On Optimal Layering and Bandwidth Allocation for Multisession Video Broadcasting

Jiangchuan Liu, *Member, IEEE*, Bo Li, *Senior Member, IEEE*, Y. Thomas Hou, *Senior Member, IEEE*, and Imrich Chlamtac, *Fellow, IEEE*

***Abstract*—For video broadcasting applications in a wireless environment, layered transmission is an effective approach to support heterogeneous receivers with varying bandwidth requirements. There are several important issues that need to be addressed for such layered video broadcasting systems. At the session level, it is not clear how to allocate bandwidth resources among competing video sessions. For a session with a given bandwidth, questions such as how to set up the video layering structure (i.e., number of layers) and how much bandwidth should be allocated to each layer remain to be answered. The solutions to these questions are further complicated by practical issues such as uneven popularity among video sessions and video layering overhead. This paper presents a systematic study to address these issues for a layered video broadcasting system in a wireless environment. The approach is to employ a generic utility function for each receiver under each video session. They cast the joint problem of layering and bandwidth allocation (among sessions and layers) into an optimization problem of total system utility among all the receivers. By using a simple two-step decomposition of intersession and intrasession optimization, they derive efficient algorithms to solve the optimal layering and bandwidth allocation problem. Practical issues for deploying the optimal algorithm in typical wireless networks are also discussed. Simulation results show that the optimal layering and bandwidth allocation improves the total system utility under various settings.**

***Index Terms*—Broadcast, rate adaptation, resource allocation, video communication.**

I. Introduction

With the proliferation of web-based services and rapid growth of wireless communication devices, layered video broadcasting is becoming an important multimedia application [1], [2]. An important advantage associated with layered video for such broadcast applications is that diverse user access devices can be easily supported; devices (e.g., cellular phone, PDA, laptop) with varying bandwidth and processing capability have the option to subscribe to an appropriate number of layers of a video program (or session) to meet their unique requirements and physical constraints. Hence, a single video session with multiple layers can simultaneously accommodate a group of users with different capacity requirements. As an example, under the *cumulative layered transmission* [2], [3], [28], a raw video is compressed into several layers. The most significant layer, called the *base layer*, contains the data representing the most important features of the video, while additional layers, called *enhancement layers*, contain data that progressively refine the reconstructed video quality. The layers are then distributed to receivers via broadcast channels by a layered transport protocol.

Recent advances in video coding have made it possible to encode video with a very flexible layering structure [12], [17]. In such coders, both the bandwidth of a layer and the number of layers can be dynamically manipulated with a fast response time. In particular, advanced video streaming standards such as the MPEG-4 delivery multimedia integration framework (DMIF) [23] are capable of performing fast layer stream setup and termination at a very low cost. Such flexibility in video coding has enabled further opportunity to deliver video contents with much improved efficiency and performance.

Although layered video has become very successful in a wireless environment, there are still several important issues that remain to be addressed for the delivery of layered video in a broadcast environment. First, there is a critical need for a bandwidth allocation mechanism to allocate bandwidth among video sessions (or programs). In a wireless environment, the total bandwidth is a constrained resource that is shared among competing video sessions [1]. A straightforward approach is to share the total system bandwidth equally among all the sessions. Such an approach, however, is not advisable since each session is of different importance (or value) and should be treated differently in terms of bandwidth allocation. For example, a popular video session attracting a large number of receivers should be allocated with more bandwidth resources (consequently, providing better perceptual quality and more revenue) than a session with few receivers.

Second, for a video session under a given bandwidth budget, it is not clear how the layering structure for this session should be organized. In particular, questions such as how many layers should be generated for this video session and how much bandwidth should be allocated for each layer remain to be answered.

Manuscript received August 7, 2002; revised December 10, 2002 and January 27, 2003; accepted January 27, 2003. The editor coordinating the review of this paper and approving it for publication is V. Leung. The work of J. Liu was supported by a grant from the Research Grants Council under Contract CUHK C001/2050312. The work of B. Li was supported in part by grants from the Research Grants Council under Contracts AoE/E-01/99, HKUST6196/02E, and HKUST6402/03E, an NSFC/RGC joint grant under Contract N_HKUST 605/02, and a grant from Microsoft Research under Contract MCCL02/03.EG01. An earlier version of this paper was presented at IEEE Infocom, San Francisco, CA, March/April, 2003.

J. Liu is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong.

B. Li is with the Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

Y. T. Hou is with the Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 USA.

I. Chlamtac is with the University of Texas at Dallas, Richardson, TX 75083 USA.

Digital Object Identifier 10.1109/TWC.2003.821216

There are several practical issues that need to be considered when addressing the above questions. The first is the *layering overhead*. Under a given session bandwidth, increasing the number of layers means smaller bandwidth for each layer and, hence, finer adaptation granularity on the receiver's side [4]. The drawback here is that more layers will bring in more overhead (for both coding and transport), which diminishes the benefits from the improved granularity in adding more layers [12], [21], [22]. Another issue is that, under a typical wireless broadcast environment, receivers' capacities generally exhibit some kind of *clustered* distribution (instead of uniform distribution). This is because receivers usually use some standard access interfaces. Therefore, if the bandwidth allocation among the layers can explore this property, the mismatch of bandwidth between a receiver's capacity and the layers can be reduced, which can translate into better performance at the receiver's end.

Motivated by these open problems, in this paper, we present a systematic study that addresses the layering and bandwidth allocation (among video sessions and layers) for video broadcasting. Our study explores the flexible and dynamic property of advanced video encoders at the source side to meet the diverse requirements from the receivers. We introduce a generic utility function for each receiver, which takes into account the receiver's physical capacity, actual received bandwidth, and number of received layers. The utility function is designed to be general enough to accommodate various performance measures, e.g., throughput, video's perceptual quality, user satisfaction, and fairness. We show that the layering and bandwidth allocation problem can be formulated into an optimization problem of maximizing the total system utilization, which is a sum of the utilities among all the receivers in the system. By using a simple two-step decomposition of *inter*session and *intra*session allocation, we derive computationally efficient (polynomial time) algorithms for both intersession and intrasession optimization problems. Furthermore, we address some important issues in practice and demonstrate that the optimal allocation algorithm can be implemented with existing layered video coders, where both the computation overhead and deployment complexity are kept at low levels.

To investigate the performance of our optimization algorithms, we conduct performance studies under various settings. Our results offer further understanding on issues such as layering overheads, perceptual video quality, and receiver capacity distribution, and provide practical guidelines on the design of video layering structure for broadcasting.

We organize the rest of the paper as follows. Section II presents some related work. In Section III, we describe the system model and introduce the notion of utility function for our investigation. Section IV formulates the problem of optimal layering and bandwidth allocation for video broadcasting. More importantly, we derive computationally efficient algorithms to solve the optimization problem. Section V discusses some implementation and computation issues. In Section VI, we present numerical results to demonstrate the performance of our optimization algorithms. Finally, Section VII concludes this paper.

II. Related Work

There has been extensive work on layered video transmission for both wired and wireless networks [3], [4], [13], [15]. McCanne *et al.* [3] proposed the first practical receiver-based adaptation algorithm for layered video multicast over the Internet. This algorithm, known as receiver-driven layered multicast (RLM), sends each video layer over a separate multicast group. A receiver periodically joins a higher layer's group to explore the available bandwidth. Since the adaptation is done only on the receiver's side, the granularity is considerably coarse, given that the number of layers is limited and the bandwidth for each layer are predetermined at the source.

To remedy this mismatch between a receiver's capacity and the bandwidth of the video layers, the use of "thin" layers [4] and dynamic layer bandwidth allocation on the sender's side [6]–[9] have been proposed in the literature. Specifically, Shacham [5] presented an optimal layer bandwidth allocation algorithm that maximizes the total utility for all the receivers. It employs an absolute utility function that depends on the received bandwidth. Optimal algorithms using relative utility functions are presented in [6] and [11]. These allocation algorithms assume end-to-end adaptation and focus only on a single session case. Kar *et al.* [8] presented an optimal algorithm that maximizes the total utility for all the receivers belonging to different sessions by employing some intermediaries. In the above optimization schemes, the number of layers is usually assumed to be predetermined. Layering overheads, in particular, the overhead that depends on the number of layers, are not considered. In addition, they are restricted to specific utility functions or have some restrictions on the utility functions that can be used, such as continuous, differentiable, strictly concave or convex.

