Optimal Stream Replication for Video Simulcasting

Jiangchuan Liu, *Member, IEEE*, Bo Li, *Senior Member, IEEE*, and Ya-Qin Zhang, *Fellow, IEEE*

***Abstract*—Video simulcasting enables a sender to generate replicated streams of different rates, serving receivers of diverse access bandwidths. As replication introduces noticeable redundancy, balancing bandwidth consumption with user satisfaction becomes a critical concern in simulcasting. This paper investigates the above issue; more explicitly, we seek answers to the following two questions: what is the number of streams that should be generated, and what is the bandwidth that should be allocated to each stream? We derive optimal and efficient solutions, and evaluate their performance under a variety of configurations. The results demonstrate that an optimal and adaptive bandwidth allocation significantly improves user satisfaction under stringent resource constraints, and an optimal choice of the stream number yields further improvements.**

***Index Terms*—Bandwidth allocation, multicast, simulcasting, stream replication.**

I. Introduction

ITH the rapid development and deployment of broadband networks, real-time video distribution is emerging as one of the most important networked applications. The multi-user nature of video programs makes multicast an efficient method for delivering video content to a large population of receivers. Such systems have also been effectively supported by existing network infrastructures, e.g., by wireless networks or the Internet with IP multicast or application-layer multicast [20], [21]. A key challenge, however, is how to handle user heterogeneity, as receivers with different platforms, such as PDAs, laptops, and PCs, or with different connection speeds, such as 1.5-Mbps ADSL or 100-Mbps Ethernet, all expect to access video services nowadays. Clearly, a single-rate transmission, though simple, is difficult to match these diverse demands.

Simulcasting has been introduced as a vehicle to mitigate the mismatches and hence improve user satisfaction [1], [2], [19]. A simulcast server maintains some replicated streams for the same video content but with different rates, and delivers each stream to a specific set of receivers through a multicast channel. Two questions arise for video simulcasting naturally: how to choose an appropriate number of streams, and how to allocate bandwidth to the streams to reduce bandwidth mismatches?

In this paper, for the first time we present a formal study on the above issues, to which we refer to as the *stream replication problem*. The key objective here is to strike a balance between bandwidth economy and user satisfaction; in other words, given some bandwidth constraint, a replication scheme should minimize the expected bandwidth mismatch for all the receivers. We first formulate the optimal replication problem for a given number of streams, and derive an efficient algorithm. This optimization follows the design nature of most existing simulcasting systems, in which the number of streams is predetermined by system operators or service providers.

Through theoretical analysis and experimental results, we then demonstrate that the number of replicated streams is a critical factor in the overall system optimization. It is necessary to choose an *optimal*, not an *ad hoc*, number of streams based on the available resources as well as receivers' requirements. In addition, recent advances in video coding have shown that a stream can be replicated in realtime by fast compression domain transcoding [3], [4]. The fast and low-cost operations for stream setup and termination have also been supported in advanced video streaming standards, such as the MPEG-4 Delivery Multimedia Integration Framework (DIMF) [1]. Given the flexibility, it is possible to adaptively regulate the number of streams to accommodate the receivers' requirements. Therefore, we further consider the use of a flexible number of video streams, and offer an algorithm that jointly optimizes the number as well as the bandwidth allocated to each stream.

We also investigate the bandwidth allocation among different video programs (sessions). We note that, for optimal replication, the expected mismatch of the receivers is a stepwise function of the session bandwidth. As a result, an equal allocation often leads to a waste of bandwidth. We thus introduce a novel mismatch-aware allocation scheme, which intelligently distributes the unused bandwidth to other sessions, yet preserving general fairness properties.

The rest of the paper is organized as follows. Section II presents some related work. The system model is described in Section III. Section IV formulates the problem of optimal steam replication for a single session, and derives the solutions. The inter-session bandwidth allocation is studied in Section V. We then present the performance results in Section VI, and conclude the paper in Section VII.

II. Related Work

A. Simulcasting Protocols

There has been a significant amount of work on simulcasting in the literature [1], [2], [5], [6], [10], [11]. A representative

Manuscript received February 4, 2003; revised November 19, 2003. The work of J. Liu was supported in part by a Canadian NSERC Discovery Grant 288325, an NSERC Research Tools and Instruments Grant, and an SFU President's Research Grant. The work of B. Li was supported in part by grants from RGC under the contracts HKUST 6196/02E, HKUST6204/03E HKUST6104/04E,, and HKUST6165/05E, a grant from NSFC/RGC under Contract N_HKUST605/02, and a grant from NSF China under Contract 60429202. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mihaela van der Schaar.

J. Liu is with the School of Computing Science, Simon Fraser University, Burnaby, BC V6B 5K3, Canada (e-mail: jcliu@cs.sfu.ca).

B. Li is with the Department of Computer Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (e-mail: bli@cs.ust.hk).

Y.-Q. Zhang is with Microsoft Corporation, One Microsoft Way, Redmond, WA 98052 USA (e-mail: yzhang@microsoft.com).

Digital Object Identifier 10.1109/TMM.2005.861279

is the destination set grouping (DSG) protocol [2]. In DSG, a source maintains a small number of video streams (say three) at different rates; a receiver subscribes to a stream that best matches its bandwidth. It also estimates the network status according to the packet loss ratios, and reports its estimation to the sender through a scalable feedback protocol. If the percentile of the congested receivers for a stream is above a certain threshold, the bandwidth of the stream is reduced by the sender. If few receivers observe packet losses, the stream bandwidth is increased. The choice of the threshold is experience-based, which is not necessarily optimal. Some heuristics are proposed in [5] for stream bandwidth adjustment. The objective is to reduce the overall bandwidth originated from the server, as well as the aggregate bandwidth, in local regions. An optimal bandwidth allocation algorithm that minimizes bandwidth mismatches for a two-stream case is proposed in [6]. The algorithm performs an exhaustive search on the receivers' expected bandwidths, based on an observation that an optimal stream bandwidth must be one of them. Due to the high complexity of exhaustive search, it is not easy to directly extend this algorithm with more streams. Simulcasting has also been used in overlay networks to match client capacities [21]; yet the choices of streaming rates are *ad hoc*.

B. Simulcasting Versus Layered Transmission

Another typical solution for multicasting to heterogeneous receivers is *layered transmission* [7]–[14], in which a sender generates multiple layers that can progressively refine the video quality, and a receiver thus can subscribe to a subset of layers commensurate with its bandwidth.

