Resource Allocation with Service Differentiation for Wireless Video Transmission

Hai Jiang, *Student Member, IEEE,* and Weihua Zhuang, *Senior Member, IEEE*

***Abstract*— The next generation wireless networks need to support video traffic. A major challenge in video services over wireless networks is quality of service (QoS) provisioning. Service differentiation is a good approach for QoS provisioning to video traffic. In this paper, we propose a cross-layer protocol stack architecture for wireless video transmission with service differentiation. In the cross-layer architecture, the application layer provides the lower link layer with the video compression information. Using the information, a dynamic-weight generalized processor sharing (DWGPS) discipline is proposed for the link layer resource allocation. The link layer tries to provide the application layer with a stringent delay bound and strong protection to high priority traffic in the case of resource shortage. Acceptable level of fairness can be achieved by DWGPS. A scheduling procedure for DWGPS is presented, which avoids complex per-packet virtual time calculation. It is shown that DWGPS can automatically adapt to multiuser diversity without many modifications. Simulation results demonstrate the effectiveness and efficiency of the link-layer DWGPS resource allocation.**

***Index Terms*— Cross-layer design, multiuser diversity, quality of service (QoS), resource allocation, wireless video transmission.**

I. Introduction

With the rapid growth of the Internet and wireless communications, the demand for fast and location-independent mobile multimedia services is steadily increasing. Typical applications include videoconferencing, video streaming, distance learning, etc. It is now widely accepted that the next generation wireless access networks are evolving into an all-IP (Internet Protocol) network (referred to as *wireless Internet*), where different wireless access networks are glued together by Mobile IP. Video communications over the wireless Internet have received significant interests from both industry and academia.

One major challenge in video services over the wireless Internet is quality of service (QoS) provisioning. Due to its real-time nature, video services typically require QoS guarantees such as a relatively large bandwidth, a stringent delay bound, and a loss rate bound. This poses significant challenges even in the current wireline Internet, where the real-time applications are supported only with best effort, far from what is desired. The case is worse in a wireless environment, due to the limited radio resources and high transmission bit error rate.

Because of the video decoder's ability to resist packet losses to a certain degree, service differentiation is a good approach for QoS provisioning to video traffic. Provided that the video streams are divided into different classes with different importance (priority) levels, the service differentiation aims at improving the service of high-priority classes and providing overall better than best-effort services. In the literature, a number of resource allocation schemes have been proposed for variable rate video delivery over the Internet, based on some kind of service differentiation. They can be categorized into three groups: channel coding/retransmission/power allocation, buffer management, and priority scheduling.

- Channel coding/retransmission/power allocation: A higher priority layer in a compressed video sequence can be protected more strongly by a more powerful automatic repeat request (ARQ) and/or forward error correction (FEC), while weaker ARQ/FEC may be applied to a lower priority layer [1]. Such a mechanism is called *unequal error protection (UEP)* [2]. UEP can be easily performed with BCH codes [3], convolutional codes [4], [5], or RS codes [2], [6], with different coding rates for different layers. Because of the simple decoder architecture, rate-compatible punctured convolutional (RCPC) codes are also popularly employed to implement UEP [7], [8]. UEP can also be implemented by means of power allocation in code-division multiple access (CDMA) systems, e.g., the transmission power can be managed so that a more important packet will experience a smaller error probability [9]. However, pure ARQ/FEC/power allocation normally deals with one video sequence, not considering resource allocation for multiplexed video traffic over a bandwidth-limited channel as in the case of the wireless Internet.
- Buffer management: This solution is originally proposed for differentiated services (DiffServ) [10], which has emerged as an efficient and scalable solution to ensure Internet QoS. In DiffServ, packets are classified into a limited number of service classes at the edge routers. In a DiffServ core router, packets from different classes are aggregately differentiated by different per-hop behaviors. To provide DiffServ to video traffic, random early detection (RED) based schemes [11] have been proposed to apply assured forwarding (AF) per-hop behaviors [12]. Different RED parameters are applied to different priority layers in video streams so that more important information is protected in case of congestion.

Manuscript received May 11, 2004; revised May 16, 2005; accepted June 6, 2005. The associate editor coordinating the review of this paper and approving it for publication was Y. Fang.

The authors are with the Centre for Wireless Communications, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: {hjiang, wzhuang}@bbcr.uwaterloo.ca).

Digital Object Identifier 10.1109/TWC.2006.04318.

When such multi-level RED schemes are used for video transmission, normally no delay bound is considered. This may bring about severe service degradation to video traffic, as packets with a large delay may be considered useless and discarded in a real-time video sequence.

- Priority scheduling: Higher priority traffic is always served before lower priority traffic [13], [14]. However, priority scheduling can lead to increased burstiness and burst packet loss. Fairness among different traffic flows needs to be considered.

In order to address the issue of provisioning acceptable QoS to multiplexed video traffic over the wireless Internet and achieving an acceptable fairness level, in this paper, a cross-layer protocol stack architecture is proposed for the wireless video transmission. In the cross-layer architecture, the application layer with video compression provides the lower layer with the compression information; on the other hand, the resource allocation in the lower layer is designed such that service differentiation can be provided to different traffic classes in the video compression. To achieve the service differentiation, we propose a dynamic-weight generalized processor sharing (GPS) discipline, which has the ability to: 1) adapt to traffic load variation; 2) make the best use of ARQ; 3) protect the more important packets during resource shortage; 4) achieve fairness among traffic flows; 5) efficiently utilize the available resources; and 6) smoothly incorporate wireless channel state information into resource allocation. As many symbols are used in this paper, Table I summarizes important ones.

The rest of this paper is organized as follows. In Section II, the system model is described and the cross-layer protocol stack architecture is introduced. Section III presents the resource allocation scheme at the link layer. In Section IV, we focus on how the channel quality information can be incorporated into the resource allocation scheme. Finally, Section V concludes this research.

TABLE I
SUMMARY OF IMPORTANT SYMBOLS USED.

Symbol	Definition
$C(k)$	Service capacity (i.e., available LL packet quota) for all video sequences at LL frame k
$C^*(k)$	Remaining service capacity in LL frame k after some packets are scheduled
$C_i^b(k)$	The allocated service capacity to batch i in LL frame k according to fluid-flow-based DWGPS
$C_l^c(k)$	Service capacity assigned to batches in class l in LL frame k
D	Wireless delay bound
F	Threshold of good/bad channel definitions
f_D	Doppler frequency shift
$g_i(k)$	Importance weight of batch i in LL frame k
L	Total class number in DWGPS
L_F	The number of packet slots in a forward link frame
L_R	The number of packet slots in a reverse link frame
$N(k)$	Number of active batches from all the video sequences at the beginning of LL frame k
$n_i^a(k)$	The actually scheduled LL packet number (an integer) in LL frame k from batch i
$n_i^e(k)$	The estimated (real) number of LL packets to be scheduled from batch i in LL frame k
$P_i^b(k)$	Backoff probability of batch i in LL frame k if in bad channel state
P_s	Probability that an LL packet is transmitted successfully
$q(i)$	The class number of batch i
$S_i(k)$	Remaining size of batch i at the beginning of LL frame k
S_l	Payload bit number in an LL packet
T_f	LL frame duration
$T_i(k)$	Remaining timer value of batch i at the beginning of LL frame k
t_m	The threshold in aggressive backoff scheme
$w_l(k)$	Importance weight for class l ($1 \le l \le L$) in LL frame k
$X(l)$	Pre-specified loss expense weight for class l ($1 \le l \le L$)
$\sigma_i(k)$	Credit value of batch i in LL frame k in DWGPS scheduling
σ^p	Credit amount in the credit pool in DWGPS scheduling
$\phi_i(k)$	Weight in DWGPS for active batch i in LL frame k

II. System Architecture

Due to the real-time nature, video delivery over the Internet normally employs unresponsive transport protocols, e.g., the Real-time Transport Protocol (RTP) [15] and/or User Datagram Protocol (UDP) [16]. RTP or UDP does not provide mechanisms to ensure timely delivery or other QoS guarantees, but relies on lower-layer services to do so. Hence, in the wireless Internet, error recovery mechanisms can be introduced at the link layer. Fig. 1 shows a scenario where, via its home base station (BS), a mobile station (MS) is in a unidirectional video transmission from a multimedia server or in bidirectional video transmissions with a correspondence node. The MS employs a cross-layer protocol stack architecture as shown in Fig. 1. In the following, we focus on the interaction between the video compression and the link layer resource allocation and the associated system performance, under the assumption of transparent RTP/UDP/IP layers as they only generate a relatively fixed amount of overhead to the overall system performance.