In mobile wireless networks, the adaptability of layered video is used to trade off the carried traffic and the bandwidth degradation, i.e., minimizing the overload probability of the system by temporally reducing some receivers' subscription levels and, at same time, ensuring the degree of fairness among receivers [13], [15], [16]. Many utility functions have been considered in existing works. However, their optimization objective is different from our problem. For example, they mainly focus on the unicast case rather than broadcast or multicast with heterogeneous receivers as we have addressed in this paper.

Our work in this paper is motivated by these previous efforts. We consider a general utility formulation, which can accommodate different measures such as throughput, video quality, user satisfaction, and fairness. It also takes into account the bandwidth overhead for layered video, as existing experimental results show that such overhead is not negligible in practice [21], [22]. We further consider the optimal allocation for multiple video sessions with heterogeneous popularity. Our optimization algorithm is also very general since it imposes very weak constraints on the utility function.

III. System Modeling and Utility Function

In this section, we describe the system model for our investigation of optimal video layering and bandwidth allocation in a broadcasting wireless environment. We also introduce the notion of a utility function for each receiver, which serves as the fundamental metric for our overall system optimization.

A. System Model

As suggested in [14] and [15], we mainly focus on the adaptation in a wireless local loop or an individual cell in a cellular network. This is because, in current networks, the wireless link bandwidth is much more valuable than the bandwidth of wired links and thus becomes a dominant factor in overall system optimization. In the last two sections, we will also present some discussions on multicell adaptation.

In such a wireless broadcast system, there is a central access point (e.g., the base station of a cell) [1]. A set of video programs (called *sessions*) $\boldsymbol{S}$ are simultaneously distributed to the receivers from this central point, which assigns a total bandwidth to all the video sessions. For each session, the video is partitioned into multiple layers.

The central broadcast point also performs management functions such as user registration and authentication. A receiver who is interested in a particular video session[1] should first send a request to the central point along with a description of its capability (i.e., access capacity). Upon admission into a video session, the receiver will subscribe to a set of cumulative layers (starting from the base layer) commensurate with its capacity. Note that a receiver cannot subscribe to a fraction of a layer. The adaptation granularity on the receiver's side is thus at the layer level, which could result in a mismatch between a receiver's capacity and the video layers if the number of layers is limited. On the other hand, since each video layer is associated with an encoding and transport overhead,[2] for a total session bandwidth budget, increasing the number of layers will lead to bandwidth inefficiency in encoding.

As mentioned earlier, the capacity of the receivers in a network typically follows some clustered distribution. Thus, we let the central point be adaptive in setting the layering structure and bandwidth allocation so as to exploit the clustering property of the receivers' capacity distribution. Specifically, this scheme adaptively determines the number of layers for each session and allocates the bandwidth among sessions as well as layers within a session to maximize the system's performance. This is a viable approach since the central point has complete knowledge of the capacity constraint for each receiver in each session.

In our system model, we use *channel* as the basic unit for bandwidth allocation. A channel in a wireless system represents a fixed unit for data transmission, e.g., a time slot in TDMA systems, a frequency in FDMA systems, or a logical allocation unit such as the logical channel in WCDMA [1]. We further assume that each video layer can occupy only an integral number of channels, and a receiver's capacity is also expressed in the number of channels.

B. Utility Function

A challenging issue for multisession video broadcasting is heterogeneity. First, each receiver has a different capacity, which imposes an upper bound of the video bandwidth it can subscribe to. Second, each video session enjoys different popularity and should thus be treated differently in bandwidth allocation. For example, a video session showing a newly released movie attracting a large number of receivers clearly should have preferential treatment in bandwidth allocation than another less popular video session with few receivers, even if both sessions use similar video coding format. To quantify such heterogeneity among receivers and video sessions, we introduce the notion of *utility function* for each receiver. The total system utility is the sum of the utilities among all the receivers for all the sessions in the system.

There are two categories of utility functions used in the literature. One category can be called "absolute" utility, referring to performance metrics being directly used as the utility function. For example, the video bandwidth delivered to the receiver [9] or the video quality perceived by the receiver [16], which can be measured by the mean-opinion-score (MOS) or the peak signal-to-noise ratio (PSNR) [24]. In general, an absolute function for a given video content is a function of the video bandwidth and, with layered coding, further depends on the number of layers delivered to the receiver [12]. The other category can be called "relative" utility, which is a "transformed" metric to represent a receiver's satisfactory given its expectation. Clearly, a relative utility not only depends on the bandwidth and number of layers delivered to the receiver, but also its expected bandwidth or its own capacity; for example, a typical relative utility function called interreceiver fairness (IRF) is given by the actual received bandwidth of a receiver normalized with respect to its capacity [10].

How a utility function should be exactly defined remains an open research issue. The choice can actually depend on a number of factors (e.g., encoding and transmission algorithms) and, more important, the design objective of the system. For example, from a network provider's perspective, an absolute utility function, such as throughput, is preferable if the revenue is proportional to the total received bandwidth of all the receivers. But such an absolute utility tends to favor broadband receivers over those narrowband receivers. In this case, a relative utility function seems more suitable from a receiver's perspective, in particular, narrowband receivers.

Instead of limiting our scope to a specific absolute or relative utility, we introduce a generic utility function which takes into account several essential parameters for layered video applications. We define the utility for a particular receiver subscribing a video session j, denoted as $\mu_j(k, r, l)$, to be a function of the receiver's capacity k, its actual received video bandwidth r, and the total number of its subscribed layers l.

There are several important advantages of the above framework for utility functions. First, by taking k, r, and l as parameters, our framework can accommodate both absolute and relative utility functions for layered video. For example, an absolute utility that characterizes the perceptual video quality for a receiver can be denoted as $\mu_j(k, r, l) = Q(r, l)$,

[1] We assume that a video program (session) guide is sent to all receivers via a dedicated broadcast channel.

[2] The overhead depends on the number of layers as well as each layer's bandwidth [12], [21], [22].

N : total number of available channels for the system;

S : the set of video sessions in the system;

$M_{j,k}$: the number of receivers that are in session j with a capacity of k channels;

$\mu_j(k,r,l)$: the utility function for a receiver in session j. k is the receiver's capacity, r is its actual received bandwidth, and l is the number of the subscribed layers corresponding to r;

R_j : *layer allocation vector* for session j, $R_j = (r_j^1, r_j^2, \ldots, r_j^{L_j})$;

L_j : the total number of layers in R_j ;

r_j^i : the cumulative bandwidth up to layer i in R_j ;

K_j : the maximum capacity among all receivers in session j;

h : the bandwidth overhead for layering (measured by channel per layer).

$Q(r,l)$: the mapping from layering structure to perceptual video quality (measured by PSNR);

$U_j(n_j)$: the utility of session j under a given session bandwidth budget of n_j channels;

$\hat{U}_j(n_j)$: the optimal utility of session j under a given session bandwidth budget of n_j channels.

Fig. 1. List of notations.

where $Q(r,l)$ is a mapping from the layering structure to the perceptual video quality. On the other hand, an extension of IRF, the *application-aware fairness index*(AFI), has the form of $\mu_j(k,r,l) = Q(r,l)/Q(k,1)$[6], which normalizes the receiver's perceived video quality by its maximum expected quality (a single-layer video with the bandwidth being equal to the receiver's capacity). Second, since the rate-quality relation for video compression depends on the video sequence and the video encoder, it is very difficult to characterize such a relation by using a closed form function for complex video coders (particularly for a layered video coder). Our utility function does not require such explicit characterization. It takes only discrete parameters that are observable from network services. Furthermore, our optimal allocation algorithms do not put any continuity or differentiable constraints on the utility function. As a result, only some sampled points of the function need to be calculated by the layered coder, and a simple table-search algorithm for the prestored values is sufficient for our optimal allocation algorithms.

We now have a 4-tuple, $(N, \boldsymbol{S}, M_{j,k}, \mu_j)$ for our system model, where N is the total number of available channels for our system, $\boldsymbol{S}$ is the set of the sessions (sharing the total bandwidth N) and $M_{j,k}$ is the number of receivers with a capacity of k channels in session $j \in \boldsymbol{S}$. The problem to solve is to find an appropriate layering structure for each session as well as bandwidth allocation among sessions and layers such that the total system utility is maximized.

IV. Optimal Layering and Bandwidth Allocation

In this section, we formally describe the utility-based optimization problem for video layering and bandwidth allocation. We also develop efficient polynomial time algorithms to solve this optimization problem.