Layered transmission also suffers from bandwidth mismatches, because adaptation on the receiver's side is at coarse-grained layer level. To minimize this mismatch, protocols using dynamic layer rate allocation on the sender's side have been proposed for cumulative layering, where the layers are subscribed cumulatively staring from a base layer [7], [8], [13]. However, the constraints and hence the optimization strategies for layer rate allocation and stream replication are quite different. For illustration, consider a single video with a given total bandwidth N and total number of replicated streams (or layers) K. As will be explained later, we assume bandwidth allocation is discrete. For simulcasting, the problem of stream rate allocation is thus to find an optimal K-partition for integer N; for cumulative layering, it is to find an optimal enumeration of K numbers with the maximum one being no more than N. In addition, it can be proved that, the more layers a layered transmission protocol generates, the smaller the mismatch that a receiver would experience (ignoring the layering overheads) [8]. Thus, the use of "thin layers" has been advocated in some existing protocols [9]. For simulcasting, this is not true, because stream replication would introduce very high redundancy with a large number of streams. As such, it is necessary to find an optimal number of streams.

It is worth noting that simulcasting and layering have been compared in many different contexts, including IP-layer multicasting [10], TCP-friendly streaming [11], and proxy-assisted streaming [17], [18]. Though it is often believed that layering achieves higher bandwidth utilization as there is no overlapping among the layers, it suffers from the high complexities as well as the structure constraints for both encoding and decoding [1]. To the contrary, simulcasting produces independent streams that can serve receivers with simpler and even heterogeneous decoding algorithms. Therefore, simulcasting remains a practical solution to address user heterogeneity.

III. System Model and Definitions

In our system, a video server distributes a set of video programs using the simulcasting technique: each video program has several replicated streams of different rates. The descriptions of the video programs are advertised to the receivers via a dedicated multicast channel. A receiver interested in a particular video program can thus subscribe to one of the streams to receive the video. We refer to a program and its receivers as a *simulcast session* (or *session*), and the bandwidth allocated to the session as *session bandwidth*, which imposes an upper bound for the total bandwidth of the replicated streams.

The status of the system can be characterized by a three-tuple, $(C, P, M_{s,t})$, where C is the maximum outbound bandwidth of the server; P is the total number of sessions, and each session has an index in $[1, \ldots, P]$; $M_{s,t}$ is the ratio of the receivers having expected bandwidth t in session s. We stress that this model captures the essentials of many existing video multicasting systems, in which the expected bandwidths of the receivers are heterogeneous and limited by their processing capabilities or access links; the video server, though having a higher output bandwidth, has to accommodate many simultaneous sessions, each with several replicated streams. Therefore, its bandwidth is also an important factor in the system optimization. Note that a receiver cannot subscribe to a fraction of a video stream. Assume a receiver's expected bandwidth is t, and the bandwidth of its subscribed stream is r, we measure the bandwidth mismatch as

$$RM(t,r) = \begin{cases} \frac{(t-r)}{t}, & 0 < r \le t \\ 1, & r > t \text{ or } r = 0 \end{cases}. \tag{1}$$

We use this relative measure (RM) instead of an absolute mismatch measure, as RM will not enlarge the impact of the mismatches perceived by wideband receivers. For a simulcast session, our objective for both sender and receiver adaptations is to minimize the expected RM for all the receivers. There could be other measures, as well as mappings from the mismatch to some application-level performance degradation. For example, different fairness measures, such as the inter-receiver fairness (IRF) function [6], or subjective/objective video quality measures, such as the peak signal-to-noise ratio (PSNR) [1]. The optimization algorithms presented in this paper are general enough, which does not impose strict constraints on the measurement function, and thus can accommodate other mismatch or fairness measures as well. Also note that we mainly focus on developing an optimization framework for stream replication in this paper; the mechanisms for receiver bandwidth estimation and report are out of the scope. These two issues have been extensively studied in the literature [8], [12], and many of the algorithms can be applied in our system.

IV. Intra-Session Optimization: Cases for Fixed and Flexible Number of Streams

A. Problem Formulation

Let $\vec{R}_s$ denote the bandwidth allocation vector for session s, $\vec{R}_s = (r_{s,1}, r_{s,2}, \ldots, r_{s,l_s})$, where l_s is the total number of the replicated streams for session s, and $r_{s,i}$ is the rate of stream i. Without loss of generality, we assume that $r_{s,1} < r_{s,2} <, \ldots, < r_{s,l_s}$; these rates, as well as other bandwidths or rates discussed in this paper, take only discrete values, for two reasons: First, given a finite number of quantizers, the output rate of a video compressor is always discrete [1]; Second, bandwidth allocation is channelized in many multicast or broadcast networks.

For a given $\vec{R}_s$, a receiver with bandwidth t should subscribe to the stream of the best-matching bandwidth $\phi(t, \vec{R}_s) = \max_{r \le t, r \in \vec{R}_s} r$, which is a relatively simple operation. The challenging problem is how to determine $\vec{R}_s$ on the server's side, to which we refer as *intra-session allocation*. The input for intra-session allocation includes the session bandwidth N_s and the receivers' bandwidth distribution $M_{s,t}$. The output is the minimum expected relative mismatch (ERM) for all the receivers in the session, together with the corresponding $\vec{R}_s$. Assume T_s is the maximum receiver bandwidth in session s. An optimal allocation clearly satisfies $r_{s,1} > 0$ and $r_{s,l_s} \le T_s$. The optimization problem thus can be formally described as follows:

$$\begin{aligned} \text{Minimize} \quad & ERM(s, N_s) = \sum_{t=1}^{T_s} M_{s,t} RM\left[t, \phi(t, \vec{R}_s)\right], \\ \text{Subject to} \quad & 0 < r_{s,1} < r_{s,2} <, \ldots, < r_{s,l_s} \le T_s, \\ & \sum_{i=1}^{l_s} r_{s,i} \le N_s. \end{aligned} \tag{2}$$

As discussed before, we consider two versions of the above optimization problem: 1) optimal bandwidth allocation for a given (fixed) number of streams (OptFN) and 2) joint optimization for the number of streams and their respective bandwidths (OptNB), which provides a general guideline for choosing the number of streams in simulcasting systems.

B. Optimization for Fixed Number of Streams (OptFN)

In this scenario, we assume the total number of streams is fixed to a given K. Since the number of valid allocations is finite for $r_{s,k} \in Z^+$ and $r_{s,K} \le T_s$, there exists an optimal bandwidth allocation vector for problem OptFN.