A. Priority in Video Compression

Consider that the video sequences are coded/decoded by an MPEG-4 compression algorithm. Layered coding has emerged as a popular and effective scheme for video transmission. In layered coding, a raw video sequence is coded into multiple layers: the base layer contains the most important features of the video and can be independently decoded to provide coarse visual quality, while the enhancement layers contain information to further improve the achieved video quality when decoded together with the base layer. In the case of bandwidth shortage, it is desired that the base layer is transmitted with a higher priority.

In MPEG-4, the compressed video frames do not have the same importance as some frames are dependent on others. Standard MPEG encoders generate three types of compressed frames (I, P, or B). An I-frame is intra-coded, having no dependence on any other frames, while P and B-frames are

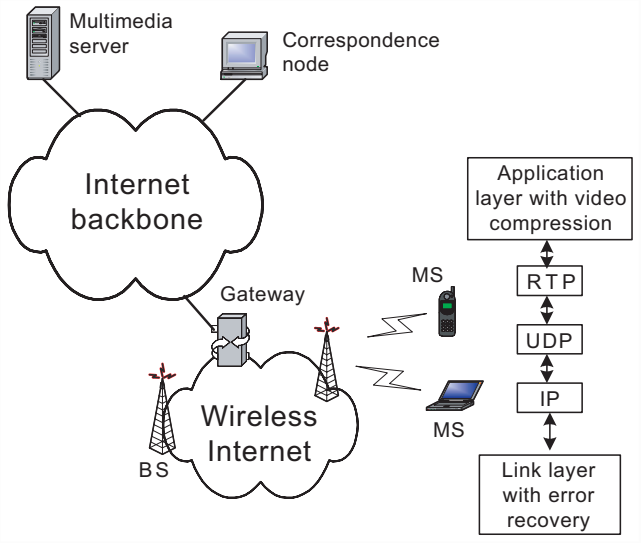

Fig. 1. Video communications over the wireless Internet.

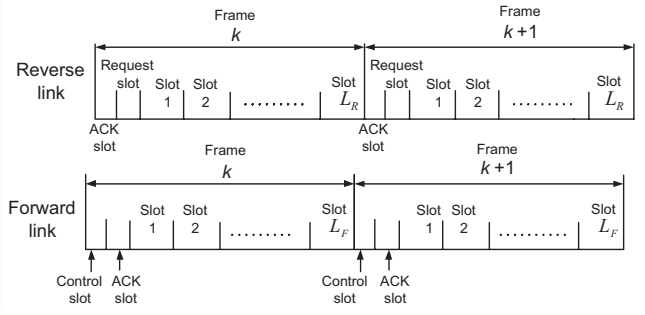

Fig. 2. Time frame structure of the FDD forward and reverse links.

coded with forward prediction and bidirectional prediction, respectively. It is clear that I-frames are the most important, followed by P-frames, and finally B-frames. During network congestion, more important frames should receive more protection.

Furthermore, the recent MPEG-4 Advanced Video Codec (AVC) recommendation (also called H.264) allows the syntax of each slice in a video frame to be separated into up to three partitions (type A, B and C partitions), depending on a categorization of syntax elements. It is shown that type A partition is of the most importance and type C partition the least importance [17].

B. Hybrid TD/CDMA Structure at the Link Layer

At the link layer, a hybrid time-division/code-division multiple access (TD/CDMA) structure is proposed. This structure has the flexibility in time scheduling and can allow simultaneous transmissions from/to multiple MSs. Time is partitioned into fixed length frames in the structure. Fig. 2 shows the time frames of the forward and reverse links operating in frequency-division duplexing (FDD) mode. The BS is responsible for resource allocation in both links. Each reverse link frame consists of an ACK slot, a request slot, and L_R packet slots, while each forward link frame consists of a control slot, an ACK slot, and L_F packet slots. In each type of slots, CDMA multiplexing is used with a fixed spreading gain. Each link layer information packet is transmitted in a packet slot. The request slot is for MSs to initiate a call. The BS responds in the control slot of the forward link. The ACK slot in both links is to send ACKs for packets successfully received in the last frame. The ACK slot in the forward link is also used by the BS to send the information of how many packets are scheduled in the next reverse link frame. The time period of the ACK slot in the forward (reverse) link is in the gap between the two consecutive packet slot clusters in the reverse (forward) link, as shown in Fig. 2. In this way, the ACK for any packet transmitted in a frame is expected to be received before any packet slot of the next frame. Hence, the frame structure allows for virtually instantaneous ACK, which has been shown to be important in traditional ARQ schemes [19]. Real-time video traffic can benefit from virtually instantaneous ACK during packet retransmissions. A similar time frame structure was proposed in [18]. However, it cannot guarantee virtually instantaneous ACK. In the following, we focus on video transmission in the reverse link, as resource allocation in the multiple-access reverse link is much more complex than that in the broadcasting forward link. However, the proposed solution also applies to the forward link video transmission.

Each compressed (base or enhancement) layer of a video frame is segmented into a batch of link layer packets (called LL packets). For a video coder with one base layer and one enhancement layer, two batches of LL packets are generated for each video frame, one for each layer. Upon the arrival of each video frame, the MS creates a transmission queue for each batch of the video frame, assigns a timer with a timeout value to each batch, and reports to the BS the batch class and the batch arrival size in LL packets, for the link layer resource allocation. The batch class and batch arrival size are determined based on the information passed from the video compression, thus our scheme is actually a cross-layer design. The timeout value in the unit of link layer time frame (called LL frame) reflects the maximum tolerant delay over the wireless link. The timer will decrease by one after every LL frame. If the timer expires, any LL packets remaining in the associated batch transmission queue are considered useless and discarded, and the batch transmission queue is deleted.

In order to reduce the LL packet loss rate seen by the higher layers, we introduce a hybrid ARQ/FEC for wireless transmission error recovery. We choose BCH code for FEC. Traditional ARQ schemes over the link layer allow for a fixed number of retransmissions. The allowed retransmission limit should be designed according to the network congestion status and the delay bound. There is also research on performance evaluation for different allowed retransmission attempts [20]. Here, we propose a new retransmission strategy in our system. If an LL packet is transmitted successfully, an ACK will be received before any packet slot of next LL frame, and the packet will be removed from the transmission queue at the MS side; otherwise, this LL packet will remain in the queue until a successful retransmission or a timeout event of the batch timer. This retransmission strategy makes better use of ARQ than

traditional strategies, because it allows the maximum number of retransmissions limited by the delay bound and available network capacity.

For the reverse link transmission, a simple power allocation strategy is applied to ensure the same target bit error rate for all the MSs. This means the same received signal bit energy to interference-plus-noise density ratio (SINR) level for all the transmitted LL packets. As a result, each LL packet is received successfully with a probability P_s, which is a function of the SINR value, the modulation scheme, and FEC mechanism, etc.

III. Link-Layer Resource Allocation

Because of the bursty nature of multimedia traffic and the limited radio resources, a flexible resource allocation scheme which can efficiently accommodate multimedia traffic flows at the link layer is required. The design of an efficient resource allocation scheme should take into account the QoS requirements (in terms of delay and packet loss), traffic priorities, resource availability, congestion management, and fairness.

For resource allocation, the well-known GPS [21] is an ideal fair scheduling discipline, originally proposed for wireline networks. The basic principle of GPS is to assign a fixed weight to each session, and allocate bandwidth for all the sessions according to their weights and traffic load. For a GPS server with N sessions and service rate capacity C, each backlogged session i is guaranteed a minimum service rate $r_i^m = \frac{\phi_i}{\sum_{j=1}^{N} \phi_j} \cdot C$, where ϕ_i is the weight for session i. At time t, the instantaneous service rate of backlogged session i is $r_i(t) = \frac{\phi_i}{\sum_{j \in B(t)} \phi_j} \cdot C$, where $B(t)$ denotes the set of the backlogged sessions at time t.

A tight delay bound can be guaranteed by the GPS server for each session if its traffic is shaped by a leaky-bucket regulator [22]. The minimum service rate and tight delay bound guarantees in GPS may seem attractive to real-time video transmission. However, as compressed video traffic is usually bursty, its peak rate is likely to be much greater than its average rate. If GPS is used, a large weight should be assigned to a video session in order to guarantee the peak rate. This means a video session will get a large portion of the total capacity whenever it has traffic to transmit, thus leading to service degradation of other sessions. On the other hand, if the peak rate cannot be guaranteed, the delay bound of video traffic cannot be guaranteed either, because of the latency in the leaky-bucket regulator. To address the problem of applying GPS to video transmission and extend it to wireless networks, we propose to use dynamic weights in GPS, and call the discipline dynamic-weight GPS (DWGPS).

