A Comparison of Layering and Stream Replication Video Multicast Schemes

Taehyun Kim and Mostafa H. Ammar
Networking and Telecommunications Group
Georgia Institute of Technology
Atlanta, GA 30332
{tkim, ammar}@cc.gatech.edu

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 of replicated video streams at different rates, 2) multicasting the video encoded in cumulative layers, and 3) multicasting the video encoded in non-cumulative layers. Even though there is a common belief that the layering approach is better than the replicated stream approach, there has 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 penalty is incurred by encoding a video stream in layers. We argue that a fair comparison of these schemes needs to take into account this penalty 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 stream replication is not as clear cut as is widely believed and that there are indeed scenarios where replication is the preferred approach.

1. 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., [1], [2], [3], [9], [13], [14]).

There are three basic approaches:

- The replicated stream approach [3], [9]
  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 obtained by encoding the source video with different compression parameters.

- The cumulative layering approach [13], [14]
  In this approach, the video is encoded in a base layer and one or more enhancement layers. The base layer can be independently decoded, but the enhancement layers can be decoded cumulatively (i.e., layer $k$ can only be decoded along with the layer 1 to the layer $k-1$). 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 and H.263 standards support the cumulatively layered encoding by defining four scalability modes: spatial scalability, temporal scalability, data partitioning, and SNR scalability [7], [8]. A combination of the scalability modes leads to the hybrid scalability consisting of a large number of layers.

- The non-cumulative layering approach [2], [6]
  In this approach, the video is encoded in two or more independent layers. Each layer is independently decodable and provides improvements to the video reception quality. Each receiver can join any subset of the video layers without joining the layer 1 multicast group.

Multiple description coding (MDC) can be used for non-cumulatively layered video multicast, although the area of MDC is still in its infancy. In MDC, each description can lead to the reconstruction of the source video, and multiple descriptions together yield a construction with the smallest distortion [4].
There is a common belief that the layering approach is better than the replicated stream approach. The main argument is that stream replication wastes 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 multicasting 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 penalty. This penalty can sometimes change the bandwidth efficiency in favor of stream replication.

An example of network heterogeneity in the video distribution is given in Figure 1. A video server is run on S and three sets of receivers, $R_1$, $R_2$, and $R_3$ are connected to the server from different domains.

Figure 1(a) shows how layered video multicasting accommodates the heterogeneity problem. Since the receivers in $R_1$ have high link capacity, they can subscribe to all layers. However, if any of the receivers in $R_2$ or $R_3$ tries to join the second enhancement layer, the 1.5Mbps link becomes congested which leads the receiver to leave the second enhancement layer. Also, the receivers in $R_1$ have to join only the base layer because of the 128kbps link. Hence, the reception rates of $R_1$, $R_2$, and $R_3$ are 10Mbps, 1.5Mbps, and 128kbps.

On the other hand, Figure 1(b) illustrates the replicated stream video multicasting scheme. Since the receivers in replicated stream multicasting subscribe to only one stream, the receivers in $R_1$ have to join one of the three streams and they will subscribe the high quality stream over the 10Mbps link. However, a receiver in $R_2$ or $R_3$ cannot receive the high quality stream because of the 1.5Mbps bottleneck link: if the receiver tries to subscribe to the high quality stream, the 1.5Mbps link will be congested. In the same way, a receiver in $R_3$ receives the low quality stream due to the limited capacity of the 128kbps link. Hence, the reception rates of $R_1$, $R_2$, and $R_3$ are 8.5Mbps, 1.37Mbps, and 128kbps.

Unfortunately, this example does not account for the video reception quality. Assume that layered video multicasting requires 20% more overhead than replicated stream video multicasting, the data rates contributing to the video quality in the layering scheme are 8Mbps, 1.2Mbps, and 102.4kbps for $R_1$, $R_2$, and $R_3$. Compared with the reception rate of the replicated stream scheme, the average data rate of the layering scheme contributing the video quality is smaller. Hence, we can expect that the average reception quality of layered video multicasting is not better than that of replicated stream video multicasting.