We now show an effective algorithm to solve this problem. Define $\alpha(n, m, k)$ as $\min_{l_s = k, r_{s,k} = m, \sum_{i=1}^{k} r_{s,i} = n} \sum_{t=1}^{T_s} M_{s,t} RM[t, \phi(t, \vec{R}_s)]$, that is, the minimum ERM when a total number of $k (\le K)$ streams are generated with a total bandwidth n, and the bandwidth of stream k is m.

For the case of only one stream ($k = 1$) that occupies all the session bandwidth (i.e., $0 < m = n \le N_s$), we have $\alpha(m, n, 1) = \sum_{t=0}^{m-1} M_{s,t} RM(t, 0) + \sum_{t=m}^{T_s} M_{s,t} RM(t, m)$. For $1 < k \le K$, one more stream is to be added based on a case of $k-1$. Without loss of generality, assume this one is stream k. The reduction of ERM, when stream k is added, depends only on the bandwidth of itself and that of stream $k-1$, because only the receivers that originally subscribe to stream $k-1$ will potentially switch to stream k. If the bandwidth stream k is m, and that of stream $k-1$ is j, the reduction is $DIFF(m, j) = \sum_{t=m}^{T_s} M_{s,t}[RM(t, j) - RM(t, m)]$ for $m \le n \le N_s$ and $k \le m \le \min\{n, T_s\}$. The minimum ERM for this k-stream case thus can be obtained by checking all possible values of $j (= 1, 2, \ldots, m-1)$, leading to the following recurrence relation:

$$\alpha(n, m, k) = \min_{1 \le j < m} \{\alpha(n-m, j, k-1) - DIFF(m, j)\}. \tag{3}$$

Clearly, the solution to problem OptFN is clearly given by $\min_{1 \le n \le N_s, 1 \le m \le T_s} \alpha(n, m, K)$. For the RM function, we have $DIFF(m, j) = \sum_{t=m}^{T_s} M_{s,t}[RM(t, j) - RM(t, m)] = \sum_{t=m}^{T_s} M_{s,t}(m-j)/t = (m-j)\sum_{t=m}^{T_s} M_{s,t} t^{-1}$. For each given m, $\sum_{t=m}^{T_s} M_{s,t} t^{-1}$ is an invariant in the algorithm; hence, its values for $m = 1, 2, \ldots, T_s$ can be pre-calculated and stored in space $O(T_s)$, and the complexity of the optimal allocation algorithm is bounded by $O(N_s^3 K)$.

C. Joint Optimization for Stream Number and Bandwidths (OptNB)

In this scenario, both the number of streams (l_s) and their bandwidth $(r_{s,i})$ are to be optimized. For discrete bandwidth allocation, there is an upper bound of l_s, given by $l_s^{\max} = \lfloor \sqrt{1 + 8N_s}/2 - 1/2 \rfloor$, which corresponds to stream bandwidth allocation $(1, 2, 3, \ldots, l_s)$ subject to $\sum_{t=1}^{l_s} t \le N_s$. Thus, a naïve solution to problem OptNB is to try l_s from 1 to $l_s^{\max}$, and invoke the algorithm for OptFN for each l_s. The complexity of this exhaustive search is $O(N_s^{3(1/2)})$.

A more efficient algorithm can be designed as follows. Let $\beta(n, m) = \min_{r_{s,l_s} = m, \sum_{k=1}^{l_s} r_{s,k} = n} \sum_{t=1}^{T_s} M_{s,t} RM[t, \phi(t, \vec{R}_s)]$, that is, the minimum ERM when the session bandwidth is n, and the bandwidth of stream l_s is m. Since there is no constraint on l_s, the solution to problem OptNB is simply given by $\min_{1 \le n \le N_s, 1 \le m \le T_s} \beta(n, m)$. Note that here $m = n$ implies $k = 1$, i.e., the same boundary condition as in the previous subsection. For $m < n \le N_s$ and $1 < m \le \min\{n, T_s\}$, we have the following recurrence relation:

$$\beta(n, m) = \min_{1 \le j < m} \{\beta(n-m, j) - DIFF(m, j)\}. \tag{4}$$

The explanation to this relation is similar to (3), expect that the index of k is omitted because there is no limit to the number of streams. As a result, calculating $\beta(n, m)$ and obtaining the optimal allocation for OptNB needs only $O(N_s^3)$ time, which is much lower than the exhaustive search algorithm and, interestingly, even lower than OptFN.

D. Remarks on the Lower Bound of ERM

Intuitively, ERM can be reduced if more session bandwidth is allocated. Yet our observation from the solutions for OptFN (fixed number of streams) is that ERM cannot be further reduced after a certain N_s. A trivial bound is $N_s^0 = \sum_{k=1}^{K}(T_s - k + 1) = K(2T_s - K + 1)/2$, corresponding to allocation $(T_s - K + 1, T_s - K + 2, \ldots, T_s)$. A tight bound N_s^{bound} can be derived from the optimal bandwidth allocation for cumulative layered multicast with K layers: If there is no constraint of session bandwidth, we can build a mapping from the cumulative layer bandwidth to the stream bandwidth: $r_{s,k} = \sum_{i=1}^{k} r'_{s,i}$, where $r'_{s,i}$ is the bandwidth of layer i [10]. In this case, the two schemes achieve the same session ERM; specifically, when the layer bandwidth allocation is optimal, this mapping gives the optimal stream bandwidth allocation with no session bandwidth constraint. As a result, N_s^{bound} is given by $\sum_{k=1}^{K}\sum_{i=1}^{k} r'_{s,i}$, which can be calculated in time $O(T_s^2 K)$ [8]. This mapping is illustrated in Fig. 1.

Nevertheless, in a bandwidth-limited case, the optimization structure for stream replication is different from that for cumulative layering, and the choice of K becomes critical. As will be shown in our numerical results, the use of a flexible number of streams can further reduce ERM for session bandwidths beyond N_s^{bound}.

V. ERM-Aware Inter-Session Bandwidth Allocation

We now consider bandwidth allocation among different sessions, with an objective of ensuring both fairness and efficiency. There are various notions of bandwidth fairness, especially in a multicast scenario [16]. Hence, rather than define a new inter-session allocation framework and claim its fairness, we try to explore the unique and useful features of our replication algorithm within existing frameworks. We observe that the session ERM for our optimal replication algorithm is a stepwise function of session bandwidth, and the ERM of OptFN even becomes flat after a certain session bandwidth (see Fig. 1 and the numerical results in Section VI-A). Thus, some bandwidth allocated to a session can be wasted, and distributing it to other sessions would be more reasonable.