A. Mathematical Formulation

Let R_j denote the *layer allocation vector* for video session j, $R_j = (r_j^1, r_j^2, \ldots, r_j^{L_j})$, where L_j is the total number of layers of the allocation vector and r_j^i is the cumulative bandwidth up to layer i. Under a given allocation vector for a session, a receiver shall subscribe to as many layers as possible, subject to its access capacity. That is, a receiver in session j with a capacity of k channels should subscribe to layers $1, 2, \ldots, l_{j,k}^*$, where $l_{j,k}^* = \max\{l | r_j^l \le k\}$. In this case, the cumulative subscription bandwidth for the receiver is $r_{j,k}^* = r_j^{l_{j,k}^*}$, and its utility is thus $\mu_j(k, r_{j,k}^*, l_{j,k}^*)$.

Let system utility be the sum of the utilities among all the receivers in the system. Our objective is to achieve the maximum system utility by properly choosing a layering structure, i.e., the number of layers for each session, and allocating the total bandwidth among the sessions and layers. The notations for this optimal layering and allocation problem are summarized in Fig. 1.

Assuming the *session bandwidth budget* for session j is n_j channels, any possible R_j should therefore satisfy $r_j^{L_j} \le n_j$. In addition, denote K_j as the maximum capacity among all the receivers in session j. Clearly, it does not help to set the cumulative bandwidth to a video layer to be higher than K_j because no receiver can subscribe to such a layer. Finally, if $\mu_j(k, r_j^l, l) \ge \mu_j(k, r_j^{l+1}, l+1)$ for $k \ge r_j^{l+1}$, then the $(l+1)$th layer allocation is not useful since subscription to layer $l+1$ will not further improve the utility to the receiver. Hence, we say that R_j is a *feasible layer allocation vector* if: 1) $L_j > 0$; 2) $0 < r_j^1 < r_j^2 < \cdots < r_j^{L_j} \le \min\{K_j, n_j\}$; and 3) $\mu_j(k, r_j^l, l) < \mu_j(k, r_j^{l+1}, l+1)$ for $k \ge r_j^{l+1}$, $l = 1, 2, \ldots, L_j - 1$.

The input to the optimization algorithm is a 4-tuple $(N, \boldsymbol{S}, M_{j,k}, \mu_j)$, and the output is the maximum system utility U^*, together with the corresponding optimal allocation vectors, $R_j, j \in \boldsymbol{S}$, which gives the bandwidth allocation n_j for each session $j \in \boldsymbol{S}$, the total number of layers (L_j) for session j, and the bandwidth allocation $(r_j^i - r_j^{i-1})$ for each layer i in session j. This optimal layering and bandwidth allocation problem can be formally stated as follows:

$$(\mathrm{OPT-SYS})$$

$$\text{Maximize } U = \sum_{j \in S} \sum_{k=1}^{K_j} M_{j,k} \mu_j(k, r_{j,k}^*, l_{j,k}^*),$$

$$\text{Subject to } R_j \text{is a feasible allocation vector, } j \in \boldsymbol{S}, \text{ and}$$

$$\sum_{j \in S} r_j^{L_j} \leq N. \tag{1}$$

B. Intrasession and Intersession Bandwidth Decomposition

Since there are a finite number of channels in the system and the rate allocations among layers and sessions are in unit of a channel, there is a finite number of feasible rate allocations vectors for the sessions. Therefore, there exists an optimal solution for *OPT-SYS*.

To solve the optimization problem, we first introduce the notion of *session utility* and present a decomposition mechanism for intrasession and intersession allocation. The session utility $U_j(n_j)$ for session j is the total utility of all the receivers in the session under a feasible layer allocation vector R_j, and the *optimal session utility*, $\hat{U}_j(n_j)$, is the maximum of $U_j(n_j)$ among all possible allocation vectors. The following lemma shows that problem *OPT-SYS* can be solved in two steps. First, we perform *optimal intrasession allocation*, which optimally sets the layering structure (i.e., the number of layers in a session) and allocates channels among the layers under each possible session bandwidth budget n_j. Second, we perform *optimal intersession allocation*, which optimally allocates the total system bandwidth N among sessions $j \in \boldsymbol{S}$ based on the results of the optimal intrasession allocation.

Lemma 1 (Decomposition Lemma): For a total number of N channels in the system, the optimal system utility is the maximum of the sum of all the optimal session utilities. That is

$$U^* = \max_{\sum_{j \in \boldsymbol{S}} n_j \leq N} \sum_{j \in \boldsymbol{S}} \hat{U}_j(n_j).$$

Proof: First, by the definition of U^*, we have $U^* \geq \max_{\sum_{j \in \boldsymbol{S}} n_j \leq N} \sum_{j \in \boldsymbol{S}} \hat{U}_j(n_j)$. Second, denote the session bandwidth allocation for U^* as $(n_1^*, n_2^*, \ldots, n_{|\boldsymbol{S}|}^*)$, and the corresponding session utilities $U_1(n_1^*),\ U_2(n_2^*),\ \ldots,\ U_{|\boldsymbol{S}|}(n_{|\boldsymbol{S}|}^*)$. We have

$$\max_{\sum_{j \in \boldsymbol{S}} n_j \leq N} \sum_{j \in \boldsymbol{S}} \hat{U}_j(n_j) \geq \sum_{j \in \boldsymbol{S}} \hat{U}_j(n_j^*) \geq \sum_{j \in \boldsymbol{S}} U_j(n_j^*) = U^*.$$

Combining the two inequalities, we have

$$U^* = \max_{\sum_{j \in \boldsymbol{S}} n_j \leq N} \sum_{j \in \boldsymbol{S}} \hat{U}_j(n_j).$$

The decomposition lemma enables us to solve problem *OPT-SYS* through separate intrasession and intersession allocations. In the following subsection, we describe these two subproblems and present efficient algorithms for each of them.

1) Intrasession Layering and Rate Allocation: For session j, assume the session bandwidth budget is given by n_j channels. The objective of optimal intrasession allocation is to find an appropriate layering structure (i.e., number of layers) and the rate allocation for each layer such that the sum of the utilities among all the receivers in this session is maximized. We formally state the optimal intrasession layering and rate allocation problem as follows:

$$\mathrm{OPT-INTRA}\ (j, n_j)$$

$$\text{Maximize } U_j(n_j) = \sum_{k=1}^{K_j} M_{j,k} \mu_j(k, r_{j,k}^*, l_{j,k}^*),$$

$$\text{Subject to } R_j \text{ is a feasible allocation vector.} \tag{2}$$

We use an iterativ algorithm to solve this problem. The key idea is as follows. Since session j's bandwidth budget is n_j channels and the maximum capacity among all receivers in this session is K_j, and the fact that rate allocation for each layer is an integral number of channels, the number of layers for session j can only take a countable number of values, i.e., $1, 2, \ldots, \min\{n_j, k_j\}$. We start with the one-layer case, i.e., there is only a single layer (base layer) for the session. In this case, the number of channels for this layer can vary from 1 to n_j, and we can easily calculate the utility for each allocation. Then, we add one more layer on top of the one-layer case and calculate the session utility for the two-layer case. In general, upon the rate allocation for the $(l-1)$th layer, the lth layer can be laid on top the $(l-1)$th layer using some remaining channels. Note that when considering an l-layer structure, only the receivers that can subscribe to layer $l-1$ in the previous step may be eligible to subscribe to a higher layer l (due to receiver capacity limitation). Therefore, given the optimal session utility for the case of $l-1$ layers, we only need to add the utility difference of these receivers but not the receivers subscribing to lower layers (1 to $l-2$).

This algorithm can be formally described with a recurrence relation. We define an auxiliary function $\pi(m, l)$ as $\max_{L_j = l,\, r_j^l = m} \sum_{k=1}^{K_j} M_{j,k} \mu_j(k, r_{j,k}^*,\ l_{j,k}^*)$ for $l = 1, 2, \ldots, \min\{n_j, K_j\}$ and $m = 1, 2, \ldots, \min\{n_j, K_j\}$, i.e., the optimal session utility when a total number of l layers are generated and the cumulative bandwidth up to layer l is m channels. The solution to the problem of intrasession allocation *OPT-INTRA* (j, n_j) is clearly given by $\max_{1 \leq l \leq \min\{n_j, K_j\}, 1 \leq m \leq \min\{n_j, K_j\}} \pi(m, l)$. Based on the above discussions, we give a recurrence relation of $\pi(m, l)$ in Fig. 2.

For $n_j > K_j$, i.e., the session bandwidth budget is higher than the maximum receiver capacity, we let $\hat{U}_j(n_j) = \hat{U}_j(K_j)$. Once the optimal session utility is obtained, the corresponding layer allocation vector can be easily obtained by applying a backtracking method on the recurrence relation for $\pi(m, l)$.

