

Bandwidth reservation for the provision of VoD services over wireless access networks using hybrid ARQ schemes

Hairuo Ma and Magda El Zarki
Department of Electrical Engineering,
University of Pennsylvania,
Philadelphia, PA 19104

Abstract In this paper, we address the bandwidth reservation issue when MPEG-2 encoded VBR video streams are transmitted over wireless access networks using hybrid ARQ schemes to combat channel errors in VoD services. In addition to the original bandwidth which is necessary for the VBR traffic, a certain amount of excess bandwidth has to be reserved for retransmissions required by hybrid ARQ schemes. Different excess bandwidth reservation algorithms are proposed and investigated via simulations. The performance results are then presented and compared.

1. MPEG-2 ENCODED VOD SERVICES OVER B-FWANS

During the past years, there has been an increasing demand for broadband video services, such as video on demand (VoD), which require higher bandwidth both in the backbone networks as well as in the local access networks. In addition, the 1996 Telecommunication De-regulation Act has also propelled the development of wireless infrastructures. On the other hand, for most of the practical broadband video services, especially VoD services, mobility is not a top-priority consideration in the near future. Therefore, fixed wireless access networks (FWANs), especially broadband FWANs (B-FWANS) have attracted a lot of attention. B-FWANS, using wireless access loops connecting a fixed customer premise or terminal to the broadband network, provide an effective solution for new competitors to capture the broadband local access market, due to their fast deployment capability, cost effective infrastructure and low maintenance cost of the overall system. One promising candidate of B-FWANS to provide broadband services to the home or business is Local Multipoint Distribution/Communication System (LMDS/LMCS).^{1,2}

Figure 1 illustrates the system architecture of providing VoD services from video servers to the end users who are connected via B-FWANS. The broadband network is assumed to be an ATM network. Since a variety of services will be provided in one infrastructure, wireless ATM (WATM) infrastructure is adopted in the B-FWAN system as the extension of the ATM network to achieve ubiquitous connections. Most broadband video services provided over B-FWANS, including VoD services, will be

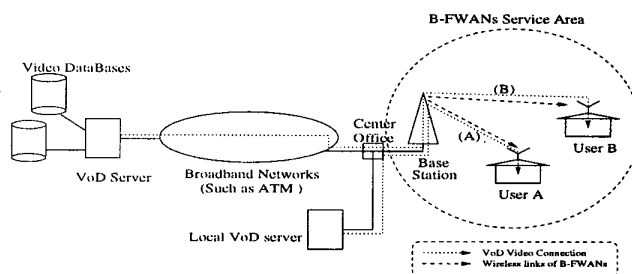


Figure 1. VoD over B-FWANS system architecture.

encoded following the MPEG-2 standard. To provide constant quality video services, MPEG-2 encoded VBR video programs are expected to be the major traffic type in VoD services.

Due to the fact that wireless channels are time-varying and have a high bit error rate (BER), error control is mandatory in order to guarantee quality of service (QoS). In the past years, hybrid ARQ schemes which combine the advantages of FEC and conventional ARQ schemes, have been proposed to combat channel errors.³⁻⁷ For VoD services, since a set-top box (STB) with a certain amount of memory (buffer) is installed on the user side, and the received video data is temporarily stored before they are removed out by the MPEG-2 decoder, hence, the real-time requirement for VoD service is not as stringent as for some other real-time interactive video services. In our previous studies, we have investigated the feasibility of applying ARQ-based error control schemes and the effectiveness of using hybrid ARQ schemes to the transmission of MPEG-2 encoded video streams over B-FWANS in VoD services.⁸⁻¹⁰

2. BANDWIDTH RESERVATION FOR HYBRID ARQ SCHEMES

2.1. Original and Excess bandwidth

Channel bandwidth reservation is necessary in a packet-switched network to ensure that VBR traffic can be transported without experiencing long delays. Many studies have been conducted related to the bandwidth reservation for VBR traffic when multiple videos are admitted and multiplexed over broadband networks. However, wireline channels, which have fairly consistent performance were

generally assumed in these studies.¹² When transmitting MPEG-2 VBR video traffic over wireless channels using hybrid ARQ schemes, the variance of the channel performance needs to be taken into consideration in the design and the implementation of the bandwidth reservation scheme. That is, a certain amount of *excess bandwidth* needs to be reserved for the retransmission traffic at the base station in addition to the *original (or normal) bandwidth* required for the video stream.