We refer to this an enhancement as *ERM-Aware Allocation* (EAA). For illustration, we use an *equal share-based allocation* (ESA) as the basic allocation framework, which uniformly distributes bandwidth among sessions. Denote $mERM(s, n)$ as the minimum ERM when bandwidth n is allocated to session s, and $\tau(s, d)$ as $[mERM(s, n_s) - mERM(s, n_s + d)]/d$, i.e., the reduction of $mERM$ per unit bandwidth when d units are added to session s. The following heuristic algorithm provides a simple EAA implementation based on ESA:

```
1: for $s = 1, 2, \ldots, P$ do
2:   $N_s \leftarrow \lfloor C/P \rfloor$;
3:   while $mERM(s, N_s) = mERM(s, N_s - 1)$ do $N_s \leftarrow N_s - 1$;
4: $\Delta \leftarrow C - \sum_{s=1}^{P} N_s$;
5: repeat
6:   $s' \leftarrow \arg_s \max_{s \in \{1,2,\ldots,P\}, d \le \Delta} \tau(s, d)$, $d' \leftarrow \arg_s \max_{s \in \{1,2,\ldots,P\}, d \le \Delta} \tau(s, d)$;
7:   if $\tau(s', d') > 0$ then $N_{s'} \leftarrow N_{s'} + d'$, $\Delta \leftarrow \Delta - d'$;
8: until $\tau(s', d') = 0$ or $\Delta = 0$.
```

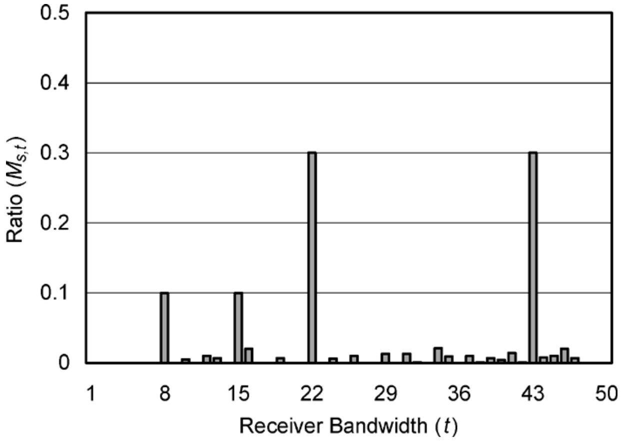


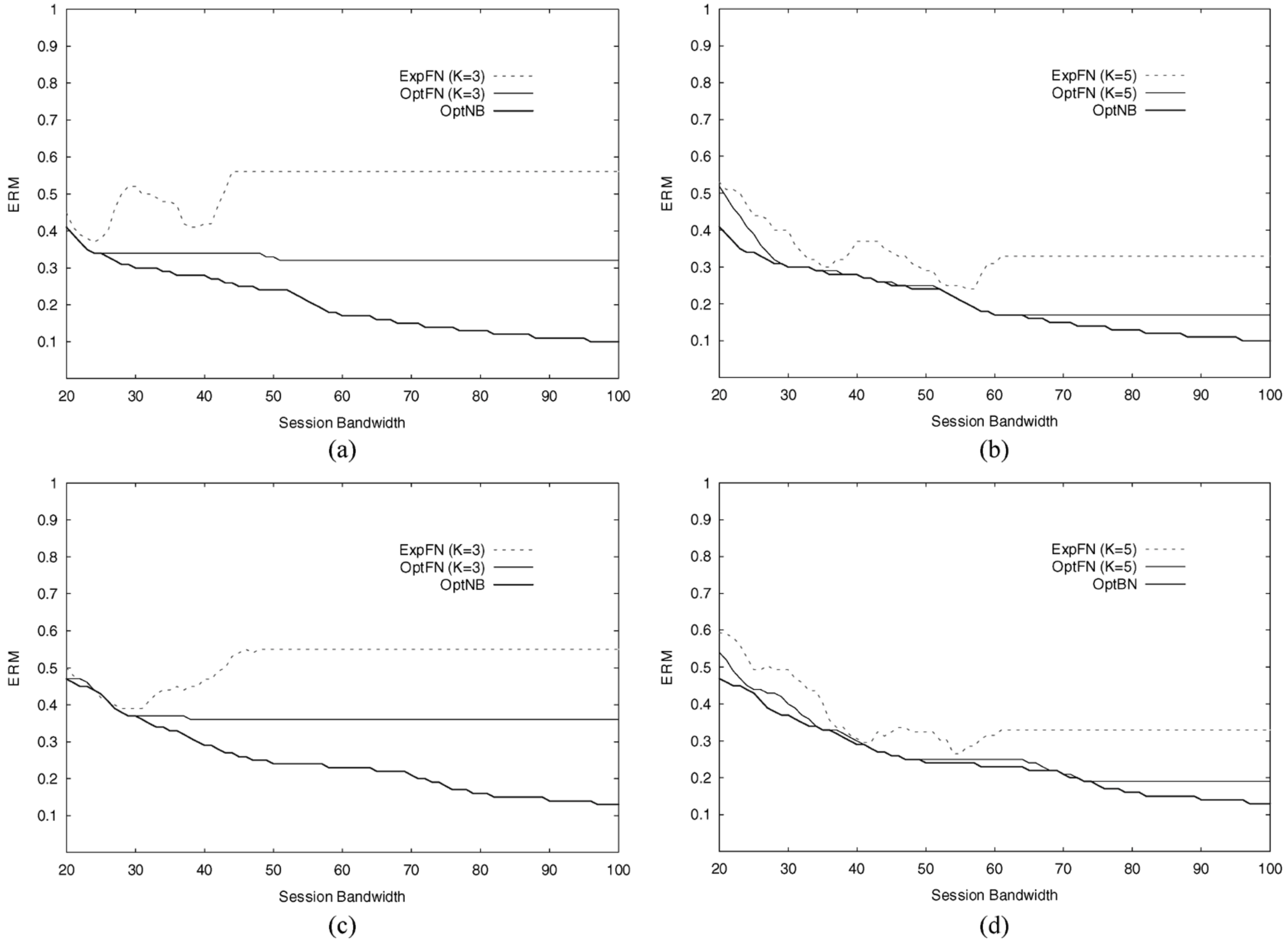
Fig. 1. In this simple example, most receivers' bandwidths are distributed at four points: 8 (10%), 15 (10%), 22 (30%), and 43 (30%). Assume the number of streams is fixed to 3. For cumulative layered multicasting, the cumulative layer bandwidth allocation (8, 22, 43) minimizes the expected mismatch, as long as the session bandwidth is no less than 43. Note that this session bandwidth is even lower than the maximum receiver bandwidth (about 47). If there is no constraint of session bandwidth, this is also gives the optimal stream bandwidth allocation for simulcasting, but the total bandwidth of the streams now is $73\ (= 8 + 22 + 43)$. The mismatch cannot be further reduced by adjusting (either increasing or decreasing) the bandwidth of any stream, even if there is extra session bandwidth. On the other hand, for a limited session bandwidth, say 70 (<73), the allocation (8, 15, 43) becomes the optimal choice for simulcasting. Actually, this is the optimal allocation for any session bandwidth between $66\ (= 8 + 15 + 43)$ and 72. In other words, the session ERM for OptFN is a stepwise function of the session bandwidth.

