduction

As MOBILE communications systems evolve into the third generation, the focus has shifted away from voice services and on to data services. In pursuit of this latest trend, many countries have adopted cdma2000 1×evolution-data-only (EV-DO), which is one of the third-generation systems, which offers a multimedia capability, support for packet-mode services, and more capabilities than second-generation systems (including peak rates of over 2 Mb/s, with an average throughput above 700 kb/s). Under the cdma2000 1×EV-DO, devices have "always-on" packet-data connections, helping to make wireless access simpler, faster, and more useful. A commercial cdma2000 1×EV-DO network has been deployed nationwide in several countries, including Korea.

Recently, work has begun, in both the Third Generation Partnership Project (3GPP) and the 3GPP2, on enhancing 3G networks to support multimedia broadcast and multicast services (BCMCS). In WCDMA, a multimedia broadcast/multicast service (MBMS) [1], [2] has been introduced for efficient support of broadcast and multicast transport in mobile networks. Also, the 3GPP2 group recently baselined the specification for a cdma2000 high-rate broadcast packet-data air interface [3], [4]. Their goal is to design a system that can deliver multimedia broadcast and multicast traffic with minimum resource usage by both the radio access and core networks. In addition, users expect low latency when joining or leaving the network, and multimedia streams must be delivered continuously as mobile users move around. A hierarchical design with localized multicasting and local servers is necessary to provide a scalable system, and an efficient air-link must also be designed to ensure that the total throughput of BCMCS is maximized. However, there has been no research on these topics in the context of cdma2000 1×EV-DO broadcast and multicast networks.

In this paper, we propose a dynamic packet scheduler, which is based on an earliest deadline first (EDF) real-time scheduling algorithm [5], to enable a scalable and adaptive service for MPEG-4 fine granular scalability (FGS) target video streams [6]–[8] across the cdma2000 1×EV-DO BCMCS [9], [10] environments. The static scheduling algorithm that is currently in the BCMCS specification cannot adapt to an environment in which content streams change dynamically. When a video flow is newly registered by mobile nodes, empty slots must be found to service this changed context. As the transmission rates of video flows differ from each other, the slot periods of these streams will also differ, causing the serving slots to overlap. Also, if there are no slots left, the existing static scheduler is unable to service a new stream because it is impossible to adjust the bit rate in a static scheme. Unlike the existing static scheduler, the proposed scheme not only efficiently adapts to dynamic contexts but also satisfies the real-time requirements of broadcast streams. Furthermore, new video flows can be admitted when all slots are already utilized, by dynamically adjusting the quality of the video flows that are currently being serviced. The average playback quality is, of course, reduced, but the average degradation across all mobile nodes is less than the one that would occur with the current static scheme. Three policies to adjust the quality of each video flow are proposed, and we compare their performance. We also show how to control the admission of packets with a utilization bound test, using the EDF scheduling algorithm.

We go on to suggest an efficient error-recovery scheme based on automatic retransmission request (ARQ) to mitigate any loss

Manuscript received July 8, 2005; revised April 11, 2006 and September 12, 2006. The review of this paper was coordinated by Prof. L. Lampe.

K. Kang, Y. Cho, and H. Shin are with the Department of Electrical Engineering and Computer Science, Seoul National University, Seoul 151-744, Korea (e-mail: ktkang@cslab.snu.ac.kr; xtg05@cslab.snu.ac.kr; shinhs@cslab.snu.ac.kr).

J. Cho is with the Department of Computer Engineering, Kyung Hee University, Youngin 449-701, Korea (e-mail: chojs@khu.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2007.899943

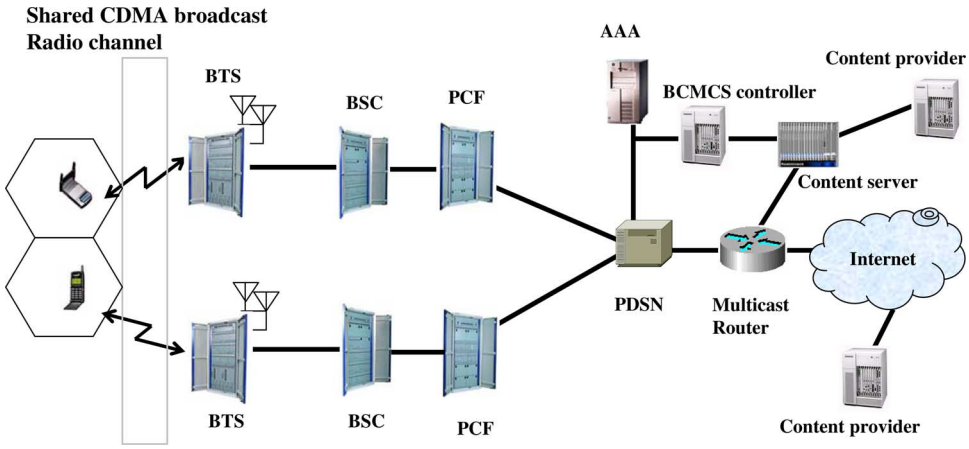

Fig. 1. cdma2000 BCMCS network architecture.

of quality, utilizing the scalability characteristics of MPEG-4 FGS video streams. Currently, BCMCS employ Reed–Solomon (RS) coding to correct errors at the cost of air resources. If the channel condition is good, many slots will be overbooked due to the space taken by parity information, which reduces the number of empty slots available for a newly registered video flow. In addition, the capacity for error recovery in the RS scheme declines suddenly as a channel condition deteriorates [11]. Thus, by adding a retransmission scheme to the EDF scheduling algorithm, we can save many slots when the condition of a channel is good and guarantee a minimum playback quality by protecting the base-layer packets within the FGS coding when a channel condition is extremely bad.

The remainder of this paper is organized as follows. In Section II, we introduce the background to our work and our research motivations. Section III describes the proposed scheme, relating transmission over the cdma2000 1×EV-DO broadcast networks to admission control and quality adjustment policies. We propose an analytic model in Section IV, and our simulation results are discussed in Section V. Finally, we conclude this paper in Section VI.

II. Background and Research Motivation

A. cdma2000 1×EV-DO BCMCS Architecture

BCMCS provide point-to-multipoint transmission of data from a single source to all users or to a group of users in a specific area. The main purpose of the BCMCS is to allow an efficient use of the cdma2000 radio interface for delivery of content streams to mobile nodes across an operator's network. A network operator can separately control each BCMCS content stream, including accounting aspects, regions of the network where the content streams are available, and the users who may receive them. Encryption of the content of multicast IP flows protects against unauthorized reception.

Fig. 1 is a high-level view of the proposed BCMCS architecture in the cdma2000 system. BCMCS content originates from the content provider and goes through the BCMCS content server. A content provider may be located within the cdma2000 serving network, in a home network, or anywhere in an IP network. If the BCMCS content provider is located in an IP network, then the business association, security association, and other related service information should be made available by that provider over the cdma2000 carrier network. The BCMCS content server is connected to the cdma2000 access network through a packet-data serving node (PDSN) that handles the BCMCS content stream, which the server makes available within an IP multicast stream. The BCMCS content server in the serving network is not necessarily the creator or source of the content. It is merely the last application-level entity in manipulating or reformatting the content before it reaches the PDSN.

The server may store and forward content from a single content provider or merge content from multiple providers. The PDSN is responsible for communication with the base station controller/packet control function (BSC/PCF), including the addition and removal of multicast IP flows. It uses IP multicast protocols in managing bearers, supporting these flows between itself and the nearest multicast router (MR), which connects back to the BCMCS content server. The way in which it treats multicast IP flows is specified by the BCMCS controller, which is a core network function that is responsible in managing and providing BCMCS session information to the PDSN, the mobile node, and the content server. The controller also performs authorization using the BCMCS user profile received from the home authentication, authorization, and accounting (H-AAA) server. In addition, the BCMCS controller performs security functions such as generating and distributing security keys to the mobile nodes [12].

Reception of a BCMCS service by a mobile node is enabled by a number of procedures. The BCMCS service announcement and discovery mechanism allow users to be informed about the services available. Users who wish to receive a service can discover its content and schedule from a BCMCS controller, which acts as a server in communication with a client application on the user's device. A user can then subscribe to one or more items of content using the BCMCS subscriber profile manager via out-of-band mechanisms. A mobile node can communicate with the BCMCS controller to acquire session-related information such as multicast IP addresses, port numbers, and transport and application protocols. Finally, the mobile node can

determine whether a particular multicast IP flow is available and can determine the corresponding BCMCS radio configuration from a base station via the overhead messages. When the mobile node has received this information, the user requests the desired IP flow via the BCMCS registration request mechanism. The first mobile node that performs BCMCS registration triggers the PDSN to join the multicast group associated with the BCMCS flow, which may, in turn, lead to the establishment of a new bearer path. When a mobile node requests a BCMCS content stream, packets are transmitted from the PDSN to the PCF, which uses a timestamp to make sure that the packet is received simultaneously at all base stations. The mobile node maintains its soft state by registering periodically. Eventually, the BCMCS is terminated if no mobile node registers to that content stream.

BCMCS bearer paths can be set up by static provisioning at any time. For dynamic broadcast services, the network sets up bearer paths when the first authorized user registers. Bearer paths from the PDSN to the MR are established using appropriate Internet engineering task force (IETF) multicast protocols.

B. Multimedia Scheduling in Current BCMCS

In a cdma2000 1×EV-DO unicast environment [13]–[16], 1×EV-DO employs a time-shared forward link that serves one user at a time in a time-multiplexed manner. The fundamental timing unit for forward-link transmissions is a 1.67-ms slot that contains the pilot and medium access control (MAC) channels, together with a data portion that may contain a traffic or control channel. When a mobile user is being served, each mobile node calculates its signal to interference plus noise ratio (SINR) at every time slot and determines the highest data rate from a list of possible rates that are supportable with the calculated SINR. The mapping between SINR and supportable data rate is given in Table 9.2.1.3.3.2-1 in the cdma2000 standard [13]. The measured data rate is reported to the home base station every 1.67 ms (i.e., every slot) by the mobile node. By using the information reported by mobile nodes, the base station schedules the slot allocation. The scheduled data on the traffic channel can be transmitted at rates between 38.4 and 2457.5 kb/s. The higher data rates are achieved through a combination of high-order modulation (QPSK, 8PSK, and 16-QAM), forward error-correction coding (the code rate is 1/5 or 1/3), and spreading. Transmission of one encoded packet can occupy 1–16 time slots. This adaptive rate control uses the full power of the base station to achieve the highest possible data rate for each user, within the constraints of changing channel conditions.

In a broadcast environment, however, a fixed transmission rate must be used because not all mobile nodes will necessarily have the same link state. BCMCS content streams are delivered to one or more mobile nodes in 1.67-ms time slots, which the current scheduler allocates statically in advance. These streams are scheduled and allocated with serving slots when they first appear in the system, which occurs when the first subscriber requests to be serviced by a content stream. The serving slots which deliver that content stream are never subsequently changed. The amount of data that can be forwarded in one slot is 256 B, supposing QPSK modulation is applied and the data rate is 1228.8 kb/s. Therefore, for example, four 220-kb/s video streams (M1, M2, M3, and M4 in Fig. 3) can be scheduled using four slots, which are interleaved as shown in Fig. 3.

C. MAC-Layer Error Recovery in Current BCMCS

Unlike the unicast cdma2000 1×EV-DO standard, error control in BCMCS is provided by forward error correction (FEC) using RS coding [3], [4]. Fig. 2 shows the structure of the MAC-layer error-recovery scheme. The broadcast framing protocol determines how higher layer packets are fragmented at the access network, the broadcast security protocol specifies the encryption of framing packets, and the broadcast MAC protocol defines the procedures used to transmit over the broadcast channel and the additional outer code which, in conjunction with the physical-layer turbo code, forms the product code. As already mentioned, RS was chosen as the outer code for cdma2000 BCMCS, and the broadcast MAC-layer packets have a fixed size of 125 B. The protocol is completed by the broadcast physical layer, and an error control block (ECB) is transported as payload on one or more subchannels of this layer. Data from multiple ECBs are multiplexed on to the broadcast physical channel, as shown in Fig. 2.