The correctness of the above recurrence relation can be proved by induction. For the base case $l = 1$, there is only one layer to be generated with bandwidth m. $\pi(m, 1)$ is thus

(Base case)

For $l=1, m\le\min\{n_j,K_j\}$,

$$\pi(m,1)=\sum\nolimits_{k=1}^{m-1}M_{j,k}\mu_j(k,0,0)+\sum\nolimits_{k=m}^{K_j}M_{j,k}\mu_j(k,m,1);$$

(Recursion)

For $1<l\le\min\{n_j,K_j\}, 1<m\le\min\{n_j,K_j\}$,

$$\pi(m,l)=\max_{1\le i<m}\left\{\pi(i,l-1)+\sum_{k=m}^{K_j}M_{j,k}\Delta(k,m,i,l)\right\},$$

where $\Delta(k,m,i,l)=\mu_j(k,m,l)-\mu_j(k,i,l-1)$;

For any other case, $\pi(m,l)$ is set to 0.

Fig. 2. Algorithm for $\pi(m,l)$ calculation.

$\sum_{k=1}^{m-1}M_{j,k}\mu_j(k,0,0)+\sum_{k=m}^{K_j}M_{j,k}\mu_j(k,m,1)$. The first term is the total utility of the receivers that cannot subscribe to the layer ($k<m$), and the second term is the total utility of all other receivers ($k\ge m$). Note that $\mu_j(k,0,0)$ can be set to a very small value or even a negative value to ensure that, under an optimal allocation, all the receivers can subscribe to at least one layer (the base layer).

For the general case $1<l\le\min\{n_j,K_j\}$, there are l layers to be generated, which can be viewed as adding a new layer to the case with only $l-1$ layers. Without loss of generality, we assume this new layer is layer l, and suppose i is the cumulative bandwidth up to layer $l-1$. All the receivers that subscribe to layer l should have capacities greater than i. Therefore, in the $(l-1)$-layer case, all such receivers should subscribe to layer $l-1$, the highest layer. The difference of the session utility when layer l is generated on top of the $(l-1)$-layer case is thus

$$\sum_{k=m}^{K_j}M_{j,k}[\mu_j(k,m,l)-\mu_j(k,i,l-1)]$$
$$=\sum_{k=m}^{K_j}M_{j,k}\Delta(k,m,i,l).$$

Since $\pi(i,l-1)$ is the optimal session utility for the $(l-1)$-layer case, $\pi(m,l)$ is simply given by

$$\max_{1\le i<m}\left\{\pi(i,l-1)+\sum_{k=m}^{K_j}M_{j,k}\Delta(k,m,i,l)\right\}.$$

2) Intersession Rate Allocation: The objective of intersession bandwidth allocation is to optimally allocate the total N channels in the system to different sessions $j\in\boldsymbol{S}$ so that the system utility is maximized. Given that the optimal session utilities, $\hat{U}_j(n_j)$, $j\in\boldsymbol{S}$, $n_j=1,2,\ldots,K_j$ have been calculated in the optimal intrasession layering and rate allocation, the intersession allocation problem can be stated as follows:

$$(\mathrm{OPT-INTER})\ \text{Maximize } U^*=\sum_{j\varepsilon s}\hat{U}_j(n_j),$$
$$\text{Subject to } n_j>0, j\in\mathbf{S},\ \text{and } \sum_{j\varepsilon s}n_j\le N. \quad (3)$$

This optimization problem can also be solved using an iterative algorithm. We define an auxiliary function $\omega(n,i)$ as $\max_{n=\sum_{j=1}^{i}n_j}\sum_{j=1}^{i}\hat{U}_j(n_j)$ for $n=1,2,\ldots,N$, and $i=1,2,\ldots,|\boldsymbol{S}|$, i.e., the maximum total utility of sessions $1,2,\ldots,i$ when a total bandwidth of n channels are allocated to these sessions. The solution to problem *OPT-INTER* is thus $\max_{1\le n\le\min\{N,\sum_{j\in\boldsymbol{S}}K_j\}}\omega(n,|\boldsymbol{S}|)$. The algorithm in Fig. 3 can be used to calculate $\omega(n,i)$.

(Base case)

$$\omega(n,i)=\begin{cases}\hat{U}_1(n), & \text{for } i=1,\ 1\le n\le\min\{N,K_1\},\\ \sum_{j=1}^{i}\hat{U}_j(K_j), & \text{for } n\ge\sum_{j=1}^{i}K_j;\end{cases}$$

(Recursion) For $1<i\le|\boldsymbol{S}|, 1\le n\le\min\left\{N,\sum_{j\in S}K_j-1\right\}$,

$$\omega(n,i)=\max_{1\le m\le\min\{K_i,n-i+1\}}\left\{\omega(n-m,i-1)+\hat{U}_i(m)\right\};$$

For all other cases, $\omega(n,i)$ is set to 0.

Fig. 3. Algorithm for $\omega(n,i)$. calculation.

C. Complexity of the Algorithms

To perform intersession bandwidth allocation, we should calculate the optimal session utilities for all possible session bandwidth budgets, i.e., solving problems *OPT-INTRA* (j,n_j) for $j=1,2,\ldots,|\boldsymbol{S}|$ and $n_j=1,2,\ldots,K_j$. Note that if $\pi(m,l)$ for $1\le m\le k_j$ and $1\le l\le k_j$ are available, all the above problems can be solved. Fortunately, these values of $\pi(m,l)$ can be obtained using the recurrence relation (see Fig. 2) in polynomial time $O[(K_j)^3\cdot E]$, where E is the time complexity for calculating $\sum_{t=m}^{K_j}M_{j,k}\Delta(k,m,i,l)$.

For the optimal intersession allocation algorithm, when all session utilities are available, its time complexity is bounded by $O(|\boldsymbol{S}|\cdot N\cdot K^{\max})$, where $K^{\max}=\max_{j\in S}K_j$.

We can employ several techniques to speed up the optimization algorithms. For example, an absolute utility function $\mu_j(k,r,l)$ depends only on the receiver's actually received bandwidth r and the corresponding number of layers l but is independent of the receiver's own capacity k. Hence, in Fig. 2, we have

$$\sum_{k=m}^{K_j}M_{j,k}\Delta(k,m,i,l)$$
$$=\Delta(m,m,i,l)\cdot\sum_{k=m}^{K_j}M_{j,k}.$$

Note that $\sum_{k=m}^{K_j}M_{j,k}$, $k=1,2,\ldots,k_j$ are invariants in the execution of the algorithm. They can be precomputed and stored with $O(K_j)$ number of entries. Therefore, E is $O(1)$ and the time complexity of the optimal intrasession allocation algorithm is simply $O[(K_j)^3]$. For a relative utility, since $\mu_j(k,r,l)$ depends on k, E is $O(K_j)$, in general. However, there are still

TABLE I
EXECUTION TIME FOR OPTIMIZATION ALGORITHMS

Setting	Execution Time (ms)		
$(N, \|S\|, K^{max})$	Intra-Session Allocation*	Inter-Session Allocation	Joint Intra-Session and Inter-Session Allocation
(64,8,12)	0.9	1.2	8.4
(128,10,15)	1.4	2.9	16.9
(512,20,30)	2.1	7.0	49.0

* Execution time for one session with the maximum receiver capacity being equal to K^{max}.

some common relative utilities functions that are of $O(1)$ complexity. For example, consider the relative utility function AFI defined as $\mu_j(k,r,l) = Q(r,l)/Q(k,1)$[6]. We have

$$\sum_{k=m}^{K_j} M_{j,k}\Delta(k,m,i,l) = \sum_{k=m}^{kj} M_{j,k}\left[\frac{Q(m,l)}{Q(k,1)} - \frac{Q(i,l-1)}{Q(k,1)}\right]$$
$$= [Q(m,l) - Q(i,l-1)] \cdot \sum_{k=m}^{K_j} \frac{M_{j,k}}{Q(k,1)}$$

where $\sum_{k=m}^{K_j} M_{j,k}/Q(k,1)$, $k = 1, 2, \ldots, K_j$ can be precomputed and stored as well, and, thus, E remains $O(1)$.

To demonstrate the efficiency of our optimization algorithms in practice, we implement the optimization algorithms using C++ on an Intel Pentium III 900 MHz PC with 256 MB memory. The execution times under different settings with the AFI utility function are listed in Table I. It can be seen that the solutions can be computed within a reasonably short period of time, which is suitable for real-time applications.

V. IMPLEMENTATION CONSIDERATIONS

In this section, we discuss some implementation issues in practice, including the choice of layered video codec and the computation overhead in an online implementation.

