A Low Bandwidth Broadcasting Protocol for Video on Demand

Jehan-François Pâris[†]
Department of Computer Science
University of Houston
Houston, TX 77204-34
paris@cs.uh.edu

Steven W. Carter Darrell D. E. Long[‡]
Department of Computer Science
Baskin School of Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
{carter, darrell}@cse.ucsc.edu

Abstract

Broadcasting protocols can improve the efficiency of video on demand services by reducing the bandwidth required to transmit videos that are simultaneously watched by many viewers. We present here a polyharmonic broadcasting *protocol that requires less bandwidth than the best extant protocols to achieve the same low maximum waiting time.*

We also show how to modify the protocol to accommodate very long videos without increasing the buffering capacity of the set-top box.

Keywords: video on demand, video broadcasting, harmonic broadcasting.

1 Introduction

Video on demand (VOD) proposes to provide subscribers who are connected through a set-top box (STB) with the possibility of ordering at any time the video of their choice and starting immediately to watch it on their television set. Despite its great potential, VOD has had a slow start. None of the companies that invested in VOD have been able to come with a single successful commercial system. The overall consensus now is that the commercial deployment of VOD will have to wait until the cost of building and maintaining the required infrastructure can be significantly lowered.

Broadcasting is one of several techniques that aim to reduce the cost of VOD [9]. It is clearly not a panacea as it only applies to videos that are likely to be watched by many viewers. Even so, the savings that can be achieved are nevertheless considerable, as it is often the case that 40 percent of the demand is for a small number, say, 10 to 20, of popular videos [2–4]. A naive broadcasting strategy would simply consist of retransmitting the same video on several distinct channels at equal time intervals. The major problem with this approach is the number of channels per video required to achieve a reasonable waiting time. Consider, for instance, the case of a video lasting two hours, which happens to be close to the average duration of a feature movie. To guarantee that no customer would ever have to wait more than five minutes we would have to broadcast twenty-four different copies of the video starting every five minutes.

Many more efficient protocols have been proposed. They include Viswanathan and Imielinski's *pyramid broadcasting protocol* [8], Aggarwal, Wolf and Yu's *permutation-based pyramid broadcasting protocol* [1], Hua and Sheu's *skyscraper broadcasting protocol* [5], Juhn and Tseng's *harmonic broadcasting protocol* [6] and its variants [7].

All these protocols share a similar organization. They divide each video into *segments* that are simultaneously broadcast on different data streams. One of these streams transmits nothing but the first segment of the video in real time. The other streams transmit the remaining segments at lower bandwidths. When customers want to watch a video, they wait first for the beginning of the first segment on the first stream. While they start watching that segment, their set-top box (STB) starts downloading enough data from the other streams so that it will be able to play each segment of the video in turn.

This approach requires an STB capable of storing a significant fraction (around 40 percent for some protocols) of each video while it is being watched. This extra cost is more than compensated by the bandwidth savings that can be achieved. While the staggered broadcasting technique we described above requires twenty-four channels to guarantee a maximum waiting time of five minutes for a two-hour video, harmonic broadcasting protocols require slightly less than four channels to achieve the same qual-

[†] This work was performed while this author was on sabbatical leave at the Department of Computer Science, University of California, Santa Cruz.

[‡] This research was supported by the Office of Naval Research under Grant N00014–92–J–1807.

ity of service.

As we will see, even greater bandwidth savings could be achieved if the STB could start downloading data from the moment the customers select the video they want to watch rather than waiting for the beginning of the first segment on the first stream. We present a new *polyharmonic protocol* that imposes the same fixed delay to all customers wanting to watch a given video. Since this delay is the same for all customers, the protocol can take advantage of it to reduce the transmissions of all segments, including the first one. As a result, the total bandwidth can be further reduced by around ten percent.