Given an equal allocation (line 2), the above EAA algorithm first reduces the bandwidth of each session as much as possible without increasing the session's current $mERM$ (line 3). It will then re-allocate the extracted bandwidth (line 4 to 8); each time a session that has the maximum ERM reduction per unit bandwidth is selected (line 6 to 7).

Assume N'_s, $s = 1, 2, \ldots, P$, are the session bandwidths allocated by ESA. It can be easily proved that $\sum_{s=1}^{P} mERM(s, N_s) \le \sum_{s=1}^{P} mERM(s, N'_s)$ and $mERM(s, N_s) \le mERM(s, N'_s)$, $s = 1, 2, \ldots, P$. Hence, EAA reduces not only the average ERM for all the sessions, but also the ERM of each session. In the worst case, EAA yields the same ERM as ESA.

VI. Performance Evaluation

In this section, we evaluate the performance of the optimal replication algorithms, and try to identify the key factors that influence their performance. For the sake of comparison, we also implement a frequently cited nonoptimal scheme: the exponential allocation with a fixed number of streams (ExpFN) [7], [8]. In ExpFN, the stream bandwidths form a geometric progression, i.e., $r_{s,i} = \lfloor \rho^{i-1} r_{s,1} \rfloor$ for $i = 2, 3, \ldots, K$, which can cover a broad dynamic range of receivers' access bandwidths with a limited number streams. To achieve a fair comparison, we assume that both the minimum and the maximum receiver bandwidths are known, and $r_{s,1}$ is set to the minimum. Given constraints $\sum_{i=1}^{K} r_{s,i} \le N_s$ and $r_K \le \eta T_s$, the spanning factor ρ can simply be determined by a bisection search. Here, $\eta < 1$



(a) (b) (c) (d)

Fig. 2. ERM as a function of session bandwidth for different allocation schemes. (a) Distribution 1, $K = 3$ for OptFN and ExpFN. (b) Distribution 1, $K = 5$ for OptFN and ExpFN. (c) Distribution 2, $K = 3$ for OptFN and ExpFN. (d) Distribution 2, $K = 5$ for OptFN and ExpFN.

is a damping factor, which ensures a reasonable portion of receivers can subscribe to stream K. In our experiments, η is set to 0.85, the same as that in [12].

A. Effect of Intra-Session Optimization

1) Intra-Session Bandwidth Distribution: We first study the performance of the stream replication schemes in a single session. To reflect the heterogeneous nature of the receivers, we assume their bandwidths are distributed in w clusters, and each cluster follows a Gaussian distribution. The results for two representative distributions, *Distribution 1* of $w = 3$ and *Distribution 2* of $w = 6$, are reported in this paper. Their minimum and maximum cluster means are 2 and 50, respectively, and the standard deviation of a cluster is 10% of its mean. Thus most bandwidth differences are within $\pm 10\%$, yet a few reach about $\pm 40\%$ or more, reflecting the fluctuation of available bandwidth and the flexibility in device design. We assume the session has 500 receivers and draw 500 samples from the model to obtain an instance of receiver bandwidth distribution. All the results presented below are averages over ten instances.

2) Effect of Session Bandwidth: In this set of experiments, we study the effect of the bandwidth allocated to a session. Fig. 2 shows the session ERM as a function of the session bandwidth. It is clear that both optimal allocation schemes significantly outperform ExpFN; at a medium to high bandwidth, the improvement of ERM is often over 0.2. For example, in Fig. 2(a), the ERM of OptNB is reduced to 0.15 with a medium session bandwidth of 75. The ERM of ExpFN, however, remains higher than 0.5, which translates into an average bandwidth utility under 50%.

More importantly, the performance of ExpFN is not necessarily improved by allocating more bandwidth to a session. In Fig. 2(b), though the ERM of ExpFN at the session bandwidth of 60 is smaller than that at 40, it is noticeably larger than that at only 50. In Fig. 2(a) and (c), the performance is even the worst for any session bandwidth greater than 50. This is because the receivers' bandwidth distribution is not taken into account in this allocation scheme, and hence some unreasonable stream bandwidth settings could occur. To the contrary, as shown in Fig. 2, the ERM of OptFN or OptNB is nonincreasing with the increase of the session bandwidth. This is can also be formally proved from recurrence relations (3) and (4).

The number of streams also influences the performance for the replication schemes. For illustration, in Fig. 2(b), the ERM of OptFN is close to that of the joint optimization scheme (OptBN) for session bandwidth between 30 and 65, implying