The amount of penalty in these schemes will of course depend on the specifics of the encoding of the layering and replicated stream data (i.e., the number and rates of the streams). A fair comparison between these schemes needs to also take into account the protocol complexity as well as the topological placement of the video source and the receivers relative to each other.

This paper is devoted to such a comparison. It is organized as follows. In Section 2, we consider the issue of layering penalty in more detail. Section 3 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 4 reports on results from simulations we have used to provide a quantitative comparison. Section 5 provides a comparison of the protocol overheads involved in both multicast schemes. The paper is concluded in Section 6.

2. OVERHEAD IN LAYERED VIDEO

In this section, we describe how layered encoding of video incurs a bandwidth penalty. Consider a video that is encoded as a single (non-layered) 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 $R_{nl,i}$, 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 non-layered 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 in three ways:

- **Information Theoretic Results** [5], [10], [15]

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

  The general result (described more formally in Appendix A) 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 non-layered encoding. While equality is possible, it requires a strict Markov condition to apply to the source and is generally not achievable. Moreover, results in the recent research show that the performance of the layered coding is not better than that of non-layered coding, even when the Markov condition holds [10]; and that the redundancy of layered coding is accumulated at each encoding stage, which leads to a significant quality degradation as the number of layers increases [17].

- **Protocol and Packetization Overhead**

  For certain scalability modes in MPEG-2 and H.263 standards, enhancement layers are designed to be syntactically independent of one another. Along with the residual information for every enhancement layer, the data stream needs to also carry syntactic data, such as picture header, start codes, GOP information, and macroblock header and information. This can incur a large amount of overhead especially at low data rates [12].

- **Experimental Evidence**

  Figure 2 shows an experimental result of the video quality versus data rates for the flower sequence by comparing MPEG-2 SNR scalability and no scalability mode. The video quality is measured as a PSNR (Peak Signal to Noise Ratio) value by varying quantization scale. A layered stream has two layers consisting of a high priority layer and a low priority layer at a quantization scale $(Q_b, Q_e)$. Both quality and data rate values are averaged over the entire video sequence.

  This result demonstrates that a non-layered stream has better video quality than a layered stream at the same data rate. The difference in data rate ranges from 0.4% at 27.7dB PSNR to 117% at 23.2dB PSNR. The difference is expected to grow as the number of layers increases, since the accumulation of the redundancy will lead to the increase of the overall distortion in layered video coding [17].

  Similar and more extensive experimental results can be found in [11]. The authors investigated the effects 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 on a scale from 1 to 5. It was also demonstrated that a non-layered stream provides better quality than a layered stream for a given average bandwidth.

For example, Figure 1(b) in [11] shows that the data rate difference between a layered stream and a non-layered stream is dependent on the quality and the contents being observed, when data partitioning is used to encode the video in two layers. The experiment shows that the difference can range from nearly 0 for the highest quality video (scoring close to 4.5) to almost 50% for fair quality video (scoring close to 3). For a score of 4 (good quality video), the overhead is around 20%. Because of the accumulation of the redundancy, 20% overhead is a relatively optimistic value for the overhead of MPEG-2 data partitioning.

### 3. Optimizing Stream Rates

To carry out a fair comparison of the layering and replication multicast schemes, we need to insure that each scheme is optimal. 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 shall present 1) the stream assignment algorithm to determine the reception rate of each receiver by aggregating the data rates of the assigned streams, and 2) the rate allocation algorithm to determine the data rate of each stream. 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, which is 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 as the reception rate of the receiver if there is no constraint from other
receivers in the same session [9]. The isolated rate can be computed by Dijkstra’s algorithm.

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

### 3.1 Cumulatively layered multicast

#### 3.1.1 Stream assignment

A cumulatively layered multicast session is defined by $\alpha = \{\alpha_i | \alpha_i \in \mathbb{R}^+, i = 1, \ldots, m\}$, where $\alpha_i$ is the data rate of a stream and $m$ is the number of layers.

