A Comparison of Heterogeneous Video Multicast Schemes: Layered Encoding or Stream Replication

Taehyun Kim and Mostafa H. Ammar, *Fellow, IEEE*

***Abstract*—The heterogeneity of the Internet's transmission resources and end system capability makes it difficult to agree on acceptable traffic characteristics among the multiple receivers of a multicast video stream. Three basic approaches have been proposed to deal with this problem: 1) multicasting the replicated video streams at different rates; 2) multicasting the video encoded in cumulative layers; and 3) multicasting the video encoded in noncumulative layers. Even though there is a common belief that the layering approach is better than the replicated stream approach, there have been no studies that compare these schemes. This paper is devoted to such a systematic comparison. Our starting point is an observation (substantiated by results in the literature) that a bandwidth overhead is incurred by encoding a video stream in layers. We argue that a fair comparison of these schemes needs to take into account this overhead, as well as the specifics of the encoding used in each scheme, protocol complexity, and the topological placement of the video source and the receivers relative to each other. Our results show that the believed superiority of layered multicast transmission relative to replicated stream multicasting is not as clear cut as is widely believed and that there are indeed scenarios where replicated stream multicasting is the preferred approach.**

***Index Terms*—Multimedia communication, video multicasting, scalable video streaming.**

I. Introduction

A SYSTEM for multicasting video over the Internet has to deal with the question of heterogeneity of the receivers capability and/or requirements. Typically, receivers and the paths leading to them will have different reception capacity. We are, therefore, faced with the problem of trying to accommodate this difference among the receivers: low capacity receivers are heavily loaded and suffer from network congestion, but high capacity receivers are lightly loaded and under-utilized. This problem has been addressed by many researchers (e.g., [2]–[4], [13], [18], and [19]). Note that the problem of multicasting video in a heterogeneous environment is important regardless of whether native network layer multicast or application layer multicast [5] are used.

There are three basic approaches.

- **The replicated stream approach** [4], [13]

  In this approach, the video source multicasts several streams with identical content but at different data rates. Each stream is multicast over its own multicast group. Receivers subscribe to the appropriate stream and may switch among streams as their capacity changes. These streams can be generated by encoding the source video with different compression parameters.
- **The cumulative layering approach** [18], [19]

  In this approach, the video is encoded in a base layer and one or more enhancement layers. The base layer can be decoded independently, but the enhancement layers can be decoded cumulatively (i.e., layer k can only be decoded along with layers 1 to $k-1$). The enhancement layers contribute to the improvement of the video quality that leads to the progressive refinement. Each layer is multicast on its own group by a source. Receivers join at least the layer 1 multicast group and add/drop other layers according to their reception capacity.

  MPEG-2, MPEG-4, and H.263 support cumulatively layered encoding by defining scalability modes: data partitioning, signal-to-noise ratio (SNR) scalability, spatial scalability, and temporal scalability [10]–[12]. Combinations of the scalability modes lead to hybrid scalability consisting of multiple layers.
- **The noncumulative layering approach** [3], [9]

  In this approach, the video is encoded in two or more independent layers. Each layer can be decoded independently and provide improvements to the video quality. Each receiver can join any subset of the video layers.

  Multiple description coding (MDC) can be used for noncumulatively layered multicasting. Each description in MDC can lead to the reconstruction of the source video, and multiple descriptions together yield smaller distortion [6]. To provide these features, a number of MDC schemes for video encoding have been developed recently [1], [21], [24]. Each scheme provides different characteristics of compression efficiency, delay, and error resilience.

There is a common belief that the layering approach is better than the replicated stream approach. The main argument is that replicated streams waste bandwidth by essentially duplicating the transmission of the content represented by the base layer (and possibly other lower layers). Even though this is a widely stated conclusion, we are not aware of any studies that have actually compared these approaches in a quantitative and systematic manner.

The goal of this paper is to compare these video multicast techniques. Our starting point is an observation (substantiated by results in the literature) that by encoding a video stream in layers, one incurs a bandwidth overhead [7], [8], [14], [17], [22]. This overhead can sometimes change the bandwidth efficiency in favor of replicated stream video multicasting. We argue that a fair comparison needs to take into account this overhead as well

Manuscript received December 3, 2001; revised July 28, 2004. This work was supported by the AFOSR MURI Grant F49620-00-1-0327, the National Science Foundation Grant ANI-9973115, and by a research grant from Sprint. This paper was presented in part at ACM NOSSDAV 2001, Port Jefferson, NY. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Harrick M. Vin.

T. Kim is with the Wireless and Mobile Systems Group, Freescale Semiconductor, Austin, TX 78735 USA (e-mail: taehyun.kim@freescale.com).

M. H. Ammar is with the Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: tkim@cc.gatech.edu; ammar@cc.gatech.edu).

Digital Object Identifier 10.1109/TMM.2005.858376

as the specifics of the encoding used in each scheme, protocol complexity, and the topological placement of a video source and receivers relative to each other.

This paper is devoted to such a comparison. It is organized as follows. In Section II, we consider the issue of layering overhead in more detail. Section III considers the question of optimizing the rate allocation to layers and to replicated streams. This optimization is necessary in order to insure that a fair comparison of the best layering scheme with the best replication scheme. Section IV and V report on results from experiments we have used to provide a quantitative comparison. Section VI provides a comparison of the protocol overheads involved in the multicast schemes. The paper is concluded in Section VII.

II. Overhead in Layered Video

In this section, we describe how layered encoding of video incurs a bandwidth overhead. Consider a video that is encoded as a single (nonlayered) stream with a given quality and that results in a data rate of R_{nl}, including all protocol/packetization overheads. Let the same video be encoded in m cumulative layers with the data rate for layer i being R_{l_i}, again including all protocol/packetization overhead. We further assume that the layered encoding of the video is such that, if a receiver receives and decodes *all* layers, the quality of the video will be the *same* as the nonlayered video stream with rate R_{nl}.

