CHANNELIZED PARTITIONING PROBLEM IN MULTI-RATE BROADCASTING OVER BANDWIDTH-CONSTRAINED NETWORKS

Jiangchuan Liu[1], Bo Li[1], Yiwei Thomas Hou[2], and Imrich Chlamtac[3]

[1] Department of Computer Science, The Hong Kong University of Science and Technology
[2] The Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061
[3] Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas

Abstract - This paper presents a formal study on the optimal bandwidth partitioning for multi-rate video broadcasting in a broadband wireless network. The formulation is generic in that it considers both inter-session and intra-session bandwidth partitions for layering as well as stream replication based broadcasting. It also takes into account the most fundamental issues associated with video transmission including encoding overhead, non-linear relationship between the receiver perceived video quality and the delivered bandwidth, as well as intra-session and inter-session fairness. Specifically, we consider the bandwidth-constrained case with channelized allocation. We show that there exist polynomial time algorithms for both inter-session and intra-session partitioning problems.

Keywords – Broadcasting, Multi-rate video

I. INTRODUCTION

With the rapid development and deployment of broadband networks, real-time video distribution is emerging as one of the most important networked applications [1]. The multi-user nature of video programs makes broadcasting a natural choice for delivering video content to a large population of receivers. It is efficiently supported by the physical infrastructure of wireless networks, and minimizes the redundancy introduced by using multiple unicast connections. However, it is envisioned that users with different wireless access enabled platforms, such as cellular phones, PDAs and laptop computers, will be able to easily access various video services in the near future [1]. Hence, a distinctive feature of a video distribution system is that the receivers (users) in a session [1] are highly heterogeneous in terms of their access bandwidths and processing capabilities. A single broadcasting rate is unlikely to satisfy the dynamic and heterogeneous requirements from different receivers, in that it might overwhelm the slow receivers and starve the fast ones [3].

It is therefore desirable to use multi-rate transmission, in which the receivers in the same broadcast session can receive video streams at different bandwidths [2,3]. The intrinsic scalability of video content makes multi-rate coding and transmission possible. There are two typical multi-rate transmission schemes: *stream replication* [3,8] and *cumulative layering* [2,3]. In stream replication, a source maintains a small number of video streams. Each stream carries the same video content but is encoded at different rates, targeting receivers with different bandwidths [8]. In cumulative layering, a raw video is compressed into a number of layers. The layer with the highest importance, called base layer, contains the data representing the most important features of the video, while additional layers, called enhancement layers, contain data that progressively refine the reconstructed video quality. A receiver can subscribe to a selected number of layers that best match its bandwidth and processing capability.

Note that a receiver cannot partially subscribe to a stream or a layer. Moreover, existing systems usually assume the video stream structure, i.e., the number of streams or layers, as well as the bandwidth of each stream or layer, is predetermined by the coding algorithm. Usually, in a practical system, only a small number of streams or layers are generated to keep the redundancy of replicating or the overhead of layering at a low level. Thus the adaptation granularity on the receiver's side is considerably coarse which causes mismatches between a receiver's expected bandwidth and the actually delivered video bandwidth. In addition, video programs are of different interest. Some 'hot' programs attract much more receivers than others [6]. When considering the overall system utilization, it is clearly inefficient to use the same coding structure for all the sessions.

There two possible methods to reduce this bandwidth mismatch. The first is to use a large number of layers or

This research was supported in part by grants from Research Grant Council (RGC) under contracts AoE/E-01/99 and HKUST6196/02E. J. Liu's work is partially supported by a Microsoft fellowship.

[1] A *broadcast session* consists of a video program and the receivers that are interested in this video.

0-7803-7589-0/02/$17.00 ©2002 IEEE

streams where the bandwidth of each one is relatively narrow. However, since layering introduces extra overheads for both coding and transport, and replication creates redundancy, the benefits from the improved granularity could be contradicted. Therefore, it is necessary to strike a balance between the adaptation granularity and overheads, or essentially, to choice an optimal number of layers or streams for a given bandwidth budget. The second is to adaptively allocate the bandwidth among layers or streams according to the receivers' expected bandwidth distribution so that the average mismatch is minimized. Evidently this is well justified in a typical broadcast environment, in which the bandwidths of the receivers in a session usually follow some clustered distribution because they use standard access interfaces. Therefore, if the layer or stream rates can be dynamically adjusted to match these clusters, the expected mismatch for a session can be reduced. In addition, in a bandwidth-limited network, it is desirable to allocate different bandwidth and employ different layering structures to the sessions according to their popularity so that the expected mismatch for the whole system can be minimized.