The remainder of the paper is organized as follows. Section 2 presents the harmonic broadcasting protocol and its variants. Section 3 introduces our new protocol and compares its bandwidth requirements to those of the harmonic broadcasting protocols. Section 4 discusses the main advantages and limitations of polyharmonic broadcasting. Section 5 presents a possible extension of polyharmonic broadcasting that would let them handle videos of arbitrary length. Section 6 contains our conclusions.

2 Harmonic Broadcasting

Harmonic Broadcasting (HB) divides a video into n equally-sized *segments*. Each segment S_i, for $1 \leq i \leq n$, is broadcast repeatedly on its own channel with a bandwidth b/i, where b is the consumption rate of the video (see Figure 1).

When a client requests a video, it must wait for the start of an instance of S_1 and then begin receiving data from every stream for the video. That means that the client and the server must be able to support a bandwidth of

$$B_{HB}(n) = \sum_{i=1}^{n} \frac{b}{i} = b \sum_{i=1}^{n} \frac{1}{i} = bH(n)$$

where $H(n)$ is the harmonic number of n.

A *slot* is the amount of time it takes for a client to consume a single segment of the video. We represent this time by d. Since the first segment is broadcast with that periodicity, d is is given by

$$d = \frac{S_1}{b} = \frac{D}{n}$$

and is also the maximum amount of time a client must wait before viewing its request.

A *subsegment* is the amount of a segment the client receives during a slot of time. The first segment only has one subsegment, the segment itself; every other segment S_i has i equal subsegments, $S_{i,1}, S_{i,2}, \ldots, S_{i,i}$.

Unfortunately HB does not always deliver all data on time. Consider the first two streams in Figure 1. If the client makes its request in time to receive the second instance of S_1 and starts receiving data at time t_0, then it will need all of the data for $S_{2,1}$ by time $t_0 + 3d/2$. However, it will not receive all of that data until time $t_0 + 2d$. It turns out HB will not work unless the client always waits an extra slot of time before consuming data.

Two variants on HB do not impose the extra waiting time. *Cautious Harmonic Broadcasting* (CHB) broadcasts the video in a similar fashion as HB. The first stream broadcasts S_1 repeatedly as before, but the second stream alternates between broadcasting S_2 and S_3. Then the remaining $n-3$ streams broadcast segments S_4 to S_n such that the stream for S_i has bandwidth $b/(i-1)$.

As before, the client will receive data from all streams for the video simultaneously. That means CHB requires a bandwidth of

$$\begin{aligned} B_{CHB}(n) &= 2b + \sum_{i=3}^{n-1} \frac{b}{i} \\ &= \frac{b}{2} + bH(n-1) \end{aligned}$$

or roughly $b/2$ more than the original HB protocol.

Quasi-harmonic Broadcasting (QHB) uses a more complex scheme to break up the video. The first segment is left intact, but then each of the remaining segments S_i, for $2 \leq i \leq n$, is divided up into $im-1$ *fragments* for some positive parameter m. Slots are also broken up into m equal *subslots*, and each subslot can be used to broadcast a single fragment. The key to QHB is that the fragments are not broadcast in order. The last subslot of each slot is used to broadcast the first $i-1$ fragments repeatedly, and the rest of the fragments are ordered such that the k^{th} subslot of slot j is used to broadcast fragment $ik + j - 1 \bmod i(m-1) + i$ (see Figure 2).

Since the above ordering adds some redundancy—each sequence of im fragments will contain one of the first $i-1$ fragments twice—each subslot of stream i will have to broadcast $1/(im-1)$ of segment S_i instead of $1/im$ as in HB. This will increase the required bandwidth for stream i from b/i to $bm/(im-1)$ for $2 \leq i \leq n$. Thus the total bandwidth required for QHB is

$$\begin{aligned} B_{QHB}(n,m) &= b + \sum_{i=2}^{n} \frac{bm}{im-1} \\ &= bH(n) + \sum_{i=2}^{n} \frac{b}{i(im-1)}. \end{aligned}$$