We assume each layer follows the same path and each receiver joins as many layers as possible within the isolated rate. Hence, the reception rate is determined as the sum of stream rates that does not exceed the isolated rate. The stream assignment algorithm is presented in Figure 3, when the set of stream rates $\alpha$ is given.

1. Compute the isolated rate
2. Assign $\alpha_i$ that does not exceed the isolated rate

Figure 3: Stream assignment algorithm for cumulatively layered multicasting

#### 3.1.2 Rate allocation

The first rate allocation algorithm for cumulatively layered video multicast was proposed in [16] by maximizing the average signal reception quality. The authors in [18] proposed an optimal receiver partitioning algorithm based on the dynamic programming method to determine the optimal stream rates. We adopt this algorithm to allocate the optimal data rates for $\alpha_i$.

### 3.2 Replicated stream multicast

#### 3.2.1 Stream assignment

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

A set of receivers assigned to the stream $i$ is denoted by $\Omega_i = \{e_j | \phi_\beta(e_j) = \beta_i\}$, where $\phi_\beta$ is the rate allocation function defined by $\phi_\beta : C \rightarrow \beta$. The set of all receivers is $\Omega = \bigcup_{i=1}^m \Omega_i$.

We have two objectives for the stream assignment.

1. The minimum reception rate for all receivers is strictly greater than zero so that there is no receiver that cannot receive any stream.
2. Maximize $Z_\beta = \sum_{i=1}^m |\Omega_i| \beta_i$ subject to $\sum_{i \in \Gamma_{e_j}} \beta_i \leq b_j$, where $\Gamma_{e_j} = \{i | e_j \in E_i\}$, $T_i = (V_i, E_i)$, and $T_i$ is a tree for a replicated stream $i$.

We develop a greedy algorithm for the stream assignment. We first allocate $\beta_i$ to all receivers to satisfy the minimum reception rate constraint, where $\beta$ is assumed to be sorted in non-decreasing order without loss of generality. A receiver is assigned a stream that has not yet been assigned and has the maximum value of group size and stream rate product. Note that this algorithm may not give a feasible solution if any receiver has the isolated rate smaller than $\beta_i$, since we cannot allocate more data rate than can be handled. We present the stream assignment algorithm in Figure 4.

1. Compute the isolated rates
2. if all isolated rates are equal or greater than $\beta_i$ then
3. $b_i \leftarrow b_i - \beta_i$, where $b_i = B(e_j)$ and $e_j \in E$
4. $\beta \leftarrow (\beta \setminus \{\beta_i\} \cup \{0\})$
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 $\|\Omega_i\| \beta_i = \max_j \|\Omega_j\| \beta_j$ ($\beta_i, \beta_j \in \beta$)
11. if $\beta_i > 0$ then
12. Assign $\beta_i$ to $\Omega_i$
13. Reduce the link capacity leading to $\Omega_i$
14. else
15. Assign $\beta_i$ to $\Omega_i$
16. endif
17. $\Omega \leftarrow \Omega \setminus \Omega_i$
18. enddo

Figure 4: Stream assignment algorithm for replicated stream multicasting

#### 3.2.2 Rate allocation

The stream rates, $\beta_i$, are allocated based on the optimal cumulative layering rates, $\alpha_i$. More specifically, $\beta_i$ 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 $\sum_{i=2}^k \alpha_i$. Recall that we already allocated the base layer to all receivers in the stream assignment algorithm. The following is the rate allocation scheme for replicated stream multicasting.

$$\beta_i = \left\{ \begin{array}{ll}
\alpha_i, & i = 1 \\
\sum_{j=2}^m \alpha_j, & 2 \leq i \leq m.
\end{array} \right.$$  