Each logical channel uses ECBs encoded with the same RS parameters (N, K, R) and has M MAC packets per ECB row (see Fig. 2). The variables N and K, respectively, represent the total number of octets and the number of security-layer octets in an RS codeword, while R is the number of parity octets. An RS decoder can recover up to R octet erasures in each codeword. RS coding is applied to the columns of the ECB, and then, the data are transferred row by row to the physical slot, where it forms one or more physical-layer packets. The ECB is designed to provide a structure such that, in the event of a physical-layer packet erasure, octets in the same position are lost from all affected RS codewords. To decode an RS codeword correctly, the broadcast MAC protocol needs to receive at least K of the N octets in that codeword. However, if all K data octets are received without errors, decoding is not needed, and the data octets which have been successfully received are simply forwarded to the upper layer of the BCMCS protocol suite.

One of the most significant environmental factors affecting channel condition is the fading effect. This is correlated with the burstiness of errors. In bad slow-moving conditions, the error pattern tends to be more bursty than in fast-moving conditions. An RS code (N, K, R) cannot recover any lost data if the corrupted portion is larger than R. For this reason, the performance of error correction will drop if the burst length of errors becomes so large that the ECB cannot interleave them sufficiently. This situation is schematically shown in Fig. 2. Thus, the burstiness of errors caused by bad channel conditions can be an important factor in selecting an appropriate data interleaving interval, which is determined by the width of the ECB, which is $M \times 125$ octets, as shown in Fig. 2. The value of M for a given ECB has to be less than or equal to 16. As the value of M increases, the time diversity also increases, and thus, a mobile node which is in a time-varying shadow environment is able to recover a substantial amount of corrupted

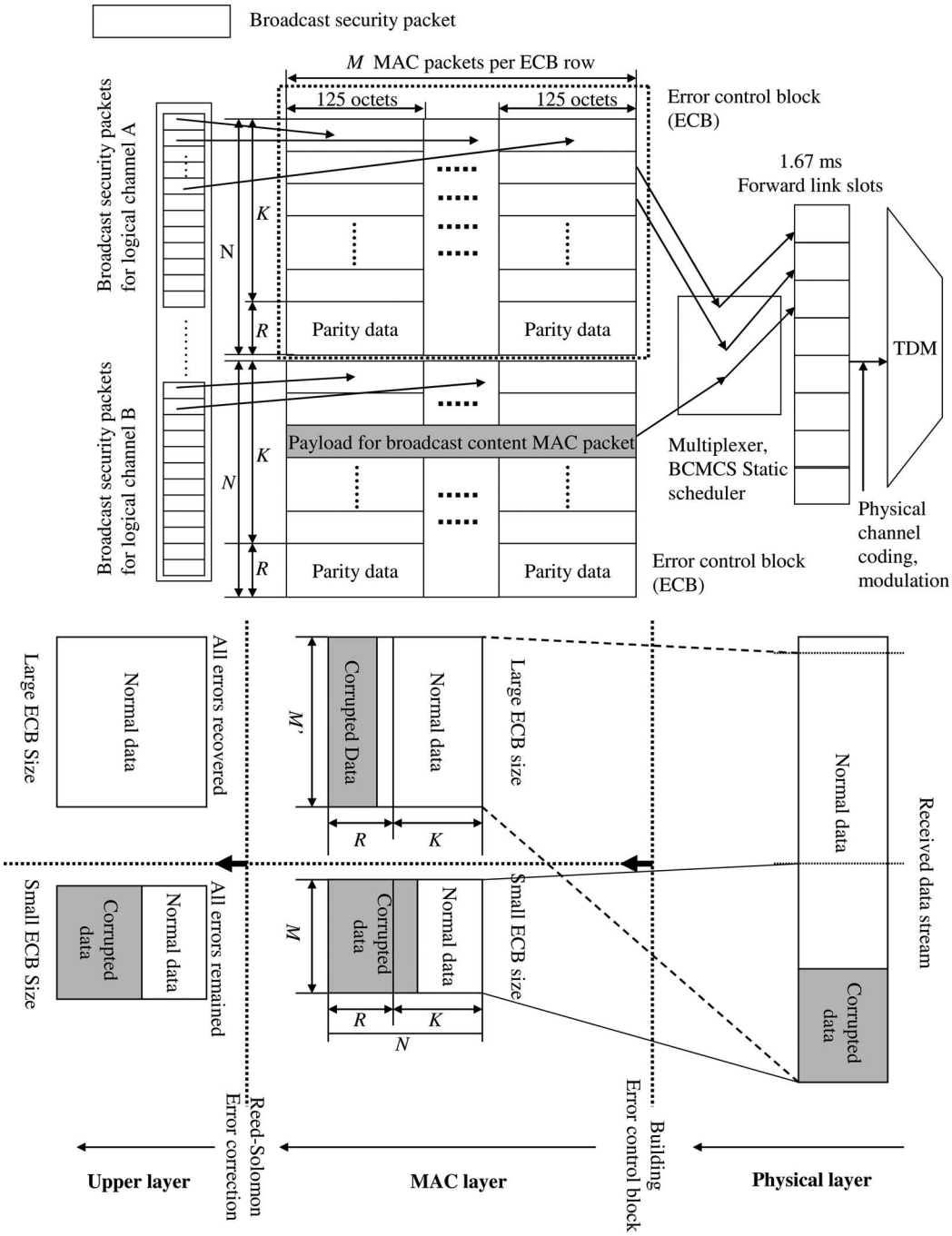

Fig. 2. Error-recovery structure for current BCMCS.

data. However, this requires additional storage at the node. Therefore, choosing an appropriate value of M is important in achieving the best error-recovery capacity for the available system resource.

D. Research Motivation

1) Drawback of the Current Static Scheduler: It is apparent from the description above that the current static scheduling algorithm cannot adapt to an environment in which new broadcast flows are requested or terminated by mobile nodes in parts of the current service area, affecting the resources available. For example, when a new video flow is added, empty slots must be found to schedule this content. If the transmission rates of video streams differ (M4 and M5 in Fig. 3), they require different slot periods. Thus, the slots used by different streams may overlap (M1 and M5, M2 and M5, and M3 and M5), which can cause problems in servicing them, as shown in case 2 of Fig. 3. It is also apparent that, if all slots are allocated to predetermined video streams, additional video streams or newly created asynchronous video streams (newsflashes or other irregular content) that are not predetermined will never be serviced, as in case 1 in Fig. 3. This inefficient slot utilization is exacerbated by the parity data required for RS coding. In addition, the current scheduling scheme makes it impossible to change the quality of a video stream in accordance with bandwidth variations, even

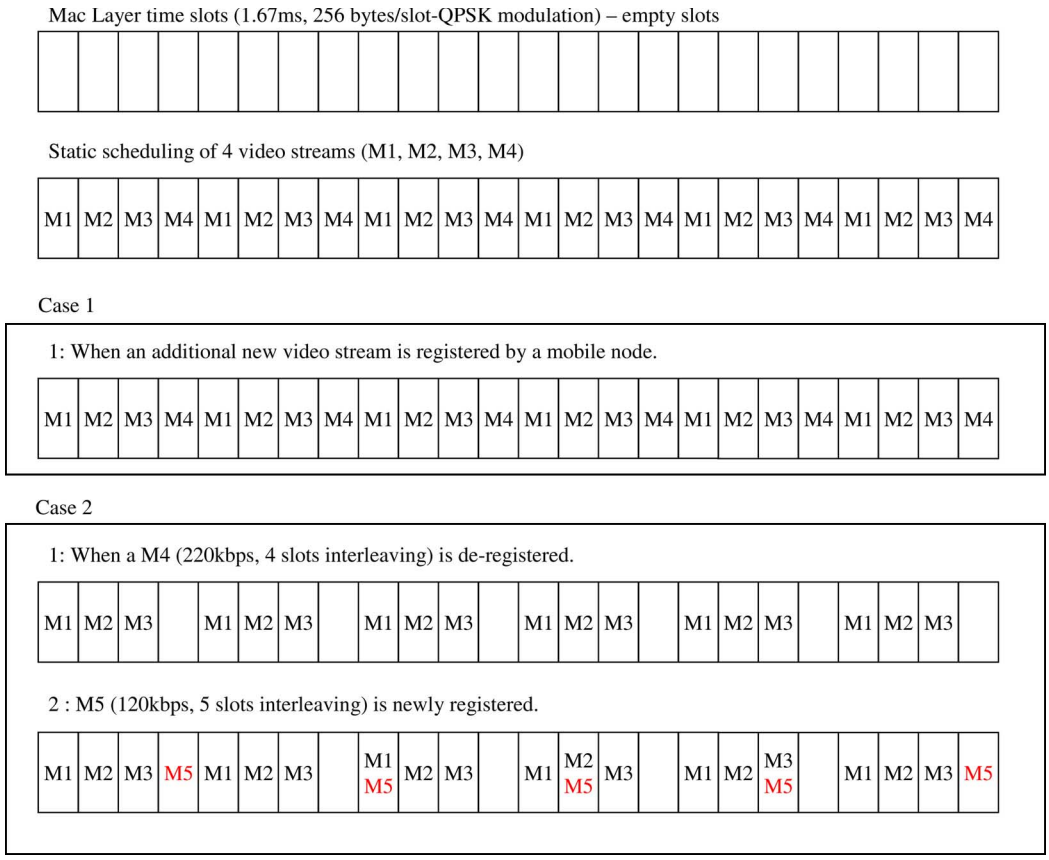

Fig. 3. Drawback of the current static scheduler.

though the video stream supports scalability. The significance of this will be explained later.

2) Drawback of the Current Error-Recovery Scheme: In general, a wireless radio channel has a higher error rate than a wired link. The unreliable and error-prone nature of the radio channel is the major challenge in servicing video streams over cdma2000 1×EV-DO broadcast networks. Wireless radio channels are affected by time-varying fading and interference conditions, which may corrupt packets. These errors tend to occur in clusters or bursts, with relatively long error-free intervals between them. Error control is mandatory when data are sent over wireless access networks to an end-user. FEC is the method most commonly used to control errors when sending a broadcast video stream over a lossy network such as a wireless radio channel. In BCMCS, the MAC protocol uses RS coding as the method of FEC, as mentioned already. A particular RS code contains a fixed amount of parity data; all of which is not needed if the channel condition is good, when a proportion of the slots are effectively wasted. In addition, using RS, the capacity for error recovery declines when the condition of a channel deteriorates, because FEC schemes become ineffective as the length of the error bursts increases. By utilizing more intelligent ARQ techniques, which are already used in cdma2000 1×EV-DO for unicast services [13], [16], to protect the transmitted video stream, we can save many slots when the condition of a channel is good, and these saved slots create flexibility. In addition, application-aware retransmission of corrupted packets, which is exploiting the characteristics of MPEG-4 FGS coded video streams, is made possible by giving more priority, and hence more chance of retransmission, to the data which are most significant for the quality of the video picture. By applying a priority-driven selective ARQ technique, we can reduce the unavoidable playback quality degradation of MPEG-4 FGS video when the channel condition is extremely bad. A detailed description of MPEG-4 FGS follows in the next section.

III. Scheduling Multimedia Streams in a Dynamic Environment

A. Scalability of MPEG-4 FGS Video Streams

A major topic in traditional multimedia research is the pursuit of efficient ways in compressing data at a given bit rate. Recently, several scalable video coding schemes have been developed in controlling stream bandwidth flexibly to meet the growing need for streaming video over packet-based networks such as the Internet, where the network channel capacity varies widely with the type of connection, the level of network traffic, and a lot of other external factors. To cope with this growing requirement for flexible stream transport, an MPEG working group has introduced a scalable video coding scheme called FGS.