with

$$\lim_{m \to \infty} \sum_{i=2}^{n} \frac{b}{i(im-1)} = 0$$

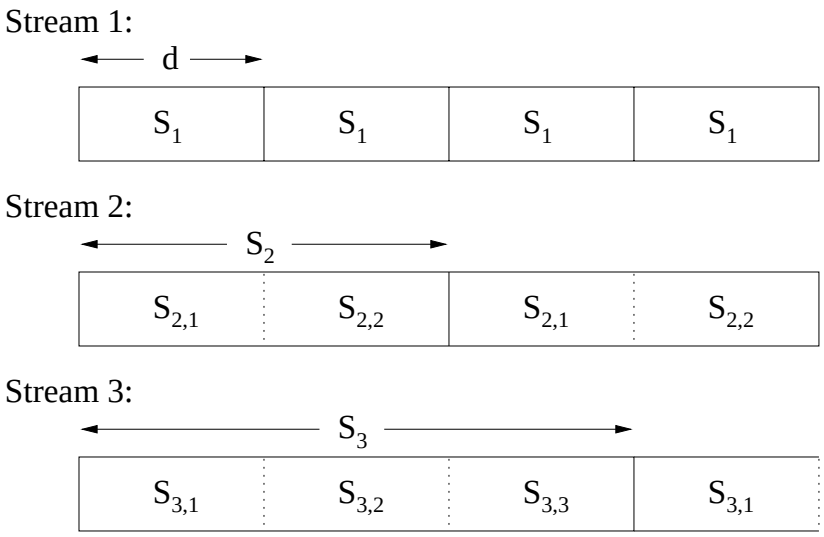

Figure 1: An illustration of the first three streams for a video under harmonic broadcasting.

S_1				S_1				S_1				S_1			
$S_{2,2}$	$S_{2,4}$	$S_{2,6}$	$S_{2,1}$	$S_{2,3}$	$S_{2,5}$	$S_{2,7}$	$S_{2,1}$	$S_{2,2}$	$S_{2,4}$	$S_{2,6}$	$S_{2,1}$	$S_{2,3}$	$S_{2,5}$	$S_{2,7}$	$S_{2,1}$
$S_{3,3}$	$S_{3,6}$	$S_{3,9}$	$S_{3,1}$	$S_{3,4}$	$S_{3,7}$	$S_{3,10}$	$S_{3,2}$	$S_{3,5}$	$S_{3,8}$	$S_{3,11}$	$S_{3,1}$	$S_{3,3}$	$S_{3,6}$	$S_{3,9}$	$S_{3,2}$

Figure 2: An illustration of the first three streams for a video under quasi-harmonic broadcasting when $m = 4$.

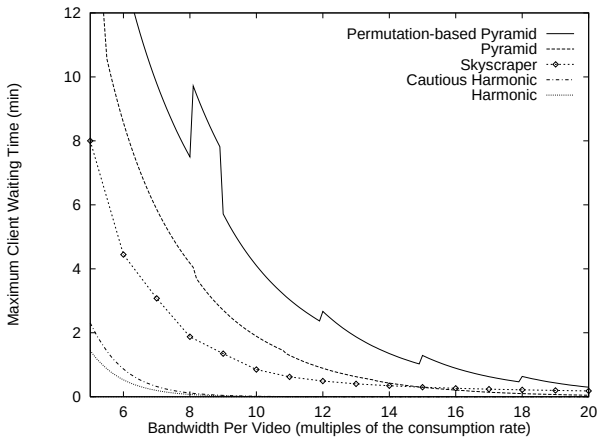

Figure 3: How harmonic broadcasting compares to other broadcasting protocols (from [7])

The major advantage of these three harmonic protocols is their low bandwidth requirements. Figure 3 shows the bandwidth versus client waiting times for the harmonic and cautious harmonic broadcasting and compares them with those of pyramid broadcasting [8], the "unconstrained" version of permutation-based pyramid broadcasting [1] and skyscraper broadcasting with a maximum width of 52 [5].