### 3.3 Non-cumulatively layered multicast

#### 3.3.1 Stream assignment

A non-cumulatively layered multicast session is defined by $\gamma = \{\gamma_i | \gamma_i \in \mathbb{R}^+, i = 1, \ldots, m\}$, where $\gamma_i$ is the data rate of a non-cumulatively layered stream and $m$ is the number of streams. A set of receivers assigned to the stream $i$ is defined by $\Omega'_i = \{e_j | e_j \in \phi_\gamma(e_j)\}$, where $\phi_\gamma$ is the stream rate function defined by $\phi_\gamma : C \rightarrow P(\gamma)$ and $P(\gamma)$ is the power set of $\gamma$. The set of all receivers is $\Omega' = \bigcup_{i=1}^m \Omega'_i$.

We have two objectives to assign the non-cumulatively layered streams.

1. The minimum reception rate for all receivers is strictly greater than zero.
2. Maximize $Z_\gamma = \sum_{i=1}^m |\Omega'_i| \gamma_i$ when $\sum_{i \in \Gamma_{e_j}} \gamma_i \leq b_j$, where $\Gamma_{e_j} = \{i | e_j \in E_i\}$, $T_i = (V_i, E_i)$, and $T_i$ is a tree for a non-cumulatively layered stream $i$.

In Figure 5, we present a greedy algorithm for non-cumulative layering to assign a stream with the maximum value of group
size and stream rate product. Note that the reception rate of a receiver is the sum of assigned stream rates.

1: Compute the isolated rates
2: if all the isolated rates are equal or greater than \( \gamma_i \) then
3: Assign \( \gamma_i \) to all receivers
4: \( b_j \leftarrow b_j - \gamma_i \), where \( b_j = B(e_j) \) and \( e_j \in E \)
5: \( \gamma \leftarrow (\gamma_1 \gamma_1) \cup \{0\} \)
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 \( \{\Omega_1^j\}_{\gamma_i} - \max_j \{\Omega_1^j\}_{\gamma_j} (\gamma_i, \gamma_j \in \gamma) \)
12: if \( \gamma_i > 0 \) then
13: Add \( \gamma_i \) to \( \Omega_1^j \)
14: Reduce the link capacity leading to \( \Omega_1^j \)
15: endif
16: \( \gamma \leftarrow \gamma \setminus \gamma_i \)
17: enddo

Figure 5: Stream assignment algorithm for non-cumulatively layered multicasting

3.3.2 Rate allocation

In non-cumulatively layered multicasting, a receiver can subscribe to any subset of layers without joining the base layer. This property provides the fine granularity for non-cumulative layering. For example, given a non-cumulatively layered stream \( \gamma = \{1, 2, 4\} \), a heterogeneity resulting in seven different isolated rates of \( \{1, 2, 3, 4, 5, 6, 7\} \) can be accommodated through selective subscription. Hence, the non-cumulatively layered stream \( \gamma \) shows the same performance as the cumulatively layered stream \( \alpha = \{\alpha_i | \alpha_i = 1, i = 1, \ldots, 7\} \). This example demonstrates that the heterogeneity caused by \( 2^m - 1 \) different link capacities can be accommodated by aggregating the reception rates of \( m \) non-cumulative layers, since \( \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{m} = 2^m - 1 \). The work in [2] describes a fine grained layered multicast scheme based on this property. The authors propose a fine grained rate adjustment scheme by using at most three join and leave operations.