A. Choice of Video Codec

In the video coding area, scalable coding typically refers to layered coding. In this paper, we are particularly interested in scalable video coders with a flexible layering structure and fine granularity in terms of rate control. Recent advances in scalable video coding have demonstrated that this is possible and can be done efficiently. A representative technique is the bit-plane coding algorithm, which uses embedded representations in compression [12]. For illustration, there are 64 (8 × 8) DCT coefficients for each video block. All the most significant bits from the 64 DCT coefficients form bitplane 0, all the second most significant bits form bitplane 1, and so forth. In the output stream, the bitplanes, not the coefficients, are placed sequentially. Hence, layers can be generated by an assembling and packetization procedure, which can truncate the embedded stream in any position to achieve a specified output rate. This post-encoding method is different from the traditional scalability tools that use a fixed layering structure and perform rate control at the source coding stage. As a result, bitplane-based scalable coding can achieve very flexible layering structure, which makes it possible to produce arbitrary number of layers and to fine tune the rate of each layer with a fast response time. Bit-plane coding has been adopted in the MPEG-4 fine granularity scalability (FGS) standard [12].

Regarding the layering overhead, there are several factors that need to be considered. For example, a packetization scheme can affect the overhead since different packetization schemes use a different amount of bits for layer identification, synchronization, and error concealment, which leads to different overheads [19]. In the experiments described in the next section, we use a wide range of settings to take into account such a layering overhead.

B. Online Implementation

There is a potential issue of computational overhead associated with online adaptation, but as we have shown earlier, our algorithm runs reasonably fast for real-time adaptation. Furthermore, since our algorithm is based on the bandwidth distribution of all the receivers (instead of the bandwidth of an individual receiver), the adaptation algorithm needs to be executed only when the distribution has changed significantly, which can be easily identified by using standard statistical methods (e.g., the Pearson's χ^2 test or the Kolmogorov–Smirnov test [29]). Finally, according to the principle of decomposition, the session utility of a particular session is independent of the receiver status of other sessions. Hence, when the status of a session changes, only its own session utility needs to be recalculated, together with one execution of intersession allocation.

VI. NUMERICAL INVESTIGATIONS

In this section, we conduct experiments to demonstrate the performance of the optimal layering and rate allocation algorithms for video broadcasting. We also compare it to commonly used nonoptimal allocation schemes to show performance improvement.

A. Simulation Settings

To show the heterogeneous nature of the receivers, we model the variation of capacity among receivers in a session with a multimodal distribution. Specifically, we observe that the access link and video decoding component of a receiver typically follows some specific standards yet some use customized software/hardware [1]. Thus, we use a mixture Gaussian model [20] to represent the capacity distribution among the receivers in a session. That is, we assume there are several clusters each following a Gaussian distribution. In our simulation, we assume

TABLE II
ALLOCATION VECTORS FOR INTRASESSION ALLOCATION WITH DIFFERENT UTILITY FUNCTIONS

Utility Function	Optimal Layer Allocation Vector
PSNR	(2,6,15,19,23,25)
AFI	(2,5,9,14,22,25)

the bandwidth of each channel is 28.8 Kb/s, with the minimum and maximum receiver capacities being 2 and 25 channels, respectively. This range covers the rate of most available wireless link access technologies. This is also the typical dynamic range of existing scalable video coders, such as the MPEG-4 PFGS coder [17]. The standard deviation of a cluster is set to 10% of the cluster mean. Therefore, most bandwidth differences are within ±10%, yet a few reach about ±40% or more, which reflects the flexibility in device design.

We use the MPEG-4 progressive fine-granularity scalable (PFGS) video encoder [17] to generate layered video streams. A standard video test sequence "Foreman (CIF)" is used in our study. The TM-5 rate control model is adopted to control the bit-rate of the base layer [24]. The number of enhancement layers as well as their respective bandwidth is allocated by an assembling and packetization module [19]. As in previous studies [21], we define the layering overhead per layer h as follows: assume L layers are generated at bandwidth r, and a single-layer stream with the same video quality has bandwidth r_0, h is given by $(r - r_0)/(L - 1)$ with a unit of channel per layer.

B. Intrasession Allocation

In this section, we focus on a single session and conduct experiments to show the performance and behavior of the optimal algorithm for intrasession layering and rate allocation.

1) Effect of Utility Functions: We have used a series of utility functions to study their impact, including typical mappings used in the literature [6], [9], [10], as well as mappings for practical layered video encoders [17]. Specifically, we present the results for an absolute utility function $\mu_s(k, r, l) = Q(r, l)$, where $Q(r, l)$ is the objective video quality measured by the PSNR (with a unit of decibels) [17] and the relative utility function AFI, defined as $\mu_s(k, r, l) = Q(r, l)/Q(k, 1)$[6].

Table II presents the optimal layer allocation vectors with the above two utility functions for a capacity distribution of six clusters. The layering overhead h is 0.5 channel per layer, which is moderate. We find that, with different utility functions, the corresponding optimal bandwidth allocation for each layer is quite different. For a system employing an absolute utility function, the rate allocation typically favors receivers with high bandwidths. This can be observed in Fig. 4, where the receiver utility under different access capacity is plotted. Under the absolute utility function (PSNR), the utility at a receiver is a nondecreasing function of the receiver access capacity. As a result, the optimal allocation tends to allocate more layers in the high capacity region so that higher utility can be obtained from this region. On the contrary, the relative utility function (AFI) does not favor high capacity receivers as the utility is normalized. Such observations confirm our arguments in Section III. For the rest of the experiments in this section, we will show results only with the AFI utility function.

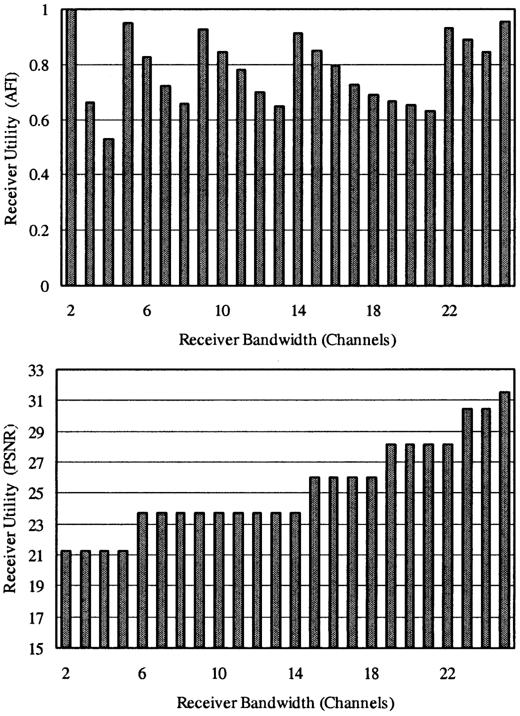


Fig. 4. Receiver utility for optimal allocation with different utility functions: (a) relative utility, application-aware fairness index (AFI) and (b) absolute utility, PSNR (decibels as unit).

2) Impact of Layering Overhead: Fig. 5 shows the optimal session utility as a function of the bandwidth allocated to the session. The layering overheads are 0, 0.2, 0.5, 1.0 channel per layer, respectively, which cover both light and heavy overhead cases. In this figure, as well as the remaining figures, the session utility (or system utility) is normalized by the number of receivers in the session (or system). Not surprisingly, all the curves are nondecreasing with the increase of the session bandwidth. The optimal session utility of $h = 0$ (no layering overhead) achieves the ideal session utility, 1, when the bandwidth budget is at least 25 channels. In this case, each receiver has a layer whose cumulative bandwidth perfectly matches the receiver's capacity. However, if the layering overhead is taken into account, the ideal session utility cannot be achieved because the overhead counteracts the benefits from increasing the number of layers. In all these cases, the session utility will converge to a steady value for bandwidth greater than 25 channels, the highest access capacity among all the receivers in the session.

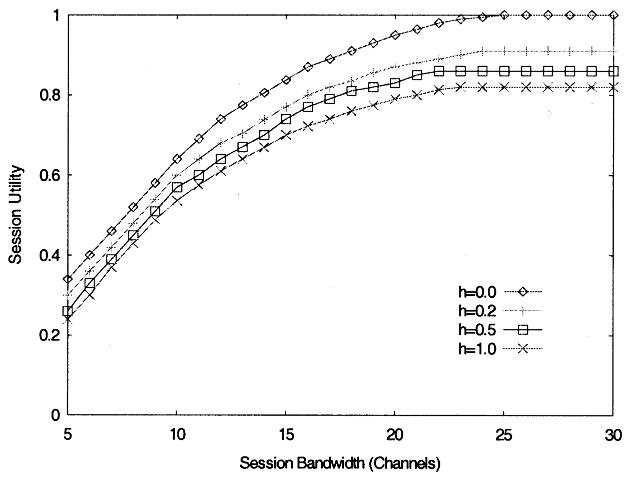


Fig. 5. Optimal session utility as a function of session bandwidth for different layering overheads (h channel per layer).

