RECEIVER-DRIVEN LAYERED OVERLAY MULTICAST FOR SCALABLE VIDEO STREAMING

Junni Zou[1],Min Wang[1], Leyang Li[1], Hongkai Xiong[2]
[1]Department of Communication Engineering, Shanghai University, Shanghai, 200079, P.R. China
[2]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, P.R. China

ABSTRACT

In this paper, we seek the optimal overlay multicast performance for scalable video streaming. To adapt to layered multicast with multiple paths, a new metric *layer stretch* is introduced, measuring video layer's dissemination latency along the overlay. Multi-path routing, network coding based layer subscription and congest control are jointly optimized, formulating a linear programming to minimize the average layer stretch over all receivers. It not only solves the issue of inter-layer dependency of scalable video coding, but allows each receiver to determine the required video quality. Moreover, a heuristic algorithm that can be implemented in a distributed manner is provided to approximate the optimal solution. Simulation results validate the network performance of the proposed scheme.

Index Terms—Scalable video coding, layered multicast, overlay, network coding, stretch

1. INTRODUCTION

Multicast is an important way for video dissemination. A critical challenge for video multicast is to provide a scalable rate control addressing receiver heterogeneity. Layered coding, such as JVT/MPEG scalable video coding (SVC) [1], provides a convenient way for rate adaption at source side. A SVC stream consists of a base layer and multiple enhancement layers with a multi-dimension layer structure. Each layer, along with all layers it depends on, forms a representation at a certain spatio-temporal resolution and quality level. On the other hand, receiver-driven layered multicast (RLM) [2] advocated receiver-based congestion control, where video layers are transmitted in different multicast groups and congestion control is employed by each receiver through subscribing to a subset of the groups.

The first optimization model for the flow control was studied by Low et al [3]. Kar et al [4] proposed an optimization approach to maximize the total utility for multirate multicast sessions. With the development of overlay networks, overlay multicast emerges a practical alternative for IP multicast. Current flow control schemes are not suitable for dynamic overlay structures, because they always use predetermined distribution trees or meshes. In this paper, we study layered multicast problem for overlay networks where each receiver can have multiple alternative paths receiving subscribed video layers.

In overlay multicast, the stretch metric is generally used to evaluate the end-to-end latency. It is defined as the ratio of path length from the source to the multicast group member along the overlay to the length of the direct unicast path [5]. Owing to its per-path definition, the stretch works well in multicast scenarios with single (non-scalable) source and single path, whereas, it fails to measure the latency of disseminating a video layer that often involves a couple of paths. In this study, we introduce a new metric called *layer stretch*. It is per-layer defined and can be directly applied to measure scalable video layer dissemination latency along the overlay.

Network coding is proven sufficient to achieve the max-flow throughput. The authors in [6] proposed a LION algorithm to improve the layered multicast throughput with network coding and multi-paths. It mainly focus on maximizing the aggregate number of video layers obtained by all receivers, less considering the inter-layer or hierarchical dependency of scalable video streaming.

In this paper, we seek to construct the optimal scalable video distribution meshes, achieving minimum end-to-end layer stretch and maximum throughput for all receivers. The inter-layer dependency of scalable video stream is imposed on the objective function to improve the received video quality. Unlike previous work, we achieve max-flow subscription level by combining multi-path scheme with network coding strategy. A heuristic algorithm that can be implemented in a distributed manner is also presented to approximate the optimal solution.

2. MOTIVATION

To distribute scalable video streaming over heterogeneous networks, two key conditions have been neglected before:

(1) Inter-layer dependency. When constructing the distribution mesh for each SVC layer, existing mathematical models are not consistent with layered decomposition of scalable video streaming. Higher layers with low latency may overwhelm lower layers. As the butterfly topology of Fig. 1 (a) with all links of unit capacity, the source generates a SVC stream with three layers, each with rate of 2. Assume that data suffer the same delay through each link. When adopting the integer linear programming (ILP) solution of

The work has been partially supported by the NSFC grants No. 60632040, No. 60736043, No. 60802019 and the National High Technology Research and Development Program of China (863 Program) (No. 2006AA01Z322) and funds from Shanghai Science and Technology Commission (No. 08220510900).