There are, however, two pre-requirements associated with the dynamic structure adjustment. First, the video coder should have a flexible layering structure. Second, the sender should know the global state of the receivers. We note that, in the coding area, several layered video encoders with flexible layering structures have been developed recently [5]. In such coders, both the bandwidth of a layer and the number of layers can be dynamically manipulated with fast responsiveness. There are also fast transcoding algorithms for stream replication presented in the literature [13]. Moreover, the fast and low-cost operations for layer stream setup and termination have been supported in advanced video streaming standards, such as the MPEG-4 Delivery Multimedia Integration Framework (DIMF). Many scalable feedback algorithms have also been presented in the networking area [11]. It is a fact that a scalable feedback loop, such as RTCP, has been embedded in many streaming video systems.

In this paper, we present a multi-granular end-to-end adaptation framework for layered video transmission. In this framework, both the sender and the receivers perform adaptation. A receiver's adaptation is at a layer or stream level and based on its individual bandwidth expectation, which is relatively simple and thus suitable for low-capability devices. The key component in the framework is the sender-based adaptation system, which jointly optimizes the number of layers or streams and the bandwidth for each layer or stream, as well as the bandwidth for each session. The allocation is based on the global status, specifically, the distributions of receiver's bandwidth and their interested videos.

Note that, for a bandwidth-constrained network, the bandwidth allocation for multiple layers or streams is essentially to partition a given bandwidth budget to a given number of parts. In this paper, we refer to this problem as the *optimal partitioning problem* for multi-rate video broadcasting, and present a formal study on it. Specifically, we formulate the optimal bandwidth partitioning problem of multi-session multi-rate video broadcasting for both the stream replication and the cumulative layering schemes, with the objective of maximizing some given revenue functions. We also show that the partitioning problem can be solved in polynomial time for the channelized partitioning case, such as wireless networks.

The rest of the paper is organized as follow. In section II, we give an overview of the framework. Sections III and IV formulate the optimal partitioning problem for layered broadcasting and replication based broadcasting, respectively, and present efficient allocation algorithms. We conclude the paper in Section V.

II. FRAMEWORK OVERVIEW

A. The Adaptation Framework

In our framework, a set of video programs are simultaneously broadcasted in a broadcast-enabled network, such as a wireless LAN. A central server allocates bandwidth among sessions as well as layers or streams within a session. A receiver interested in a particular video program can subscribe to the appropriate stream or layers that best match its bandwidth and processing capability. It cannot subscribe to a fraction of a stream or a layer. The basic bandwidth allocation unit is a *channel*, which can be a fixed transmission unit, such as a time slot in TDMA systems or a frequency in FDMA systems [1], or a logical allocation unit, such as the logical channel in WCDMA [1]. A video layer or stream can occupy multiple channels. In the following of this paper, we use the number of channels as the bandwidth measure, and use channel allocation and bandwidth allocation interchangeably.

This model can be formally characterized by a 4-tuple, $(N, P, M_{s,t}, W_s)$. Here, N is the total number of available channels in the network; P is the total number of video programs (also the number of sessions). Each program has an index in [1,..,P]; $M_{s,t}$ is the number of receivers with expected bandwidth t in session s; and W_s is a revenue function of video s.

B. Objective Functions

In the context of broadcast or multicast, the commonly used optimization objective for content distribution is to maximize certain revenue function(s) [3,7,8]. This function maps the service received into some performance level delivered to the end user. However, the exact form of the revenue function is still an open research topic, and is really depends on the encoding and transmission algorithms and, more important, the design objective of the system. Rather than choose a specific function, claim that it accurately represents the truth about application revenue, and optimize it accordingly, our objective is to extract the essential properties of the revenue functions, e.g., the parameters they depend on and their most common behaviors with different parameter settings, and to devise optimization algorithms only rely on these essential properties.

We note that, in the literature, the widely used revenue functions can be classified into two categories:

1. *Absolute revenue*. Usually used to characterize the throughput of the system. For example, the bandwidth actually delivered to a receiver.

2. *Relative revenue*. Usually used to characterize the receivers' satisfaction. For example, the fairness function [8], which is define as the bandwidth delivered to a receiver over its expected bandwidth.

Note that, in the context of video transmission, the delivered bandwidth and the perceived video quality exhibit a somewhat non-linear relation [5,7]. Therefore, a mapping from the bandwidth measure to the corresponding video quality measure should also be performed. This mapping depends on the video coding algorithm and the transmission bandwidth. It is also affected by the overhead of layering. In generally, the more the layers are generated for a target bandwidth, the higher the overhead incurred.