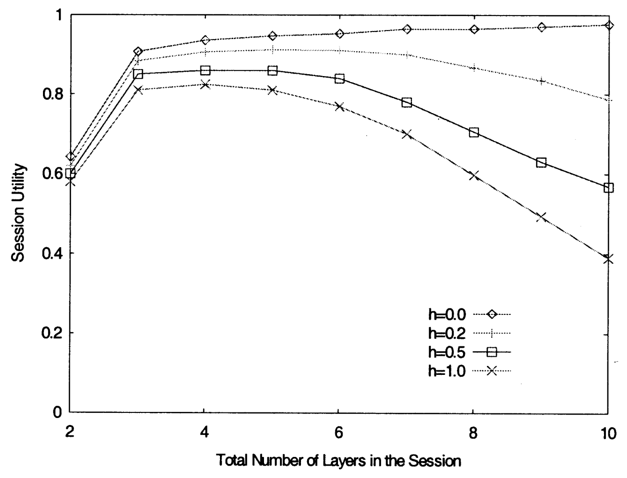
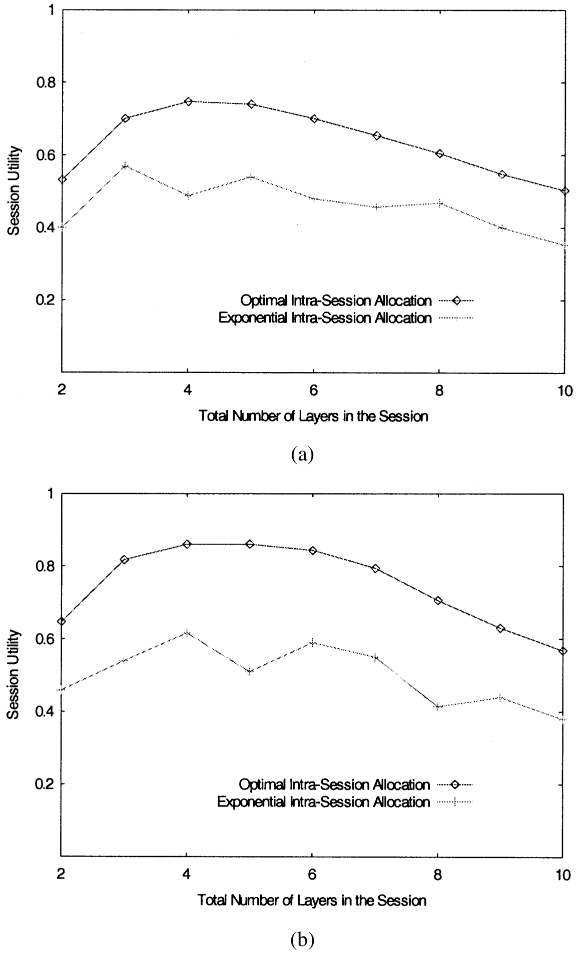


Fig. 6. Optimal session utility for a given number of layers with different layering overheads. Session bandwidth is 25 channels.

Fig. 6 shows the optimal session utility as a function of the number of layers for different layering overheads. It can be seen that when $h = 0$, the session utility is nondecreasing with the number of layers for the given session bandwidth budget. However, if $h > 0$, the session utility is no longer nondecreasing and has a maximum value at a certain number of layers (in this example, about four to six layers, depending on the overheads). Intuitively speaking, below that number of layers, the adaptation granularity on the receiver's end is somewhat coarse; above which, more layering overhead is incurred.

The above results clearly demonstrate that if the layering overhead is not considered, the use of "thin" layers, i.e., generating more layers for a given session bandwidth budget, is preferable. On the other hand, if the layering overhead is considered, there exists an optimal number of layers such that the session utility is maximized. This optimal number can be found using our optimal intrasession layering and rate allocation algorithm.

Note that, regardless of whether layering overhead is considered or not, the session utility (as well as the corresponding allocations) under different session bandwidth budget is different, even if the same number of layers are generated. For example, in Fig. 7, for the 15-channel case, the maximum session utility is 0.78. However, in the 25-channel case, it can reach 0.86. In a bandwidth-limited network, it is thus necessary to use an intersession allocation scheme to optimally allocate the available bandwidth to different sessions.



(a)

(b)

Fig. 7. Session utility as a function of the number of layers for the optimal and exponential allocation schemes. (a) Session bandwidth is 15 channels, layering overhead is 0.5 channel/layer. (b) Session bandwidth is 25 channels, layering overhead is 0.5 channel/layer.

3) Optimal Versus Nonoptimal Allocations: In this experiment, we compare the performance of our optimal allocation scheme and a scheme employing a fixed layering structure. Again, we focus our study on a single session. In the literature, a widely recommended fixed allocation scheme is the exponential allocation, in which the cumulative layer rates are exponentially spaced by a constant factor $\alpha > 1$, i.e., $r_j^{i+1} = \alpha r_j^i$. This is the scheme adopted in the original receiver-driven layered multicast (RLM) protocol [3] and many other experiments [6], [7]. Given the session bandwidth budget $n_j (\leq K_j)$, the lower bound of the base layer bandwidth n_b and the number of layers L_j, we can can calculate α as $\sqrt[(L_j-1)]{n_j/n_b}$. In this experiment, we assume that L_j is fixed to five layers for the exponential allocation. For $n_j > K_j$, we assume the allocation is the same as that for $n_j = K_j$.

Fig. 8 shows the session utility as a function of session bandwidth for the optimal allocation and the exponential allocation.

Clearly, the session utility under the optimal allocation is greater than that under the exponential allocation. In particular, under the optimal intrasession allocation, the session utility is nondecreasing, while under the exponential allocation, the behavior of the session utility is somewhat unpredictable. This is because the exponential allocation is not aware of the receivers' bandwidth distribution for the session. It may allocate the layer bandwidth to be a level with few receivers, and hence aggravates the bandwidth mismatches. We also show the results with different numbers of layers for the two allocation schemes in Fig. 7. Again, there are significant gaps between the two schemes even if the numbers of layers are the same. These results reaffirm that the optimal choice of the number of layers must be used in conjunction with the optimal bandwidth allocation for each layer and vice versa.

C. Joint Intrasession and Intersession Allocation

We also study the effect of joint intrasession and intersession layering and bandwidth allocation and try to identify the respective contribution of the optimal intrasession and intersession allocations to the total system utility, specifically in the case where the sessions have uneven populations.

We assume that the demand probabilities for different video sessions follow a Zipf distribution [25]. This distribution has been widely used in the literature [26] and captures the difference in terms of popularity for the video sessions. The Zipf distribution is expressed as

$$p_j = \frac{(\frac{1}{j})^\theta}{\sum_{j=1}^{|S|} (\frac{1}{j})^\theta}, \quad j = 1, 2, \ldots, |S| \tag{4}$$

where θ is a parameter called *skew factor*. For $\theta = 0$, the Zipf distribution is reduced to a uniform distribution with $p_j = 1/|\boldsymbol{S}|$. However, the distribution becomes increasingly "skewed" as θ increases, i.e., a few popular video sessions attract many more receivers than the others. In other words, the session popularities are differentiated.

We consider all four possible combinations of the intra–inter session allocation: 1) OptIntra-OptInter, where both intrasession and intersession allocations are optimal; 2) OptIntra-UniInter, where only intrasession allocation is optimal and intersession allocation is a uniform allocation; 3) ExpIntra-OptInter, where only intersession is optimal and intrasession is exponential allocation; and 4) ExpIntra-UniInter, where both are nonoptimal. In the experiments, we assume that there are 500 receivers belonging to ten sessions. We vary the skew factor θ for session popularity distribution from 0 to 1. The number of clusters for the receiver capacity distribution in a session is uniformly distributed from 2 to 9. We then draw 500 samples from the above model to obtain a receivers' status distribution for the whole system.