A. DWGPS

In DWGPS, a "session" is defined as an active batch in the transmission queue. So, a video sequence may have multiple "sessions" simultaneously. The service capacity of DWGPS is defined as available LL packet quota (i.e., the total number of LL packets that can be scheduled) in an LL frame for all the video sequences, denoted by $C(k)$ for LL frame k. Assume $N(k)$ active batches at the beginning of LL frame k from all the video sequences. At LL frame k, an active batch i is assigned a DWGPS weight

$$\phi_i(k) = g_i(k)\frac{S_i(k)}{T_i(k)}, \quad 1 \le i \le N(k) \tag{1}$$

where $g_i(k)$ is the importance weight of batch i at LL frame k, $S_i(k)$ and $T_i(k)$ are the remaining size and remaining timer value of batch i at the beginning of LL frame k, respectively. Equation (1) is based on two facts: the larger a batch's remaining size, the more capacity it requires; and the smaller the batch's timer value, the more urgent the batch's delivery [18].

When generated, each batch is classified into one of total L classes (numbered from 1 to L) according to its importance, where class L has the highest priority and requires the maximum protection. In LL frame k, the weight $g_i(k)$ is determined based on the batch's class, chosen from $\{w_1(k), w_2(k), ..., w_L(k)\}$, where $w_1(k) \le w_2(k) \le ... \le w_L(k)$. Generally speaking, $g_i(k) = w_l(k)$, if batch i is in class l. From (1), it is clear that a batch from a higher priority class will be assigned a relatively larger weight in DWGPS, corresponding to a relatively higher transmission rate of this batch than those of the lower priority classes, so as to better protect higher priority classes during capacity shortage.

In (1), $\frac{S_i(k)}{T_i(k)}$ is in fact the average service capacity amount required by batch i in each subsequent frame before the batch times out. With the weight proportional to its average required capacity portion, each batch is expected to be served smoothly rather than in burst during the delay bound. If a batch is expected to transmit all backlogged LL packets before it times out, all other batches in the same class are expected to do the same. Similarly, if a batch is expected to lose a portion of LL packets, all other batches in the same class are expected to have the same share of packet loss. By the means, fairness can be achieved among different traffic flows. In this work, fairness means that all batches in the same class deliver successfully a similar portion of their arrival traffic (thus leading to a similar packet loss rate).

B. Weight Selection in DWGPS

In the weight definition (1) for batch i, $S_i(k)$ and $T_i(k)$ are known to the scheduler at the BS. Here we need to determine the importance weight vector $\mathcal{W}(k)$=$(w_1(k), w_2(k), ..., w_L(k))$ for $g_i(k)$. A simple method is to assign a constant value to each class. However, taking into account the bursty nature of multimedia traffic and the time-varying characteristic of available service capacity, a constant importance weight vector is not appropriate to different traffic load scenarios. It is desired to determine the importance weight vector based on the available service capacity and traffic backlog status in each LL frame. In DWGPS implementation, we use an optimization approach to obtain the importance weight vector.

For resource allocation, a lost packet of higher class traffic costs more than that of lower class traffic. Accordingly, we specify a loss expense weight $X(l)$ for each batch class l $(1 \le l \le L)$, where $X(1) \le X(2) \le ... \le X(L)$. For LL frame k, a cost function for batch i is defined as

$$F_i(k) = X(q(i)) \cdot \max\{0, S_i(k) - C_i^b(k) \cdot P_s \cdot T_i(k)\} \tag{2}$$

where $q(i)$ is the class number of batch i, and $C_i^b(k)$ is the allocated service capacity to batch i in LL frame k based on the DWGPS discipline. Here we assume that the service capacity is infinitely divisible. We use $C_i^b(k)$ to estimate the average allocated service capacity in the following $T_i(k)$ LL frames before the batch times out. If positive, $S_i(k) - C_i^b(k) \cdot P_s \cdot T_i(k)$ is in fact the estimated LL packet losses when batch i times out. The above estimation accuracy is affected by two factors: 1) batch leaving — if some batches transmit all their LL packets successfully before they time out, their capacity share in the remaining LL frames before timeout should be distributed to other batches; and 2) batch arriving — new arrival batches in later LL frames require their capacity share. Even so, the above estimation should have reasonable accuracy, taking into account that: 1) as the DWGPS weight definition is to make each batch be served smoothly rather than in burst during the delay bound, a batch leaving or arriving does not affect much the available resources for other batches; and 2) the effects of batch leaving and arriving alleviate each other.

Based on each batch's cost function, an optimization problem is formulated in order to minimize the total cost of all the batches in the resource allocation:

$$\underset{\mathcal{W}(k)}{\text{Minimize}} \sum_{i=1}^{N(k)} X(q(i)) \cdot \max\{0, S_i(k) - C_i^b(k) \cdot P_s \cdot T_i(k)\} \tag{3}$$

subject to

$$\begin{cases} C_i^b(k) = \frac{w_{q(i)}(k)\frac{S_i(k)}{T_i(k)}}{\sum_{j=1}^{N(k)} w_{q(j)}(k)\frac{S_j(k)}{T_j(k)}} \cdot C(k), & 1 \le i \le N(k) \\ w_l(k) \ge w_{l-1}(k) \cdot (1+\delta_l) \ge 0, & 1 < l \le L \\ \delta_l \ge 0, & 1 < l \le L. \end{cases} \tag{4}$$

The constant value δ_l should be set to positive if we require an absolute better service (in terms of delay or packet loss rate) for class l than that for class $l-1$, independent of network congestion.

In (4), the expression for $C_i^b(k)$ can be rewritten as

$$C_i^b(k) = \frac{\frac{S_i(k)}{T_i(k)}}{\sum_{j:\ q(j)=q(i)} \frac{S_j(k)}{T_j(k)}} \cdot \frac{w_{q(i)}(k) \sum_{j:\ q(j)=q(i)} \frac{S_j(k)}{T_j(k)}}{\sum_{l'=1}^{L} w_{l'}(k) \sum_{j:\ q(j)=l'} \frac{S_j(k)}{T_j(k)}} \cdot C(k). \tag{5}$$

For class l, the service capacity assigned to batches in this class in frame k can be represented as

$$C_l^c(k) = \frac{w_l(k) \sum_{j:\ q(j)=l} \frac{S_j(k)}{T_j(k)}}{\sum_{l'=1}^{L} w_{l'}(k) \sum_{j:\ q(j)=l'} \frac{S_j(k)}{T_j(k)}} \cdot C(k), \quad 1 \le l \le L. \tag{6}$$

Also, it is clear that

$$\sum_{l=1}^{L} C_l^c(k) = C(k). \tag{7}$$

Equation (5) can be further rewritten as

$$C_i^b(k) = \frac{\frac{S_i(k)}{T_i(k)}}{\sum_{j:\ q(j)=q(i)} \frac{S_j(k)}{T_j(k)}} \cdot C_{q(i)}^c(k). \tag{8}$$

Thus, the optimization problem (3) is equivalent to

$$\underset{(C_1^c(k), C_2^c(k), ..., C_L^c(k))}{\text{Minimize}} \sum_{i=1}^{N(k)} X(q(i)) \cdot \max\{0, S_i(k) - C_i^b(k) \cdot P_s \cdot T_i(k)\} \tag{9}$$

subject to

$$\begin{cases} \sum_{l=1}^{L} C_l^c(k) = C(k) \\ C_i^b(k) = \frac{\frac{S_i(k)}{T_i(k)}}{\sum_{j:\ q(j)=q(i)} \frac{S_j(k)}{T_j(k)}} \cdot C_{q(i)}^c(k), \quad 1 \le i \le N(k) \\ C_l^c(k) \ge 0, \ 1 \le l \le L \\ C_l^c(k) \ge (1+\delta_l) \cdot \frac{\sum_{j:q(j)=l} \frac{S_j(k)}{T_j(k)}}{\sum_{j:q(j)=l-1} \frac{S_j(k)}{T_j(k)}} \cdot C_{l-1}^c(k), \ 1 < l \le L \\ \delta_l \ge 0, \ 1 < l \le L. \end{cases} \tag{10}$$

We constrain the $C_l^c(k)$ values to integers. If taking away the constraints of an absolute better service for class l than class $l-1$ independent of network congestion (i.e., the last two constraints in (10)), this optimization problem can be solved by a dynamic programming approach. The computation complexity is $O(C^2(k))$, assuming the computation complexity of the cost calculation for all the classes is $O(1)$ [23]. If taking into account the last two constraints in (10), we can use an exhaustive search. The total number of searches is equal to the number of ways to put $C(k)$ balls into L bins (where a bin may be assigned 0 or a positive number of balls), $\binom{C(k)+L-1}{L-1}$. However, with the last two constraints in (10), many searching rounds can be skipped. A loose upper bound of the computation complexity for the exhaustive search is $O(\binom{C(k)+L-1}{L-1})$. Normally, only a limited number of traffic classes are considered, e.g., 3 video frame types in a standard MPEG codec, and 3 partition modes in MPEG-4 AVC. So the search procedure complexity in the dynamic programming or the exhaustive search approach is limited.