In this section, we propose a rate allocation algorithm for non-cumulatively layered streams. The stream rates \( \gamma_i \) are allocated based on \( \alpha_i \), as the optimum values of \( \alpha_i \) are provided in Section 3.1 and the aggregated reception rates of \( \gamma_i \) can be determined by \( \alpha_i \), such that

\[
\gamma_1 = \alpha_1, \\
\gamma_2 = \alpha_1 + \alpha_2, \\
\gamma_1 + \gamma_2 = \alpha_1 + \alpha_2 + \alpha_3, \\
\gamma_3 = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4, \\
\ldots \\
\gamma_2 + \gamma_3 + \cdots + \gamma_m = \alpha_1 + \alpha_2 + \cdots + \alpha_{2^m-2}, \\
\gamma_1 + \gamma_2 + \gamma_3 + \cdots + \gamma_m = \alpha_1 + \alpha_2 + \cdots + \alpha_{2^m-2} + \alpha_{2^m-1}.
\]

We simplify this relationship by using a matrix form: \( AX = BY \), where \( A \) is a binary counting matrix, \( B \) is a lower triangular matrix, \( X \) is a vector of the allocated data rates of non-cumulative layering, and \( Y \) is a vector of the optimal data rates of cumulative layering, such that

\[
A = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{pmatrix}, \\
B = \begin{pmatrix} 1 & \cdots & 1 & 1 \\ 0 & \cdots & 0 & 0 & 1 \\ 0 & \cdots & 0 & 0 & 1 & 1 \\ \vdots \\ 0 & \cdots & 0 & 1 & 1 & 1 & 1 \\ 1 & \cdots & 1 & 1 & 1 & 1 & 1 \end{pmatrix}, \quad Y = \begin{pmatrix} \alpha_{2^m-1} \\ \alpha_{2^m-2} \\ \vdots \\ \alpha_2 \\ \alpha_1 \end{pmatrix}.
\]

However, it is not generally possible to determine the data rate \( \gamma_i \) for given \( \alpha_i \), since the number of equations exceeds that of unknown variable \( \gamma_i \). We develop an approximate rate allocation scheme by minimizing the mean square error, \( Z = (AX - BY)^T(AX - BY) \). The allocated data rates for non-cumulatively layered multicast streams are given by \( X = (A^T A)^{-1} A^T BY \), since

\[
\nabla_X Z = \nabla_X (AX - BY)^T(AX - BY) \\
= \nabla_X (X^T A^T AX - Y^T B^T BX - X^T A^T BY - Y^T B^T BY) \\
= 2A^T AX - 2A^T BY \\
= 0,
\]

where \( \nabla_X \) is the gradient operator for the vector \( X \).

4. SIMULATION RESULTS

We compare the performance of the video multicasting schemes by simulation. We first generate 100 different transit-stub graphs forming the network topology by using GT-ITM [19]. The graphs consist of 1,640 nodes with 10 transit domains, 4 nodes per transit domain, 4 stubs per transit node, and 10 nodes in a stub domain. The available link bandwidth is chosen randomly in the range of 1%-80% of the full capacity. We assign 2.4Gbps to transit-to-transit edges; 10Mbps and 1.5Mbps to stub-to-stub edges; and 155Mbps, 45Mbps, and 1.5Mbps to transit-to-stub edges. The amount of overhead incurred by cumulative and non-cumulative layering is assumed to be 20%, consistent with the results observed in [11]. For the set of experiments we report, the number of layers for cumulatively layered video multicasting and replicated stream multicasting are 8, and that for non-cumulatively layered video multicasting is 4.

In the experiment, we compare the performance by using the following measures: 1) the average reception rate which is the average data rate received by a receiver, 2) 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, and 3) 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.

We define efficiency as follows:
Hence, the efficiency can be considered as a measure of delivered data rate contributing to the video quality for each unit of bandwidth used in the network.

4.1 **Topology 1: receivers are widely distributed**

In the first simulation, we randomly select a server and receivers from the set of nodes in the graph.