The FGS video coding scheme in MPEG-4 not only provides an effective method of video compression but also adapts its bit

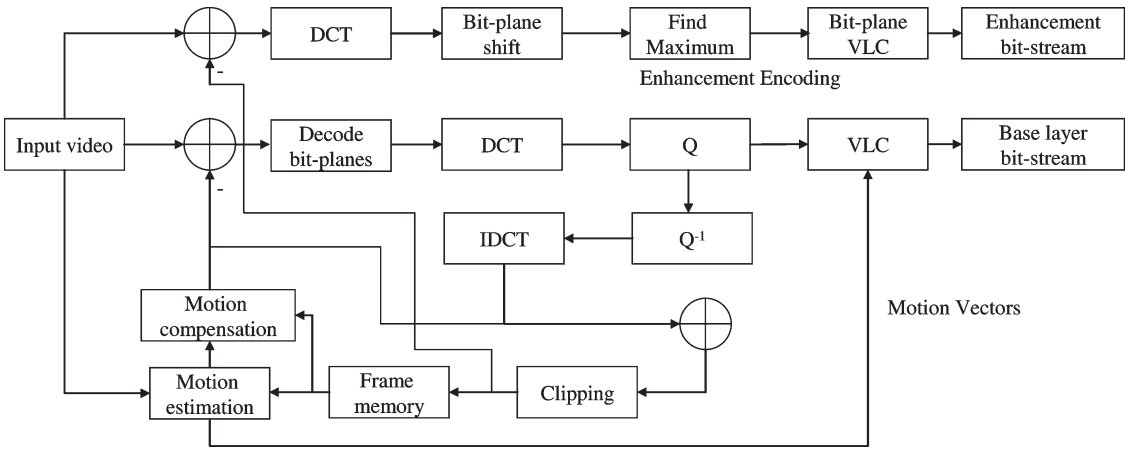

Fig. 4. Structure of an MPEG-4 FGS encoder.

rate flexibly to changing network conditions. The FGS scheme encodes video frames into two layers, which are distinguished by the priority of information that they contain, using bit-plane coding of discrete cosine transform (DCT) coefficients [17]. The layer containing essential information for the decodability of the whole video stream is called the base layer. It requires a modest bit rate and produces a lower bound on picture quality. The second layer is called the enhancement layer and includes a more detailed information that is used in enhancing the quality of the decoded video. As shown in Fig. 4, the enhancement layer is obtained by a bit-plane DCT coefficient coding of the differences between the original picture and the picture degraded by image transformations such as DCT and quantization. Because a bit-plane coding method considers each quantized DCT coefficient as a binary rather than a decimal number, each bit plane of the enhancement data stream has its own level of significance, from the most to the least significant bits. Thus, the enhancement layer can be truncated to any reduced number of bits. After receiving the whole base layer, the quality of the video proportionally increases as more of the enhancement layer is received. Thus, the FGS scheme provides a very efficient method of scalable video transmission since the overhead of layer separation accounts for at most 2 dB in the peak signal-to-noise ratio (PSNR) [18].

However, the base-layer stream is very sensitive to channel errors. If the decoder finds any errors in this layer, the enhancement layer of the current frame is discarded, whether it is correct or not. If they go undetected, errors in the base layer will propagate to the start of the next group of pictures and cause serious drifting problems in the enhancement layers of the following frames. While accuracy in the base layer is essential in decoding video streams, the enhancement layer is more tolerant, as described above. Errors in the enhancement layer cannot degrade the video quality below the lower bound provided by the base layer.

To effectively utilize this layered approach, the more important base layer should be provided with more protection against errors than the less important enhancement layer. However, the error-recovery scheme in the current BCMCS standard makes no allowance for the different value of the two layers, and each is likely to be equally damaged during transmission. We aim to overcome this defect by introducing the following features.

1) Graceful quality degradation: When the network has to drop packets, it does so with awareness of the video that the data represents. As a result, perceptual quality gracefully degrades, even under severe channel conditions.
2) Efficiency: When there is an excess bandwidth, it is used efficiently in a way that maximizes perceptual quality or revenue.
3) Fairness: Resources can be shared to enhance the average playback quality. We will propose three policies to share slot resources efficiently.
4) Guaranteed delivery of the base layer: The base-layer stream is guaranteed by giving it more opportunities for retransmission to avoid abrupt drops in service quality.

B. Scheduling Using EDF and the Proposed Error-Recovery Scheme

The weaknesses of the current scheduling scheme, which are mentioned in Section II, motivate us to deploy a dynamic scheduling algorithm. Specifically, we propose to apply the EDF real-time scheduling algorithm, which endeavors to deliver packets before their deadlines, together with retransmission to maximize the slot utilization in BCMCS [19]. We also propose the use of FGS scalable video coding, as mentioned earlier, to support quality adjustment. Thus, our scheduling algorithm can find the resource for new video streams, even when no more slots can be made available, by dynamically reducing the quality (bit rate) of the streams which are currently being serviced. Consequently, our scheme can adapt to an environment where the resource requirements of video streams change dynamically.

When many video servers are providing multimedia streams, each BCMCS flow is split into packets that are continually waiting to be served. While serving these packets over a wireless channel, a mobile user can experience great variations in multipath fading, past loss from distance attenuation, shadowing by obstacles, and interference from other users. Thus, a wireless radio channel has a higher error rate than a wired link. This

unreliable and error-prone nature of radio channels is the major challenge in serving video streams over cdma2000 broadcast networks. However, using RS, the capacity for error recovery declines when the condition of a channel deteriorates. Bursts of errors are inevitable in a wireless channel, and FEC schemes become increasingly ineffective as the length of these bursts increases. However, by exploiting the FGS characteristics of the MPEG-4 Part 2 standard, we can avoid abrupt playback quality degradation, even when the channel condition is extremely bad, by protecting the base layer using a retransmission technique. If a mobile node cannot receive a packet, that packet is reported as lost or damaged to the base station, via the reverse acknowledgement (ACK) channel [15]. Because successful receipt of a broadcast packet will give satisfaction to more users than receipt of a retransmitted packet, the scheduler gives a higher priority to broadcast packets than to retransmitted packets, and base-layer packets are given a higher priority than enhancement-layer packets. Thus, broadcast packets will be transmitted earlier, giving them more chance of retransmission in case of loss. Then, if any slots are still available, base-layer packets are retransmitted, and finally enhancement-layer packets, ordered by their deadline.

When using this scheme for error recovery, if one user in bad channel conditions has to have their base-layer packets retransmitted, the quality for all the other users will be reduced. However, it is important to guarantee a minimum quality for every user, even if the service to the bulk of users, who are already receiving high-quality video, is somewhat degraded. This is the key idea behind our scheme. In the worst case, all users may, indeed, be serviced at minimum quality, but we aim to guarantee quality to all users while improving on it when conditions permit.

Other approaches are possible. For example, our scheduler could be modified to reflect the channel condition of each mobile in a way that increases throughput while guaranteeing fairness. The proportionally fair (PF) scheduler is well-suited in balancing cell throughput against user fairness, while considering the channel condition [20], [21]. The 1×EV-DO unicast system includes a dynamic data-rate transmission technique that supports an efficient data service with a PF scheduler, which allows a user to receive data when their channel condition is better than average. In addition to providing good throughput and fairness, this type of scheduler can weigh the retransmission of packets that are requested by many users. Because the retransmissions are scheduled as unicasts, PF scheduling could be combined with our scheduler to increase throughput while guaranteeing fairness.

One limitation of our scheduling algorithm is that it requires modification to support multiple applications such as simultaneous streaming and Internet applications. However, in general, broadcast service receives a dedicated frequency allocation in order to guarantee the quality of service (QoS) of broadcast streaming applications, which should not be affected by other applications. In our scheme, the base station runs the proposed algorithm independently under the assumption that there are dedicated channels for streaming applications.

1) Handoff and Admission Control in a Dynamic Environment: We also consider handoff between different service areas [22]. In current BCMCS, soft combining is allowed on the forward link during a high-speed broadcast service because multiple sectors may simultaneously transmit the same data. However, using a dynamic packet scheduler, the packets are individually scheduled at each base station, and they may be transmitted in a different order in each service area. Consequently, soft combining is impossible. We, therefore, propose an alternative method of preventing the degradation of playback quality during the handoff of a node by the retransmission of lost packets, which are identified by the sequence numbers in their headers.

TABLE I
PARAMETERS USED IN THIS PAPER

Parameter	Description
τ_i	Periodic broadcast video flow
$\ddot{\tau}_j$	Newly created video flow
$\bar{\tau}_k$	Aperiodic retransmitted video packets
b_{τ_i}	Bit-rate of each BCMCS IP flow τ_i (kb/s)
p_i	Period of τ_i in units of slots
$\ddot{p}_j$	Period of $\ddot{\tau}_j$ in units of slots
R_i	Buffer requirement of each BCMCS IP flow
D_i	Relative deadline slot for the periodic multicast video flow τ_i
$\bar{D}_k$	Deadline slot for retransmitted packets $\bar{\tau}_k$
F_i	Size of transmitted data during one period for each video stream
e_i	Number of slots required by τ_i during one period
$\ddot{e}_j$	Number of slots required by $\ddot{\tau}_j$ during one period
$\bar{e}_k$	Number of slots required to retransmit lost packets of $\bar{\tau}_k$
ϕ_i	Number of bytes that can be forwarded in one slot (data payload)
U_P	Utilization of periodic video flows
$U_{present}$	Current slot utilization
$\bar{A}_k$	Arrival time of $\bar{\tau}_k$
N_{total}	Total number of mobile nodes in a cell
N_{τ_i}	Number of mobile nodes registered for video flow τ_i

Admission control and resource allocation are tightly related to the packet-scheduling algorithm. In our scheme, the admission of newly created video flows and retransmitted packets is allowed if sufficient resources are available for the adequate delivery of data to a mobile node. In the EDF scheduling algorithm, resource availability is checked by a utilization bound test [23], [24].

In our dynamic scheduling environment, packets could be lost during handoff, as explained above, or corrupted during transmission. Mobile nodes will request the retransmission of lost packets, and their admission into the system can be controlled using our utilization bound test as follows: Suppose that there are n BCMCS flows currently being serviced in the cell and that the bit rate of each BCMCS flow is b_i (in kilobits per second). The scheduler must periodically transmit at b_i (in kilobits per second) since a BCMCS flow corresponds to a periodic task in a real-time system. If each flow has the period p_i, the buffer requirement (R_i) that is needed to guarantee a continuous playback is $K * p_i$, which determines the relative deadline of τ_i (Table I). By using the transmission scheme in cdma2000 1×EV-DO, e_i is equal to $\lfloor b_i \times p_i/\phi_i \rfloor$. Therefore, the admission of a new multicast video flow $(\ddot{\tau}_j)$ is decided using the following equation [5]:

$$U_P = \sum_{i=0}^{n-1} \frac{e_i}{\min(p_i, D_i)} + \frac{\ddot{e}_j}{\ddot{p}_j} \leq 1. \tag{1}$$

The admission of newly generated packets for retransmission can be decided in the following way: Suppose that the aperiodic task set $Q(t) = \{\bar{\tau}_0, \bar{\tau}_1, \bar{\tau}_2, \ldots, \bar{\tau}_k, \ldots\}$ and the periodic task set $P = \{\tau_0, \tau_1, \tau_2, \ldots, \tau_{n-1}\}$ are to be scheduled by the EDF algorithm. For $\bar{\tau}_k \in Q(t)$, $\bar{e}_k(t)$ is defined as the remaining execution time of $\bar{\tau}_k$ at time t, and we define the set S to be the union of $Q(t)$ and P. For all $\bar{\tau}_k \in Q(t)$, $S(t,\infty)$ is schedulable if and only if $\forall k$, $U_{S(t,\infty)}(\bar{\tau}_k) \leq 1$, where

$$U_{S(t,\infty)}(\bar{\tau}_k) = \frac{\sum_{i=0}^{n-1} e_i(t) + \sum_{\forall \bar{\tau}_r | \bar{\tau}_r \in Q(t,\infty) \cap \mathrm{HP}(\bar{D}_k)} \bar{e}_r(t)}{\bar{D}_k - t}. \tag{2}$$

The term $\mathrm{HP}(\bar{D}_k)$ represents "higher priority": If $\bar{\tau}_k$ is included in the base-layer packets, then $\bar{\tau}_r$ are tasks which are included in the base-layer packets and have an earlier deadline than $\bar{D}_k$. If $\bar{\tau}_k$ is included in the enhancement-layer packets, then all packets for retransmission that correspond to the base layer or the enhancement layer, with deadlines earlier than $\bar{D}_k$, are included in $\mathrm{HP}(\bar{D}_k)$. Thus, base-layer packets have a higher priority than enhancement-layer packets. If two packets correspond to the same layer, then the packet with the earlier deadline has the higher priority. By using (2), we might now expect to be able to determine the feasibility of a new retransmission request. However, this test is actually impossible because the infinite number of potential tasks included in set $Q(t,\infty) \cap \mathrm{HP}(\bar{D}_k)$ would have to be considered. Therefore, we use the following recursive method to calculate an upper bound on the utilization.