The basic conclusion that we reach is that

$$R_{nl} \leq R_l = \sum_{i=1}^{m} R_{l_i}.$$

Results in the literature indicate that the equality above is rarely achieved and that R_l can be as much as 20%–30% higher than R_{nl}.

We substantiate this conclusion as follows.

- **Information theoretic results**

  These results are derived in terms of the *rate distortion function*, $R(P, \Delta)$, which describes the required rate to encode a memoryless source at a maximum distortion of Δ. The distortion is a measure of the quality degradation represented by the encoding of the source.

  The general result is that, for the same source and the same distortion, a *successively refined* (i.e., layered) encoding requires at least as much data rate as a nonlayered encoding [15]. While equality is possible, it requires a strict Markovian condition to apply to the source and is generally not achievable. Moreover, the result in [14] shows that the performance of the layered encoding is not better than that of nonlayered encoding for a finite-length block code, even if the Markovian condition holds.
- **Packetization overhead**

  For certain scalability modes, enhancement layers are designed to be syntactically independent of one another. Along with the residual information, a data stream needs to also carry syntactic data, such as picture header, start codes, group of pictures (GOP) information, and macroblock header. This can incur a large amount of overhead especially at low data rates [17].
- **Experimental evidence**

  Fig. 1 shows an experimental result of the video quality versus data rates for the *flower* sequence by comparing MPEG-2 SNR scalability and nonscalability mode. The video quality is measured in peak signal-to-noise ratio (PSNR) by varying quantization step size. A layered stream has two layers consisting of a base layer and an enhancement layer at (Q_b, Q_e), where Q_b is the quantization step size of the base layer and Q_e is that of the enhancement layer. Both PSNR and data rate are averaged over the entire video.

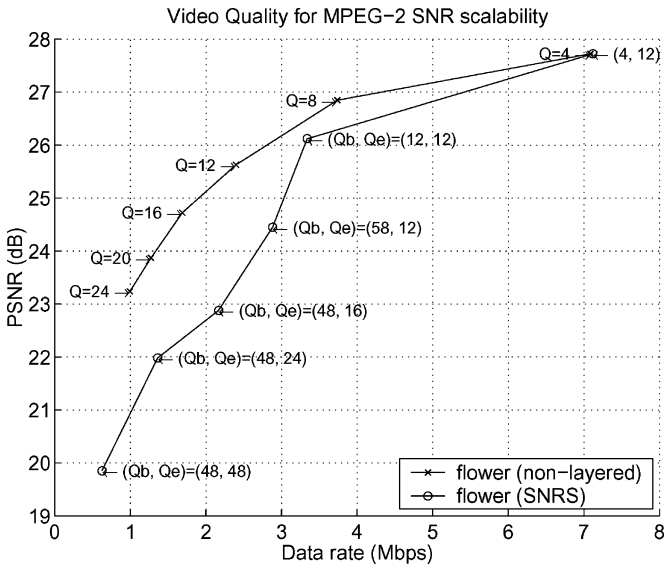

Fig. 1. Performance comparison of MPEG-2 SNR scalability and nonscalability mode. The layering overhead ranges from 0.4% to 117%.

This result demonstrates that a layered stream requires more data rate than a nonlayered stream to provide the same quality. The difference ranges from 0.4% at 27.7 dB to 117% at 23.2 dB. Note that the difference is expected to grow as the number of layers increases, since the accumulation of the redundancy leads to the increase of the overall distortion in layered video encoding [25].

Similar and more extensive experimental results can be found in [16]. The authors investigated the impact of the number of layers, bit rates, and packet loss on the perceptual video quality as determined by subjects scoring the quality of the video, when MPEG-2 data partitioning and SNR scalability are employed. The experimental results showed that the difference ranges from nearly 0 for the highest quality video (scoring close to 4.5) to 57% for fair quality video (scoring close to 3). For a score of 4 (good quality video), the overhead varies from 2% (the *flower* sequence) to 49% (the *basket* sequence). Recent research efforts have focused on developing an efficient scalable video encoding. One of the most significant results is MPEG-4 fine-grained scalability (FGS). Extensive experimental results can be found in [20]. The authors compared the performance of FGS with the traditional SNR scalability, and showed that FGS provides better performance than SNR scalability. However, it was found that FGS suffers coding inefficiency if a video sequence has a high temporal correlation, or if the data rate of base layer is small. Experimental results show that there exists a coding overhead, and the overhead ranges from 0 in a high degree of motion to 50% in low degree of motion.

- **Protocol overhead**
  The nature of the subscription to multiple layers in layered video multicasting may cause additional overhead, as the receiver needs to manage these multiple subscription. For example, in the context of receiver-driven layered multicast (RLM) [19], the probing mechanism of available bandwidth depends on join experiments. However, join experiments incur bandwidth overhead since they require to send a join message and to multicast a message for shared learning. The amount of bandwidth overhead is increased, as the group size of a multicast group grows. Also the subscription of multiple layers requires more buffer size and better synchronization capability than replicated stream video multicasting. More discussions and experimental results are presented in Section VI.

III. Optimizing Stream Rates

To carry out a fair comparison of the layered and replicated stream multicast schemes, we need to insure that each scheme is optimal. We also need to insure that a similar set of rate choices is available for all schemes. The question here is how to determine the number and rates for the set of replicated streams and for the layers.

In this section, we present: 1) a *rate allocation* algorithm to determine the data rate of each stream and 2) a *stream assignment* algorithm to determine the reception rate of each receiver by aggregating the data rates of the assigned streams in layered and replicated stream multicasting. The goal of these algorithms is to maximize the bandwidth utilization by each scheme for a given network, a particular set of receivers, and given available bandwidth on the network links.