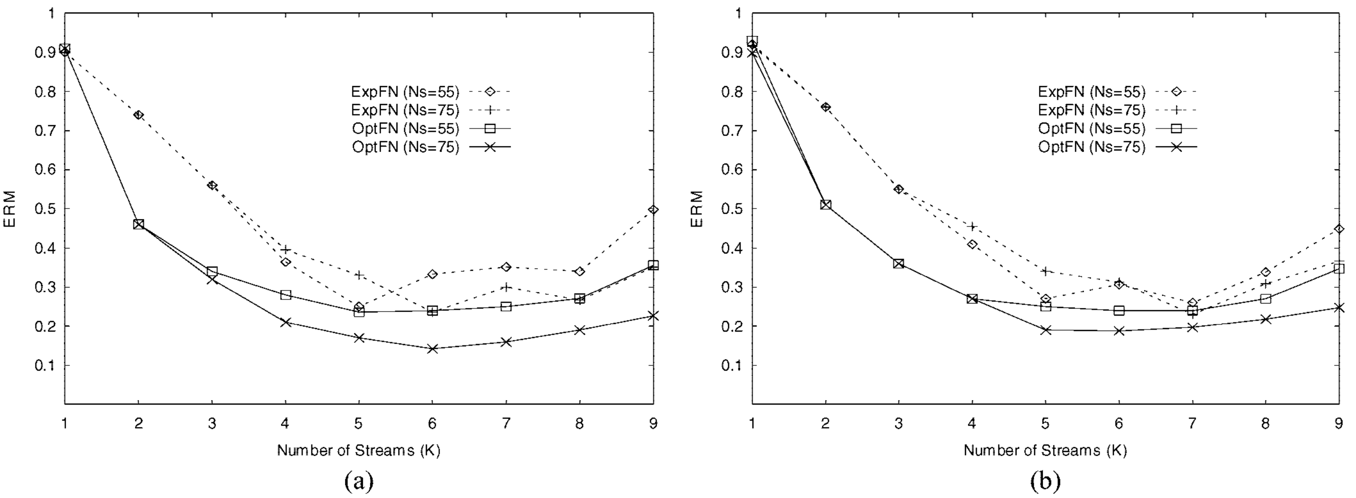


Fig. 3. ERM as a function of the number of streams for OptFN and ExpFN. (a) Distribution 1. (b) Distribution 2.

that five-stream is the best choice for this interval. However, with lower or higher session bandwidth, it is no longer optimal, and the gaps of ERM can be 0.05 or even larger. In Fig. 2(a), OptFN is close to OptNB only for session bandwidth between 20 and 25, and the ERM gaps generally exceed 0.15 for other bandwidths. Moreover, the ERM of OptFN becomes flat for session bandwidths greater than 52, suggesting that the optimal allocation for the three streams has been reached (as illustrated in Fig. 1, this corresponds to the optimal allocation for cumulative layered multicast). On the contrary, OptNB can generate more streams to enjoy the extra bandwidth and hence to further reduce ERM. As a result, with a high session bandwidth (>95), the ERM of OptNB is less than 0.1, which is much lower than that of OptFN. Similar observations can also be made from Fig. 2(c) and (d).

3) Impact of the Number of Streams: To study the impact of the number of streams (K) in OptFN and ExpFN, we vary K from 1 to 9. Fig. 3 shows the results when bandwidths 55 and 75 are allocated to the session, respectively. It can be seen that the variations of the ERMs with different numbers of streams are as large as 0.3 (excluding the single stream case, $K = 1$) for both OptFN and ExpFN. For OptFN, obviously there is an optimal setting for K, at which the session ERM is minimized. Intuitively speaking, if K is small, the receivers' choice is limited and the adaptation is not flexible; an extreme case is $K = 1$, the single-rate transmission. On the other hand, if K is large, the redundancy of replication contradicts the benefit of improved adaptability.

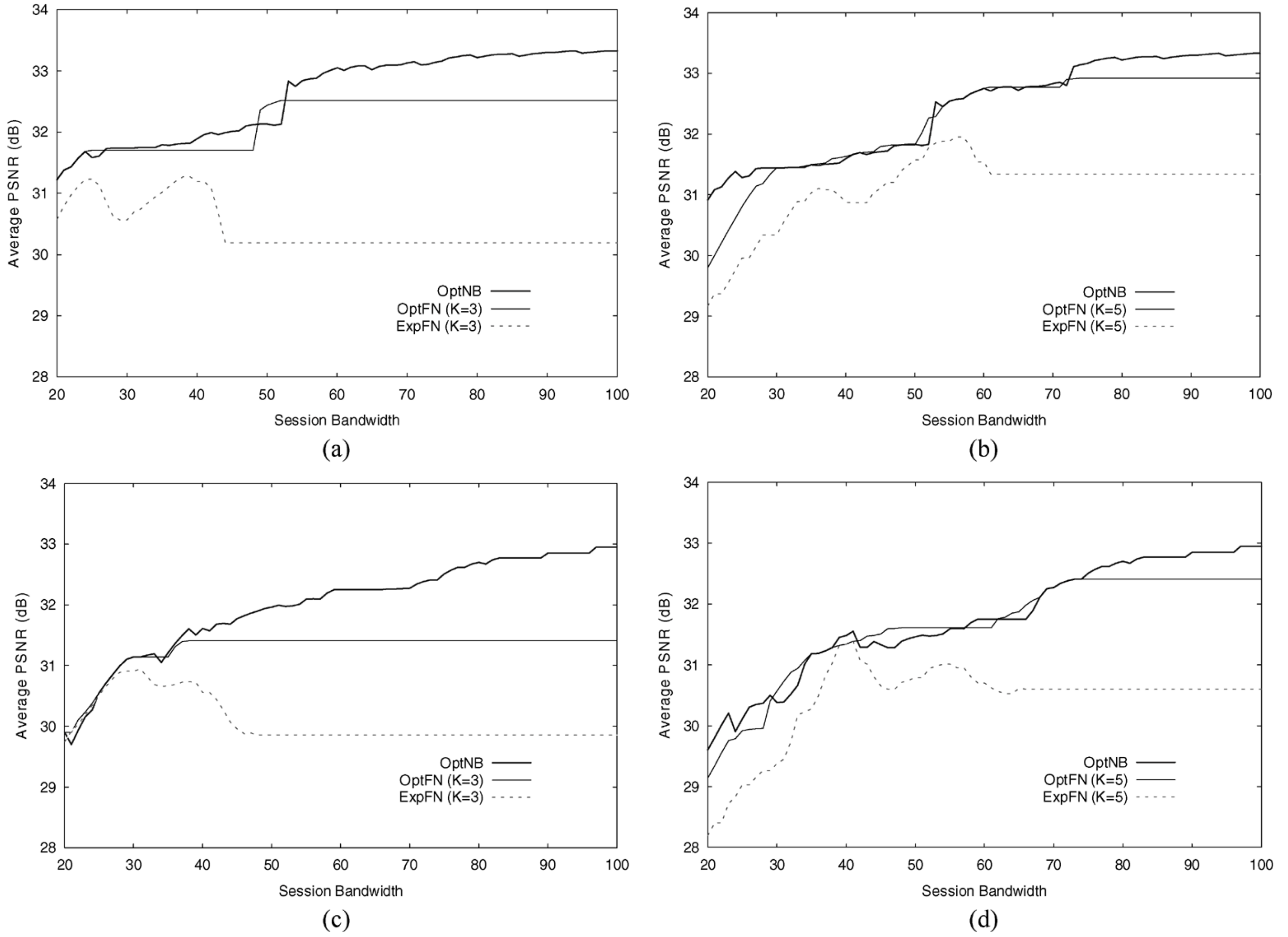
The exact optimal setting of K can be found using the OptNB algorithm. Note that this optimal setting in Fig. 3(a) (Distribution 1) is different from that in Fig. 3(a) (Distribution 2) for the same session bandwidth. Moreover, as shown in Fig. 3, when the session bandwidths are different, the optimal settings change as well, even for the same distribution. Therefore, there is no universal choice for the optimal number of streams, suggesting an adaptive setting of the number of stream and their rates.