Note that DWGPS does not require deterministic or statistical knowledge of the incoming traffic. It can adapt to different incoming traffic load scenarios. This is particularly attractive to resource allocation in future all-IP wireless networks, where different application types (e.g., voice, data, video) are multiplexed together, and where it is difficult to predict in advance the arrival traffic rate and available resource capacity for a specific application type.

In addition, the weight selection in DWGPS is quite flexible, depending on the system QoS objectives. The above discussion shows an example to minimize system cost. The weights in DWGPS can also be configured in order to achieve a target ratio of LL packet loss rates of different classes. DWGPS can be extended to serve heterogeneous voice/video/data traffic with different delay bound requirements, where a ratio of LL packet loss rates of voice/video/data traffic is required, or a cost value should be assigned to the loss of LL packets from each traffic type, so as to formulate an optimization problem.

C. DWGPS Scheduling Procedure

One drawback of the GPS principle comes from its two assumptions: infinitely divisible traffic in fluid-flow traffic

model and simultaneously multiple transmissions from multiple sessions. The ideal GPS principle is not realizable in practical systems, especially in a time-division multiple access (TDMA) system where no parallel transmissions are allowed. Some packet-based versions of GPS are proposed for wireline or wireless TDMA systems [21] [24], which use complex per-packet virtual time calculation to determine the transmission order. Furthermore, the virtual time-based GPS discipline is not designed for a network supporting packet retransmission. In the hybrid TD/CDMA structure of our system, multiple transmissions are allowed. We propose a packet-based implementation of DWGPS with low computation complexity and small signaling overhead to approximate ideal fluid-flow-based DWGPS server. It is worth noting that the proposed packet-based implementation is not specific to DWGPS, but can be applied to traditional GPS disciplines over a hybrid TD/CDMA structure.

Consider packet-based DWGPS in the hybrid TD/CDMA, where $C(k)$ service capacity is available for all video MSs in LL frame k and $N(k)$ active batches are from all the MSs at the beginning of LL frame k. Recall that $C_i^b(k)$ is the service capacity amount assigned to batch i in LL frame k according to the fluid-flow-based DWGPS, where each batch may transmit a non-integer number of LL packets in each frame. However, in packet-based DWGPS, each batch can only transmit an integer number of LL packets. We let $n_i^e(k)$ denote the estimated (real) number of LL packets from batch i scheduled to be transmitted in LL frame k. Different from $C_i^b(k)$, when $n_i^e(k)$ is used, we need to compensate for service gains or losses of batch i due to the integer scheduled packet number constraint in previous LL frames. Let $n_i^a(k)$ denote the actually scheduled LL packet number (an integer) in LL frame k from batch i, and $C^*(k)$ the remaining service capacity (i.e., available LL packet quota, an integer) in LL frame k after some packets are scheduled. A credit value $\sigma_i(k) = n_i^e(k) - n_i^a(k)$ is also defined to indicate how much capacity is sacrificed by batch i in LL frame k.

For the first LL frame, k = 1, according to the fluid-flow-based DWGPS discipline, the estimated service capacity for batch i ($1 \le i \le N(1)$) is

$$n_i^e(1) = \frac{\phi_i(1)}{\sum_{j=1}^{N(1)} \phi_j(1)} \cdot C(1) \ . \tag{11}$$

As $n_i^e(1)$ is very likely not an integer, the actually scheduled LL packet number is temporarily set to $n_i^a(1) = \min\{S_i(1), \lfloor n_i^e(1) \rfloor\}$ ($\lfloor \cdot \rfloor$ being the floor function) with a credit $\sigma_i(1) = n_i^e(1) - n_i^a(1)$. The remaining available LL packet quota in LL frame 1 is

$$C^*(1) = C(1) - \sum_{i=1}^{N(1)} n_i^a(1). \tag{12}$$

The $C^*(1)$ LL packet quota is assigned one by one to the $N(1)$ batches based on the descending order of their credit values, and the credit value of each beneficiary batch is decreased by one correspondingly. For LL frame k ($\ge$ 2), the same procedure applies except for the following:

- At the beginning of each LL frame, a batch becomes inactive if its timer value is equal to 0 or all its LL packets are received by the BS successfully. The scheduler sets a credit pool to collect the remaining credit values of the batches that become inactive at the beginning of each frame. The credit amount in the credit pool (denoted σ^p) is distributed to all the active batches, each with an amount proportional to its weight;
- $n_i^e(k) = \frac{\phi_i(k)}{\sum_{j=1}^{N(k)} \phi_j(k)} \cdot C(k) + \sigma_i(k-1) + \frac{\phi_i(k)}{\sum_{j=1}^{N(k)} \phi_j(k)} \cdot \sigma^p$;
- $n_i^a(k)$ is first temporarily set to $\max\{0, \min\{S_i(k), \lfloor n_i^e(k) \rfloor\}\}$;
- If $C^*(k) = C(k) - \sum_{i=1}^{N(k)} n_i^a(k) < 0$, $|C^*(k)|$ LL packet quota is taken away one by one from the batches with positive $n_i^a(k)$ values, based on the ascending order of the corresponding $\sigma_i(k)$ values.

The scheduling procedure for LL frame k is illustrated in Fig. 3. It can be seen that usually $\sigma_i(k)$ is kept in the range $(-1,\ 1)$, which means that the difference of scheduled LL packet number between fluid-flow-based DWGPS and packet-based DWGPS is small.

To implement the scheduling procedure, upon each batch arrival, the MS reports to the BS the batch class and batch size. The report message is transmitted in the reverse link request slot, using a more powerful channel coding technique to avoid transmission collision, or piggybacked at the end of the transmitted reverse link LL packets (if any) to reduce contention in the request slot. Using the batch information collected from the MSs, the BS decides how many LL packets can be scheduled from each batch according to the above procedure, and announces to the MSs via the ACK slot in the forward link. It can be seen that per-batch (rather than per-packet) information is exchanged, kept, and used in the packet scheduler.

D. Performance Evaluation

To evaluate the performance of the proposed DWGPS scheme, it is desired to compare it with other resource allocation schemes. However, in the open literature, there is only limited work providing multiple service priorities in a multiplexing environment. Here we consider the comparison with the multi-level random early detection (MRED) mechanism proposed in [11] and the priority scheduling. MRED refers to an RED configuration with multiple sets of RED parameters to multiple classes such that a higher priority class achieves a lower packet dropping probability. Normally MRED is applied to the IP layer. As a transparent IP layer is assumed in our system, we apply MRED to the link layer instead. For the reverse link transmission, a virtual MRED queue is kept at the BS. In the priority scheduling, packets with higher priority are always served before those with lower priority. Among the same priority traffic class, earliest deadline first (EDF) is applied.

Consider a single cell environment. For the simplicity of presentation, time is measured in the unit of LL frame, each LL frame having a duration of T_f = 10 ms. In the cell, 30 video test sequences (with IDs ranging from 1 to 30) are transmitted from 30 MSs to their correspondence nodes. Each raw video test sequence is in Quarter Common Intermediate Format (QCIF) with a duration of 3000 LL frames at a rate of