Fig. 9 shows the system utilities with different skew factors for all the four combinations. It is clear that the optimal intra/inter allocation scheme outperforms all other schemes. Comparing these curves, specifically the curves of OptIntra-UniInter and ExpIntra-OptInter, we find that the contribution of the optimal intersession allocation becomes more important as the skew factor increases. Note that a higher skew factor means that some popular video programs attract much more receivers than others. It is thus advisable to allocate more channels to these sessions. Specifically, in Fig. 9, for a total system bandwidth of 128 channels, when $\theta > 0.5$, ExpIntra-OptInter outperforms OptIntra-UniInter. This reaffirms our claim that the optimal intersession allocation should be considered.

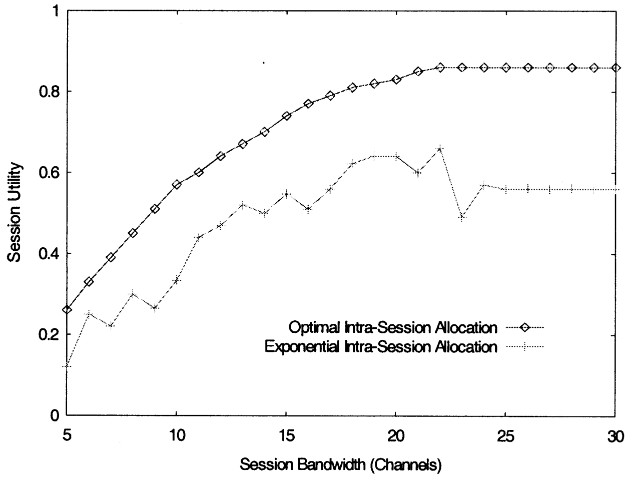


Fig. 8. Session utility as a function of session bandwidth for the optimal and exponential allocation schemes. Layering overhead h is 0.5 channel/layer.

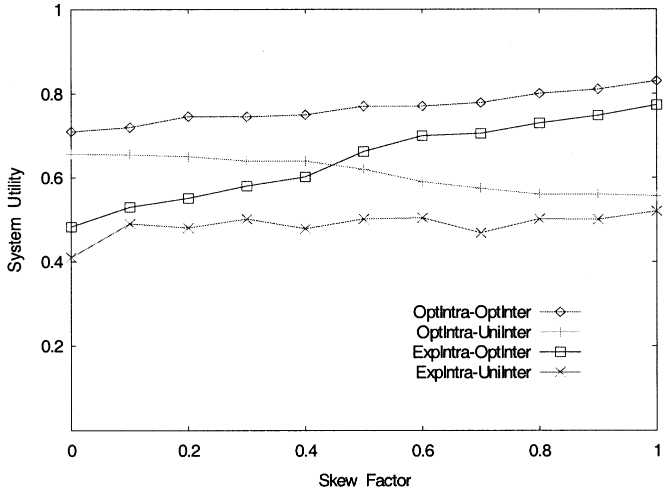


Fig. 9. Total system utility for different skew factors. Total system bandwidth N is 128 channels.

D. Bandwidth Difference in Handoff

So far, we have focused on adaptation in a single cell. In a multicell network, receivers would move across cells. If the receiver distribution is nonuniform in the network, there could be some bandwidth variations during handoff and smooth handoff thus becomes a crucial issue. More explicitly, we define the bandwidth difference for each receiver during handoff as $|r_0 - r_1|/r_0 \times 100\%$, where r_0 is the receiver's original subscription bandwidth and r_1 is its subscription bandwidth in the new cell; if this difference is large, it would cause noticeably video quality fluctuation.

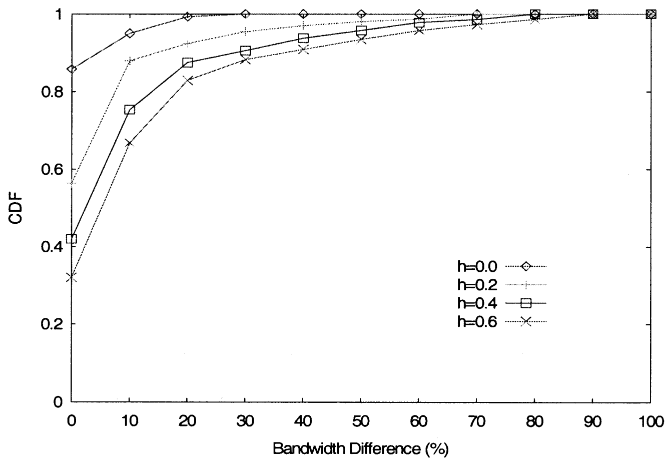


Fig. 10. CDF of the bandwidth difference during handoff.

To investigate such variance, we generate 50 cell instances with cell population (the number of receivers in the cell) being uniformly distributed from 10 to 100. All these receivers belong to the same session and the number of clusters for the receiver capacity distribution in a cell is uniformly distributed in 2 through 9. For each cell pair, we assume 10% of the receivers move from one cell to another and then calculate the bandwidth difference under the optimal allocation. The cumulative distribution function (CDF) of the bandwidth differences is shown in Fig. 10. We can see that the subscription bandwidth for a receiver usually does not change drastically during handoff. Most times, the bandwidth differences are less than 20%, which often can be effectively masked on the receiver's side by seamless transition techniques for video streams [18], [27].The difference is particularly small for low layering overheads because more layers are generated in this case and, hence, finer adaptation granularity can be achieved.

VII. Conclusion and Future Work

In this paper, we presented a systematic study for addressing the important problem of optimal layering and bandwidth allocation (among sessions and layers) for video broadcasting in a wireless environment. We employed a generic utility function for each receiver under each video session. We cast the joint problem of layering and bandwidth allocation into an optimization problem of total system utility among all receivers. By using a simple two-step decomposition of intersession and intrasession allocations, we derived efficient algorithms to solve the optimization problem. Numerical results showed that the optimal layering and bandwidth allocation can improve the total system utility substantilly. Practical issues for deploying the algorithm in typical wireless networks were also discussed. We demonstrated that our algorithm can be efficiently supported by the recently developed scalable video codecs, such as FGS or PFGS, and that the overall system complexity can be kept at a reasonably low level.

It is worth noting that our optimal layering and bandwidth allocation algorithm is also applicable to other broadcast- or multicast-capable networks. In a multicell network with the optimal allocations, the subscription bandwidth for a receiver usually does not change drastically during handoff, and thus could be masked on the receiver's side by using seamless stream transition techniques. Since a global optimization with cell collaborations is usually of high complexity and incurs extra overheads for information exchange among cells [15], we suggest the cells simply perform allocation independently. This is also well supported by FGS or PFGS coding, since their layer partitioning and rate control are performed as a postencoding process, which can be implemented easily at each access point, without generating replicated streams from the video source.

References

[1] Y.-B. Lin and I. Chlamtac, *Wireless and Mobile Network Architectures*. New York: Wiley, 2001.

[2] D. Wu, Y. T. Hou, and Y.-Q. Zhang, "Scalable video coding and transport over broadband wireless networks," *Proc. IEEE*, vol. 89, pp. 6–20, Jan. 2001.

[3] S. McCanne, V. Jacobson, and M. Vetterli, "Receiver-driven layered multicast ," in *Proc. ACM SIGCOMM'96*, Standford, CA, Aug. 1996.

[4] L. Wu, R. Sharma, and B. Smith, "Thinstreams: An architecture for multicast layered video," in *Proc. Workshop Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'97)*, St. Louis, MO, May 1997.

[5] N. Shacham, "Multipoint communication by hierarchically encoded data," in *Proc. IEEE INFOCOM'92*, Florence, Italy, May 1992.

[6] J. Liu, B. Li, and Y.-Q. Zhang, "An end-to-end adaptation protocol for layered multicast using optimal rate allocation," *IEEE Trans. Multimedia*, vol. 6, pp. 87–102, Feb. 2004.

[7] S. Gorinsky and H. Vin, "The utility of feedback in layered multicast congestion control," in *Proc. Workshop Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'01)*, Port Jefferson, New York, June 2001.

[8] K. Kar, S. Sarkar, and L. Tassiulas, "Optimization based rate control for multirate multicast sessions," in *Proc. IEEE INFOCOM'01*, Anchorage, AK, Apr. 2001.

[9] B. Vickers, C. Albuquerque, and T. Suda, "Source adaptive multi-layered multicast algorithms for real-time video distribution," *IEEE/ACM Trans. Networking*, vol. 8, no. 6, pp. 720–733, Dec. 2000.

[10] T. Jiang, E. W. Zegura, and M. H. Ammar, "Inter-receiver fair multicast communication over the internet," in *Proc. Workshop Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'99)*, Basking Ridge, NJ, June 1999.