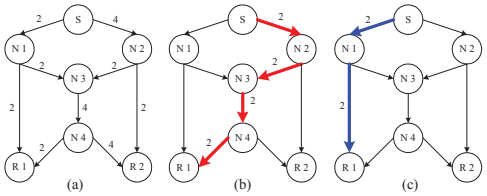

Fig. 1. Layered flow distribution sample by LION model, where (b) the base layer mesh, (c) the first enhancement layer mesh.

LION model [6], we can obtain the distribution mesh of R_1 shown in Fig.1 (b) and (c). Note that the solid line labels the routing path with the allocated bandwidth. It is observed that the base layer data will traverse four links, while the first enhancement layer will traverse two links. The cross-layer synchronization of SVC decoding in R_1 will be greatly influenced by the reversed latency, resulting in heavy buffer management and decoder burden.

(2) Heterogeneous QoS requirements. Most work in literature contributes to a uniform objective over all receivers. In practical video dissemination, receivers have heterogeneous QoS requirements. Fig. 2 displays a SVC-based layered multicast scenario, where each class is associated with different QoS requirements. The broadband receivers of class 1 are supposed to access high-fidelity video where the throughput (video quality) is critical. The mobile receivers of class 3 concern a smooth play instead of throughput gain, where the delay is a main constraint.

3. PROBLEM STATEMENT

3.1. Network Model

Overlay network is modeled as a directed graph $G(V, E)$. The node set V has two disjoint subsets S(source) and R (receivers). Each link (i, j) is associated with two weights $C(i, j)$ and $D(i, j)$, denoting link capacity and propagation delay along the link respectively. Suppose the SVC stream is encoded into a set M of layers, with each layer m distributed over a multicast group at rate B_m. Let a_m^r denote whether receiver r subscribes to layer m. If it does, $a_m^r = 1$; else $a_m^r = 0$. Also let $f_m(i, j)$ represent the bandwidth consumed on link (i, j) in layer m.

Assume there exists multiple alternative paths $P(r)$ from the source to receiver r. For receiver r, we use a matrix $\mathbf{Z}^r$ to reflect the relationship between its paths and related links. $\mathbf{Z}^r$ is defined as:

$$z_k^r(i,j) = \begin{cases} 1, & \text{if link } (i,j) \text{ is included in path } k; \\ 0, & \text{otherwise.} \end{cases}$$

We use $x_{m,k}^r(i,j)$ to denote the bandwidth consumed on link (i, j) by receiver r along path k in layer m. Let U_m^r and D_m^r denote layer m's latency for receiver r along the direct unicast paths and the overlay paths, respectively.

In overlay multicast, the stretch is used to evaluate the end-to-end latency of single path. To adapt to scalable video dissemination where a video layer may be delivered through multiple paths, we introduce a new metric layer stretch noted as s_m. It is defined as the ratio of layer m's end-to-end latency along the overlay to that of the direct unicast transmission in the IP-layer. Since layer stretch reflects the relative increase of a layer's latency as compared with IP-layer multicast, it is critical to minimize the layer stretch so as to approximate IP-layer multicast performance. Clearly, the stretch is designed for non-scalable and single path video distribution, while the layer stretch is a more generalized metric and well-suited for multi-layer multi-path cases. If the layer stretch is used in single-layer single-path multicast scenario, it degrades to an essential stretch metric.

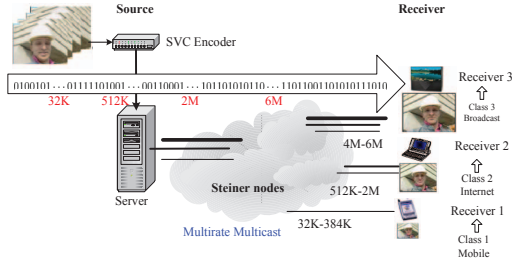

Fig. 2. A practical SVC-based layered multicast scenario.