For a task set $S(t,\infty) = P \cup Q(t)$, $Q(t) = \{\bar{\tau}_1, \bar{\tau}_2, \bar{\tau}_3, \ldots, \bar{\tau}_{k-1}, \bar{\tau}_k, \ldots\}$ is sorted by priority. Thus, $\bar{\tau}_{k-1}$ has a higher priority than $\bar{\tau}_k$, and an aperiodic task $\bar{\tau}_k \in Q(t)$ is schedulable when the upper bound on the utilization demand $(\bar{U}(\bar{\tau}_k))$ of $\bar{\tau}_k$ is less than one. The virtual finish time f_k is defined as the sum of the execution times (number of slots required) of all periodic tasks and of the aperiodic tasks that have a higher priority than $\bar{\tau}_k$ during the time period $[\bar{A}_k, \bar{D}_k]$. Thus

$$\begin{aligned} U(\bar{\tau}_k) &\leq \frac{\max\{f_{k-1} - \bar{A}_k, 0\} + \bar{e}_k + U_P * (\bar{D}_k - \bar{D}_{k-1})}{\bar{D}_k - \bar{A}_k} \\ &= \bar{U}(\bar{\tau}_k) \end{aligned} \tag{3}$$

$$f_{k-1} = \bar{A}_{k-1} + U(\bar{\tau}_{k-1}) * \bar{D}_{k-1}. \tag{4}$$

By using these equations while maintaining the value of the utilization demand $U(\bar{\tau}_k)$ and f_k, the feasibility of a schedule can be analyzed within a constant time. The time complexity of this test is $O(n)$ if the number of aperiodic packets in $Q(t)$ is n. Hence, we can ignore the admission control overhead.

2) Adjusting the Quality of Video Streams: When the admission of a new video stream fails (no slots are left), the current scheduler simply fails to deliver the new stream. However, our proposed dynamic scheduler can permit a new video stream by sharing the slot resource [25]. The scheduler adjusts the quality of existing video streams in order to permit the new stream. We consider three policies in adjusting the quality of each video stream, experimentally comparing their performance with respect to the PSNR [26] value.

1) Fair degradation (FD): The appearance of a new video stream produces equal quality degradation in all existing flows.
2) Victim degradation (VD): Victim flows are selected for adjustment to permit the newly requested video flow.
3) PF degradation (PFD): The degradation in quality of each existing flow is inversely proportional to the number of subscribers that it has.

We applied all the above policies to broadcast streams in our experiments. Besides the broadcast packets, there may exist packets for retransmission, generating aperiodic tasks which increase slot utilization in proportion to their number. Our proposed scheme considers these aperiodic packets and reserves the minimum number of slots required for retransmitting them. If the current slot utilization is U_{present}, then the average slot utilization by retransmitted packets for one mobile node will be as follows:

$$U_{\text{retransmitted packets/node}} = \frac{U_{\mathrm{present}} - U_P}{N_{\mathrm{total}}}. \tag{5}$$

When a request to service a new video flow $(\ddot{\tau}_j)$ is made, the bit rates of existing flows have to be adjusted. At this time, a reserved capacity of U_{reserved} must be provided for retransmitted packets, where

$$U_{\mathrm{reserved}} = \frac{(U_{\mathrm{present}} - U_P)}{N_{\mathrm{total}}} * (N_{\mathrm{total}} + N_{\tau_j}). \tag{6}$$

This leaves $1 - U_{\mathrm{reserved}}$ of utilization to be allocated to periodic broadcast flows by application of one of the policies just discussed. By using the first policy, when m new video flows are requested but insufficient slots are available, the bit rate of flow i is degraded in an egalitarian manner. The degraded bit rate is given as follows:

$$b'_{\tau_i} = \left\lfloor b_{\tau_i} * \frac{1 - U_{\mathrm{reserved}}}{U_{\mathrm{present}} + \sum_{j=0}^{m-1} \frac{\ddot{e}_j}{\ddot{p}_j}} \right\rfloor. \tag{7}$$

By using the second policy, victim flows are selected for degradation. If the bit rate of the first victim flow drops below the base-layer bit rate, then another victim is selected from among these flows which are not already victims: We choose the one with the smallest number of subscribers. The bit rates of additional victim flows are degraded in a similar manner to the first policy. By using the third policy, the bit rate is degraded in proportion to the number of subscribers. If the bit rate of the flow with the lowest subscriber count falls below the base-layer bit rate, then the bit rate of that flow is fixed at the minimum

value, and the other flows are appropriately scaled down. The adjusted bit rate $b_{\tau_i}^{'}$ of video flow i is given as follows:

$$\text{Degradation} = \left(\sum_{i=0}^{n-1} b_{\tau_i} + \sum_{j=0}^{j=m-1} b_{\ddot{\tau}_j}\right) \times \left(1 - \frac{1 - U_{\text{reserved}}}{U_{\text{present}} + \sum_{j=0}^{m-1} \frac{\ddot{e}_j}{\ddot{p}_j}}\right) \quad (8)$$

$$b_{\tau i}^{'} = \text{Degradation} * w_i \quad (9)$$

where w_i is a weight that is inversely proportional to the mobile node count of video flow i, which is selected from among the values $N_{\tau_i}/N_{\text{total}}$, $N_{\ddot{\tau}_j}/N_{\text{total}}$ $(0 \le i \le n-1, 0 \le j \le m-1)$.

Whichever policy is used, the bit rate of the base layer is not affected by the adjustment: A new video flow that would require a reduction of the base-layer bit rate is not permitted. When a flow begins or ends, the admission control algorithm calculates e_i for every flow i. However, if the channel condition of a mobile node changes, there is no need to run the entire algorithm again; instead, the solution can be updated for the new value of e_i. Because the base station does not run our admission algorithm every time that the channel condition of a mobile changes, we do not expect the computational overhead to be significant.

3) Service Delay Incurred by the Proposed Scheme: The transport stream system target decoder model of an MPEG system provides a guideline for managing timing constraints and buffers in the MPEG transport stream decoding process [6], [27]. In an MPEG system, the delay should not exceed 100 ms from arrival to the system decoder, and this requirement can be satisfied by using an adequate size of buffer if the delay can be bounded.

Thus, a major challenge in achieving the necessary QoS over a network is the implementation of a bounded delay service: that is, a communication service with deterministically bounded delays for all packets. The admission control functions in a network with a bounded delay service require schedulability conditions. Our new scheduler exploits EDF, which is known to provide the optimal delay performance at a single switch, in the sense that, for any set of connections, it can support the same delay bounds as any other packet-scheduling method, assuming that the maximum packet transmission time is known for all the connections [28], [29]. Optimality is defined in terms of the schedulable region associated with the scheduling policy. It has been shown [28], [29] that EDF has the largest schedulable region of all scheduling disciplines. Ferrari and Verma [30] presented sufficient schedulability conditions for EDF scheduling of a bounded delay service.

Our admission control algorithm has a complexity of $O(n)$, where n is the number of packets to be retransmitted, as mentioned above, and the maximum delay is bounded by restricting the size of n. As a result, the delay bound of a service is guaranteed in the same way as previous authors have assured the deterministic delay bound of an EDF scheduler.

4) Handling of Uplink Bottlenecks: Unlike pure RS, our method of error recovery leads to an increase in the utilization of the uplink channel as the number of mobile nodes increases, because of the growing number of retransmission requests. Also, more packets request retransmission as the channel condition deteriorates. As a result, the uplink may become a bottleneck when many mobile nodes simultaneously send ACK/NACK signals to the base station. However, just as our scheme supports scalable multimedia streaming by controlling the transmission rate at the base station, the possibility of an ACK/NACK storm can be countered by restricting the number of NACK signals from mobile nodes using information about the traffic on the uplink channel, which is forwarded via the overhead messages from the base station. When it knows there is a traffic problem, a mobile node simply drops further retransmission requests when the allowable number of requests has been exceeded. For example, if the uplink becomes a bottleneck, a base station might limit or eliminate the retransmission of enhancement-layer packets. Such a policy would restore the capacity of the uplink channel for important retransmission requests. A more scalable approach to an efficient utilization of the uplink channel is also possible: The number of retransmission requests permitted to each mobile node could be continuously adjusted to take account of the utilization of the uplink channel, which, in turn, depends on the channel conditions of all mobile node, making this a kind of channel-aware mechanism. Alternatively, we could independently build channel awareness into each mobile so that, for instance, a mobile does not transmit NACKs for enhancement-layer packets if its local channel condition exceeds some threshold of badness, or the number of allowable requests could be scaled continuously to reflect the local channel condition.

IV. Analysis of the Proposed Scheme

A. Context of the Analysis

In our experiments, handoff events are generated with a Poisson distribution. Mobile nodes arrive in, and depart from, the service area with a mean interval of $1/\lambda$. Packets that are deemed to have been lost wait in a retransmission queue to be scheduled. The information as to whether or not a video packet is received successfully at a mobile node is determined by an error-generation module that uses the simple threshold model suggested by Zorzi *et al.* [31], [32] to simulate the error sequences generated by data transmission on a correlated Rayleigh fading channel. A first-order two-state Markov process can simulate the error sequences generated by data transmission on a correlated Rayleigh fading channel: These errors occur in clusters or bursts with relatively long error-free intervals between them. By choosing different values for the physical-layer packet error rate and $f_d N_{\text{BL}} T$ (the Doppler frequency normalized to the data rate with block size N_{BL}, where f_d is the Doppler frequency and is equal to the mobile velocity divided by the carrier wavelength [33]), we can model different degrees of correlation in the fading process. The value of $f_d N_{\text{BL}} T$ determines the correlation properties, which are related to the mobile speed for a given carrier frequency. When $f_d N_{\text{BL}} T$ is small, the fading process is strongly autocorrelated, which means long bursts of errors (slow fading).

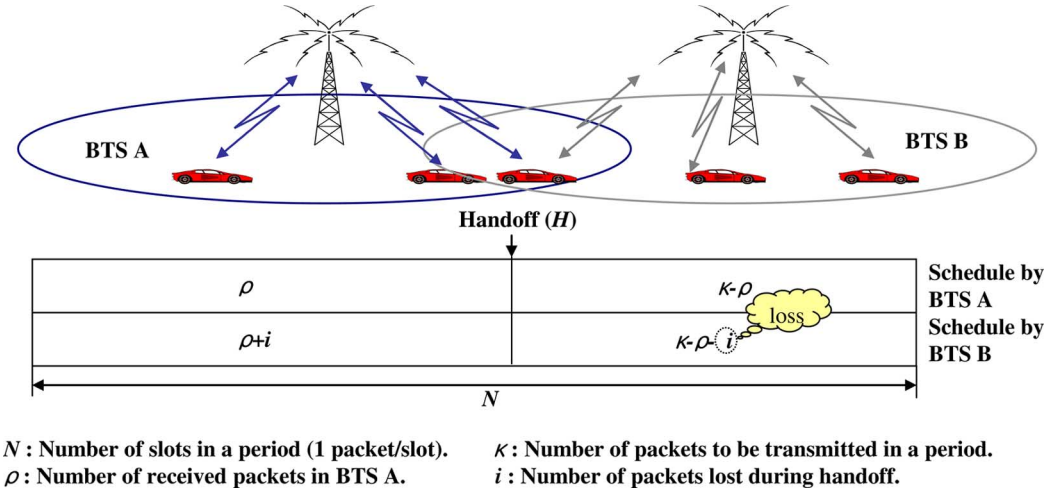

Fig. 5. Packet loss during handoff.

Conversely, the errors are weekly auto-correlated for large values of $f_d N_{\mathrm{BL}} T$ (fast fading). In our experiments, we set $f_d N_{\mathrm{BL}} T$ to 0.01 to simulate fast fading and to 0.02 to simulate moderate fading. The block size N_{BL} is the packet length.