4) Perceived Video Quality: Since our target application is video distribution, we also examine the video quality achieved by different replication schemes. We use the standard MPEG-4 video encoder with TM-5 rate control to generate replicated video streams at different rates. The average video quality of all the receivers for a standard test sequence "Foreman (CIF)" is presented in Fig. 4, where the quality is measured by the peak signal-to-noise ratio (PSNR) of the Y channel [1]. It can be seen that the optimal replication algorithms generally improve the perceptual video quality. With medium and high session bandwidths, the gaps of PSNR between OptNB and OptFN are about 0.5 to 1 dB, and that between OptNB and ExpFN are usually larger than 2 dB; both are noticeable from the video coding/decoding point of view. This is consistent with our observations on the relationship between ERM and session bandwidth. Since PSNR often has a nonlinear relationship with transmission bandwidth, Fig. 4 is not simply an inverse and rescaled version of Fig. 2. In particular, the PSNR for OptNB is not necessarily nondecreasing; see, for example, Fig. 4(d), with session bandwidths from 20–50.

B. Effect of Inter-Session Bandwidth Allocation

Finally, we study the effects of inter-session bandwidth allocation. We assume that the demand probabilities for different video programs follow a Zipf distribution of a skew factor 0.271, as suggested by movie rental statistics [15]. The number of clusters for each session is uniformly distributed in between 2 and 7. In the experiments, we assume that there are 2500 receivers belonging to 15 sessions, and draw 2500 samples from the above model.

The different combinations of the intra- and inter-session allocation schemes are compared in Fig. 5. The results are consistent with our previous observations in intra-session allocation: OptNB generally outperforms OptFN when the same inter-session allocation scheme is employed. However, the impact of different inter-session allocation schemes is also nonnegligible. It can be seen that, the ERM-Aware Allocation (EAA) consistently outperforms the Equal Share Allocation (ESA), both with OptNB and with OptFN. At low or medium bandwidths, the performance gaps can be larger than 0.1. More interestingly, for bandwidth around 500, the ERM of OptFN+EAA is quite close to that of OptNB+EAA, and is better than that of OptNB plus ESA. This is because the preset number of streams



(a)

(b)

(c)

(d)

Fig. 4. Average PSNR as a function of session bandwidth for different allocation schemes. (a) Distribution 1, $K = 3$. (b) Distribution 1, $K = 5$. (c) Distribution 2, $K = 3$. (d) Distribution 2, $K = 5$.

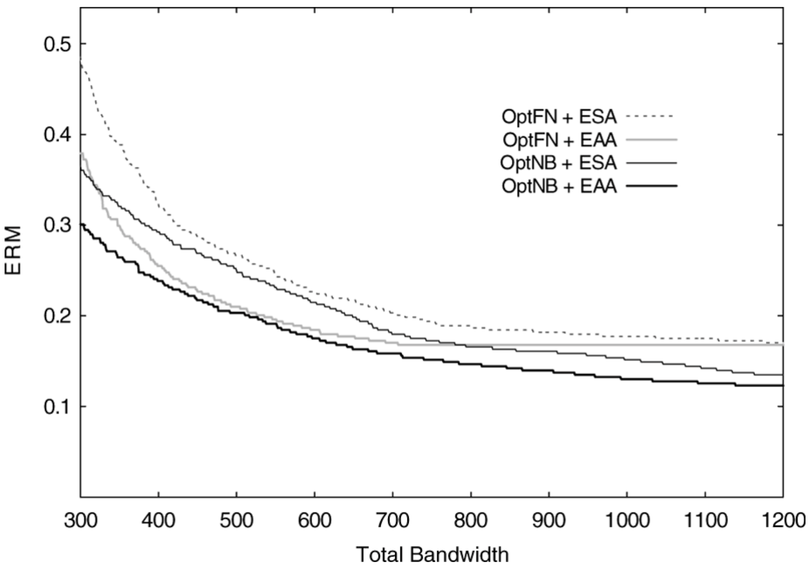


Fig. 5. ERM as a function of the total bandwidth of all the sessions for different combinations of the intra-session and inter-session allocation schemes.

(five in our study) is likely to be the optimal choice for such low to medium bandwidths (refer to Fig. 2), and the choice of inter-session allocation thus has stronger influence on the ERM. To conclude, EAA is particularly suitable for the cases where the stream number is fixed and the bandwidth resource is relatively scarce.

VII. Conclusions

This paper presented a formal study on the problem of stream replication for video simulcasting. We derived efficient algorithms to optimize both the number of streams and the bandwidth for each stream. Numerical results under various configurations demonstrated that an optimal and adaptive bandwidth allocation significantly improves user satisfaction under stringent resource constraints, and an optimal choice of the stream number yields further improvements.

References

[1] Y. Wang, J. Ostermann, and Y.-Q. Zhang, *Video Processing and Communications*. Englewood Cliffs, NJ: Prentice-Hall, Sep. 2001.

[2] S. Cheung, M. H. Ammar, and X. Li, "On the use of destination set grouping to improve fairness in multicast video distribution," in *Proc. IEEE INFOCOM'96*, Mar. 1996, pp. 553–560.

[3] J. Youn, J. Xin, and M.-T. Sun, "Fast video transcoding architectures for networked multimedia applications," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS'00)*, May 2000.

[4] P. Kuhn, T. Suzuki, and A. Vetro, "MPEG-7 transcoding hints for reduced complexity and improved quality," in *Proc. PacketVideo'01*, Apr. 2001.

[5] X. Li and M. H. Ammar, "Bandwidth control for replicated-stream multicast video distribution," in *Proc. HPDC'96*, Aug. 1996.

[6] T. Jiang, E. W. Zegura, and M. H. Ammar, "Inter-receiver fair multicast communication over the Internet," in *Proc. NOSSDAV'99*, Jun. 1999.

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

[8] J. Liu, B. Li, and Y.-Q. Zhang, "A hybrid adaptation protocol for TCP-friendly layered multicast and its optimal rate allocation," in *Proc. IEEE INFOCOM'02*, Jun. 2002.

[9] L. Wu, R. Sharma, and B. Smith, "Thinstreams: an architecture for multicast layered video," in *Proc. NOSSDAV' 97*, May 1997.