3.2. Optimization Problem

Using multi-path and network coding based routing, we aim at constructing distribution meshes with minimum aggregate layer stretch and maximum throughput. Mathematically, the problem can be formulated as a linear programming (LP):

P1: (1)

$$\text{minimize} \quad \sum_{m\in M} p_m \times s_m =$$

$$\sum_{m\in M} p_m \times \frac{\sum_{\{r|r\in R \cap a_m^r>0\}} D_m^r \Big/ \sum_{\{r|r\in R \cap a_m^r>0\}} a_m^r}{\sum_{\{r|r\in R \cap a_m^r>0\}} U_m^r \Big/ \sum_{\{r|r\in R \cap a_m^r>0\}} a_m^r}$$

s.t. 1)

$$\sum_{\{j|(i,j)\in E\}} \sum_{k\in P(r)} z_k^r(i,j) x_{m,k}^r(i,j) - \sum_{\{j|(j,i)\in E\}} \sum_{k\in P(r)} z_k^r(i,j) x_{m,k}^r(j,i)$$

$$= \begin{cases} a_m^r \cdot B_m, & \text{if } i \in S; \\ -a_m^r \cdot B_m, & \text{if } i \in R; \\ 0, & \text{otherwise.} \end{cases}$$

$\forall \{(m,r)|m \in M, r \in R, a_m^r > 0\}$;

2) $$\sum_{\{j|(i,j)\in E\}} z_k^r(i,j) x_{m,k}^r(i,j) = \sum_{\{j|(j,i)\in E\}} z_k^r(i,j) x_{m,k}^r(j,i),$$

$\forall \{i|i \notin S, i \notin R\}, \{k|k \in P(r)\}$, and $\{(m,r)|m \in M, r \in R, a_m^r > 0\}$;

3) $$\sum_{k\in P(r)} z_k^r(i,j) x_{m,k}^r(i,j) \le f_m(i,j),$$

$\forall \{(m,r)|m \in M, r \in R, a_m^r > 0\}$ and $\{(i,j)|(i,j) \in E\}$;

4) $$\sum_{m\in M} f_m(i,j) \le C(i,j), \quad \forall \{(i,j)|(i,j) \in E\};$$

5) $$\frac{x_{m,k}^r(i,j)}{C(i,j)} \le b_{m,k}^r(i,j),$$

$\forall \{(m,r)|m \in M, r \in R, a_m^r > 0\}$ and $\{k|k \in P(r)\}$;

6) $\sum_{\{(i,j)|(i,j)\in E} b_{m,k}^r(i,j)D(i,j) \le D_m^r$,

$\forall \{(m,r)|m \in M, r \in R, a_m^r > 0\}$ and $\{k|k \in P(r)\}$.

The objective function is to minimize the aggregate layer stretch for all receivers. $\sum_{\{r|r\in R \cap a_m^r>0\}} D_m^r \Big/ \sum_{\{r|r\in R \cap a_m^r>0\}} a_m^r$ represents layer m's average latency along the overlay among all receivers that join layer m. $\sum_{\{r|r\in R \cap a_m^r>0\}} U_m^r \Big/ \sum_{\{r|r\in R \cap a_m^r>0\}} a_m^r$ denotes the corresponding latency over the IP-layer unicast paths.

p_m is the weights associated with layer m, satisfying $p_1 + p_2 + \cdots + p_M = 1$. p_m denotes the impact of layer m's stretch on the aggregate layer stretch, therefore, layers with higher priority or of more importance should have larger weights to guarantee smaller stretch. For scalable video streaming, we can let the stretch for each layer in an incremental order and have $p_1 > p_2 > \cdots > p_M$.

Each receiver is allowed to choose the number of video layers it requires as long as its subscription level does not exceed the max-flow throughput. For example, if receiver r has the max-flow capacity f^r that meets

$$\sum_{m=1}^{J} B_m \le f^r \le \sum_{m=1}^{J+1} B_m \qquad (2)$$

it can subscribe to at most J video layers.

Constraints 1) are information flow balance equations for source node, steiner nodes and sink nodes. Constraints 2) are information flow balance equations for paths in each layer. Constraints 3) represent the relationship between information flow rate and physical flow rate on each link where network coding is applied to information flows of the same video layer. Constraints 4) specify that the aggregate physical flow rates of different layers over each link do not exceed the link capacity. Constraints 5) determine the value of $b_{m,k}^r(i,j)$. If $x_{m,k}^r(i,j) > 0$, $b_{m,k}^r(i,j)$ is set to 1, otherwise it is 0. Note that $b_{m,k}^r(i,j)$ is used to check whether link (i,j) on path k of receiver r is related to layer m. Constraints 6) specify that layer m's end-to-end latency for receiver r is equal to the maximum latency among all paths used by the receiver to deliver that layer.