In the following equations, α is the probability that the ith block of a packet is corrupted, given that the $(i-1)$th block was transmitted successfully, and β is the probability that the ith block of a packet arrives successfully, given that the $(i-1)$th block was corrupted. Fading in the air channel is assumed to have a Rayleigh distribution. The steady-state error rate ε is then obtained as follows:

$$\varepsilon = \frac{\alpha}{\alpha + \beta}. \tag{10}$$

If the Rayleigh fading margin is F, the average physical-layer packet error rate can be expressed as

$$\varepsilon = 1 - e^{-\frac{1}{F}}. \tag{11}$$

By using (8) and the equations which follow, we can now derive values for α and β. If F is the fading margin, then the average length of packet errors is given by $1/\beta$, where

$$\beta = \frac{Q(\theta, \rho\theta) - Q(\rho\theta, \theta)}{e^{\frac{1}{F}} - 1}, \qquad \theta = \sqrt{\frac{2/F}{1-\rho^2}}. \tag{12}$$

The term ρ is the correlation coefficient of two samples of the complex Gaussian fading process and is expressed as $\rho = J_0(2\pi f_d N_{\mathrm{BL}} T)$, where $j_0(\cdot)$ is the Bessel function of the first kind and is of the zeroth order. In addition

$$Q(x, y) = \int_y^{\infty} e^{-\frac{x^2+w^2}{2}} I_0(xw) w dw \tag{13}$$

is the Marcum-Q [34] function. Thus, the relationship between the physical-layer packet error rate and the Markov parameter can be represented as

$$\beta = \frac{1-\varepsilon}{\varepsilon} \left[Q(\theta, \rho\theta) - Q(\rho\theta, \theta) \right] \tag{14}$$

where

$$\theta = \sqrt{\frac{-2\log(1-\varepsilon)}{1 - J_0^2(2\pi f_d N_{\mathrm{BL}} T)}}.$$

B. Analysis of Packet Error Rate

In a dynamic packet-scheduling environment, a mobile node which is moving toward a neighboring cell may lose packets, even when there is no handoff delay because all packets are independently scheduled in each service area, as shown in Fig. 5. When a mobile node enters a new service area and handoff occurs at time H, packets that have not been received in the previous cell (BTS A) may already have been transmitted in the new cell (BTS B). The expected number of packets lost (E_{lost}) during handoff can be calculated as follows:

$$E_{\mathrm{lost}} = \sum_{H=0}^{N} \sum_{\rho=\mathrm{Max}(0,\kappa-N+H)}^{\mathrm{Min}(\kappa,H)} \sum_{i=\mathrm{Max}(0,\kappa-N+H-\rho)}^{\mathrm{Min}(\kappa-\rho,H-\rho)} \frac{1}{N} \times \left[\frac{{}_H C_\rho \times {}_{N-H} C_{\kappa-\rho}}{{}_N C_\kappa} \times \frac{{}_H C_{\rho+i} \times {}_{N-H} C_{\kappa-\rho-i}}{{}_N C_\kappa} \times i \right]. \tag{15}$$

The error-recovery capacity of the proposed scheme is dependent on the current slot utilization. Assuming that handoff requests by mobile nodes follow a Poisson distribution with an arrival and departure rate of λ requests per second, the maximum slot utilization in the current service area is as follows:

$$\max(U_{\mathrm{present}}) = U_P + \sum_{\psi=1}^{\text{node count}} \left(\frac{e_\psi}{p_\psi} \right) \sum_{\kappa=1}^{\mathrm{retry}_{\mathrm{Max}}} (\varepsilon_\psi)^\kappa + U_{\mathrm{handoff}} \tag{16}$$

where e_ψ represents the number of slots required in a period p_ψ for the video stream which is being received by mobile node ψ, ε_ψ is the average steady-state physical-layer packet error rate of node ψ, and a mobile node is able to make up to $\mathrm{retry}_{\mathrm{Max}}$ retransmission requests. U_{handoff} is the average utilization factor achieved by packets during handoff. As the

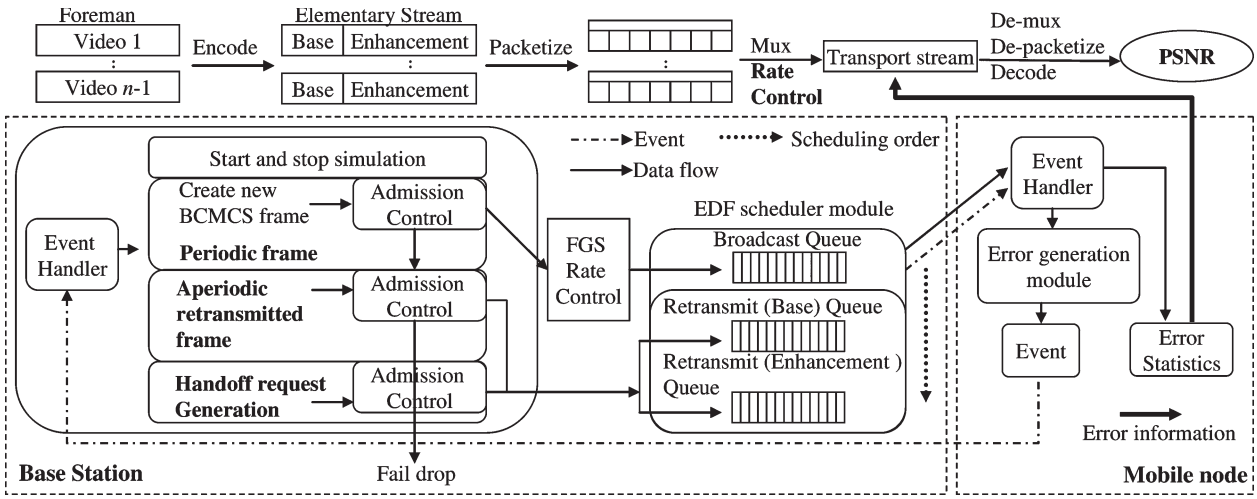

Fig. 6. Experimental structure.

mean interarrival time is $1/\lambda$ (s), U_{handoff} can be expressed as follows:

$$U_{\text{handoff}} = \frac{\lambda E_{\text{lost}}}{600}, \quad (600 \text{ slots/s}). \tag{17}$$

P_{drop} is the probability that a packet τ waiting for retransmission is not admitted into the wait queue and is, therefore, not scheduled again, and this probability can be expressed as

$$P_{\text{drop}} = \begin{cases} 0 & \text{if } \tau \text{ is admitted by the utilization bound test.} \\ 1 & \text{if } \tau \text{ is not admitted to the system.} \end{cases}$$

Thus, the expected value of P_{drop} (which is denoted as E_{drop}) can be expressed as

$$E_{\text{drop}} = \left(\frac{|1 - U_{\text{present}}| - (1 - U_{\text{present}})}{2|1 - U_{\text{present}}|} \right) \left(1 - \frac{1}{U_{\text{present}}} \right). \tag{18}$$

Until the current slot utilization is less than one, retransmission of all lost packets is attempted. However, when the slot utilization is above unity, some packets must be dropped. However, only the lower priority enhancement-layer packets are affected at this stage. In our scheme, the average packet error rate after the $\text{retry}_{\text{Max}}^{\text{th}}$ retransmission, which we call the packet error rate in the upper layer and denote as $\varepsilon_{\text{upper}}$, satisfies the following inequality:

$$\varepsilon_{\text{upper}} \leq \varepsilon_{\text{physical}} \times [E_{\text{drop}} + (1 - E_{\text{drop}})(\varepsilon_{\text{physical}})]^{(\text{retry}_{\text{Max}})} \tag{19}$$

where $\varepsilon_{\text{physical}}$ is the average steady-state physical-layer packet error rate of all mobile nodes.

V. Simulation Results

A. Experimental Environment

Fig. 6 shows the structure of our experimental study. We used the Foreman testbench video sequences streamed at 30 frames/s, with a total of 10 000 frames. Results are presented for the average physical-layer packet error rates of all mobile nodes (denoted by $\varepsilon_{\text{physical}}$) of 1%–10%. Each video stream is handled with our reference MPEG-4 FGS codec, which is derived from the framework of the European ACTS Project Mobile Multimedia Systems [35], but modified for our experiments. The streams consist of a 100-kb/s base-layer bit rate and a 120-kb/s enhancement-layer bit rate with 120 levels (1-kb/s step). All experiments were performed with a mean handoff interarrival time $(1/\lambda)$ of 1 s. All video streams are sent through channel coding and packetized before being transferred via a cdma2000 1×EV-DO physical slot. In these experiments, we used QPSK modulation with a 1228.8-kb/s data-rate forward channel, and thus, the values of ϕ_i and N_{BL} are 256 B. In our proposed error-recovery scheme, all corrupted packets may be retransmitted up to three times $(\text{retry}_{\text{Max}} = 3)$.

At a mobile node, intermittent receipt of video packets is simulated by an error-generation module, using the threshold model explained in Section IV-A. When the event-generation module determines that an error has occurred, the mobile may subsequently request retransmission of packets by means of a retransmission request. The affected packets are then inserted into a queue by an event handler in the base station, where they wait to be served by the scheduler.

We compared our scheme with the original FEC-based static scheme employed in BCMCS. In our comparison, RS is used as an erasure code not as an error-correcting code; specifically, we use the (16, 12, 4) code, which is capable of recovering up to four octet erasures in each RS codeword, and 16 MAC packets per ECB row. These settings give the best error-recovery capacity with the current scheme. To evaluate the two error-recovery schemes, errors (from the error-generation module) are injected into the original transport stream of a target video sequence, and the PSNR of the resulting video stream is calculated to estimate the difference in quality between a reconstructed image and an original image.

B. Performance Evaluation

Fig. 7 shows the average number of packets lost during handoff as the handoff position H changes. The two curves represent

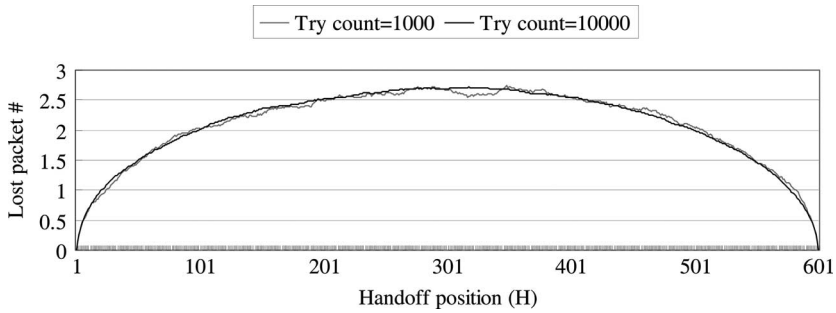

Fig. 7. Number of lost packets during handoff.

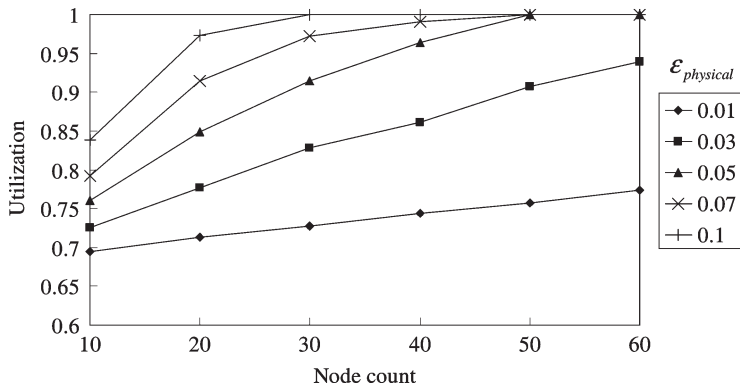

Fig. 8. Utilization of slots.

the results of 1000 and 10 000 trials; the video is streamed at 220 kb/s. With more trials, the graph stabilizes to a smooth curve. The expected number of lost packets during handoff is 2.107 and 2.111 for 1000 and 10 000 trials, respectively, which is nearly in agreement with the theoretical value of 2.094, which is derived from (15). From this result, we can see that the extent to which slot utilization increases due to handoff requests by mobile nodes is negligible.

We also investigated the relationship between the average value of $\varepsilon_{\rm upper}$ and M, which is the number of MAC packets per ECB. Fig. 9(a) shows that, for all values of $\varepsilon_{\rm physical}$ and $f_d N_{\rm BL} T$, the average $\varepsilon_{\rm upper}$ of all the mobile nodes when using RS coding is inversely proportional to the value of M. We are using $\varepsilon_{\rm upper}$ to denote the upper layer packet error rate of the data (not including parity bits) after the corrupted packets have been partly recovered, either by current RS or by the proposed error-recovery scheme. As M increases, bursts of errors are sufficiently dispersed between neighboring codewords, and error recovery is achieved by reconstructing data in columns, which increases the number of errors that can be corrected. Fewer packets are corrupted by fast fading because the error bursts are shorter than in slow-moving conditions. Bursts of errors can be sufficiently interleaved in an ECB by increasing the value of M, and even a small increment considerably reduces $\varepsilon_{\rm upper}$.