[10] T. Kim and M. H. Ammar, "A comparison of layering and stream replication video multicast schemes," in *Proc. NOSSDAV'01*, Jun. 2001.

[11] 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. Packet Video Workshop*, Apr. 2001.

[12] D. Sisalem and A. Wolisz, "MLDA: a TCP-friendly congestion control framework for heterogeneous multicast environments," in *Proc. IEEE/IFIP IWQoS'00*, Jun. 2000.

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

[14] W. Tan and A. Zakhor, "Video multicast using layered FEC and scalable compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 373–386, Mar. 2001.

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

[16] A. Legout, J. Nonnenmacher, and E. W. Biersack, "Bandwidth allocation policies for unicast and multicast flows," *IEEE/ACM Trans. Netw.*, vol. 9, no. 4, pp. 464–478, Aug. 2001.

[17] F. Hartanto, J. Kangasharju, M. Reisslein, and K. W. Ross, "Caching video objects: layers vs versions?," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME'02)*, Aug. 2002.

[18] J. Liu, X. Chu, and J. Xu, "Proxy cache management for fine-grained scalable video streaming," in *Proc. IEEE INFOCOM'04*, Hong Kong, Mar. 2004.

[19] J. Liu, B. Li, and Y.-Q. Zhang, "Adaptive video multicast over the Internet," *IEEE Multimedia*, vol. 10, no. 1, pp. 22–31, Jan./Feb. 2003.

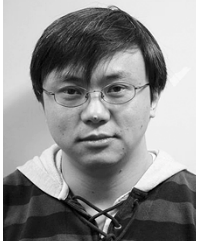
[20] S. Banerjee and B. Bhattacharjee, "A comparative study of application layer multicast protocols," Univ. Maryland, College Park, Tech. Rep., 2002.

[21] X. Zhang, J. Liu, B. Li, and T.-S. P. Yum, "CoolStreaming/DONet: a data-driven overlay network for peer-to-peer live media streaming," in *Proc. IEEE INFOCOM'05*, Miami, FL, Mar. 2005.

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

He is currently an Assistant Professor in the School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, and was an Assistant Professor at The Chinese University of Hong Kong from 2003 to 2004. He is a co-inventor of one European patent and two U.S. patents. His research interests include Internet architecture and protocols, media streaming, wireless ad hoc networks, and service overlay networks.

Dr. Liu was a recipient of Microsoft research fellowship (2000), a recipient of *Hong Kong Young Scientist Award* (2003). He won first-class honors in several regional and national programming contests. He serves as TPC member for various networking conferences, including IEEE INFOCOM'04 and '05. He was TPC Co-Chair for The First IEEE International Workshop on Multimedia Systems and Networking (WMSN'05), Information System Co-Chair for IEEE INFOCOM'04, and a guest-editor for ACM/Kluwer *Journal of Mobile Networks and Applications* (MONET), Special Issue on Energy Constraints and Lifetime Performance in Wireless Sensor Networks. He is an editor of *IEEE Communications Surveys and Tutorials*, a member of the IEEE Communications Society, and an elected member of Sigma Xi.



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

Between 1993 and 1996, he worked on high performance routers and ATM switches at the IBM Networking System Division, Research Triangle Park, NC. Since 1996, he has been with the Department of Computer Science, Hong Kong University of Science and Technology. Since 1999, he has also held an adjunct researcher position at the Microsoft Research Asia (MSRA), Beijing. His current research interests are on adaptive video multicast, peer-to-peer streaming, resource management in mobile wireless systems, across layer design in multihop wireless networks, content distribution and replication. He has published 80 journal papers and held several patents in above areas. He

Dr. Li received the Oversea Outstanding Young Scientist Investigator Award from Natural Science Foundation of China (NSFC) in 2004. He has been on editorial board for IEEE Transactions on Wireless Communications, IEEE Transactions on Mobile Computing,IEEE Transactions on Vehicular Technology, ACM/Kluwer *Journal of Wireless Networks* (WINET), IEEE Journal of Selected Areas in Communications (JSAC)—Wireless Communications Series, ACM *Mobile Computing and Communications Review* (MC2R), SPIE/Kluwer *Optical Networking Magazine* (ONM), Elsevier *Ad Hoc Networks*, SPIE/Kluwer *Optical Networking Magazine*(ONM), KICS/IEEE *Journal of Communications and Networks* (JCN). He served as a guest editor for *IEEE Communications Magazine* Special Issue on Active, Programmable, and Mobile Code Networking (April 2000), 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), IEEE Journal on Selected Areas in Communications Special Issue on Protocols for Next Generation Optical WDM Networks (October 2000), Special Issue on Recent Advances in Service-Overlay Network (January 2004), and Special Issue on Quality of Service Delivery in Variable Topology Networks (September 2004), and ACM/Kluwer *Journal of Mobile Networks and Applications* (MONET) Special Issue on Energy Constraints and Lifetime Performance in Wireless Sensor Networks (2nd Quarter of 2005). In addition, He has been involved in organizing over 40 conferences, including IEEE INFOCOM since 1996. He was the Co-TPC Chair for Infocom'2004.

Ya-Qin Zhang (S'87-M'90-SM'93-F'97) received the B.S. and M.S. degrees in electrical engineering from the University of Science and Technology of China (USTC) in 1983 and 1985, respectively, and the Ph.D. degree in electrical engineering from George Washington University, Washington D.C. in 1989. Healso had executive business training from Harvard University, Cambridge, MA.

He is currently Corporate Vice President of the Microsoft Corporation, Redmond, WA, responsible for the product development in the mobility and embedded devices. He was the co-founder and Managing Director of Microsoft Research Asia, Microsoft's basic research arm in Asia-Pacific region from 2000 to 2003. He was the Director of Multimedia Technology Laboratory at Sarnoff Corporation in Princeton, NJ (formerly David Sarnoff Research Center, and RCA Laboratories). He has been engaged in research and commercialization of MPEG2/DTV, MPEG4/VLBR, and multimedia information technologies. He was with GTE Laboratories Inc. in Waltham, MA, and Contel Technology Center in Chantilly, VA, from 1989 to 1994. He has authored and co-authored over 300-refereed papers in leading international conferences and journals. He has been granted over 50 U.S. patents in digital video, Internet, multimedia, wireless and satellite communications. Many of the technologies he and his team developed have become the basis for start-up ventures, commercial products, and international standards. He serves on the Board of Directors of five high-tech IT companies.