Both harmonic broadcasting protocols emerge as clear winners as none of the three other protocols even approaches their performance. One may then wonder whether harmonic broadcasting does not provide the minimum bandwidth required to guarantee a given maximum waiting time. As we will see in the next section, this is not the case. The same maximum waiting times can be achieved at a lower cost by switching to a fixed wait policy and having the STB download data from the moment the customer selects the video.

3 Polyharmonic Broadcasting

Polyharmonic broadcasting (PHB) is a new broadcasting protocol aiming at reducing the bandwidth cost of providing a given maximum waiting time. To achieve its goal, polyharmonic broadcasting introduces two major changes. First, it requires that the client STB starts downloading data from the moment a customer requests a specific video instead of waiting until the customer begins watching the beginning of the first segment. Second, polyharmonic broadcasting uses a fixed wait policy. Under harmonic broadcasting and its variants, customers have to wait for the beginning of an instance of the first segment of a video. Polyharmonic broadcasting requires all customers to wait exactly the same amount of time regardlessly of the timing of their request.

Like harmonic broadcasting, polyharmonic broadcasting breaks a video into n segments of duration $d = D/n$ where D is the duration of the video. If b represents again the video consumption rate, the total size of the video S will be equal to bD and the size of each segment equal to S/n.

The protocol will allocate n distinct broadcasting streams to these n segments. Each stream i will repeatedly show segment S_i. Under polyharmonic broadcasting, no client can start consuming the first segment of the video before having downloaded data from all n streams during a time interval of duration $w = md$ where m is some integer $m \geq 1$. As a result, segment S_i will not be consumed until $(m + i - 1)d$ time units have elapsed from the moment the client started downloading data from the server. Ensuring that segment S_i will be entirely broadcast over this time interval would suffice to guarantee that all the contents of segment S_i will be already loaded in the STB before the customer starts viewing that segment. This can be achieved by retransmitting segment S_i at a transmission rate $b_i = \frac{b}{m+i-1}$. The total bandwidth B_{PHB} required by the polyharmonic broadcasting protocol is given by

$$\begin{aligned} B_{PHB}(n,m) &= \sum_{i=1}^{n} b_i \\ &= b \sum_{i=1}^{n} \frac{1}{m+i-1} \\ &= b(H(n+m-1) - H(m-1)) \end{aligned} \quad (1)$$

where $H(k)$ represents again the harmonic number of k.

Since the whole contents of segment S_i will be received by the STB before the customer starts viewing that segment, these data can be received in any arbitrary order. There is thus no need to wait as before for the beginning of a transmission of the first segment of the video. Hence the minimum waiting time w required by the protocol is also the *maximum* time a customer will ever have to wait.

Whenever the number of segments n is a multiple k of m, equation 1 can be rewritten as

$$B_{PHB}(k,m) = b(H((k+1)m - 1) - H(m-1))$$

Since $w = md$ and $d = D/n$, the waiting time w is linked with the duration of the video D by the relation

$$w = D/k.$$

In other words, all combinations of the two parameters n and m keeping the ratio n/m constant, will achieve the same waiting time w.