To this end, we model the network by a graph $G = (V, E)$, where V is a set of vertices representing routers and hosts. E is a set of edges representing connection links defined over $V \times V$. A set of receivers is defined by $C = \{c_i | c_i \in V, i = 1, \ldots, n\}$, where n is the number of receivers.

An *isolated rate* for each receiver is defined by the data reception rate of the receiver if there is no constraint from other receivers in the same session [13]. The isolated rate can be computed by the Dijkstra's algorithm.

A bandwidth function $B : E \rightarrow \mathbf{R}^+$ is defined on E with $b_j = B(e_j)$, where $\mathbf{R}^+$ is the set of positive real numbers. The bandwidth function is considered as a measure of the residual bandwidth available on the link e_j.

A. Cumulatively Layered Multicasting

1) Rate Allocation: A cumulatively layered multicast session is defined by $\alpha = \{\alpha_i | \alpha_i \in \mathbf{R}^+, i = 1, \ldots, m\}$, where α_i is the data rate of layer i and m is the number of layers.

The first rate-allocation algorithm for cumulatively layered video multicasting was proposed in [23] by maximizing the average signal reception quality. The authors in [26] proposed an optimal receiver partitioning algorithm to determine the optimal stream rates using dynamic programming. We adopt the latter algorithm to determine the optimal data rates of α_i: we define the group utility function by $U(\{j+1, \ldots, i\}) = (i - j) f(r_{j+1})$, where r_{j+1} is the isolated rate of the receiver $j+1$ and $f(r_{j+1})$ is an effective rate allocation function. The *effective rate allocation function* is defined by $f : \mathbf{R}^+ \rightarrow \mathbf{R}^+$ describing the *effective reception rate*, which is the data reception rate contributing to video quality. By applying the optimal receiver partitioning algorithm with this group utility function, we can maximize the overall effective reception rate.

1: Compute the isolated rates
2: Assign $\Sigma_i \alpha_i$ that does not exceed the isolated rate

Fig. 2. Stream assignment algorithm for cumulatively layered multicasting. A receiver can subscribe to as many layers as possible within its capability.

2) Stream Assignment: The stream assignment algorithm is presented in Fig. 2, given the stream rates α_i. Note that we assume each layer is routed over the same path and each receiver can join as many layers as possible (line 2). Hence, the reception rate is determined by the sum of stream rates that does not exceed the isolated rate.

B. Replicated Stream Multicasting

1) Rate Allocation: A replicated stream multicast session is defined by $\beta = \{\beta_i | \beta_i \in \mathbf{R}^+, i = 1, \ldots, m\}$, where β_i is the data rate of a replicated stream and m is the number of replicated streams.

We determine the stream rates based on α_i, determined in Section III-A1. Specifically, β_1 corresponds to the base layer of cumulative layering and the other stream rates are determined in a cumulative manner: if a receiver can join up to k layers in cumulative layering, the receiver has the capability to join a replicated stream of data rate $\beta_k = \sum_{i=1}^{k} \alpha_i$.

2) Stream Assignment: In the stream assignment algorithm, it is required that the data rate of every receiver is strictly greater than zero so that there is no receiver that cannot receive any stream. This requirement can be satisfied by multicasting β_1 to all receivers, and therefore a receiver can subscribe to either the base layer stream or a higher rate stream.

We define δ_i, such that

$$\delta_i = \begin{cases} \beta_1, & i = 1 \\ \beta_i - \beta_1, & 1 < i \leq m. \end{cases}$$

The stream assignment algorithm for replicated stream multicasting determines the data reception rate of a receiver, given the set of data rates $\delta = \{\delta_i | \delta_i \in \mathbf{R}^+, i = 1, \ldots, m\}$. We define Ω_i to be a set of receivers, such that $\Omega_i = \{c_j | \phi_\delta(c_j) = \delta_i\}$, where ϕ_δ is the rate-allocation function defined by $\phi_\delta : C \rightarrow \delta$.

We set up two objectives for stream assignment.

1) The minimum reception rate of all receivers is strictly greater than zero.
2) Maximize $Z_\delta = \sum_{i=1}^{m} |\Omega_i| \delta_i$ subject to $\sum_{i \in \Gamma_{e_j}} \delta_i \leq b_j$, where $\Gamma_{e_j} = \{i | e_j \in E_i\}$, $T_i = (V_i, E_i)$, and T_i is a multicast tree for a replicated stream i.

We develop a greedy algorithm to achieve the requirements. The algorithm is described in Fig. 3. We first allocate δ_1 to all receivers to satisfy the minimum reception rate constraint (lines 2–7). Next, a receiver is assigned a stream that has not yet been assigned and has a maximum product of the group size and the effective reception rate until *every receiver* is assigned to at least one stream (lines 8–18). In stream selection, we assign an identity function to the effective rate allocation function for replicated stream multicasting, since a nonlayered stream is not supposed to incur layering overhead. Therefore, every receiver subscribes to either a high-quality stream (line 12) or the base layer stream (line 15).

1: Compute the isolated rates
2: **if** all isolated rates are greater than δ_1 **then**
3: $b_j \leftarrow b_j - \delta_1$, where $b_j = B(e_j)$ and $e_j \in E$
4: $\delta \leftarrow (\delta \backslash \delta_1) \cup \{0\}$
5: **else**
6: There is no feasible solution
7: **endif**
8: **while not** $\Omega = \emptyset$ **do**
9: Compute the isolated rates
10: Select a stream i with $|\Omega_i|\delta_i = \max_j |\Omega_j|\delta_j$ $(\delta_i, \delta_j \in \delta)$
11: **if** $\delta_i > 0$ **then**
12: Assign δ_i to Ω_i
13: Reduce the link capacity leading to Ω_i
14: **else**
15: Assign δ_1 to Ω_i
16: **endif**
17: $\Omega \leftarrow \Omega \backslash \Omega_i$
18: **enddo**

Fig. 3. Stream assignment algorithm for replicated stream multicasting. A receiver can subscribe to either the base layer stream or high quality stream.