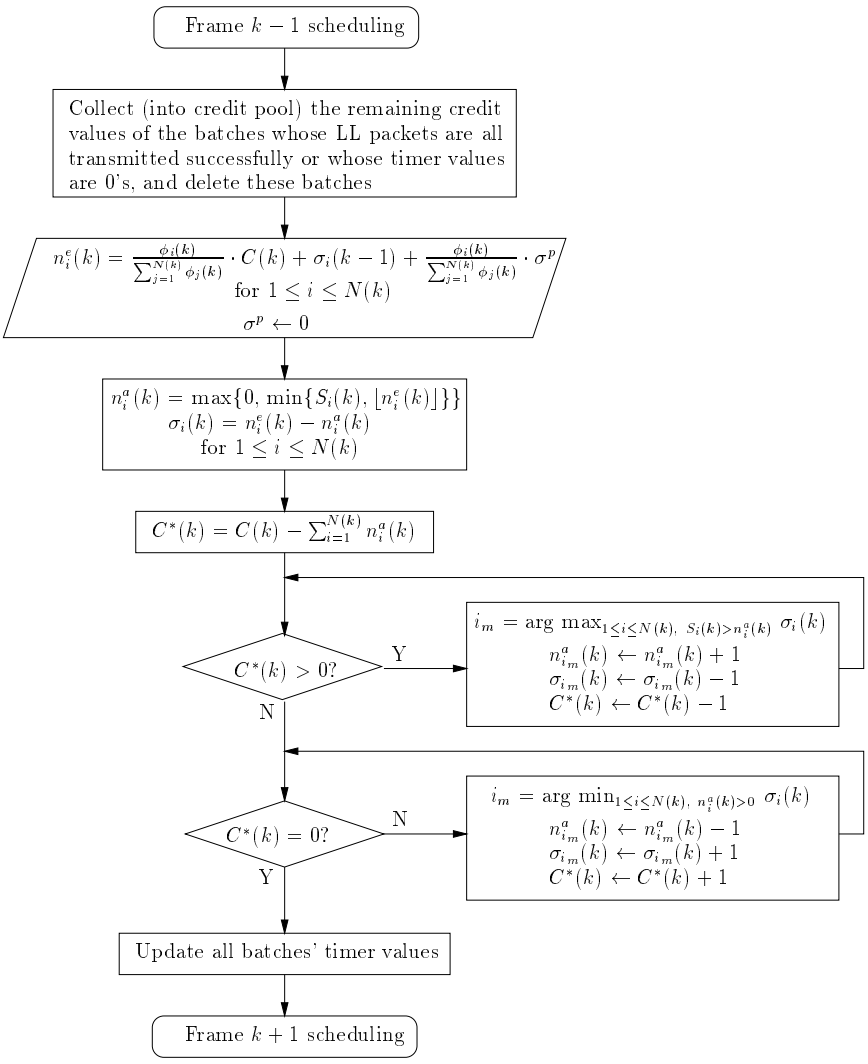

Fig. 3. The DWGPS scheduling procedure for LL frame k.

10 video frames per second, and is compressed by an MPEG-4 coder with a base layer and an enhancement layer. In the base layer, only I-frame and P-frame are used. B-frame is not used due to the additional delay involved in its video compression and decompression process. Hence, there are 3 classes of batches from each video sequence: I-frame batch in the base layer (called IB batch), P-frame batch in the base layer (called PB batch), and batch in the enhancement layer (called E batch). IB batch is class 3 (with the highest priority), while PB and E batches are class 2 and class 1, respectively. In the simulation, we obtain the trace of the encoded IB, PB and E batches' sizes (from the MPEG-4 encoder) of 30 raw video sequences, and configure the traffic arrivals to the MRED, DWGPS, and priority scheduling systems according to the trace record.

At the MS side, each batch is packetized into RTP packets. After the packetizing procedure in the UDP and IP layer, a batch is finally segmented into a number of fixed-size LL packets. Each LL packet is BCH$(224, 192)$ coded for FEC, which means that there are $S_l = 192$ payload bits in an LL packet. A perfect power control strategy is applied such that each LL packet is transmitted with a success probability $P_s = 0.9$. Any LL packet not received successfully is retransmitted repeatedly until an ACK is received or the timer expires. The wireless delay bound is set to $D = 15$ LL frames, which is also the initial value of each timer. The wireline part of the end-to-end transmission is assumed to be error-free with a fixed delay.

In the simulations, the MRED queue size Q is set to be $Q = \max\{C(k)\} \cdot P_s \cdot D$ in order to efficiently utilize the available resources. The MRED parameters ($\min_i$, $\max_i$, $P_{\max_i}$), $i \in \{\text{IB, PB, E}\}$, are $(0.6Q, 0.8Q, 0.025)$, $(0.4Q, 0.6Q, 0.05)$, and $(0.2Q, 0.4Q, 0.1)$ for IB, PB, and E batch classes (as used in [11]), respectively. If the average queue length $\bar{q}$ is less than $\min_i$, all the arrival packets from batch class i are accepted into the queue; if $\bar{q}$ is greater than $\max_i$, all arrival packets from batch class i are rejected; and if $\bar{q}$ is between $\min_i$ and $\max_i$, an arrival packet from batch class i is dropped with probability $P_{\max_i} \cdot (\bar{q} - \min_i)/(\max_i - \min_i)$. In the priority scheduling, an LL packet remains in the transmission queue if it is not transmitted successfully. The sender discards LL packets not successfully transmitted/retransmitted within the delay bound.

The traffic arrival pattern in the simulations is as follows.

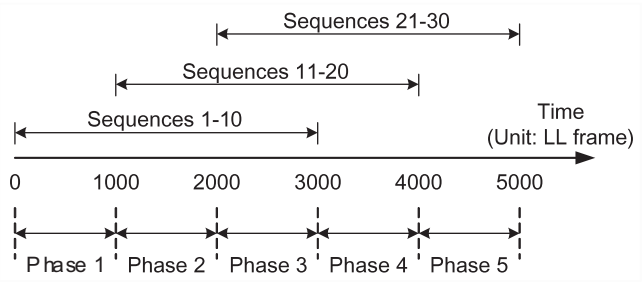

Fig. 4. The traffic load phases in the simulations.

Video sequences 1 - 10, 11 - 20, and 21 - 30 are admitted into the system at LL frame 0, 1000, and 2000, respectively. All the simulations are finished at LL frame 5000. The 5000 simulated LL frames can be divided into 5 phases as shown in Fig. 4. The total average arrival rates of the 3 traffic classes in each phase are shown in Table II, which also summarizes the parameters used in the simulations. Note that DWGPS does not need to know the total average arrival rates. We list them here only for the convenience of the performance evaluation.

For the resource availability, we consider the case with varying available system capacity: All the 30 video sequences are allowed to transmit up to 60, 60, 48, 32, 36 LL packets in each LL frame in phase 1 - 5, respectively. The available capacity $(C(k) \cdot P_s \cdot S_l/T_f)$ for the 5 phases in the unit of kbps is 1037, 1037, 829, 553, and 622, respectively. It can be seen that phase 1 is overprovisioned, phases 2 and 5 are moderately underprovisioned, while phases 3 and 4 are severely underprovisioned.

Table III shows the simulated LL packet loss rates of IB, PB, and E classes for the 5 phases. It can be seen that, MRED, DWGPS and priority scheduling can all guarantee that IB class traffic receives a better service than PB class traffic, which receives a better service than E class traffic. However, in phases 2 - 5, DWGPS and priority scheduling provide much lower LL packet loss rates to IB and PB traffic classes than MRED. From Table III, it can be shown that DWGPS and priority scheduling have the similar packet loss performance while priority scheduling slightly outperforms DWGPS in terms of protection to higher priority traffic. Indeed, priority scheduling provides the best protection to higher priority traffic as it always serves higher priority traffic before lower priority traffic. However, recent work shows that priority scheduling leads to increased burst packet loss [14], thus resulting in resources being unfairly shared by traffic flows.

For resource allocation to video traffic with different priority classes, the fairness can be measured by Fairness Index [25] defined as

$$\mathrm{FI}^j(t) = \frac{(\sum_{i=1}^{n} \rho_i^j)^2}{n \cdot \sum_{i=1}^{n} (\rho_i^j)^2} \qquad j \in \{IB, PB, E\} \qquad (13)$$

for class j at time t, where n is the number of active video sequences at time t, ρ_i^j is the portion of the successfully delivered LL packets of the nearest j-class batch arrived before time t in video sequence i. The higher the Fairness Index value, the better the fairness performance. The upper bound

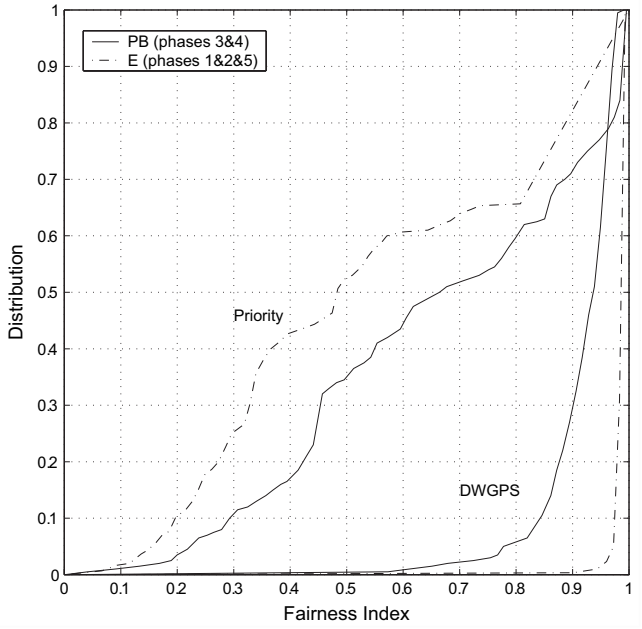

Fig. 5. The Fairness Index distributions of class PB (at phases 3&4) and E (at phases 1&2&5).

of the Fairness Index is 1, which is achieved when ρ_i^j is independent of i.