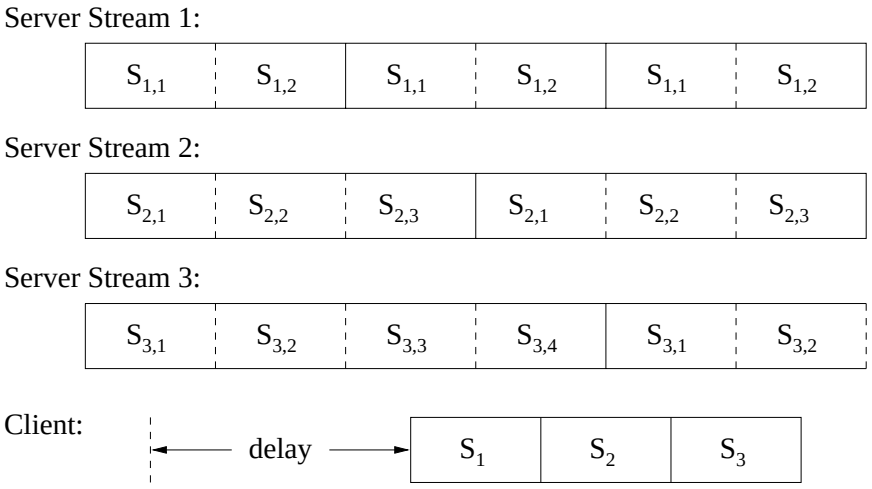

Figure 4: An illustration of the first three streams for a video under polyharmonic broadcasting with $m = 2$.

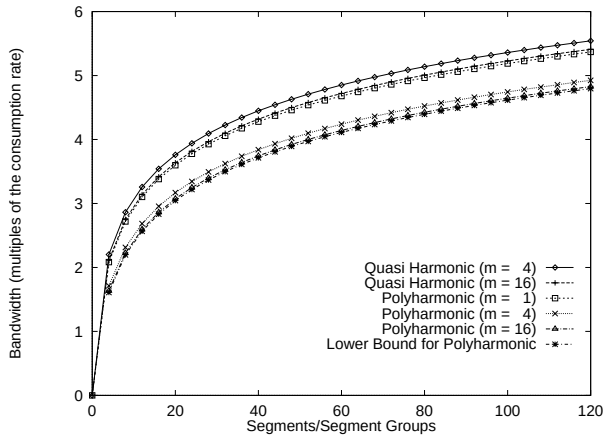

Figure 5: Bandwidth requirements of quasi-harmonic and polyharmonic broadcasting. The numbers on the x-axis represent number of segments for QHB and number of groups of m segments for PHB.

When $m = 1$, k becomes equal to n and equation 1 degenerates into

$$B_{PHB}(k,1) = b(H((k) - H(0)) = bH(n).$$

The polyharmonic protocol with $m = 1$ requires thus the same bandwidth as the harmonic protocol, while guaranteeing a maximum waiting delay equal to D/n, which the harmonic protocol cannot achieve.

Observing that

$$H((k+1)(m+1)-1) - H(m) = H((k+1)m-1) + \frac{1}{(k+1)m} + \cdots + \frac{1}{(k+1)(m+1)-1} - H(m-1) - \frac{1}{m}$$

and that

$$\frac{1}{(k+1)m} + \cdots + \frac{1}{(k+1)(m+1)1-1} < \frac{1}{m}$$

for all $k \geq 1$ and $m \geq 1$, we obtain

$$H((k+1)(m+1)-1) - H(m) < H((k+1)m-1) - H(m-1)$$

and

$$B_{PHB}(k,m+1) < B_{PHB}(k,m)$$

for all $k \geq 1$ and $m \geq 1$.

In other words, increasing m and n while keeping k constant will always result in a reduction of the total bandwidth. Selecting the optimum m for a given broadcast will thus be a trade-off between minimizing the overall bandwidth by increasing m and keeping the total number of streams $n = km$ manageable.

To derive a lower bound for the total bandwidth required by the polyharmonic broadcasting protocol, we need to compute the limit of $B_{PHB}(k,m)$ when m and n go to infinity while k remains constant.

$$\begin{aligned}\lim_{m\to\infty} B_{PHB}(k,m) &= \lim_{m\to\infty}\sum_{i=1}^{n}\frac{b}{m+i-1}\\ &= \int_0^D \frac{b}{w+t}dt\\ &= \log\frac{w+D}{w}\\ &= \log(k+1). \qquad (2)\end{aligned}$$