The feasibility of the stream assignment algorithm is guaranteed, if the data rate of a replicated stream multicasting session is determined by the rate allocation algorithm in Section III-B-1. This is because the data rate of the base layer is originally determined by the optimal partitioning algorithm. Otherwise, this algorithm may not provide a feasible solution when any receiver has an isolated rate smaller than δ_1.

Note that the rate allocation/stream assignment scheme may reduce the data reception rate of a receiver compared to cumulatively layered multicasting. However, this does not always guarantee the effective reception rate of cumulatively layered multicasting is greater than that of replicated stream multicasting, when we take the layering overhead into account.

C. Noncumulatively Layered Multicasting

1) Rate Allocation: A noncumulatively layered multicast session is defined by $\gamma = \{\gamma_i | \gamma_i \in \mathbf{R}^+, i = 1, \ldots, m\}$, where γ_i is the data rate of a noncumulatively layered stream and m is the number of streams. A set of receivers assigned to the stream i is defined by $\Omega_i' = \{c_j | \gamma_i \in \phi_\gamma(c_j)\}$, where ϕ_γ is the stream rate function defined by $\phi_\gamma : C \rightarrow P(\gamma)$ and $P(\gamma)$ is a power set of γ. The set of all receivers is $\Omega' = \cup_{i=1}^m \Omega_i'$.

In noncumulatively layered multicasting, a receiver can subscribe to any subset of layers. This property provides a fine granularity for noncumulatively layered multicasting. For example, given a noncumulatively layered stream $\gamma = \{1, 2, 4\}$, a heterogeneity resulting from seven different isolated rates of {1, 2, 3, 4, 5, 6, 7} can be accommodated through selective subscription. Hence, the noncumulatively layered session γ shows the same performance as the cumulatively layered session $\alpha = \{\alpha_i | \alpha_i = 1, i = 1, \ldots, 7\}$. This example demonstrates that the heterogeneity caused by $\sum_{i=1}^m \binom{m}{i} = 2^m - 1$ different link capacities can be accommodated by aggregating the reception rates of m noncumulative layers.

In this section, we propose a rate allocation algorithm for noncumulatively layered streams. The stream rates γ_i are allocated based on α_i as follows:

$$\begin{aligned}
\gamma_1 &= \alpha_1 \\
\gamma_2 &= \alpha_1 + \alpha_2 \\
\gamma_1 + \gamma_2 &= \alpha_1 + \alpha_2 + \alpha_3 \\
\gamma_3 &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 \\
&\cdots \\
\gamma_2 + \gamma_3 + \cdots + \gamma_m &= \alpha_1 + \alpha_2 + \cdots + \alpha_{2^m - 2} \\
\gamma_1 + \gamma_2 + \gamma_3 + \cdots + \gamma_m &= \alpha_1 + \alpha_2 + \cdots + \alpha_{2^m - 2} + \alpha_{2^m - 1}
\end{aligned}$$

where the optimum α_i can be determined in Section III-A-1.

We simplify this relationship in a matrix form: $\mathbf{AX} = \mathbf{BY}$, where $\mathbf{A}$ is a binary counting matrix, $\mathbf{B}$ is a lower triangular matrix, $\mathbf{X}$ is a vector of the allocated data rates of noncumulative layering, and $\mathbf{Y}$ is a vector of the optimal data rates of cumulative layering [15].

Note that it is not generally feasible to determine the data rate γ_i for given α_i, since the number of equations exceeds that of unknown variable γ_i. We develop an approximate rate allocation scheme by minimizing the mean-square error $Z = (\mathbf{AX} - \mathbf{BY})^T(\mathbf{AX} - \mathbf{BY})$. Hence, the allocated data rates of noncumulatively layered multicast streams are determined by $\mathbf{X} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{BY}$.

2) Stream Assignment: We have two objectives to assign the noncumulatively layered streams.

1) The minimum reception rate of all receivers is strictly greater than zero.
2) Maximize $Z_\gamma = \sum_{i=1}^m |\Omega_i'| f(\gamma_i)$ when $\sum_{i \in \Gamma_{e_j}} \gamma_i \leq b_j$, where $\Gamma_{e_j} = \{i | e_j \in E_i\}$, $T_i = (V_i, E_i)$, T_i is a multicast tree for a noncumulatively layered stream i, and $f(\gamma_i)$ is the effective rate allocation function.

1: Compute the isolated rates
2: **if** all isolated rates are greater than γ_1 **then**
3: Assign γ_1 to all receivers
4: $b_j \leftarrow b_j - \gamma_1$, where $b_j = B(e_j)$ and $e_j \in E$
5: $\gamma \leftarrow (\gamma \backslash \gamma_1) \cup \{0\}$
6: **else**
7: There is no feasible solution
8: **endif**
9: **while not** $\gamma = \emptyset$ **do**
10: Compute the isolated rates
11: Select a stream i with $|\Omega_i'| f(\gamma_i) = \max_j |\Omega_j'| f(\gamma_j)$ $(\gamma_i, \gamma_j \in \gamma)$
12: **if** $\gamma_i > 0$ **then**
13: Add γ_i to Ω_i'
14: Reduce the link capacity leading to Ω_i'
15: **endif**
16: $\gamma \leftarrow \gamma \backslash \gamma_i$
17: **enddo**

Fig. 4. Stream assignment algorithm for noncumulatively layered multicasting in which a receiver can subscribe to multiple streams. The data rate of the aggregated streams leads to the minimum distortion.

In Fig. 4, we present a greedy algorithm for noncumulatively layered multicasting to assign a stream with a maximum value of the product of the group size and the effective reception rate. We first allocate the base layer stream, γ_1, to all receivers (lines 2–8). A receiver is assigned every layer which has not yet been multicast to any receiver and maximizes the product of group size and effective reception rates until *every stream* is assigned (line 11). The data reception rate of a receiver is the sum of data rates of all assigned streams, since the aggregated data rate of noncumulatively layered streams leads to the minimum distortion.