Our method of error recovery leads to an increase in slot utilization, in contrast with RS, as the number of mobile nodes increases because of the growing number of retransmission requests. Also, more packets become candidates for retransmission as the channel condition deteriorates. When $\varepsilon_{\rm physical}$ is 1% or 3%, the utilization falls below unity, even when the number of mobile nodes reaches 60, which means that slots are still available after corrupted packets have been retransmitted up to $\rm retry_{Max}$ times. However, when $\varepsilon_{\rm physical}$ is 5%, 7%, or 10%, the utilization arrives at unity as the number of mobile nodes exceeds 50, 40, and 30, respectively, as shown in Fig. 8. After these saturation points, the average value of $\varepsilon_{\rm upper}$ in our scheme increases abruptly as the number of mobiles increases and finally exceeds the average $\varepsilon_{\rm upper}$ of the current scheme because of the lack of slots for additional retransmission. This is shown in Fig. 9(b), which demonstrates how average $\varepsilon_{\rm upper}$ changes as the number of mobile nodes increases, for values of $\varepsilon_{\rm physical}$ of 3%, 5%, 7%, and 10%. As the number of mobiles increases and the channel condition deteriorates, average $\varepsilon_{\rm upper}$ increases with our scheme, and its performance, simply in terms of average $\varepsilon_{\rm upper}$, becomes inferior to that of the current scheme. However, our scheme still shows improved playback quality (PSNR) because it protects the base layer by retrying transmission repeatedly. Figs. 10 and 11 show the average $\varepsilon_{\rm upper}$ of the base layer for the proposed and current schemes. These results demonstrate that the error rate in the base layer reduces dramatically compared to the current scheme by giving packets more chances of retransmission, even though the node count increases.

In all cases, with either scheme, the value of $f_d N_{\rm BL} T$ affects the error-recovery capacity. When the value of $f_d N_{\rm BL} T$ is small (slow fading), the errors are more bursty, which reduces the error-recovery capacity, as shown in Fig. 9.

1) Video Quality: Fig. 9(c) shows the average PSNR for different numbers of mobile nodes, with values of $\varepsilon_{\rm physical}$ between 3% and 10%, and $f_d N_{\rm BL} T = 0.02$. Fig. 9(d) presents

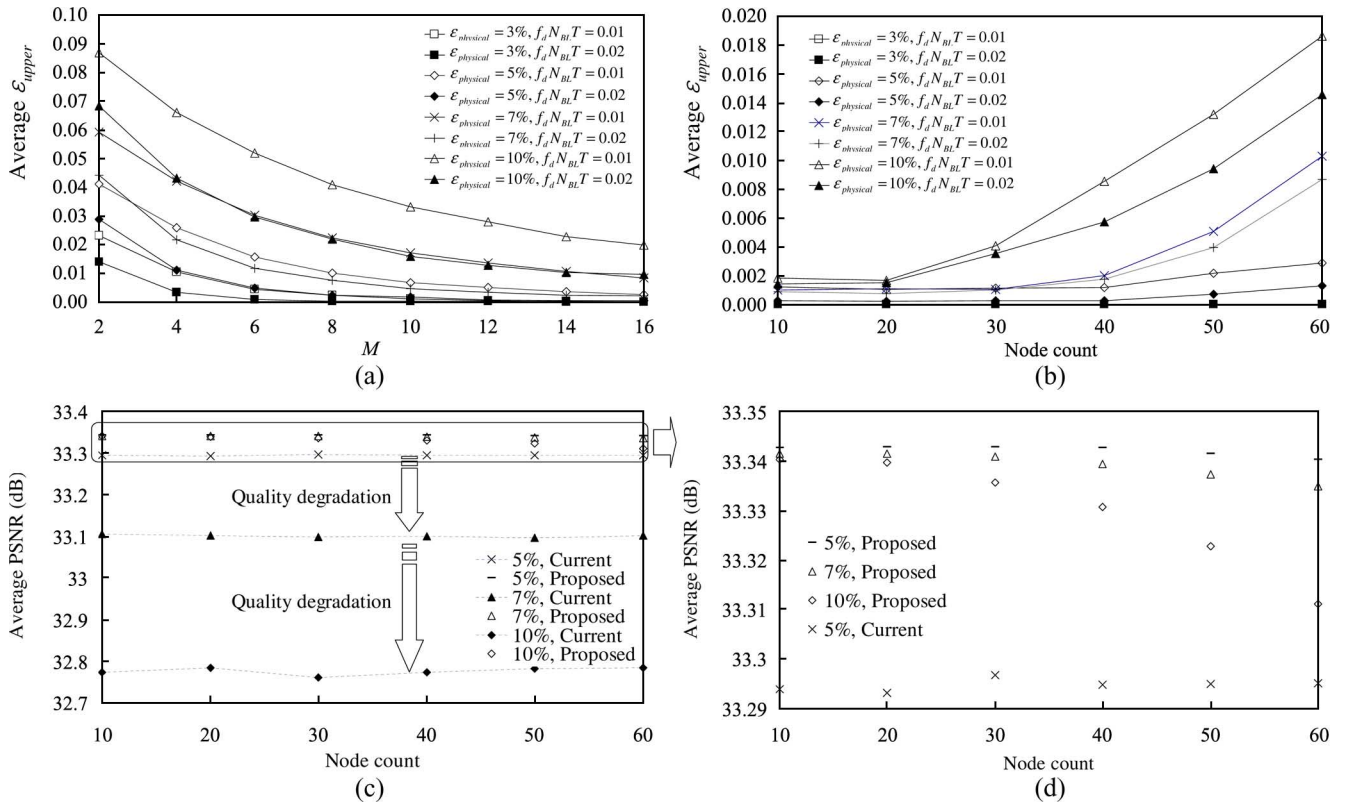

Fig. 9. Experimental values of average $\varepsilon_{\rm upper}$ and PSNR. (a) Average $\varepsilon_{\rm upper}$ in the current scheme. (b) Average $\varepsilon_{\rm upper}$ in the proposed scheme. (c) Average video quality against channel condition and node count $(M = 16)$. (d) Detailed performance of the proposed scheme.

the results achieved by our scheme in detail. With the current scheme, error recovery is most effective when the size of the ECB is set to the maximum $(M = 16)$. As the channel condition deteriorates, the error rate increases, even after error recovery, both with the current scheme and with ours. However, the average drop in quality is more severe for the current scheme.

When $\varepsilon_{\rm physical}$ is 10% and the number of mobile nodes is 60, the average video quality drops to 33.31 dB, but the average PSNR value is still higher than that achieved by the current scheme, which is 32.78 dB. As the number of mobile nodes increases, average $\varepsilon_{\rm upper}$ increases with our scheme, eventually exceeding the value for the current scheme; however, the average PSNR remains lower, and so, the relative advantage of our scheme is larger as $\varepsilon_{\rm physical}$ increases. Also, the gap in video quality between our scheme and the current scheme grows when a different RS code, such as (16, 14, 2), or a smaller ECB is used. From these results, we can see the importance of protecting the base layer of an MPEG-4 FGS video stream to prevent a significant drop in quality. If delivery of the base layer can be guaranteed, we can present the user with a watchable video stream, even when the channel conditions are very poor.

Fig. 12 compares the average playback quality of our scheme and the current BCMCS scheme when the average value of $\varepsilon_{\rm physical}$ is 5% and the number of mobile nodes is 50. We sampled from frame 1 to frame 1000. From this graph, we see that the average PSNR of the current scheme more significantly fluctuates than it does with our scheme (see the circled portions of the graph in Fig. 12). The bold fluctuations shown in Fig. 12(a) come from the corruption of base-layer packets, which is an effect that is exacerbated as the value of $\varepsilon_{\rm physical}$ increases. The results of this corruption can clearly be seen in the video quality at an individual mobile node. In case of node A, the PSNR drops to 3.76 dB, at which point a base-layer packet is corrupted, producing a significant distortion of the picture, as shown in the captured image in Fig. 12(c). Better images are obtained when our scheme restricts the occurrence of errors to the enhancement layer by selective retransmission [Fig. 12(d)]. This confirms the effectiveness of incorporating awareness of the characteristics of the application. In Fig. 12(b), we omitted the graph for node A because it is so similar to that for an average node. In our scheme, almost all corrupted base-layer packets are recovered by giving them more chances of retransmission during error recovery, and packet loss is restricted to the enhancement layer. This means that the loss of quality can be ignored, even when unrecovered enhancement-layer packets are forwarded to the application layer.

2) Effect of Adjusting Quality: We also simulated the operation of our scheme when the sum of the required slot utilizations of all video streams is above unity. When the bandwidth resource is insufficient, some video streams can never be serviced using the current BCMCS scheme, which of course causes considerable damage to the average playback quality. However, in our scheme, the available bandwidth is shared across all video streams, and each mobile node experiences only a small loss of quality. Even when the admission of new video streams brings about a shortage of physical-slot resource, admission can still be permitted by dropping the bit rate of the existing

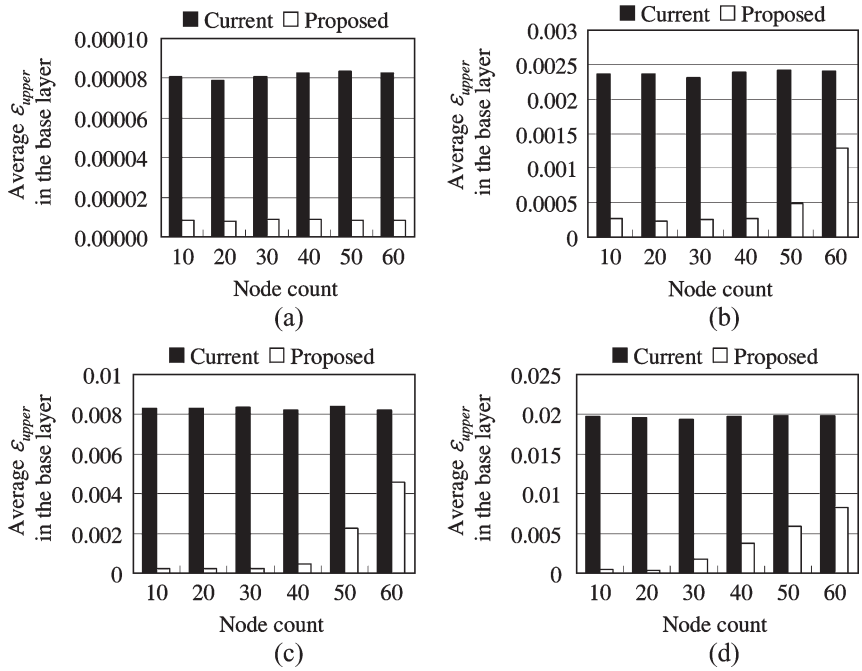

Fig. 10. Average $\varepsilon_{\mathrm{upper}}$ in the base layer when $f_d N_{\mathrm{BL}} T = 0.01$. (a) $\varepsilon_{\mathrm{physical}} = 3\%$. (b) $\varepsilon_{\mathrm{physical}} = 5\%$. (c) $\varepsilon_{\mathrm{physical}} = 7\%$. (d) $\varepsilon_{\mathrm{physical}} = 10\%$.

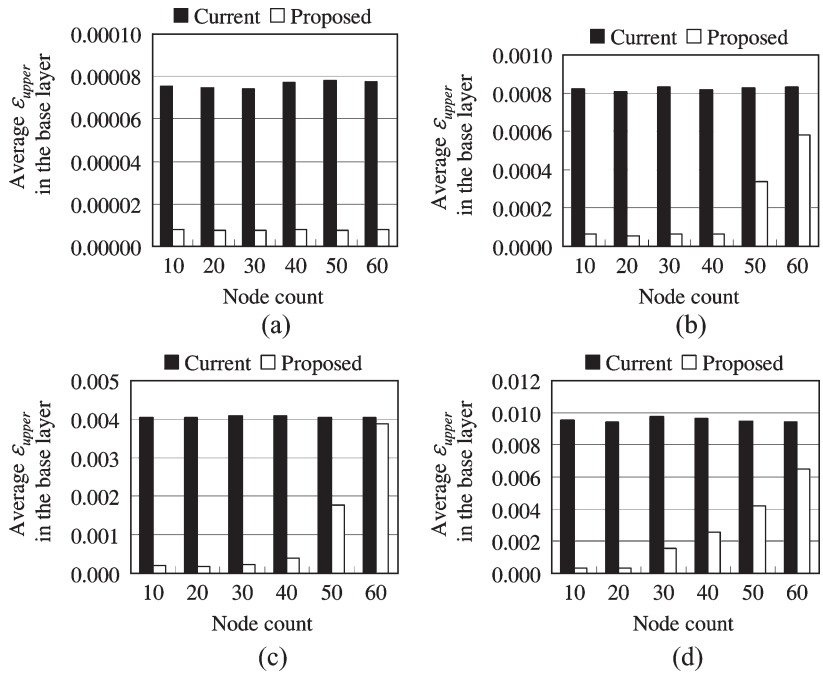

Fig. 11. Average $\varepsilon_{\mathrm{upper}}$ in the base layer when $f_d N_{\mathrm{BL}} T = 0.02$. (a) $\varepsilon_{\mathrm{physical}} = 3\%$. (b) $\varepsilon_{\mathrm{physical}} = 5\%$. (c) $\varepsilon_{\mathrm{physical}} = 7\%$. (d) $\varepsilon_{\mathrm{physical}} = 10\%$.

video streams. Adjusting the bit rate with fine granularity using the MPEG-4 FGS coding makes the proposed scheme very effective.