3.3. Practical Network Coding

Network coding is proven to enhance network robustness in overlay content distribution. In the context of layered video dissemination, network coding across different video layers may further improve the overall throughput. However, such coding scheme lacks of scalability and becomes highly complex, especially when the number of layers becomes large. In this work, we limit network coding operation within the same layer, thereby the original data of each layer can be recovered separately. For each video layer m, its stream is separated into multiple segments $g_m^1, g_m^2, \cdots$, each corresponding to a playback period. Each segment g_m^i is further divided into n equal-sized blocks. To reduce computational complex and decoding delay, we choose XOR operation [7] to substitute for linear operation in Galois Field $GF(2^q)$. In this way, a receiver may easily recover a segment from decoding any n innovative blocks belong to that segment.

3.4. Distributed Heuristic Algorithm

The proposed LP scheme requires global information to obtain the optimal solution, which is costly and infeasible. In this section, we provide a heuristic algorithm that can be implemented in a distributed manner to approximate the optimal solution. Inheriting the basic idea of the LP algorithm, the heuristic method also selects paths with less delay to transfer lower layers. It ensures that the more important video layers reach the receivers with lower latency and smaller stretch. Specifically, each receiver will select paths in order with path latency. If receiver r has a demand for the base layer, its path k with minimum latency d_k^r will be chosen. Once this path is detected not to fully accommodate the base layer traffic, paths with sub-minimum latency will be searched to carry the remainder flow. After rate allocation of the base layer among all receivers with $a_1^r = 1$, the available capacity of the overlay is updated. In succession, it is iterated to construct higher layers' distribution mesh. The pseudocode of rate allocating of layer m for receiver r, is summarized as follows:

```
if a_m^r = 0, stop accessing to layer m;
else
{
  unallocated flow rate R_un = B_m;
  while (R_un > 0)
  { choose path k with minimum latency in P(r);
    if (c_k^r ≥ R_un)
      f_k^r = R_un; /* flow rate allocated on path k */
      c_k^r = c_k^r − f_k^r; /* update the path capacity */
    else
      f_k^r = c_k^r; /* flow rate allocated on path k */
      c_k^r = 0;
      remove path k from P(r);
      R_un = R_un − f_k^r;
  }
}
```

4. SIMULATION RESULTS

We evaluate performance of the proposed scheme over a random overlay topology in Fig. 3 (c). The capacity (kbps) and the delay (ms) are marked on each link. We adopt Joint Scalable Video Model 9_10 reference codec of H.264/AVC extension, with two standard test video sequences: "Foreman" and "Mobile" with a frame rate of 30 fps, CIF (352x288) resolution, and a GOP-length of 32 frames. They are encoded with 256 kbps on the base layer, and 384kbps, 512kbps and 1024kbps on enhancement layers using FGS coding. Assume R_6, R_7, and R_8 has respectively subscribed to 2, 3 and 4 video layers in terms of their max-flow rate. And $R_1 \sim R_5$ act as the steiner nodes. The weights of each layer is set to $p_1 = 0.6$, $p_2 = 0.27$, $p_3 = 0.1$, $p_4 = 0.03$.

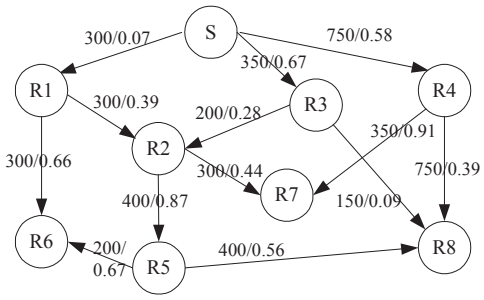

Fig. 3. a random overlay network topology

Table I compares the layer stretch of the shortest path algorithm, the LION-based ILP algorithm, the proposed LP algorithm and the distributed heuristic algorithm. It can be seen that our LP algorithm achieves the average minimum layer stretch in an incremental order. The LION scheme is not efficient for practical SVC multicast, because receivers would access lower layers with an unreasonable higher latency. The shortest path scheme with single-path routing fails to effectively utilize the network resource. Fig.4 compares the achievable throughput, where our LP and the LION algorithms can meet the subscription demands. In the shortest path scheme, receivers can access to at most one video layer. The performance of the heuristic algorithm is not as good as the LP algorithm.