Figure 6(a) shows the average reception rate for a receiver. The reception rates of cumulative layering, non-cumulative layering, and replicated stream multicasting are around 92%, 84%, and 73% of the isolated rates, respectively. Note that the performance of non-cumulative layering is not so good as that of cumulative layering, since the number of layers in cumulative layering is twice as many as that of non-cumulative layering. In this case, non-cumulative layering shows, at best, the same performance as cumulative layering, when the mean square error in Section 3.3.2 is zero. The total bandwidth usage is presented in Figure 6(b). The result shows that the bandwidth consumption of cumulatively layered multicasting is the largest and that of replicated stream multicasting is the smallest. In Figure 6(c), we demonstrate the efficiency of the video multicast schemes. Although cumulative and non-cumulative layering have good bandwidth scalability and high reception rates, only 80% of them contributes to improving the video quality. This makes replicated stream video multicasting more efficient. Figure 6(d) shows the average effective reception rate for each receiver. Note that the video quality of replicated stream video multicasting and cumulatively layered video multicasting is expected to be nearly the same, since the difference of the effective reception rates of two schemes is very small.

Figure 7 demonstrates the effect of overhead in cumulatively layered video multicasting and replicated stream video multicasting. The number of receivers is fixed to 40% of all nodes. Figure 7(a) shows the ratio of the effective reception rate of replicated stream video multicasting to that of cumulatively layered video multicasting. As the overhead increases, the effective reception rate of cumulatively layered video multicasting gets worse. The effective reception rate is equal when the overhead is 22% and the effective reception rate of replicated stream video multicasting is better than cumulatively layered video multicasting after that point. In Figure 7(b), we investigate how the overhead affects the efficiency by measuring the ratio of the efficiency of stream

![Figure 6: Wide distribution result](image-url)

\[
\text{efficiency} = \frac{\text{total effective reception rate}}{\text{total bandwidth usage}}.
\]
replication to that of cumulative layering. For this measure, layering overhead of more than 7% tends to favor the replicated stream approach.

For example, Figures 9 and 10 illustrate the impact of network topology on the performance. In Figure 9, the receivers $R_1$, $R_2$, and $R_3$ are placed on different domains and no link is shared by them. We can find that both schemes consume the same amount of bandwidth and the reception rates of $R_1$, $R_2$, and $R_3$ are 10Mbps, 1.5Mbps, and 128kbps. Hence, replicated stream multicasting has an advantage if we consider the overhead of layered video multicasting.

On the other hand, Figure 10 presents an example that the receivers $R_1$, $R_2$, and $R_3$ are placed on the same domain and they share the 10Mbps bottleneck link. In this topology, the reception rates in Figure 10(a) is larger than that of Figure 10(b) by 1.63Mbps.

4.2 Topology 2: receivers are distributed in only one domain

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

Figure 7: Effect of overhead

Figure 8 shows the performance as the number of layers and consequently the number of replicated streams changes. Figure 8(a) shows that the effective reception rates of the two schemes are nearly the same when the number of layers is five. When the number of layers is larger than six, the effective reception rate of cumulatively layered video multicast is the larger. When the number of layers is smaller than three, the effective reception rate of stream replication is larger than that of cumulative layering. Hence, cumulatively layered video multicasting is not desirable when the number of layers is small. In Figure 8(b), the efficiency of stream replication is always greater than that of cumulative layering. From the result, it is expected that the effect of the number of layers on efficiency is not so significant as the amount of layering overhead.

Figure 8: Effect of the number of layers
5. PROTOCOL COMPLEXITY

In this section, we consider 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 (Receiver-driven Layered Multicast) [14]. 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 summarize below our protocol complexity analysis. In cumulatively layered video multicasting, the average group size is given by $\frac{1}{m} \sum_{k=1}^{m} k|\Omega_k|$, since a receiver can join multiple groups and $\frac{1}{m} (\sum_{k=1}^{m} |\Omega_k| + \sum_{k=2}^{m} |\Omega_k| + \cdots + \sum_{k=m}^{m} |\Omega_k|) = \frac{1}{m} \sum_{k=1}^{m} k|\Omega_k|$, where $m$ is the number of layers and $\Omega_k$ is the set of receivers that can accommodate up to layer $k$. 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 2 unicast messages and 1 multicast message to $\frac{1}{m} \sum_{k=1}^{m} k|\Omega_k|$ receivers.