In these experiments, the following sets of data were used: Set 1 is $\{(\tau_i, N_{\tau_i}) | (\tau_0, 2), (\tau_1, 5), (\tau_2, 12), (\tau_3, 20), (\ddot{\tau}_4, 1 \sim 10)\}$ ($\ddot{\tau}_4$ is a new video flow), and set 2 is $\{(\tau_i, N_{\tau_i}) \mid (\tau_0, 2), (\tau_1, 5), (\tau_2, 12), (\tau_3, 20), (\ddot{\tau}_4, 1), (\ddot{\tau}_5, 1)\}$ ($\ddot{\tau}_4$ and $\ddot{\tau}_5$ are new video flows). Before creating the new video flows, the slot utilization of the current BCMCS scheme is unity, which means that there are no slots available. However, the slot utilization of the proposed scheme is about 0.75 for the same video streams because 25% of the slots are now saved by reclaiming the parity data overhead. These saved slots are used to recover lost packets by retransmission.

In Fig. 13, the three policies proposed in Section III-B are compared with each other and with the current scheme using the experimental data set 1 to show what happens as the number of subscribers to the new video flows increases, when $\varepsilon_{\mathrm{physical}}$ is

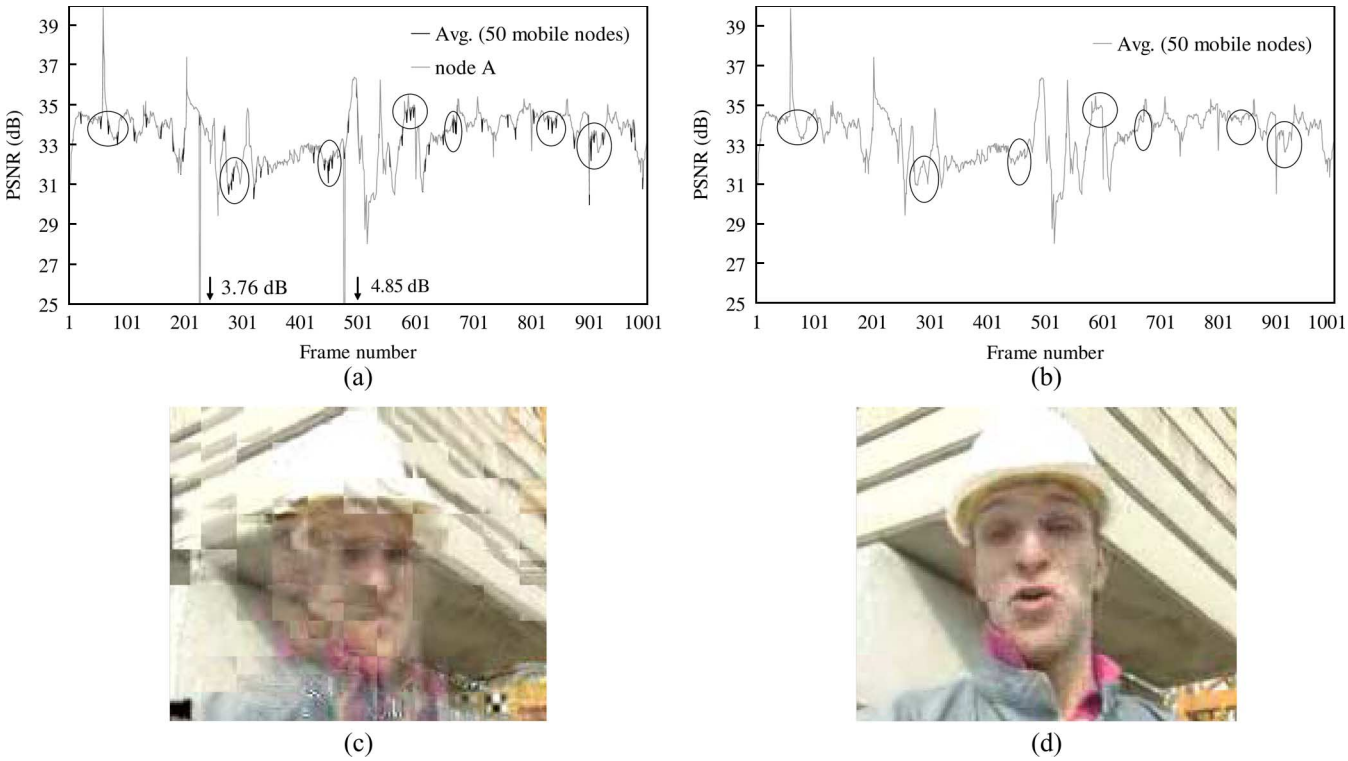

Fig. 12. Comparison of playback quality when $\varepsilon_{\text{physical}} = 5\%$ and $f_d N_{\text{BL}} T = 0.01$. (a) Average PSNR of 50 mobile nodes and of node A in the current scheme. (b) Average PSNR of 50 mobile nodes in the proposed scheme. (c) Captured image using the current scheme when errors occurred. (d) Captured image using the proposed scheme when errors occurred.

1%, 3%, 5%, and 7%, and $f_d N_{\text{BL}} T = 0.02$. As a new video flow is admitted, the average PSNR per flow is somewhat degraded because existing flows share their slot resource with the new flow. However, the overall average is much better than it is for the current scheme, which will never permit a new flow when there are no slots left. When $\varepsilon_{\text{physical}}$ is 5% and the PFD scheme is used, the overall average drops to 33.26 dB for subscriber 5. However, this is still better than the current scheme, which achieves 30.2 dB with the same number of subscribers. The degradation of playback quality in the current scheme grows as the number of subscribers to the new video flows increases and finally reaches 26.5 dB. Similar results are shown in Fig. 13 when $\varepsilon_{\text{physical}}$ is 1%, 3%, and 7%. It is noticeable that the average PSNR is largely unaffected by the value of $\varepsilon_{\text{physical}}$. The effect of not permitting new video flows dominates the overall average performance, and the effect of channel condition is secondary, showing how important it is to share slot bandwidth, as much as possible, so as to prevent precipitous reductions in playback quality when requests for new video flows are made. Fig. 13 also shows a comparison of the three policies.

Although it is a relatively small effect, we can improve the average playback quality by considering the distribution of the mobile nodes which are subscribing to each video flow. The performance of the three policies depends on the dispersion of subscriber numbers to each video flow and on the shape of the curve relating bit rate to PSNR, which is an observation that can be used to optimize the resource allocation to maximize the average playback quality. In the FD scheme, the bit rates of each video stream are adjusted when a request to service a new flow is received; however, subsequently, they are never changed, even if the number of subscribers to that new stream increases. Thus, the resulting average PSNR values are independent of the number of subscribers.

In the VD scheme, the result depends on whether the new video flow is selected to be a victim flow or not. When a new video stream is admitted and the number of subscribers to that stream is small, it is selected as a victim flow, and its bit rate is reduced. Thus, the average PSNR degrades as the number of subscribers increases. However, when the number of subscribers to that new flow exceeds a certain threshold, it will no longer be a victim flow, and then, the average PSNR recovers as the number of subscribers increases, as shown in Fig. 13(b) and (c). In Fig. 13(c), the average PSNR decreases continuously, with no tailing-off. This is because, when $\varepsilon_{\text{physical}} = 7\%$, the new video flow is always selected to be a victim. As the error rate in the channel increases, many reserved slots are required for the retransmission of packets, and thus, the number of slots available for broadcast packets is diminished. This means that more victims, including new video flows, are needed to share their slot resource with the broadcast video flows.

Finally, in the PFD scheme, there is also a threshold beyond which the average PSNR bottoms out as the number of subscribers to a new video flow increases. To service a new video flow with a small number of subscribers (for example, $1 \sim 6$ in $\ddot{\tau}_4$), the bit rate of existing flows with a large number of subscribers (for example, τ_0, τ_1, τ_2, or τ_3) must be cut, reducing the average PSNR. However, as the number of subscribers to the new video flow exceeds the threshold, the effect on average PSNR is reversed.

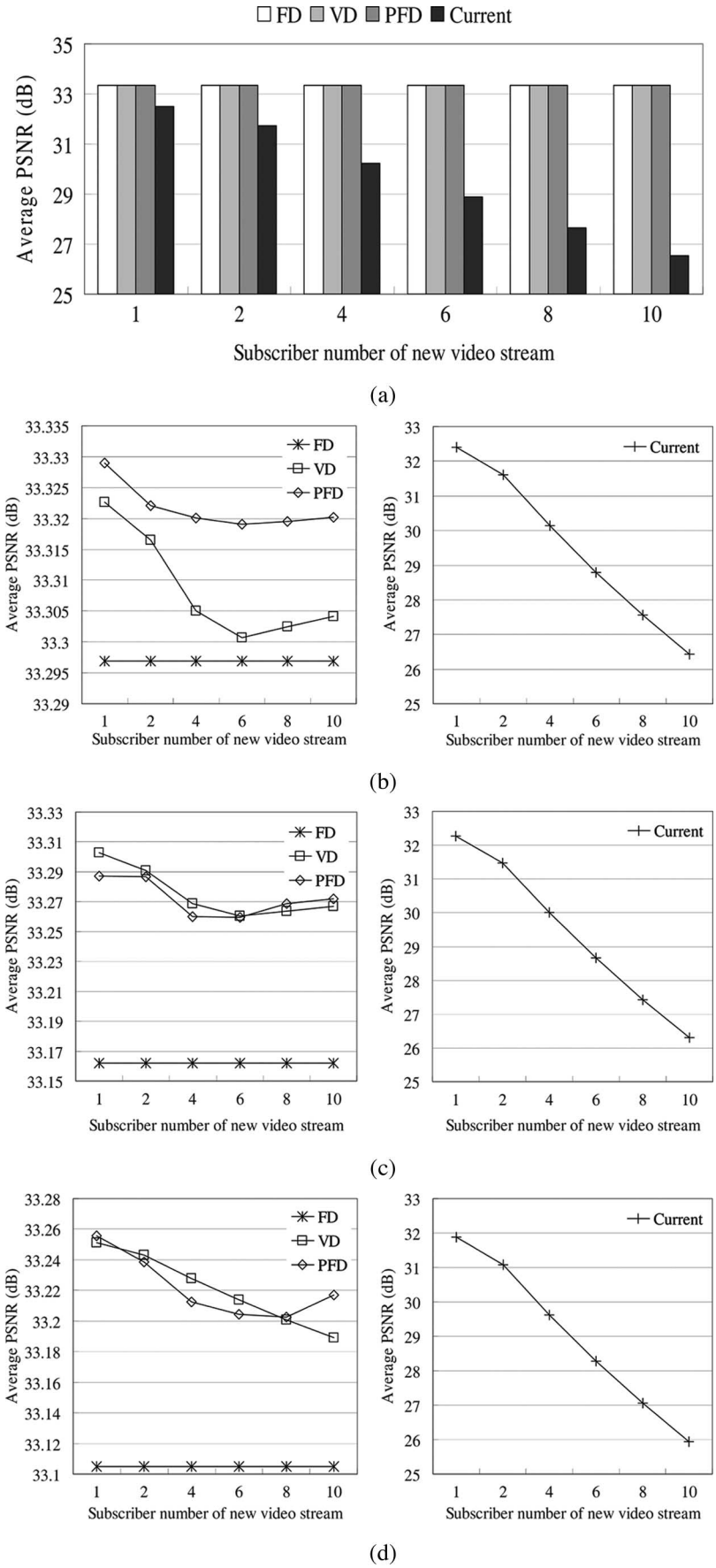

Fig. 13. Effect of adjusting quality as the number of subscribers varies for data set 1. (a) $\varepsilon_{\text{physical}} = 1\%$. (b) $\varepsilon_{\text{physical}} = 3\%$. (c) $\varepsilon_{\text{physical}} = 5\%$. (d) $\varepsilon_{\text{physical}} = 7\%$.