When priority scheduling or DWGPS is applied, we calculate the Fairness Index for different classes over each LL frame from frame 0 to frame 5000. We find that, the Fairness Index values for IB class at all time or for PB class at phases 1, 2 and 5 (i.e., the overprovisioned and moderately underprovisioned phases) are quite close to 1, as the two classes are protected very well in the resource allocation. Fig. 5 shows the distributions of the Fairness Index values for PB class at phases 3&4 and E class at phases 1&2&5. In phases 3&4, almost all E class traffic is dropped. Hence, it is meaningless to evaluate the fairness level. Here the distribution means the percentage of time that the Fairness Index is smaller than the x-axis coordinate. From Fig. 5, it can be seen that, in the selected phases, approximately half of the Fairness Index values fall into $(0.3, 0.8)$ in priority scheduling, while more than 70% of the Fairness Index values are above 0.9 in DWGPS. That is, DWGPS can achieve much better fairness than priority scheduling.

IV. Adaptation To Multiuser Diversity

So far, we base our model on the perfect power control to combat the wireless channel fading. The transmitted power from each MS is adjusted such that the received SINR is the pre-defined constant value. In the system, the reverse link video transmission can be more efficient if an MS stops transmission when its channel quality is poor, because of three reasons: 1) The total power supply of an MS is limited. If it transmits when the channel quality is poor, the power consumption is large; 2) With power-consuming video compression and decompression, it is desired that an MS uses less power in transmission; 3) With bad channel quality, an MS uses large power in order to achieve the target received

TABLE II
PARAMETERS USED IN SIMULATIONS.

Parameter	Value
Number of video sequences	30
Batch class types	IB, PB, E
Importance weights in LL frame k	$w_{IB}(k) \geq w_{PB}(k) \geq w_E(k)$
LL frame duration (T_f)	10 ms
Number of packet slots per frame in the reverse link (L_R)	8
Wireless delay bound D	15 LL frames
Payload bit number in each LL packet (S_l)	192
Link layer transmission success probability (P_s)	0.9
Traffic load phase	1, 2, 3, 4, 5
Total average arrival rate (in kbps) of IB class in phase 1 - 5	220, 443, 655, 351, 171
Total average arrival rate (in kbps) of PB class in phase 1 - 5	93, 233, 397, 295, 192
Total average arrival rate (in kbps) of E class in phase 1 - 5	653, 1335, 1999, 1345, 721
Capacity value $C(k)$ for LL frame k in phase 1 - 5	60, 60, 48, 32, 36
Capacity (in kbps) for phase 1 - 5	1037, 1037, 829, 553, 622
The same velocity (kilometer/hour) for all MSs	0.5, 15
Doppler frequency shift (f_D) at carrier frequency 2 GHz	0.9 Hz, 27.8 Hz
Channel coherence time	1080 ms, 36 ms
Normalized fading rate ($f_D T_f$)	0.009 (slow fading), 0.278 (fast fading)

TABLE III
LL PACKET LOSS RATES OF IB, PB, AND E CLASSES FOR PHASE 1-5 FOR MRED[11]/DWGPS/PRIORITY SCHEDULING.

	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
$P_e(IB)$ (%)	0/0/0	7.9/0.7/0	22.5/1.6/0	64.8/1.2/0	21.4/0.8/0
$P_e(PB)$ (%)	0/0/0	12.9/1.3/0	58.9/49.9/45.7	82.4/45.0/39.4	30.5/1.7/0.2
$P_e(E)$ (%)	1.2/0/0	69.1/71.2/71.9	99.7/98.2/99.5	99.4/97.4/99.6	70.4/59.9/60.3

SINR at its BS, which leads to large interference to neighbor cells, thus reducing the overall system capacity in a multi-cell environment. Although the time-varying characteristic of a wireless channel is traditionally viewed as a source of unreliability, it can be exploited to address the above issue, similar to what multiuser diversity mechanisms do in a multiple access network.

A. Multiuser Diversity

Multiuser diversity mechanisms are actually a kind of cross-layer design between the link and physical layers. The principle of multiuser diversity is that, for a cellular system with multiple MSs having independent time-varying fading channels, it is very likely that there exists an MS with instantaneous received signal close to its peak value. The overall resource utilization is maximized by providing service at any time only to the MS with the highest instantaneous channel quality [26]. This ideal multiuser diversity strategy aims at maximizing system throughput. To exploit multiuser diversity in a practical communication system, two concerns should be addressed: fairness and delay. The recently proposed wireless fair queuing schemes [24], [27], [28] try to exploit the multiuser diversity mechanism and at the same time provide long-term fairness and delay assurance. In these schemes, all the MSs are divided into two sets: bad channel state MSs if the channel gain is F dB less than the average value, and good channel state MSs otherwise. F is called the good/bad threshold. The scheduler keeps track of the obtained services and channel states of all the MSs. Each MS in a bad channel state stays dormant (i.e., relinquishes its service share to those in a good channel state), and gets compensated when it has a good channel state. This mechanism can work well with fair queuing schemes and obtain long-term fairness based on the fact that an MS in a bad channel state will experience good channel quality in the future, at the cost of a reduced diversity benefit compared with the ideal multiuser diversity mechanism.

In the above multiuser diversity, the channel utilization is improved at the expense of short-term fairness. Although some of them [24], [28] may provide a delay bound assurance, using the GPS-based delay bound in them is not suitable for video delivery as described in Section III. These limitations pose a challenge on applying multiuser diversity to wireless video transmission, where a stringent delay bound is required and short-term fairness should be obtained within the bound. In multiuser diversity, if a video MS is in a bad channel state for a relatively long period, its packets will be discarded as it has to postpone transmission until the channel changes to a good channel state. Consequently, it is not reasonable to always keep an MS dormant when experiencing a bad channel.

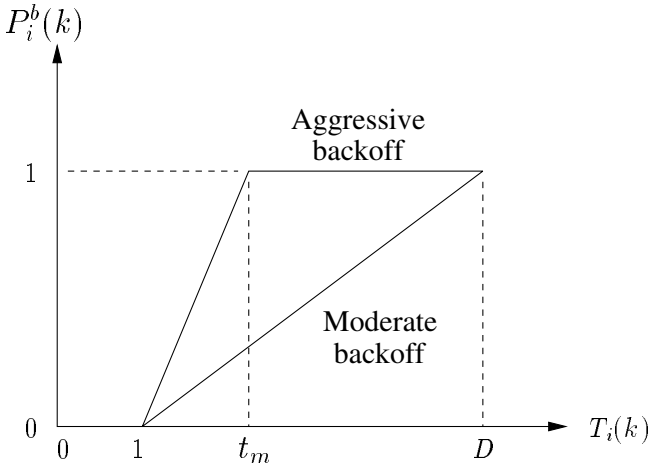

Fig. 6. The backoff probability of moderate and aggressive backoff schemes.

B. Modification to DWGPS for Multiuser Diversity

As the hybrid TD/CDMA supports parallel transmissions, we do not need to constrain our system to schedule only one MS at a time as the traditional multiuser diversity schemes do. To address the short-term fairness and delay issues in traditional multiuser diversity schemes, we propose that, in DWGPS, each batch i with a bad channel at LL frame k is kept dormant with a probability $P_i^b(k)$ (called *backoff probability*) at this frame. For the simplicity of presentation, if an MS is in a bad channel state, we say all of its batches are in a bad channel state. Intuitively the smaller a batch's timer value, the more urgent the batch's transmission, and the smaller the backoff probability for this batch should be. It is worth noting that the relation of backoff probability versus timer should depend on the channel fading rate: if the channel fades fast, we should use relatively large backoff probabilities in a bad channel state with the expectation that the channel quality will get better soon and the affected batches will be compensated soon. Furthermore, the good/bad threshold F should be determined carefully. This threshold affects the probability of a batch being considered in a bad channel state, and indirectly affects the performance of the backoff scheme.

In DWGPS, we consider both a moderate and an aggressive backoff schemes when batch i is in a bad channel state at LL frame k, as shown in Fig. 6.

- Moderate backoff scheme: The backoff probability is set to 1 when the timer value is equal to the wireless delay bound D, and linearly decreased to 0 at timer value 1, i.e.,

$$P_i^b(k) = \frac{T_i(k) - 1}{D - 1}, \quad 1 \le T_i(k) \le D. \tag{14}$$

- Aggressive backoff scheme: The backoff probability is kept at 1 until the timer value decreases to a threshold t_m, from which point the backoff probability is linearly decreased to 0 at timer value 1, i.e.,

$$P_i^b(k) = \begin{cases} 1, & t_m \le T_i(k) \le D \\ \frac{T_i(k)-1}{t_m - 1}, & 1 \le T_i(k) < t_m. \end{cases} \tag{15}$$

These two schemes are to be evaluated for different fading scenarios.

C. Implementation

To implement multiuser diversity in DWGPS, we need to address two issues: 1) how to compensate a batch if it sacrifices access to the resources in the previous LL frames; and 2) the fact that the wireless channel state is not known to the scheduler in advance and cannot be predicted accurately.