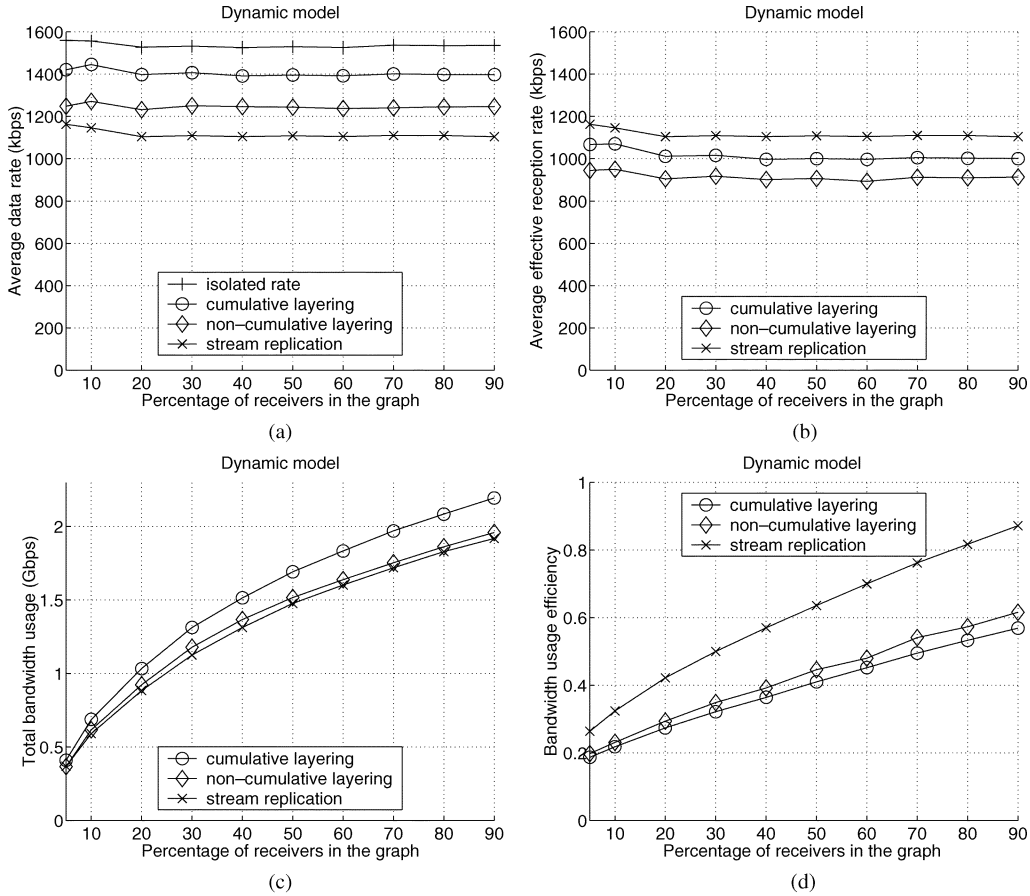

Fig. 5. Experiment results in the dynamic model under the random receiver distribution: (a) Reception rate, (b) effective reception rate, (c) total bandwidth usage, and (d) efficiency. Replicated stream multicasting shows the largest effective reception rate in (b) and the best bandwidth usage efficiency in (d).

IV. Models in Experiments

We compare the performance of the video multicast schemes by experiment. The main goal in the experiment is to evaluate the impact of the parameters, such as the amount of layering overhead and the topological placement of receivers, on the video reception quality. All schemes use the rates and stream assignment as determined in Section III.

A. Network Model

We use GT-ITM [27] to generate 100 different transit-stub graphs representing hierarchical Internet topologies. The graphs consist of 1,640 nodes including ten transit domains, four nodes per transit domain, four stubs per transit node, and ten nodes in a stub domain (i.e., the number of nodes is $1640 = 10 \cdot 4 \cdot (1 + 4 \cdot 10)$). We assign 2.4 Gbps to transit-to-transit edges; 10 Mbps and 1.5 Mbps to stub-to-stub edges; and 155 Mbps, 45 Mbps, and 1.5 Mbps to transit-to-stub edges. The available link bandwidth is chosen uniformly randomly in the range 1%–80% of the full capacity of the edge.

B. Layering Overhead Models

Unless otherwise mentioned, the number of cumulatively layered video streams and the number of replicated video streams are 8, and the number of noncumulatively layered video streams is 4, as discussed in Section III-C-1. The amount of overhead incurred by cumulative and noncumulative layering is modeled as follows.

In our preliminary work in [15], we assumed that the amount of layering overhead is affected proportionally by the amount of the data reception rate. However, the amount of layering overhead depends on many parameters in practice, such as the specifics of encoder, the degree of motion, and/or the dynamic range of the bandwidth (i.e., $[R_1, R_l]$, where R_1 is the data rate of base layer and R_l is the sum of the data rates of all layers).

In this paper, we consider a dynamic overhead model. The dynamic overhead model captures the notion of the dynamically varying nature of the layering overhead. The model is based on the experimental results in [20]. The authors showed that the *stefan* sequence exhibiting high temporal correlation between frames (i.e., low degree of motion) incurs bandwidth overhead, since FGS exploits the temporal information only at the base layer.

The amount of layering overhead increases when the data rate of base layer is small (i.e., wide dynamic range). Based on the experimental results, we model the layering overhead by linear interpolation: the amount of layering overhead is given by $520 - 1.6 \cdot R_{l_1}$ (kbps), where R_{l_1} is the data rate of the base layer and $R_{l_1} \leq 325$ kbps. Note that the layering overhead is assumed zero, when $R_{l_1} > 325$ kbps.