In conclusion, given a layered coding algorithm, a general revenue function for a receiver in session *s* is a function of its expected video quality and the video quality delivered to it, which further depends on: t, its expected bandwidth, *b*, its actually received bandwidth, and *l*, the number of layers generated for this bandwidth. Hence, the function can be denoted as $W_s\ (t,b,l)$. According to the basic properties of practical layered video coders, the revenue function should satisfy: 1) $W_s\ (t,b,l) \le W_s\ (t,b',l), b < b'$. That is, for two allocations, if a receiver subscribes to the same number of layers, then the allocation with a higher cumulative bandwidth to this layer delivers a higher revenue. 2) $W_s\ (t,b,l) \le W_s\ (t,b,l'), l > l'$. That is, for two allocations, if the bandwidths delivered to a receiver are the same, then the allocation with a smaller number of layers for this bandwidth delivers a higher revenue.

The revenue function for stream replication scheme can be viewed as a special case of layered coding with *l*=1. For simplicity, we denote it as $W_s\ (t,b)$ [2].

III. THE CHANNELIZED PARTITIONG PROBLEM

In this section, we first consider the problem of bandwidth partition for layered broadcasting. Let $\vec{R}_s$ denote the channel partition for multi-rate video program *s*, $\vec{R}_s = (r_{s,1}, r_{s,2}, ..., r_{s,l_s})$, where l_s is the total number of layers of this partition, and $r_{s,i}$ is the cumulative bandwidth up to layer *i*. A valid $\vec{R}_s$ should satisfy: 1) $l_s > 0$; 2) $0 < r_{s,1} < r_{s,2} <, ..., < r_{s,l_s} \le N$.

For receiver adaptation, we consider a generic case that a receiver tries to subscribe to the best-matching layers to maximize its individual revenue [2,3,7,8]. Assume a receiver with bandwidth t joins session s, the best-matching video stream bandwidth is given by $\alpha(t,\vec{R}_s) = \max_{r \le t, r \in \vec{R}_s} r$, and the corresponding index of the highest layer is $\beta(t,\vec{R}_s) = \arg_i \{r_{s,i} = \alpha(t,\vec{R}_s)\}$. The receiver should subscribe to layers 1,2,…, $\beta(t,\vec{R}_s)$, and its revenue is thus given by $W_s\ [t, \alpha(t,\vec{R}_s), \beta(t,\vec{R}_s)]$.

In our model, receiver adaptation is trivial given the one hop nature. Thus, we focus on sender adaptation, i.e., channel allocation at the server's side. Note that, given a specific allocation, there could be mismatches between receivers' expected bandwidths and the received bandwidths as the adaptation unit at the receiver's side is a layer. To minimize these mismatches, a centralized algorithm attempts to partition the available channels among different sessions as well as among layers within a session. The input of the algorithm is the system state, $(N, P, M_{s,t}, W_s)$, and the output is the maximum total revenue U^* of all the receivers in the network, together with the corresponding partitions, $\vec{R}_s, s = 1,2,...,P$. This yields not only the bandwidth allocation for each layer ($r_{s,i}$), but also the number of layers (l_s) required, and the bandwidth for each session (r_{s,l_s}).

The optimal partitioning problem for layered broadcasting is formally stated as follows:

(*OPT-PARTITION-LYR*)

Maximize $U = \sum_s^P \sum_{t=1}^{T_s} M_{s,t} W_s\ [t, \alpha(t,\vec{R}_s), \beta(t,\vec{R}_s)]$.

Subject to $\vec{R}_s$ is valid, $s = 1,2,...,P$,

$$\sum_s^P r_{s,\ l_s} \le N .$$

[2] Note that there is extra bandwidth consumption for control such as packet and stream identifications. However, they are either fixed or depend on the transmission bandwidth or the number of layers; thus can be easily incorporated into the revenue function.

Here, $T_s = \max_{M_{s,t}>0} t$; that is, the maximum bandwidth of the receivers in session s.

The basic framework of stream replication based broadcasting is similar to that of layered broadcasting. The difference is that, in the channel partition vector $\vec{R}_s = (r_{s,1}, r_{s,2}, ..., r_{s,l_s})$, l_s is the total number of streams and $r_{s,i}$ is the bandwidth of stream i. Without loss of generality, we assume that $0 < r_{s,1} < r_{s,2} <, ..., < r_{s,l_s} \le N$. A receiver with bandwidth t should subscribe to stream $\beta(t, \vec{R}_s)$, and its revenue is thus given by $W_s\,[t, \alpha(t, \vec{R}_s)]$.