Wireless fair queuing schedulers in the literature keep track of excess services obtained by MSs in a good channel state, and compensate for MSs (previously sacrificing service) from the previous beneficiaries. This means that the compensation is not memoryless. However, in DWGPS, a memoryless compensation can be achieved, resulting in a less complex compensation procedure. If batch i keeps dormant in LL frame k when it is experiencing a bad channel state, its remaining size will not change after LL frame k, i.e., $S_i(k+1) = S_i(k)$. In the subsequent LL frames $k' (\ge k+1)$, the batch will have a relatively larger $S_i(k')$ (and thus a relatively larger DWGPS weight $\phi_i(k')$ according to (1)) than that if it did not keep dormant in LL frame k, taking into account the fact that a non-dormant batch's remaining size will decrease. This means that the DWGPS scheme automatically compensates for service degradation of the batches having experienced bad channel quality, resulting in memoryless compensation. The scheduler only needs to determine which batches are to be dormant based on the wireless channel conditions and the backoff probability, and follow the same procedure as that in Section III-C.

In general, the prediction for the wireless channel state is a challenging task. Some proposed approaches make the prediction based on the finite state Markov channel model. For these prediction approaches, there exists a concern about the validity of the mathematical model [24]. As a result, in DWGPS, we adopt a measurement-based prediction, which is model independent. As shown in Fig. 2, each MS transmits ACKs (for forward link transmission) to the BS in the ACK slot of the reverse link frame, which is located at the beginning of each reverse link frame k. The BS measures the wireless channel state (good or bad) of each MS in the ACK slot, and uses this state information to estimate the channel quality when the MS is transmitting in the packet slots of the same reverse link frame k. Also, the BS schedules (the scheduling decision is announced via the ACK slot of forward link frame k, see Fig. 2) the active video MSs to transmit in the first several packet slots of each reverse link frame k, in order to make the estimation more accurate. With a relatively short LL frame length (e.g., T_f = 10 ms), the duration between the ACK slot of the reverse link and each MS's transmission packet slot can be only several milliseconds, less than the channel coherence time (as shown in Table II). It is very likely that the channel good/bad state will not change in this duration, leading to accurate channel state prediction.

D. Performance Evaluation

To evaluate the performance improvement achieved by multiuser diversity, we run computer simulations in the same simulation environment as that described in Section III-D. Other parameters related to MS mobility are given in Table II. All MSs are moving with the same velocity. Assume that the scheduler has knowledge of each MS channel gain due to path

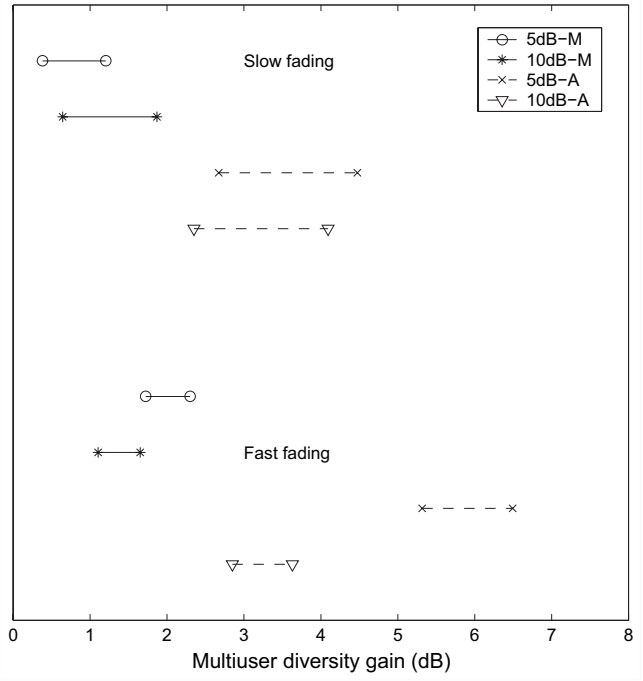

Fig. 7. The 90% confidence interval for mean multiuser diversity gain, where 'M' and 'A' represent moderate and aggressive backoff schemes, respectively.

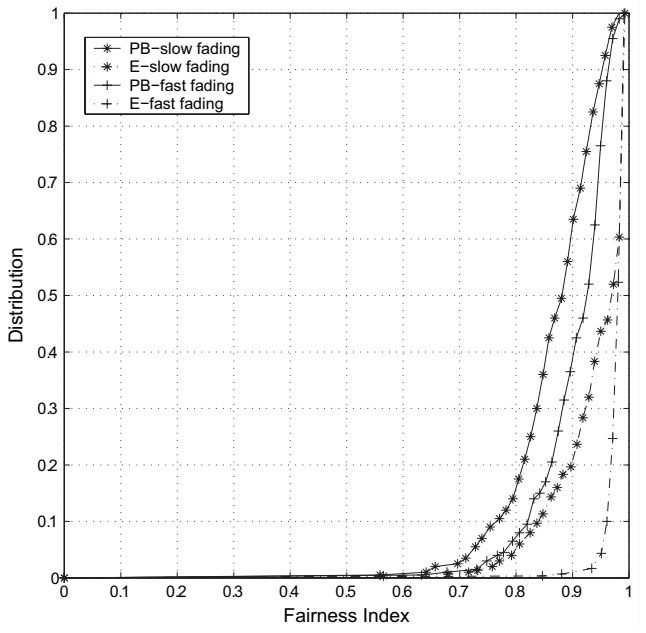

Fig. 8. The Fairness Index distributions of class PB (at phases 3&4) and E (at phases 1&2&5) in slow and fast fading cases with $F = 10$ dB and aggressive backoff.

loss and shadowing effect, which is the local average channel gain for a Rayleigh fading channel. If the scheduler estimates that the instantaneous Rayleigh fading channel gain is F dB less than the average channel gain, the scheduler considers the MS in a bad channel state; otherwise, in a good channel state. We consider two possible F values: $F = 5$ dB or $F = 10$ dB. If an MS is in a bad channel state, the scheduler uses the moderate or aggressive backoff scheme to determine the probabilities to keep the batches of this MS dormant. So, from the same MS, some batches may be dormant, and others may not. In the aggressive backoff scheme, we choose the threshold $t_m = D/3$.

Define a multiuser diversity gain for an MS in a diversity DWGPS scheme as the ratio of average transmission power for an LL packet from the MS in the non-diversity DWGPS scheme to that in the diversity DWGPS scheme. Fig. 7 shows the 90% confidence interval of the mean multiuser diversity gain for $F = 5$ dB/10 dB and moderate/aggressive backoff schemes in the slow fading and fast fading environments. Table IV shows the LL packet loss rates of the IB/PB/E batch classes in the diversity DWGPS environments, in comparison with those in non-diversity DWGPS. It can be seen that the diversity DWGPS does not provide as strong protection to high priority class (i.e., IB and PB) traffic as non-diversity DWGPS does. This is because if a high priority batch keeps dormant when having bad channel quality, part of its sacrificed resources will be allocated to low priority classes. This part of resources cannot be claimed back if most of low priority class traffic is deemed to be lost in the non-diversity DWGPS. Even though this imposes on subsequent LL frames more capacity requirements for high priority traffic, the available capacity may not be sufficient to satisfy the high priority traffic. In comparison with $F = 10$ dB, $F = 5$ dB brings about obvious service degradation in terms of LL packet loss rates to high priority batch classes. This is because $F = 5$ dB leads to a larger probability of an MS being considered in a bad channel state, resulting in a larger probability of a batch being kept dormant, which imposes more capacity requirements on later LL frames. Note that the probability of a channel 10 dB less than average quality is 10%, while the probability of a channel 5 dB less than average quality is 27%. At $F = 10$ dB, the aggressive backoff scheme outperforms the moderate backoff scheme in terms of mean multiuser diversity gain, at the cost of negligible (or non-negligible) service degradation in terms of LL packet loss rates to high priority classes in the fast (or slow) fading environment.

The fairness performance in the fast fading cases are not affected much, as compared with the non-diversity case. This is because in fast fading, sacrificed services can be compensated quickly (likely to be within the delay bound). In the slow fading cases, the fairness performance is slightly degraded. As an example, Fig. 8 shows the distribution of Fairness Index values for the slow fading and fast fading cases at $F = 10$ dB with aggressive backoff.

V. Conclusion

In this paper, we present a cross-layer protocol stack architecture for transmitting compressed video traffic over the wireless Internet. In the cross-layer architecture, the proposed link layer DWGPS resource allocation benefits from the application layer information such as the batch class and batch arrival size; and in return, the link layer tries to provide the application layer with a stringent delay bound and a strong protection to high priority traffic classes in the case of resource shortage. DWGPS also shows good fairness performance. A scheduling

TABLE IV

LL PACKET LOSS RATES OF THE DIVERSITY DWGPS SCHEMES IN THE SLOW AND FAST FADING ENVIRONMENTS WITH PARAMETERS: I) $F = 5$ dB, MODERATE BACKOFF; II) $F = 10$ dB, MODERATE BACKOFF; III) $F = 5$ dB, AGGRESSIVE BACKOFF; AND IV) $F = 10$ dB, AGGRESSIVE BACKOFF.