C. Performance Measures

In the experiment, we compare the performance by using the following measures:

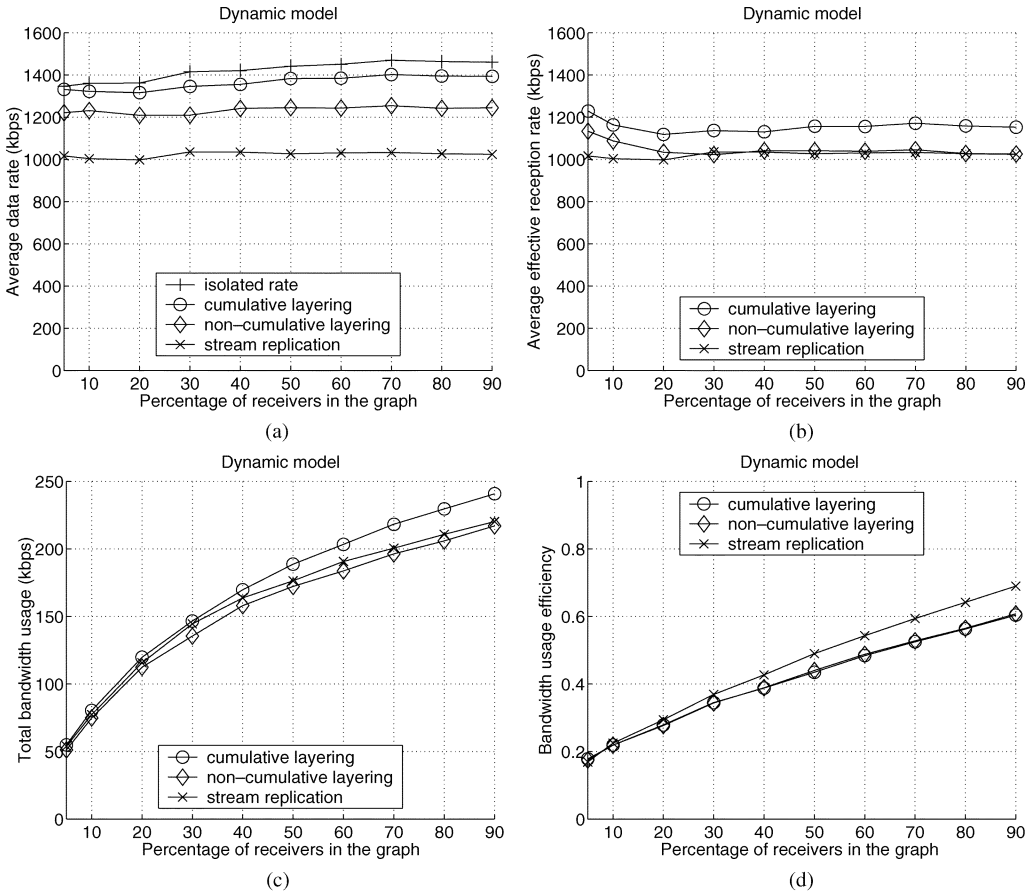

Fig. 6. Experiment results in the dynamic model under the clustered receiver distribution: (a) Reception rate, (b) effective reception rate, (c) total bandwidth usage, and (d) efficiency. Both cumulatively and noncumulatively layered video multicasting achieves greater data reception rate and greater effective reception rate than that of replicated stream multicasting.

- the *average reception rate* which is the average data rate received by a receiver;
- the *average effective reception rate* where the effective reception rate at a receiver is defined by the amount of data received less the layering overhead;
- the *total bandwidth usage* calculated by adding the total traffic carried by all links in the network for the multicast session—including all layers and all replicated streams;
- the *efficiency* defined by

$$\text{efficiency} = \frac{\text{total effective reception rate}}{\text{total bandwidth usage}}.$$

The efficiency is a measure of delivered data rate contributing to the video quality for each unit of bandwidth used in the network.

V. Experiment Results

A. Random Distribution

In the first experiment, we randomly select a server and receivers from a set of nodes in the graph. Receivers are selected from all domains which results in random distribution of receivers. We investigate the performance of the video multicast schemes by varying the number of receivers.

Fig. 5 shows the experiment results of the video multicast schemes under the dynamic overhead model. In Fig. 5(a), the average reception rate of cumulative layering, noncumulative layering, and replicated stream multicasting are given by 91%, 81%, and 73% of the isolated rate. However, the effective reception rate of replicated stream video multicasting is the largest in Fig. 5(b). Therefore, we can expect that the average video quality of replicated stream video multicasting is the best. The efficiency of replicated stream video multicasting is also the best in Fig. 5(d).

B. Clustered Distribution

The layered video multicast schemes achieve better bandwidth efficiency when multiple streams share the bottleneck link. When the receivers are placed in one domain, it is more probable that many of the receivers share a bottleneck link. Hence, the layered video multicasting would be more efficient than replicated stream video multicasting.

Fig. 6 shows the experiment results under the clustered receiver distribution, where receivers are chosen within only one transit domain and a sender is selected from another domain. Compared with Fig. 5, the performance of layered video multicasting is improved but that of replicated stream video multicasting is degraded. Even though some amount of layering overhead is incurred, the effective reception rate of cumulatively layered video multicasting is always greater than that of replicated stream multicasting: 11% in Fig. 6(b). The decrease of replicated stream video multicasting accounts for the decrease of the efficiency. Therefore, we can expect the performance characteristics are changed in favor of layered video multicasting, when receivers are clustered in a small number of domains.

VI. Protocol Complexity

In this section, we consider the protocol complexity of cumulatively layered multicasting and replicated stream multicasting, since no existing protocol supports all three schemes. We compare the protocol complexity by using RLM [19]. In RLM, a receiver has the capability to decide whether to drop an additional layer or not. The decision is made by performing a join experiment. Join experiments incur a bandwidth overhead, since a receiver carrying out the experiment sends a join message and multicasts a message identifying the experimental layer to the group. In addition, the shared learning mechanism requires each receiver to maintain significant amount of state information even if it is not necessary.