[11] Y. Yang, M. Kim, and S. Lam, "Optimal partitioning of multicast receivers," in *Proc. IEEE ICNP'00*, Nov. 2000.

[12] W. Li, "Overview of the fine granularity scalability in MPEG-4 video standard," *IEEE Trans. Circuits Syst.*, vol. 11, pp. 301–317, Mar. 2001.

[13] A. K. Talukdar, B. R. Badrinath, and A. Acharya, "Rate adaptation schemes in networks with mobile hosts," in *Proc. ACM/IEEE MOBICOM'98*, Dallas, TX, Oct. 1998.

[14] M. Naghshineh and M. Willebeek-LeMair, "End-to-end QoS provisioning in multimedia wireless/mobile networks using an adaptive framework," *IEEE Commun. Mag.*, vol. 35, pp. 72–81, Nov. 1997.

[15] T. Kwon, Y. Choi, and S. K. Das, "Bandwidth adaptation algorithms for adaptive multimedia services in mobile cellular networks," *Wireless Personal Commun.*, vol. 22, no. 3, pp. .337–357, 2002.

[16] G. Bianchi, A. T. Campbell, and R.-F. Liao, "On utility-fair adaptive services in wireless networks," in *Proc. 6th Int. Workshop Quality of Service (IEEE/IFIP IWQOS'98)*, Napa, CA, May 1998.

[17] S. Li, F. Wu, and Y.-Q. Zhang, "Experimental results with progressive fine granularity scalable (PFGS) coding,", ISO/IEC JTC1/SC29/WG11, MPEG 99/m5742, 2000.

[18] X. Sun, F. Wu, S. Li, W. Gao, and Y.-Q. Zhang, "Seamless switching of scalable video bitstreams for efficient streaming," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS'02)*, Scottsdale, AZ, May 2002.

[19] H. Cai, G. Shen, Z. Xiong, S. Li, and B. Zeng, "An optimal packetization scheme for fine granularity scalable bitstream," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS'02)*, Scottsdale, AZ, May 2002.

[20] D. Titterington, A. Smith, and U. Makov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.

[21] P. de Cuetos, D. Saparilla, and K. W. Ross, "Adaptive streaming of stored video in a TCP-friendly context: Multiple versions or multiple layers," in *Proc. Int. Packet Video Workshop*, Kyongju, Korea, Apr. 2001.

[22] T. Kim and M. H. Ammar, "A comparison of layering and stream replication video multicast schemes," in *Proc. Workshop Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'99)*, Basking Ridge, NJ, June 2001.

[23] Delivery Multimedia Integration Framework, DMIF (ISO/IEC 14 496–6), ISO/IEC/SC29/WG11, Feb. 1999.

[24] J.-R. Ohm, "Description of Core Experiments in MPEG-4 Video,", ISO/IEC JTC1/SC29/WG11, N2554, 1998.

[25] G. Zipf, *Human Behavior and the Principle of Least Effort*. Reading, MA: Addison-Wesley, 1949.

[26] A. Dan, D. Sitaram, and P. Shahabuddin, "Scheduling policies for an on-demand video server with batching," in *Proc. ACM Multimedia''94*, San Francisco, CA, Oct. 1994.

[27] X. Sun, S. Li, F. Wu, G. Shen, and W. Gao, "Efficient and flexible drift-free video bitstream switching at predictive frames," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME'02)*, Lausanne, Switzerland, Aug. 2002.

[28] B. Li and J. Liu, "Multirate video multicast over the internet: An overview," *IEEE Network Special Issue on Multicasting: An Enabling Technology*, vol. 17, Jan./Feb. 2003.

[29] A. Alan, *Introduction to Categorical Data Analysis*. New York: Wiley, 1996.

Jiangchuan Liu (S'01–M'03) received the B.S degree (*cum laude*) from Tsinghua University, Beijing, China, in 1999 and the Ph.D. degree at the Hong Kong University of Science and Technology, Hong Kong, in 2003, both in computer science.

He is currently an Assistant Professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong. In the summers of 2000–2002, he had internships with Microsoft Research, Asia, working in video multicast over the Internet and service discovery in variable topology networks. He is a recipient of a Microsoft research fellowship and a co-inventor of two U.S. patents (pending). His research interests include multicast protocols, streaming media, wireless *ad hoc* networks, and service overlay networks.

Dr. Liu received a Young Scientist Award from Hong Kong Institute of Science and won first-class honors in several regional and national programming contests. He was a TPC Member and Information System Co-Chair for IEEE Infocom'04, a TPC member for IEEE ICCCN'04 and Infocom'05, and a Guest Editor for the *ACM/Kluwer Journal of Mobile Networks and Applications* Special Issue on Wireless Sensor Networks.

Bo Li (S'89–M'92–SM'99) received the B.S. (*summa cum laude*) and M.S. degrees in computer science from Tsinghua University, Beijing, China, in 1987 and 1989, respectively, and the Ph.D. degree in computer engineering from University of Massachusetts, Amherst, in 1993.

Between 1994 and 1996, he worked on high performance routers and ATM switches in IBM Networking System Division, Research Triangle Park, NC. Since 1996, he has been with Computer Science Department, the Hong Kong University of Science and Technology, Hong Kong. He is also an Adjunct Researcher in Microsoft Research Asia. His current research interests are on multimedia communications, wavelength-routed networks, resource management in wireless cellular systems, and service overlay networks. He has published over 130 papers in journal and conferences.

Dr. Li has been on the editorial board for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *ACM Journal of Wireless Networks* (WINET), IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS, *ACM Mobile Computing and Communications Review* (MC2R), *SPIE/Kluwer Optical Networking Magazine* (ONM), KICS/IEEE JOURNAL OF COMMUNICATIONS AND NETWORKS. He served as a Guest Editor for *IEEE Communications Magazine Special Issue on Active, Programmable, and Mobile Code Networking* (April 2000), IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS Special Issue on Protocols for Next Generation Optical WDM Networks (October 2000), an *ACM Performance Evaluation Review Special Issue on Mobile Computing* (December 2000), and *SPIE/Kluwer Optical Networks Magazine Special Issue on Wavelength Routed Networks: Architecture, Protocols and Experiments* (January/February 2002). Currently, he is co-guest editing two special issues for the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS. In addition, he has been involved in organizing over 30 conferences, especially the IEEE Infocom since 1996. He is a co-TPC chair for IEEE Infocom 2004.

Y. Thomas Hou (S'91-M'98–SM'04) obtained the B.E. degree (*summa cum laude*) from the City College of New York, New York, in 1991, the M.S. degree from Columbia University, New York, in 1993, and the Ph.D. degree from Polytechnic University, Brooklyn, New York, in 1998, all in electrical engineering.

From 1997 to 2002, he was a Research Scientist and Project Leader at Fujitsu Laboratories of America, IP Networking Research Department, Sunnyvale, CA. He is currently an Assistant Professer at Virginia Tech, Blacksburg, VA. His current research interests include wireless video sensor networks, multimedia delivery over wireless networks, scaleable architectures, protocols, and implementations for differentiated services Internet, and service overlay networking. He has published numerous papers in the above areas.

Dr. Hou is a Co-Recipient of the 2002 IEEE International Conference on Network Protocols (ICNP) Best Paper Award and the 2001 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Paper Award. He is a member of ACM.

Imrich Chlamtac (M'86-SM'86-F'93) received the Ph.D degree in computer science from the University of Minnesota, Minneapolis.

Since 1997, he has been the Distinguished Chair in Telecommunications, University of Texas, Dallas. He also holds the titles of the Sackler Professer at Tel Aviv University, Israel, The Bruno Kessler Honorary Professer at the University of Trente, Italy, where he is currently on sabbatical, and University Professer at the Technical University of Budapest, Hungary. He has published close to 300 papers in refereed journals and conferences and is the co-author of *Local Area Networks* (New York: Lexington Books, 1981, 1982, 1984) and of *Mobile and Wireless Networks Protocols and Services* (New York: Wiley, 2000), an IEEE Network Magazine's 2000 Editor's Choice.

Dr. Chlamtac serves as the founding Editor-in-Chief of the *ACM/URSI/Kluwer Wireless Networks* (WINET), the *ACM/Kluwer Mobile Networks and Applications* (MONET) journals, and the *SPIE/Kluwer Optical Networks Magazine*. He is a Fellow of the ACM, a Fulbright Scholar, and an IEEE Distinguished Lecturer. He is the winner of the 2001 ACM Sigmobile annual award and the IEEE ComSoc TCPC 2002 Award for contributions to wireless and mobile networks and of multiple best paper awards in wireless and optical networks.