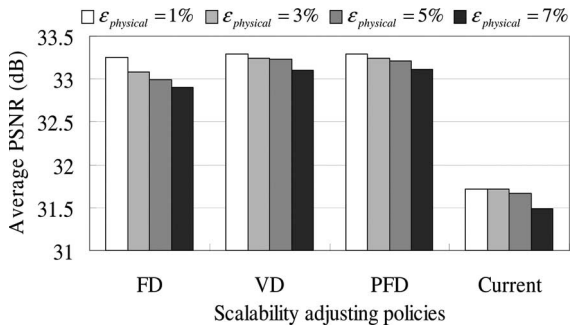

Fig. 14. Effect of adjusting quality for data set 2.

Fig. 14 shows the average PSNR of all mobile nodes for the current and proposed schemes, applying the three scaling policies to data set 2. In this data set, the admission of two additional new video flows to the system is requested. By using the current scheme, the new video flows are not admitted, which means that the subscribers will never receive them. This results in a significant drop in the total throughput, damaging the average PSNR. Although the bit rate of each existing video flow drops more than it does in the case of set 1, the average PSNR is still better than that achieved by the current scheme. The difference in average PSNR between the proposed and current schemes also grows as the number of subscribers to the new video flows increases, showing that it is better to share physical slots than to fail to service a new flow. If the bandwidth requested by video flows fluctuates frequently, our dynamic scheme is decidedly more efficient.

VI. Concluding Remarks

The architecture of BCMCS over 3G cellular systems based on cdma2000 is focused on providing point-to-multipoint transmission of multimedia data, including text, audio, images, and video, from a single source to all users on the network or to a group of users in a service area. However, 2.5G and 3G cellular networks already offer text, audio, images, and short video clips. We therefore focus on video streaming, which is effectively the main contribution of BCMCS. In pursuit of an efficient multimedia streaming broadcast service, we have proposed a dynamic scheduling algorithm based on EDF to handle the situation where broadcast streams are dynamically started and terminated. Our dynamic scheduler cooperates with an error-recovery scheme based on priority-driven ARQ and exploits the scalability characteristics of MPEG-4 FGS. We do not use the RS coding that the current BCMCS employs to conceal errors. This makes it unnecessary to add parity information to the original content stream, and we can make flexible use of the slots we have saved, in particular, to retransmit broadcast packets of higher priority, such as base-layer packets, and to prevent abrupt playback degradation. We preserve playback quality by awareness of the characteristics of an MPEG-4 FGS video stream. The proposed retransmission scheme allocates a higher priority to the base-layer packets of a video stream, giving them more chances of retransmission and improving the likelihood of delivery. As a result, the abrupt distortion of video images at a mobile node can be avoided. It also becomes possible to use the slot resource more efficiently by sharing it with other video streams to achieve better average playback quality at the mobile nodes than the current BCMCS. Extensive simulation has shown the efficiency of our dynamic scheme compared with the current static scheme.

Our scheme follows the specifications for unicast and broadcast services in cdma2000 networks, and our scheduler is simply using them more efficiently to improve performance. In WCDMA, an MBMS with an architecture functionally similar to BCMCS has been introduced for efficient support of broadcast and multicast transport in mobile networks. Therefore, an essentially similar approach can be applied to the WCDMA systems. However, other approaches in improving the performance of WCDMA have been suggested. The packet downlink ACK/NACK (PDAN) mode combines RLC/MAC with an ARQ scheme, and blind repetition uses RLC/MAC without ARQ. In the PDAN scheme, session feedback is required from up to 16 terminals in a given cell. We could apply our scheme to WCDMA, but more uplink capacity would be required.

References

[1] A. Boni, E. Launay, T. Mienville, and P. Stuckmann, "Multimedia broadcast multicast service—Technology overview and service aspects," in *Proc. IEE Int. Conf. 3G Mobile Commun. Technol.*, 2004, pp. 634–638.
[2] M. Bakhuizen and U. Horn, "Mobile broadcast/multicast in mobile networks," *Ericsson Rev.*, vol. 82, no. 1, pp. 6–13, Oct. 2005.
[3] *CDMA2000 High Rate Broadcast–Multicast Packet Data Air Interface Specification,* Rev. 1.0, 3GPP2 Std. C.S0054, Mar. 2005.
[4] P. Agashe, R. Rezaiifar, and P. Bender, "CDMA2000 high rate broadcast packet data air interface design," *IEEE Commun. Mag.*, vol. 42, no. 2, pp. 83–89, Feb. 2004.
[5] F. Cottet, J. Delacroix, C. Kaiser, and Z. Mammeri, *Scheduling in Real-Time Systems*. Hoboken, NJ: Wiley, Oct. 2002.
[6] *Coding of Audio-Visual Objects-Part2: Visual,* ISO/IEC Std. 14 496-2, May 2004.
[7] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 301–317, Mar. 2001.
[8] F. Wu, S. Li, and Y. Q. Zhang, "A framework for efficient progressive fine granularity scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 332–344, Mar. 2001.
[9] *Broadcast–Multicast Services (BCMCS) Framework Draft Document,* Rev. 0.1.3, 3GPP2 Std. X.P0019, Aug. 2003.
[10] J. Wang, R. Sinnarajah, T. Chen, Y. Wei, and E. Tiedemann, "Broadcast and multicast services in cdma2000," *IEEE Commun. Mag.*, vol. 42, no. 2, pp. 76–82, Feb. 2004.
[11] W. J. Ebel and W. H. Tranter, "The performance of Reed–Solomon codes on a bursty-noise channel," *IEEE Trans. Commun.*, vol. 43, no. 234, pp. 298–306, Feb./Mar./Apr. 1995.
[12] *Broadcast–Multicast Services Security Framework,* Rev. 1.0, 3GPP2 Std. S.R0083, Oct. 2003.
[13] *CDMA2000 High Rate Packet Data Air Interface Specification,* Rev. 3.0, 3GPP2 Std. C.S0024, Dec. 2001.
[14] R. Parry, "cdma2000 1×EV-DO [for 3G communications]," *IEEE Potentials*, vol. 21, no. 4, pp. 10–13, Oct./Nov. 2001.
[15] E. Esteves, "The high data rate evolution of the cdma2000 cellular system," in *Multiaccess, Mobility and Teletraffic for Wireless Communications*, vol. 5. Norwell, MA: Kluwer, Dec. 2000, pp. 61–72.
[16] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "Cdma/hdr: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, vol. 38, no. 7, pp. 70–77, Jul. 2000.
[17] W. Li, "Bit-plane coding of DCT coefficients for fine granularity scalability," in *Proc. Contribution 45th MPEG Meeting*, Atlantic City, NJ, Oct. 1998. m3989.
[18] W. Li and Y. Chen, "Experiment result on fine granularity scalability," in *Proc. Contribution 46th MPEG Meeting*, Seoul, Korea, Mar. 1999. m4473.

[19] K. Kang, J. Cho, and H. Shin, "Dynamic packet scheduling for cdma2000 1×EV-DO broadcast and multicast services," in *Proc. IEEE Wireless Commun. and Netw. Conf.*, Mar. 2005, vol. 4, pp. 2393–2399.
[20] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA–HDR: A high efficiency—High data rate personal communication wireless system," in *Proc. IEEE Veh. Technol. Conf.*, May 2000, vol. 3, pp. 1854–1858.
[21] V. K. N. Lau, "Proportional fair space-time scheduling for wireless communications," *IEEE Trans. Commun.*, vol. 53, no. 8, pp. 1353–1360, Aug. 2005.
[22] K. Kang, Y. Cho, and H. Shin, "Performance analysis of dynamic packet scheduling within a cdma2000 broadcast networks," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2005, vol. 4, pp. 2586–2590.
[23] M. Spuri and G. Buttazzo, "Scheduling aperiodic tasks in dynamic priority systems," *Real-Time Syst.*, vol. 10, no. 2, pp. 179–210, Mar. 1996.
[24] B. K. Choi, D. Xuan, R. Bettati, W. Zhao, and C. Li, "Utilization-based admission control for scalable real-time communication," *Real-Time Syst.*, vol. 24, no. 2, pp. 171–202, Mar. 2003.
[25] K. Kang, J. Cho, Y. Cho, and H. Shin, "Dynamic scheduling for scalable media transmission over cdma2000 1×EV-DO broadcast and multicast networks," *Lecture Notes in Computer Science*, vol. 3452, pp. 968–979, Apr. 2005.
[26] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Reading, MA: Addison-Wesley, 1992.
[27] *Generic Coding of Moving Pictures and Associated Audio Information,* ISO/IEC Std. 13 818, 2004.
[28] L. Georgiadis, R. Guerin, and A. Parekh, "Optimal multiplexing on a single link: Delay and buffer requirements," *IEEE Trans. Inf. Theory*, vol. 43, no. 5, pp. 1518–1535, Sep. 1997.
[29] J. Liebeherr, D. Wrege, and D. Ferrari, "Exact admission control for networks with a bounded delay service," *IEEE/ACM Trans. Netw.*, vol. 4, no. 6, pp. 885–901, Dec. 1996.
[30] D. Ferrari and D. C. Verma, "A scheme for real-time channel establishment in wide-area networks," *IEEE J. Sel. Areas Commun.*, vol. 8, no. 3, pp. 368–379, Apr. 1990.
[31] M. Zorzi and R. R. Rao, "On the statistics of block errors in bursty channels," *IEEE Trans. Commun.*, vol. 45, no. 6, pp. 660–667, Jun. 1997.
[32] M. Zorzi, R. R. Rao, and L. B. Milstein, "Error statistics in data transmission over fading channels," *IEEE Trans. Commun.*, vol. 46, no. 11, pp. 1468–1477, Nov. 1998.
[33] W. C. Jakes, *Microwave Mobile Communications*. Hoboken, NJ: Wiley, 1974.
[34] J. G. Proakis, *Digital Communications*, 2nd ed. New York: McGraw-Hill, 1989.
[35] A. Pearmain, A. Carvalho, A. Hamosfakidis, and J. Cosmas, "The MoMuSys MPEG-4 mobile multimedia terminal," in *Proc. 3rd ACTS Mobile Summit Conf.*, Jun. 1998, pp. 224–229.

Kyungtae Kang (M'02) received the B.S. and M.S. degrees in computer engineering from Seoul National University, Seoul, Korea, in 1999 and 2001, respectively, where he is currently working toward the Ph.D. degree in the School of Electrical Engineering and Computer Science.

His research interests include packet scheduling, error control, quality-of-service provision, and energy minimization issues in next-generation wireless/mobile networks. In particular, he is researching the performance and energy requirements of 3G cellular broadcast services such as broadcast and multicast services and multimedia broadcast and multicast service.

Yongwoo Cho (S'06) received the Premedical degree from the College of Medicine, University of Ulsan, Seoul, Korea, in 1997, the B.S. degree in computer science from Korea National Open University, Seoul, in 2004, and the M.S. degree in electrical engineering and computer science from Seoul National University, Seoul, in 2006, where he is currently working toward the Ph.D. degree in the School of Electrical Engineering and Computer Science.

He worked as a Researcher with the Dooin Corporation and as a General Manager with Bluecord Technology, Inc. His primary interests include multimedia systems, digital broadcasting, next-generation wireless/mobile networks, error control, real-time computing, and low-power design.

Jinsung Cho (M'01) received the B.S., M.S., and Ph.D. degrees in computer engineering from Seoul National University, Seoul, Korea, in 1992, 1994, and 2000, respectively.

He is currently an Assistant Professor with the School of Electronics and Information, Kyung Hee University, Youngin, Korea. His research interests include mobile computing and communications, embedded systems, and sensor networks.

Heonshik Shin (M'88) received the B.S. degree in applied physics from Seoul National University, Seoul, Korea, in 1973 and the Ph.D. degree in computer engineering from the University of Texas, Austin, in 1985.

He has actively involved himself in research on various topics, ranging from real-time computing and distributed computing to mobile systems and software. He is currently a Professor with the School of Computer Science and Engineering, Seoul National University.