We present below our protocol complexity analysis. The network is modeled by a graph $G = (V, E)$, where V is a set of vertices and E is a set of edges. A set of receivers is defined by $C = \{c_i | c_i \in V, i = 1, \ldots, n\}$, where n is the number of receivers. A set of video stream is defined by $R = \{r_i | r_i \in \mathbf{R}^+, i = 1, \ldots, m\}$, where r_i is the data rate of a video stream and m is the number of streams. A set of receivers assigned to stream i is defined by $\Omega_i = \{c_j | \phi(c_j) = r_i\}$, where ϕ is the rate allocation function defined by $\phi : C \rightarrow R$. Since we determine the data rate of replicated streams by adding the data rate of the cumulative layering, the receivers in Ω_i can subscribe to the replicated stream β_i, where $\beta_i \leq r_i$, or can accommodate the cumulatively layered stream up to layer i, such as $\alpha_1 + \alpha_2 + \cdots \alpha_i \leq r_i$.

In cumulatively layered video multicasting, the average group size is given by $(1/m)\sum_{k=1}^{m} k|\Omega_k|$, since a receiver can join multiple groups and $(1/m)(\sum_{k=1}^{m}|\Omega_k| + \sum_{k=2}^{m}|\Omega_k| + \cdots + \sum_{k=m}^{m}|\Omega_k|) = (1/m)\sum_{k=1}^{m} k|\Omega_k|$. In a join experiment, cumulatively layered video multicasting requires a receiver send one join message and multicast a message reporting a join experiment to the receivers in the same group. When the link capacity does not change for a long period, the receiver will return to the previous state after a detection time and it has to send a leave message. Hence, the average number of messages in a join experiment is two unicast messages and one multicast message to $(1/m)\sum_{k=1}^{m} k|\Omega_k|$ receivers.

On the other hand, the average group size in replicated stream video multicasting is $(1/m)\sum_{k=1}^{m}|\Omega_k|$, since every receiver joins only one group. In replicated stream video multicasting, a receiver sends one leave message, one join message, and one multicast message reporting each join experiment. When the link capacity is in the steady state, it has to return to the previous group that involves another one join and one leave messages. The average number of messages in a join experiment is four unicast messages and one multicast message to $(1/m)\sum_{k=1}^{m}|\Omega_k|$ receivers. Therefore, the protocol overhead in cumulatively layered multicasting and replicated stream multicasting consists of two or four unicast messages and one multicast message. The cost of a multicast message is dominant when the number of receivers is large.

Fig. 7 presents the experimental results of the average group size and the average number of groups joined by a receiver, when the receivers are randomly distributed. In Fig. 7(a), the group size in cumulatively layered video multicasting is twice as large as that in stream replication. Hence, layered video multicasting requires more bandwidth to multicast a message reporting the join experiment and more memory to keep state information. In Fig. 7(b), we can expect that a receiver in a cumulatively layered video multicast session requires more buffer size and better synchronization capability than replicated stream video multicasting, since a receiver in cumulative layering subscribes to more than five layers on average whereas a receiver in stream replication subscribes to only one stream.

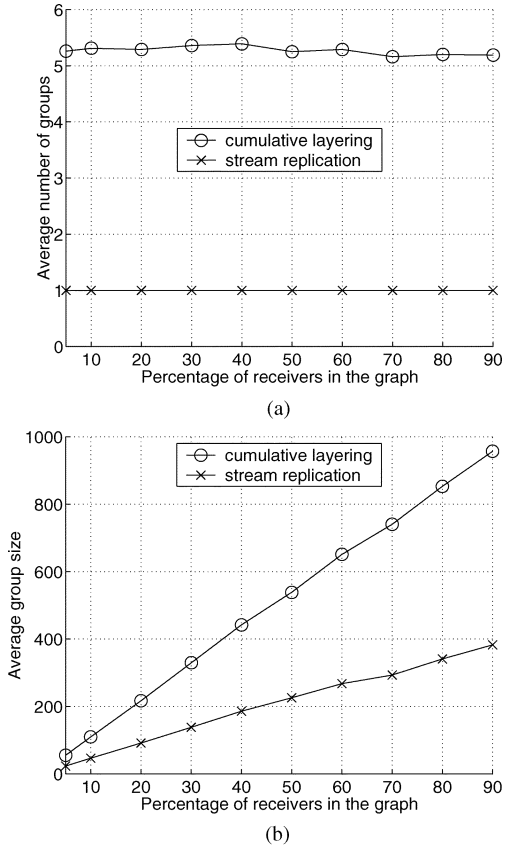

Fig. 7. Experiment results for protocol complexity: (a) Average group size and (b) average number of groups. The nature of joining multiple multicast groups in cumulative layering leads to twice larger group size than replicated stream multicasting in (b).

VII. Conclusion

In this paper, we undertake a comparison between the cumulative/noncumulative layering and replicated stream video multicast schemes. These schemes have been proposed for multicasting to a set of receivers with heterogeneous reception capabilities. While it has been generally accepted that layering is superior to stream replication, this does not appear to be based on a systematic and quantitative comparison of these schemes. We undertake such a comparison here. We first argue that a fair comparison needs to be taken into account: 1) the layering bandwidth overhead; 2) the specifics of the encoding of the layers or replicated streams; 3) the complexity of the protocol required to allow receivers to join and leave the appropriate streams; and 4) the topological placement of receivers relative to each other and relative to the video source. Our results demonstrate the effect of these dimensions on the relative performance of three schemes.

They also show the conditions under which each scheme is superior.

Our work has focused on video multicasting applications. Layering and replication of multicast transmission has also been proposed for bulk-data multicast applications. The layering overhead does not apply in those circumstances, however, the other comparison dimensions will still come into play.