On the other hand, the average group size in replicated stream video multicast is $\frac{1}{m} \sum_{k=1}^{m} |\Omega_k|$, since every receiver joins only one group. In replicated stream video multicast, a receiver sends one leave message, one join message, and one multicast message reporting each join experiment. Since the link capacity is assumed to be stable, it has to return to the previous group that involves another one join and one leave messages. Therefore, the average number of messages in a join experiment is 4 unicast messages and 1 multicast message to $\frac{1}{m} \sum_{k=1}^{m} |\Omega_k|$ receivers.

The bandwidth overhead in cumulatively layered multicasting and replicated stream multicasting consists of 2/4 unicast messages and 1 multicast message. The cost of a multicast message is dominant when the number of receivers is large enough.

Figure 12 presents the simulation result of the average group size and the average number of groups joined by a receiver, when the receivers are widely distributed. In Fig-
6. CONCLUSION

In this paper, we undertake a comparison between the cumulative/non-cumulative layering and replicated stream video multicasting 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 take into account 1) the layering bandwidth penalty, 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 layered encoding penalty does not apply in those circumstances, however, the other comparison dimensions will still come into play.

7. REFERENCES

APPENDIX

A. MORE ON THE INFORMATION THEORETIC LIMITS

Consider a lossy data compression code consisting of one base layer and one enhancement layer. Let \( \{X_i\}_{i=1}^{\infty} \) be a \( X \) valued discrete memoryless source with probability mass function, \( P \), and \( \mathcal{Y}_1 \) be a finite reproduction alphabet. \( d_1 \) is defined by a non-negative valued mapping indicating a distortion measure, such that \( d_1 : X \times \mathcal{Y}_1 \rightarrow \mathbb{R}^+ \).

Given a positive real value \( \Delta_1 \) specifying the expected distortion, a rate distortion function \( R(P, \Delta_1) \) that characterize the minimum achievable rate for a non-layered code is given by

\[
R(P, \Delta_1) = \inf_{P_X = P, \ E_d(X,Y_1) \leq \Delta_1} I(X;Y_1),
\]

where \( I(X;Y_1) \) is the mutual information.

In the same way, the rate distortion function can be extended to the layered coding. Let \( \mathcal{Y}_2 \) be a finite reproduction alphabet and \( d_2 : X \times \mathcal{Y}_2 \rightarrow \mathbb{R}^+ \) be a non-negative valued mapping representing a distortion measure. Suppose that a coding is done in the base layer with the rate \( R_1 \geq R(P, \Delta_1) \) and the distortion \( \Delta_2 \) is given, the minimum achievable rate in the enhancement layer is given by

\[
R(P, R_1, \Delta_1, \Delta_2) = \inf_{P_X = P, \ E_d(X,Y_1) \leq \Delta_1, \ E_d(X,Y_2) \leq \Delta_2, \ I(X;Y_1) \leq R_1} I(X;Y_1Y_2),
\]

where \( 0 < \Delta_2 < \Delta_1 \).

By definition, the rate distortion for a layered code is no smaller than that of a non-layered code. Hence, the following relationship generally holds [15].

\[
R(P, \Delta_2) \leq R(P, R_1, \Delta_1, \Delta_2). \tag{1}
\]

The condition under which equality holds requires that \( X, Y_1, \) and \( Y_2 \) satisfy a Markov condition [5]:

\[
P_{XY_1Y_2}(x, y_1, y_2) = P_X(x)P_{Y_2|X}(y_2|x)P_{Y_1|Y_2}(y_1|y_2),
\]

where \( x \in X, y_1 \in Y_1, \) and \( y_2 \in Y_2 \).

From (1), the minimum data rate of non-layered stream is equal or smaller than that of layered stream. However, it is difficult to design a layered code satisfying Markov condition in general. Moreover, it was shown that, even if the Markov condition holds, the performance of the layered code is inferior to that of non-layered code in terms of the error exponent measure [10].