Figure 5 displays the bandwidth requirements of polyharmonic broadcasting and quasi-harmonic broadcasting. Since polyharmonic broadcasting requires m times as many segments as quasi-harmonic broadcasting to achieve the same maximum waiting time, we had to compare the bandwidths required by PHB with n segments with those required by QHB with m times less segments. Hence the numbers on the x-axis represent numbers of segments for QHB and numbers of groups of m segments for PHB. To eliminate the factor b representing the bandwidth of a standard full speed channel, all quantities on the y-axis are expressed in standard channels, that is, taking the bandwidth of a standard channel as unit of measurement. As one can see, the bandwidth required by polyharmonic broadcasting becomes significantly lower than that required by quasi-harmonic broadcasting for values of m as small as 4. The graph further indicates that very large values of m will not significantly reduce the bandwidth as polyharmonic broadcasting with $m = 16$ are virtually identical to the theoretical lower bound given by equation 2.

Figure 6 displays the bandwidth needed by all four harmonic broadcasting protocols to guarantee a given maximum waiting time. To eliminate the factor D representing the length of the video, the maximum waiting times on the x-axis are expressed as percentages of the video length. As in Figure 5, all quantities on the y-axis are expressed in standard channels, that is, taking the bandwidth of a standard channel as unit of measurement.

To compute the storage requirements of polyharmonic broadcasting, we can follow the approach as Juhn and Tseng in their analysis of the harmonic broadcast protocol [6]. Let R_i be the amount of data the client STB receives during time slot i. Then we can compute it as

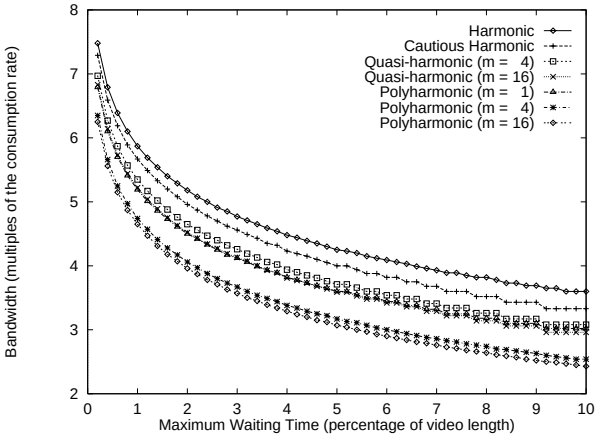

Figure 6: Bandwidth versus maximum waiting times for the quasi-harmonic and polyharmonic broadcasting protocols.

$$R_i = \begin{cases} db\displaystyle\sum_{j=1}^{n}\frac{1}{m+j-1} & 1 \le i \le m \\ db\displaystyle\sum_{j=i-m+1}^{n}\frac{1}{m+j-1} & m < i \le n+m-1 \\ 0 & i = n+m \end{cases}$$

We can also compute the amount of data consumed during a time slot as

$$C_i = \begin{cases} 0 & 1 \le i \le m \\ db & m < i \le n+m \end{cases}$$

Finally, we can define B_i as the amount of data the client has in its buffer after each time slot i, and calculate it as

$$B_i = B_{i-1} + R_i - C_i$$

where $B_0 = 0$. The maximum B_i gives the storage requirements for the protocol.

Figure 7 represents the storage requirements of the polyharmonic broadcasting protocol for videos having up to 200 segments at selected values of the parameter m. Since polyharmonic broadcasting stores every segment before broadcasting it, its storage requirements are particularly high when the video is subdivided into a small number n of segments with the worse case being $n = 1$. It is however very unlikely that polyharmonic broadcasting would be used in this context as it would be much simpler then to rebroadcast the video at normal speed on a single channel.