Traffic load phase	LL packet loss rate	Non-diversity	Diversity in slow fading				Diversity in fast fading			
			I	II	III	IV	I	II	III	IV
Phase 1	$P_e(IB)$ (%)	0	0.1	0	1.9	2.0	0	0	0	0
	$P_e(PB)$ (%)	0	0.1	0	1.0	0.3	0	0	0.1	0
	$P_e(E)$ (%)	0	0	0	6.1	0.9	0	0	0.1	0
Phase 2	$P_e(IB)$ (%)	0.7	1.1	0.8	5.1	2.4	0.9	0.7	1.5	0.7
	$P_e(PB)$ (%)	1.3	3.5	1.8	15.4	4.0	2.5	1.7	6.9	2.1
	$P_e(E)$ (%)	71.2	70.7	71.1	67.5	70.3	71.1	71.2	70.0	71.1
Phase 3	$P_e(IB)$ (%)	1.6	4.1	2.4	12.2	4.2	3.1	2.0	6.8	2.8
	$P_e(PB)$ (%)	49.9	52.1	50.9	52.3	51.2	52.8	51.4	55.5	52.0
	$P_e(E)$ (%)	98.2	97.0	97.8	94.4	97.1	97.2	97.8	95.5	97.4
Phase 4	$P_e(IB)$ (%)	1.2	4.2	2.1	13.9	4.6	2.3	1.7	5.8	2.2
	$P_e(PB)$ (%)	45.0	47.0	45.2	46.2	46.4	47.3	45.9	49.3	47.0
	$P_e(E)$ (%)	97.4	96.2	97.2	93.9	96.3	96.7	97.2	95.4	96.8
Phase 5	$P_e(IB)$ (%)	0.8	2.0	1.1	6.3	2.4	1.4	1.1	3.2	1.5
	$P_e(PB)$ (%)	1.7	3.4	2.1	10.6	4.3	2.1	1.9	3.8	2.2
	$P_e(E)$ (%)	59.9	58.8	59.5	56.6	58.7	59.3	59.7	58.8	59.4

procedure for DWGPS is proposed, which uses only per-batch information and avoids complex per-packet virtual time calculation. The scheduling procedure can also provide helpful insights to the design of a traditional GPS server over a hybrid TD/CDMA system. The DWGPS resource allocation can automatically adapt to multiuser diversity without many modifications. With careful design of the backoff strategy and good/bad threshold, the multiuser diversity gain can be achieved with negligible service degradation in terms of packet loss rates and fairness levels.

ACKNOWLEDGEMENTS

This work was supported by the Premier's Research Excellence Award (PREA) from the Ontario Government and by the Strategic Project (STPGP 257682 - 02) of the Natural Science and Engineering Research Council (NSERC) of Canada.

The authors would like to thank the anonymous reviewers for their constructive comments which improve the presentation of this paper.

REFERENCES

[1] M. Khansari, A. Jalali, E. Dubois, and P. Mermelstein, "Low bit-rate video transmission over fading channels for wireless microcellular systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 1–11, Feb. 1996.

[2] C. Dubuc, D. Boudreau, and F. Patenaude, "The design and simulated performance of a mobile video telephony application for satellite third-generation wireless systems," *IEEE Trans. Multimedia*, vol. 3, pp. 424–431, Dec. 2001.

[3] H. Liu and M. E. Zarki, "Transmission of video telephony images over wireless channels," *Wireless Networks*, vol. 2, pp. 219–228, Aug. 1996.

[4] R. N. Gajaweera, R. M. A. P. Rajatheva, and K. M. Ahmed, "Low bit rate video transmission over correlated frequency selective channel," in *Proc. IEEE GLOBECOM 2001*, pp. 2065–2069.

[5] H. Lee, P. K. Varshney, and W. Ye, "Image transmission with multiuser detection over DS/CDMA channels," in *Proc. IEEE VTC Spring 2001*, pp. 2071–2075.

[6] Q. Zhang, W. Zhu, and Y.-Q. Zhang, "Resource allocation for multimedia streaming over the Internet," *IEEE Trans. Multimedia*, vol. 3, pp. 339–355, Sept. 2001.

[7] J. Hagenauer and T. Stockhammer, "Channel coding and transmission aspects for wireless multimedia," *Proc. of the IEEE*, vol. 87, pp. 1764–1777, Oct. 1999.

[8] L. P. Kondi, S. N. Batalama, D. A. Pados, and A. K. Katsaggelos, "Joint source-channel coding for scalable video over DS-CDMA multipath fading channels," in *Proc. International Conf. Image Processing 2001*, pp. 994–997.

[9] I.-M. Kim and H.-M. Kim, "Efficient power management schemes for video service in CDMA systems," *Electron. Lett.*, vol. 36, pp. 1149–1150, June 2000.

[10] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," IETF RFC 2475, Dec. 1998.

[11] A. Ziviani, J. F. de Rezende, O. C. M. B. Duarte and S. Fdida, "Improving the delivery quality of MPEG video streams by using differentiated services," in *Proc. 2nd European Conference on Universal Multiservice Networks 2002*, pp. 107–115.

[12] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, "Assured forwarding PHB group," IETF RFC 2597, Jun. 1999.

[13] K. L. E. Law, "The bandwidth guaranteed prioritized queuing and its implementations," in *Proc. IEEE Globecom'97*, pp. 1445–1449.

[14] T. Ferrari and P. F. Chimento, "A measurement-based analysis of expedited forwarding PHB mechanisms," in *Proc. IWQOS 2000*, pp. 127–137.

[15] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: a transport protocol for real-time applications," IETF RFC 3550, July 2003.

[16] J. Postel, "User datagram protocol," IETF RFC 768, Aug. 1980.

[17] S. Wenger, "H.264/AVC over IP," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 645–656, July 2003.

[18] I. F. Akyildiz, D. A. Levine, and I. Joe, "A slotted CDMA protocol with BER scheduling for wireless multimedia networks," *IEEE/ACM Trans. Networking*, vol. 7, pp. 146–158, Apr. 1999.

[19] R. Fantacci, "Queuing analysis of the selective repeat automatic repeat request protocol wireless packet networks," *IEEE Trans. Veh. Technol.*, vol. 45, pp. 258–264, May 1996.

[20] A. Chockalingam and G. Bao, "Performance of TCP/RLP protocol stack on correlated fading DS-CDMA wireless links," *IEEE Trans. Veh. Technol.*, pp. 28–33, Jan. 2000.

[21] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Trans. Networking*, vol. 1, pp. 344–357, June 1993.

[22] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the multiple

node case," *IEEE/ACM Trans. Networking*, vol. 2, pp. 137–150, Apr. 1994.

[23] A. Stamoulis, N. D. Sidiropoulos, and G. B. Giannakis, "Time-varying fair queueing scheduling for multicode CDMA based on dynamic programming," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 512–523, Mar. 2004.

[24] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. Networking*, vol. 7, pp. 473–489, Aug. 1999.

[25] R. Jain, A. Durresi, and G. Babic, "Throughput fairness index: an explanation," ATM Forum Document Number: ATM_Forum/99-0045, Feb. 1999.

[26] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[27] P. Bhagwat, P. Bhattacharya, A. Krishna, and S. K. Tripathi, "Enhancing throughput over wireless LANs using channel state dependent packet scheduling," in *Proc. IEEE INFOCOM '96*, vol. 3, pp. 1133–1140.

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

Hai Jiang (S'04) received the B.S. degree in 1995 and the M.S. degree in 1998, both in electronics engineering, from Peking University, Beijing, China. He is currently working toward his Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada. His current research interests include quality-of-service provisioning and resource management for multimedia communications in all-IP wireless networks.

Weihua Zhuang (M'93-SM'01) received the B.Sc. and M.Sc. degrees from Dalian Maritime University, Liaoning, China, and the Ph.D. degree from the University of New Brunswick, Fredericton, NB, Canada, all in electrical engineering. Since October 1993, she has been with the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada, where she is a full professor. She is a co-author of the textbook *Wireless Communications and Networking* (Prentice Hall, 2003). Her current research interests include multimedia wireless communications, wireless networks, and radio positioning. Dr. Zhuang is a licensed Professional Engineer in the Province of Ontario, Canada. She received the Premier's Research Excellence Award (PREA) in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions. She is an Associate Editor of *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, and *EURASIP Journal on Wireless Communications and Networking*.