We vary the playout deadline for "Mobile" stream from 200 ms to 400 ms, fixing the background traffic load to 10%. Comparison of average video quality is shown in Table II, where receivers can obtain the best video quality by our LP algorithm. In the LION scheme, the base layer packets for receiver 6 and 7 are dropped due to the high latency. Once the deadline approximates 400ms, PSNR performances are similar within LION, heuristic scheme, and the LP algorithm. Fig. 5 shows the impact of background traffic on resource allocation, we choose "Foreman" sequence and fix playout deadline at 400ms. As the amount of background traffic increases, the average video quality of the proposed algorithm degrades much more slowly than other schemes.

Table I. Layer stretch metric for "Foreman" sequence.

	Layer 1	Layer 2	Layer 3	Layer 4
Shortest path	1.36	-	-	-
LION algorithm	3.53	1.77	2.05	1.62
LP algorithm	1.45	2.05	3.53	3.96
Heuristic algorithm	1.45	2.75	2.05	-

Table II. Received average video quality (PSNR)

"Mobile" sequence, Playout deadline=200ms			
Average PSNR (dB)	R6	R7	R8
Shortest path	29.63	29.63	0
LION algorithm	0	0	33.77
LP algorithm	30.71	30.71	31.82
Heuristic algorithm	29.63	30.71	31.82
"Mobile" sequence, Playout deadline=400ms			
Average PSNR (dB)	R6	R7	R8
Shortest path	29.63	29.63	0
LION algorithm	30.71	31.82	33.77
LP algorithm	30.71	31.82	33.77
Heuristic algorithm	30.71	31.82	31.82

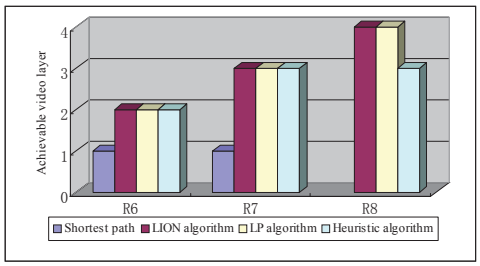

Fig. 4. Comparison of achievable throughput.

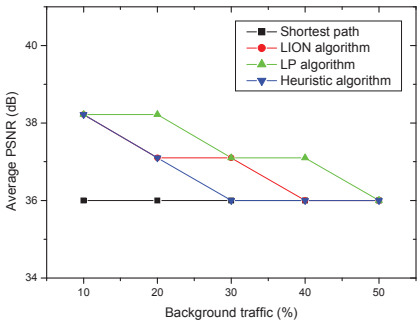

Fig. 5. Received average PSNR for receiver 7.

5. MAJOR HEADINGS

This work investigates network performance optimization for layered overlay multicast. We employ a joint optimization, including multi-path and network coding based routing and receiver-driven flow control, minimizing average layer stretch as well as maximizing aggregate throughput over all receivers. Each receiver can adapt to its subscription level by evenly assigning the layered access in the overlay. We develop a decentralized heuristic algorithm to approximate the optimal solution.

6. REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of H.264/AVC," IEEE Trans. CAS for Video Tech., vol. 17, no. 9, pp. 1103-1120, Sept. 2007.

[2] S. McCanne, V. Jacobson, and M. Vetterli, "Receiver-Driven Layered Multicast," In Proceedings of ACM SIGCOMM '96, volume 26(4), pages 117-130, Stanford, CA, August 1996.

[3] S. Low and D. E. Lapsley, "Optimization flow control, I: Basic algorithm and convergence," IEEE/ACM Trans. Networking, vol. 7, pp. 861–874, Dec. 1999.

[4] K. Kar, S. Sarkar, and L. Tassiulas,"Optimization based rate control for multirate multicast sessions," Proc. IEEE INFOCOM 2001, April 2001.

[5] Y.-H. Chu, S. G. Rao, and H. Zhang, "A Case for End System Multicast," In Proceedings of ACM SIGMETRICS, June 2000.

[6] J. Zhao, F. Yang, Q. Zhang, et.al., "LION: Layered overlay multicast with network coding," IEEE Transactions on Multimedia, vol. 8, no. 5, pp. 1021-1032, Oct. 2006.

[7] Katti, S. Rahul, H. Wenjun, et.al., "XORs in the Air: Practical Wireless Network Coding ," IEEE/ACM Trans. Networking, vol.16, no.3,pp.497-510, June, 2008.