More reasonable values of n, say $n > 20$ and $m \ge 2$ lead to storage requirements below 50 percent of the video

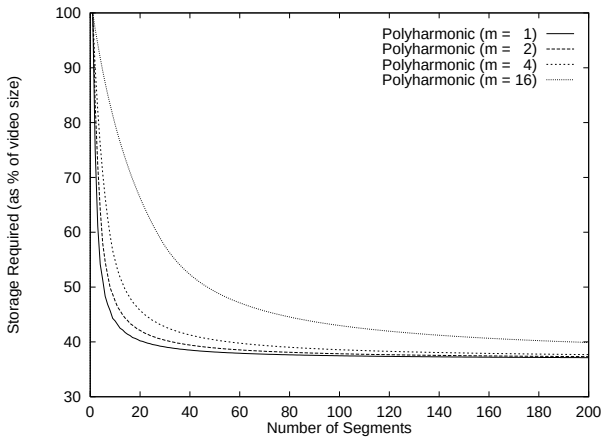

Figure 7: Storage requirements of polyharmonic broadcasting.

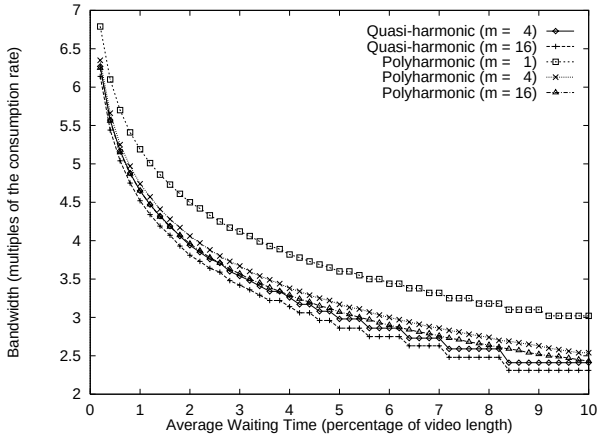

Figure 8: Bandwidth versus average waiting times for polyharmonic and quasi-harmonic broadcasting.

size for all values of $m \leq 4$. Higher values of m are not likely to be used since they do not provide significantly lower bandwidths than $m = 4$.

4 Discussion

As one can see, polyharmonic broadcasting requires less bandwidth than the best extant harmonic broadcasting protocol to guarantee a given maximum response time. It does not require the complex encoding scheme of quasi-harmonic broadcasting and, unlike harmonic broadcasting, it never fails to deliver the data on time. Despite these major advantages, our new broadcasting protocol presents two limitations that need to be addressed.

First, polyharmonic broadcasting requires m times more streams than other harmonic broadcasting protocols. This will require additional bookkeeping from the server and the client and will undeniably complicate their tasks.

Second, polyharmonic broadcasting forces all customers to wait for the maximum waiting delay while other harmonic protocols only require few customers to wait that long. If we were indeed comparing their mean waiting times as in Figure 8, polyharmonic broadcasting would be no better than extant harmonic protocols. The real issue here is customer response to service delays. We believe the average waiting time is not the best performance indicator for the quality of the service being provided because it assumes that the customer is insensitive to the variance of the service time. This is not true. Most customers are more annoyed when experiencing unusually long waiting times than they are elated when the experience a very fast service. A fixed waiting time has the advantage of being predictable. Many providers of video on demand services will probably use this delay to let their customers watch some previously downloaded announcements such as trailers for coming attractions and stern warnings to potential copyright violators. This would not be very different of what is already done on most video cassettes even though the customer would not have the option to "fast forward" until the beginning of the video itself.

5 Handling Long Videos

All efficient broadcasting protocols require enough buffer space in the user set-top box to store about 40 percent of each video. Hence they cannot be used to broadcast videos whose duration exceeds the capacity of the set-top box. The following extension eliminates this problem.

Let us assume without loss of generality that the set-top box buffer cannot hold more than l segments of the video. Then the client can operate in the following fashion:

1. It captures and stores in its buffer the l first segments of the video—that is, segments S_1 to S_l;
2. It plays these segments in sequence;
3. When it has finished playing a segment S_j with $1 \leq j \leq n - l$, it starts capturing segment S_{l+j} using the buffer that was previously used by S_j for that purpose.