Most video programs in VoD services are pre-recorded, information related to the original traffic becomes available, as soon as the request for a video program is granted and a virtual connection is setup. Therefore, it is reasonable to assume that this information is available at the base station. If we let $X(i)$ denote the original bandwidth for the i th video frame, and $\hat{x}(i)$ and $x(i)$ denote the predicted excess bandwidth and the excess bandwidth actually needed respectively, then the total bandwidth reserved ($\hat{X}(i)$) and actually used ($\bar{X}(i)$) for the i th frame, are respectively

$$\hat{X}(i) = X(i) + \hat{x}(i), \quad \bar{X}(i) = X(i) + x(i).$$

Hence, the residual of the reserved bandwidth for the frame, $e(i)$, is

$$e(i) = \bar{X}(i) - \hat{X}(i) = x(i) - \hat{x}(i).$$

It indicates that the issue of bandwidth reservation for pre-recorded video traffic is equivalent to the excess bandwidth reservation or prediction. In the remainder of this section, different prediction algorithms are proposed for excess bandwidth reservation when a type-I hybrid ARQ scheme is applied.

2.2. Fixed reservation

The most straight forward approach is to reserve the excess bandwidth for each frame at a fixed rate, which is estimated based on the average retransmission ratio over a relatively long range. If p_r denotes the average retransmission probability, and L is the retransmission limit, then the average retransmission ratio (\hat{P}_r) and the reservation rate for the excess bandwidth for each frame, b , can be estimated as

$$\hat{P}_r = \sum_{j=1}^L p_r^j, \quad b = \hat{x}(i) = \alpha \times \hat{P}_r \times \frac{1}{N} \times \sum_{j=1}^N X(j),$$

where α is a scaling factor close to 1 which is used for fine tuning, and N is the number of frames over which we are averaging.

2.3. Static reservation

Intuitively, the excess bandwidth needed for each frame is proportional to the original bandwidth of the frame. Therefore, in the static reservation algorithm,

$$\hat{P}_r = \sum_{j=1}^L p_r^j, \quad \hat{x}(i) = \alpha \times \hat{P}_r \times X(i).$$

2.4. Dynamic reservation

In the previous reservation algorithms, the retransmission ratio \hat{P}_r is calculated by using the average performance of the underlying wireless channel. To take the variation of the channel performance into consideration, in the dynamic reservation algorithm, the retransmission ratio for the i th frame $\hat{P}_r(i)$ is predicted dynamically based on the channel performance for the preceding d frames instead. That is, if $p_r(i, d)$ represents the average retransmission probability for the d frames preceding the i th frame, then

$$\hat{P}_r(i) = \sum_{j=1}^L p_r(i, d)^j, \quad \hat{x}(i) = \alpha \times \hat{P}_r(i) \times X(i).$$

Intuitively, a smaller value for d will give a more accurate prediction.

2.5. Adaptive reservation

In the dynamic reservation algorithm, $p_r(i)$ is calculated as an average of the retransmission probability for the previous d frames. If instead, an adaptive algorithm with a d -order linear predictor is applied, in which the retransmission ratio for the current frame is estimated as a linear combination of the previous d values, then,

$$\hat{P}_r(i) = \sum_{k=1}^d w(k) P_r(i-k), \quad \hat{x}(i) = \hat{P}_r(i) \times X(i),$$

where $w(k)$, for $k = 1, 2, \dots, d$, are the linear predictor filter coefficients (or weights), and $P_r(i-k)$ is the actual retransmission ratio for the frame that is k frames before the i -th frame.