REFERENCES

[1] J. Apostolopoulos, "Reliable video communication over lossy packet networks using multiple state encoding and path diversity," in *Proc. VCIP 2001*, San Jose, CA, Jan. 2001.

[2] J.-C. Bolot, T. Turletti, and I. Wakeman, "Scalable feedback control for multicast video distribution in the internet," in *Proc. ACM SIGCOMM '94*, London, U.K., Sept. 1994.

[3] J. Byers, M. Luby, and M. Mitzenmacher, "Fine-grained layered multicast," in *Proc. IEEE INFOCOM 2001*, Anchorge, AK, Apr. 2001.

[4] S. Y. 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*, San Francisco, CA, Mar. 1996.

[5] Y.-H. Chu, S. G. Rao, and H. Zhang, "A case for end system multicast," in *Proc. ACM SIGMETRICS 2000*, Santa Clara, CA, Jun. 2000.

[6] T. Cover and A. El Gamel, "Achievable rates for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 6, pp. 851–857, Nov. 1982.

[7] 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 2001*, Kyongju, Korea, Apr. 2001.

[8] W. H. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 269–275, Mar. 1991.

[9] V. K. Goyal, J. Kočević, R. Arean, and M. Vetterli, "Multiple description transform coding of images," in *Proc. ICIP '98*, Chicago, IL, Oct. 1998.

[10] ISO/IEC, Generic coding of moving pictures and associated audio information, in ISO/IEC 13 818-2, 1995.

[11] ISO/IEC, Coding of audio-visual objects—Part 2: Visual, in ISO/IEC 14 496-2, 1999.

[12] ITU, Video coding for low bit rate communication, in ITU-T Recommendation H.263, 1998.

[13] T. Jiang, M. H. Ammar, and E. W. Zegura, "Inter-receiver fairness: a novel performance measure for multicast ABR sessions," in *Proc. ACM SIGMETRICS '98*, Madison, WI, Jun. 1998.

[14] A. Kanlis and P. Narayan, "Error exponents for successive refinement by partitioning," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 275–282, Jan. 1996.

[15] T. Kim and M. H. Ammar, "A comparison of layering and stream replication video multicast schemes," in *Proc. ACM NOSSDAV 2001*, Port Jefferson, NY, Jun. 2001.

[16] J.-I. Kimura, F. A. Tobagi, J. M. Pulido, and P. J. Emstad, "Perceived quality and bandwidth characterization of layered MPEG-2 video encoding," in *Proc. SPIE Int. Symp. Voice, Video, and Data Communications*, Boston, MA, Sept. 1999.

[17] L. P. Kondi, F. Ishtiaq, and A. K. Katsaggelos, "On video SNR scalability," in *Proc. ICIP '98*, Chicago, IL, Oct. 1998.

[18] X. Li, S. Paul, and M. H. Ammar, "Layered video multicast with retransmissions (LVMR): evaluation of hierarchical rate control," in *Proc. IEEE INFOCOM '98*, San Francisco, CA, Mar. 1998.

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

[20] H. M. Radha, M. van der Schaar, and Y. Chen, "The MPEG-4 fine-grained scalable video coding method for multimedia streaming over IP," *IEEE Trans. Multimedia*, vol. 3, no. 1, pp. 53–68, Mar. 2001.

[21] A. Reibman, H. Jafarkhani, Y. Wang, M. Orchard, and R. Puri, "Multiple description coding for video using motion compensated prediction," in *Proc. ICIP '99*, Kobe, Japan, Oct. 1999.

[22] B. Rimoldi, "Successive refinement of information: characterization of the achievable rates," *IEEE Trans. Inform. Theory*, vol. 40, no. 1, pp. 253–259, Jan. 1994.

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

[24] V. Vaishampayan and S. John, "Interframe balanced multiple description video compression," in *Proc. ICIP '99*, Kobe, Japan, Oct. 1999.

[25] G. Voronov and M. Feder, "The redundancy of successive refinement codes and codes with side information," in *ISIT 2000*, Sorrento, Italy, Jun. 2000.

[26] Y. R. Yang, M. S. Kim, and S. S. Lam, "Optimal partitioning of multicast receivers," in *Proc. ICNP 2000*, Osaka, Japan, Nov. 2000.

[27] E. W. Zegura, K. L. Calvert, and M. J. Donahoo, "A quantitative comparison of graph-based models for internetworks," *IEEE Trans. Netw.*, vol. 5, no. 6, pp. 770–783, Dec. 1997.

Taehyun Kim received the B.S. degree in control and instrumentation engineering from Seoul National University, Seoul, Korea, in 1992, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Taejon, in 1994, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, in 2005.

From 1994 to 1999, he was a Research Scientist with the KBS Technical Research Institute, Seoul. He joined the Wireless and Mobile Systems Group, Freescale Semiconductor, Austin, TX, in 2005. His research interests include video coding, video streaming, video multicasting, contents distribution networks, and video telephony.

Dr. Kim is the recipient of the Rotary Ambassadorial Scholarship in 1999 and the Nortel Fellowship in 2003.

Mostafa H. Ammar (F'02) received the B.S. and M.S. degrees from the Massachusetts Institute of Technology, Cambridge, in 1978 and 1980, respectively, and the Ph.D. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 1985.

During 1980–1982, he was with Bell-Northern Research (BNR), Ottawa, ON. He is currently a Regents' Professor with the College of Computing at Georgia Institute of Technology, Atlanta, where he has been since 1985. His research interests are in the area of computer network architectures, protocols, and services.

Dr. Ammar was the co-recipient of the Best Paper Awards at the Seventh World Wide Web conference for the paper "Interactive Multimedia Jukebox" and the 2002 Parallel and Distributed Simulation (PADS) Conference for the paper "Updateable Network Simulation." He served as the Editor-in-Chief of the IEEE/ACM TRANSACTIONS ON NETWORKING from 1999 to 2003. He is a Fellow of the ACM.