The major drawback of this solution is the additional bandwidth. The client will have to capture segment S_{l+j} during the $l - 1$ time frames occurring within the time interval between the time when its has finished playing segment S_j and the time when it must start playing segment S_{l+j}. As a result, all segments S_{l+j} with $0 < j \leq n - l$ will have to be transmitted at a minimum bandwidth $\frac{b}{l-1}$ instead of $\frac{b}{m+l+j-1}$. Instead of having a bandwidth of

$bH(n+m-1) - bH(m-1)$, the new protocol would require a bandwidth of

$$bH(m+l-1) - bH(m-1) + \frac{b(n-l)}{l-1}$$

and the bandwidth overhead of the method would be given by

$$\frac{b(n-l)}{l-1} - bH(m+n-1) + bH(m+l-1)$$

These results are better illustrated in an example. Consider a video lasting four hours and assume we want to achieve a maximum waiting time of two minutes. With $m = 4$, that would require $4 \times 240/2 = 480$ segments and a total bandwidth equal to 4.925 times the video consumption rate. If we do not want to have more than one half of these segments simultaneously stored in the STB, the total server bandwidth would increase to 5.243 times the video consumption rate, that is still 10 percent less than the total bandwidth that cautious broadcasting would require to provide the same response time without restricting the number of stored segments. Since the client never has to capture more then 240 streams at the same time, the total client bandwidth will be somewhat lower and never exceed 4.239 times the video consumption rate.

6 Conclusions

Video broadcasting protocols can improve the efficiency of video on demand services by reducing the bandwidth required to transmit videos that are simultaneously watched by many viewers. Some of the newest broadcasting protocols to be proposed, harmonic broadcasting and its variants require much less bandwidth than other broadcasting protocols to guarantee the same maximum waiting time.

We have presented a new broadcasting protocol that provides the same maximum waiting time as the harmonic broadcasting protocol while consuming significantly less bandwidth. We also have shown how to modify the protocol to accommodate very long videos without increasing the buffering capacity of the set-top box.

More work needs to be done to investigate the possible existence of a theoretical lower bound for the bandwidth required to achieve a given maximum waiting time under any feasible broadcasting protocol.

References

[1] C. C. Aggarwal, J. L. Wolf, and P. S. Yu. A permutation-based pyramid broadcasting scheme for video-on-demand systems. In *Proceedings of the International Conference on Multimedia Computing and Systems*, pages 118–26, Hiroshima, Japan, June 1996. IEEE Computer Society Press.

[2] D. Clark. Oracle predicts interactive gear by early 1994. *The Wall Street Journal*, November 10, 1993.

[3] A. Dan, D. Sitaram, and P. Shahabuddin. Scheduling policies for an on-demand video server with batching. In *ACM Multimedia*, pages 15–23, San Francisco, California, Oct. 1994.

[4] A. Dan, D. Sitaram, and P. Shahabuddin. Dynamic batching policies for an on-demand video server. *Multimedia Systems*, 4(3):112–121, June 1996.

[5] K. A. Hua and S. Sheu. Skyscraper Broadcasting: a new broadcasting scheme for metropolitan video-on-demand systems. In *SIGCOMM 97*, pages 89–100, Cannes, France, Sept. 1997. ACM.

[6] L. Juhn and L. Tseng. Harmonic broadcasting for video-on-demand service. *IEEE Transactions on Broadcasting*, 43(3):268–271, Sept. 1997.

[7] J.-F. Pâris, S. W. Carter, and D. D. E. Long. Efficient broadcasting protocols for video on demand. In 6^{th} *International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS '98)*, pages 127–132, July 1998.

[8] S. Viswanathan and T. Imielinski. Metropolitan area video-on-demand service using pyramid broadcasting. *Multimedia Systems*, 4(4):197–208, Aug. 1996.

[9] J. W. Wong. Broadcast delivery. *Proceedings of the IEEE*, 76(12):1566–1577, Dec. 1988.