The optimal partitioning problem in this case is formally stated as follows:

(*OPT-PARTITION-REP*)

$$\text{Maximize } U = \sum_{s}^{P} \sum_{t=1}^{T_s} M_{s,t} W_s\,[t, \alpha(t, \vec{R}^s)],$$

$$\text{Subject to } \vec{R}_s \text{ is valid}, s = 1,2,...,P,$$

$$\sum_{s}^{P} \sum_{i=1}^{l_s} r_{s,i_s} \le N .$$

IV. OPTIMAL PARTITIONGING ALGORITHMS

Note that the number of valid allocations is finite and the revenue functions are well defined for each valid allocation. Hence, there exist optimal solutions for problems *OPT-PARTITION-LYR* or *OPT-PARTITION-REP*.

Assume the revenue of session *s*, $U_s(n)$, is the total revenue of all the receivers in the session, when *n* channels are allocated to this session, and the maximum session revenue, $\hat{U}_s(n)$, is the maximum of $U_s(n)$. We have $U^* = \max_{\sum_{s=1}^{P} n_s \le N} \sum_{s=1}^{P} \hat{U}_s(n_s)$, which implies that the partitioning problems can be solved by two steps. First, optimal intra-session partitioning that optimally partitions channels for the layers in a session for any possible channel budget of this session; Second, optimal inter-session partitioning that optimally partitions channels among sessions for a total channel budget of the network. We next briefly describe the allocation algorithms and present their time complexities.

A. Intra-Session Partitioning for Layered Broadcasting

In the cumulative layering case, the sub-problem of optimal intra-session partitioning for session *s*, given n_s channels assigned to this session, is stated as follows:

$(INTRA-LYR_{s,n_s})$

$$\text{Maximize } U_s(n_s) = \sum_{t=1}^{T_s} M_{s,t} W_s\,[t, \alpha(t, \vec{R}_s), \beta(t, R_s)],$$

$$\text{Subject to } \vec{R}_s \text{ is valid, and } r_{s,l_s} = n_s .$$

We have devised an iterative algorithm to solve this problem. Note that, it is useless to allocate more than T_s channels to session *s* in the cumulative layering case. Let $\sigma(m,l) = \max_{l_s=l,\ r_{s,l}=m} \sum_{t=1}^{T_s} M_{s,t} W_s\,[t, \alpha(\vec{R}_s, t), \beta(\vec{R}_s, t)]$, *l*=1,2,…, T_s , and *m*=1,2,…, T_s . That is, the optimal session revenue when totally *l* layers are generated and the cumulative bandwidth up to layer 1 is *m*. The solution to problem $INTRA-LYR_{s,n_s}$ is clearly given by $\max_{1\le l\le n_s, 1\le m\le n_s} \sigma(m,l)$. The iterative algorithm calculates $\sigma(m,l)$ from a boundary condition where *l*=1, and in each iteration, one more layer is added. Since the subscription policy is cumulative, only the receivers that can subscribe to the previous layer have the potential of subscribing to the current layer; the algorithm thus can find local optimal solution for each *l* and finally obtain the global solution. Note that, after T_s iterations, problems $INTRA-LYR_{s,n_s}$, n_s=1,2,…,T_s, all can be solved because all the values of $\sigma(m,l)$ are available, and in the worst case, the time complexity is bounded by $O[(T_s)^4]$.

B. Inter-Session Partitioning for Layered Broadcasting

The sub-problem of optimal inter-session allocation is stated as follows:

(*INTER-LYR*)

$$\text{Maximize } U^* = \sum_{s=1}^{P} \hat{U}_s(n_s),$$

$$\text{Subject to } n_s > 0, s = 1,2,...,P, \text{ and } \sum_{s=1}^{P} n_s \le N.$$

Let $\tau(n,p) = \max_{\sum_{s=1}^{p} n_s = n} \sum_{s=1}^{p} \hat{U}_s(n_s)$, *n*=1,2,…,*N*, and *p*=1,2,…,*P*. That is, the maximum total revenue of sessions 1, 2, …, *p*, when totally n channels are allocated to these sessions. The solution to problem *INTER-LYR* is clearly given by $\max_{1 \le n \le N} \tau(n, P)$. We have devised an iterative algorithm to calculate $\tau(n, p)$. The algorithm starts from p=1, i.e., only session 1 is considered. In each iteration, one more session is added, and all possible bandwidth allocations (from 1 to T_s) to this session are checked. This can be done in time $O(T_s)$ because the session's revenue is independent of other sessions. The time complexity of this iterative algorithm is bounded $O(P \cdot N \cdot T)$, where $T = \max_{1 \le s \le P} T_s$.

C. Stream Replication Case

The partitioning problem for the stream replication case can also be solved the decomposition mechanism. The sub-problem of optimal intra-session allocation for session s, given the number of channels assigned to this session, $n_s \in [1..N]$, is stated as follows:

$$(INTRA - REP_{s,n_s})$$

$$\text{Maximize } U_s(n_s) = \sum_{t=1}^{T_s} M_{s,t} W_s\,[t, \alpha(t, \vec{R}_s)]\,,$$

$$\text{Subject to } \vec{R}_s \text{ is valid, and } \sum_{i=1}^{l_s} r_{s,i} = n_s\,.$$

This problem can be solved by an iterative algorithm with time complexity $O[N(T_s)^2]$. Note that, once the optimal session revenue are obtained, the optimal inter-session allocation for stream replication is similar to that of layered broadcasting, except that the maximum bandwidth allocated to a session is not bounded by T_s, but by N. As a result, its time complexity is $O(PN^2)$.

V. CONCLUSION

This paper presents a formal study on the optimal bandwidth partitioning for multi-rate video broadcasting in a broadband wireless network. The formulation is generic in that it considers both inter-session and intra-session bandwidth partitions. It also takes into account the most fundamental issues associated with video transmission including encoding overhead, non-linear relationship between the receiver perceived video quality and the delivered bandwidth, as well as intra-session and inter-session fairness.

It is worth pointing out that the mechanism we discuss and analyze in this paper is independent of the actual transport network. For networks that can allocation bandwidth in a continuous manner, the formulation and solutions presented this paper remain valid by some minor modifications. Specifically, note that, a practical video coder has only a finite set of admissible quantizers; therefore, there are only a finite number of possible rates for any given source. These discrete outputs can be used to emulate the channelized allocation. Therefore, our framework is generally applicable to other systems that are broadcast- or multicast-capable, and can provide reasonably fast responses.

REFERENCES

[1] L. Hanzo, P. Cherriman, and J. Streit, *Wireless Video Communications: Second to Third Generation Systems and Beyond*, IEEE Press, 2001.

[2] S. McCanne, V. Jacobson, and M. Vetterli, "Receiver-driven Layered Multicast," in *Proceedings of ACM SIGCOMM 96*, pp.117-130, August 1996.

[3] X. Li, M. Ammar, and S. Paul, "Video Multicast over the Internet," *IEEE Network Magazine*, Vol. 13, No. 2, pp. 46-60, April 1999.

[4] D.-P. Wu, Y.-T. Hou, W. Zhu, Y.-Q. Zhang, and J. Peha, "Streaming Video over the Internet: Approaches and Directions," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 11, No. 1, February 2001.

[5] S. Li, F. Wu and Y.-Q. Zhang, "Experimental Results with Progressive Fine Granularity Scalable (PFGS) Coding," *ISO/IEC JTC1/SC29/WG11*, *MPEG99/m5742,* March 2000.

[6] G. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley Press, 1949.

[7] S. Bajaj, L. Breslau, and S. Shenker, "Uniform versus Priority Dropping for Layered Video," in *Proceedings of ACM SIGCOMM' 98*, September 1998.

[8] T. Jiang, E. Zegura, and M. Ammar, "Inter-Receiver Fair Multicast Communication Over the Internet," in *Proceedings of NOSSDAV'99*, June 1999.

[9] D. Titterington, A. Smith, and U. Makov, *Statiscal Analysis of Finite Mixture Distributions*, Wiley Publishers, NY, 1985.

[10] K. Almeroth and M. Ammar, "On the Use of Multicast Delivery to Provide a Scalable and Interactive Video-on-Demand Service," *IEEE Journal on Selected Areas in Communications*, Vol. 14, No. 50, p. 1110-1122, August 1996.

[11] T. Turletti, S. Parisis, and J. Bolot, "Experiments with a Layered Transmission Scheme over the Internet," *Technical Report*, INRIA, N'3296, November 1997.

[12]P.Kuhn, T.Suzuki, and A.Vetro, "MPEG-7 Transcoding Hints for Reduced Complexity and Improved Quality," in *Proceeding of PacketVideo '01*, April 2001.

[13] J. Youn, J. Xin, and M.-T. Sun, "Fast Video Transcoding Architectures for Networked Multimedia Applications," in *Proceedings of IEEE International Symposium of Circuits and Systems (ISCAS'00)*, May 2000.