The predictor coefficients $w(k)$ are adjusted adaptively based on the reservation residuals. The reservation residual is given as,

$$e(i) = x(i) - \hat{x}(i) = X(i) [P_r(i) - \sum_{k=1}^d w(k) P_r(i-k)],$$

where $P_r(i)$ is the actual retransmission ratio for frame i . Since the residual $e(i)$ is proportional to $X(i)$, the relative residual $\tilde{e}(i)$ is a more convenient feedback for adjusting $w(k)$, i.e.,

$$\tilde{e}(i) = \frac{e(i)}{X(i)} = P_r(i) - \sum_{k=1}^d w(k) P_r(i-k).$$

Thus, the square error ϵ of $\tilde{e}(i)$ and its gradient $\nabla\epsilon$ are

$$\epsilon = E\{\tilde{e}(i)^2\}, \quad \nabla\epsilon = \nabla E\{\tilde{e}(i)^2\} = -2E\{\tilde{e}(i)\mathbf{P}_r\},$$

where $\mathbf{P}_r = [P_r(i-1), P_r(i-2), \dots, P_r(i-d)]$.

The adaptive algorithm can be briefly summarized as¹¹

- Start with an initial array of predictor coefficients, $w(k)$, $k = 1, 2, \dots, d$.
- For each frame, compute $\nabla\epsilon$. In practice, $\nabla\epsilon$ can be estimated as $-2\tilde{e}(i)\mathbf{P}_r$.
- Update \mathbf{w} by $\mathbf{w} = \mathbf{w} - 0.5\mu\nabla\epsilon = \mathbf{w} + \mu\tilde{e}(i)\mathbf{P}_r$, where μ is the step size which determines the speed of convergence, and \mathbf{w} is the vector of the predictor coefficients $[w(1), w(2), \dots, w(d)]$. The choice of μ is a trade off between convergence speed and fluctuation. A larger μ results in a faster convergence, but causes a larger fluctuation.
- In order to reduce the sensitivity of the convergence on μ , a normalized update equation can be used, $\mathbf{w} = \mathbf{w} + \frac{\mu\tilde{e}(i)\mathbf{P}_r}{\|\mathbf{P}_r\|^2}$, where $\|\mathbf{P}_r\|^2 = \mathbf{P}_r\mathbf{P}_r^T$.

2.6. Best-effort reservation

Once the excess bandwidth for a frame is predicted, for example, by any of the algorithms discussed above, a best-effort (or enhanced) reservation algorithm may be achieved by using the knowledge of the previous residuals or the current status of the queue. If the residual of the excess bandwidth for the previous frame $e(i-1)$ is used, then the reservation for the i th frame can be chosen as

$$\hat{x}(i) = \begin{cases} \hat{x}(i) + e(i-1), & \text{if } e(i-1) > 0 \\ \hat{x}(i), & \text{if } e(i-1) \leq 0 \end{cases}$$

If $e(i-1)$ is positive, i.e., the reserved value for the previous frame was less than actually required (under-estimation). Then, the buffered residual that was not transmitted will be scheduled in the following frame interval for transmission in addition to the predicted excess traffic for the next frame.

3. PERFORMANCE EVALUATION

3.1. Simulation system

To investigate the performance of different bandwidth reservation algorithms when a type-I hybrid ARQ scheme is applied to transmit MPEG-2 video over B-FWAnS, the fixed wireless channel was modeled as the combination of an ad hoc finite state Markov model⁸ as shown in Figure 2 and the K-factor Rician fading model.¹³ Basically, the K-factor Rician fading model captures the short-term variation of the channel performance, while the Markov

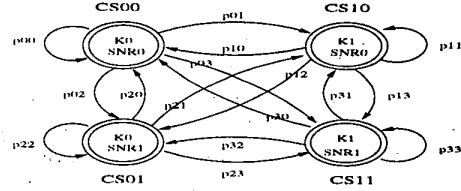


Figure 2. A finite state Markov model of fixed wireless channels.

model emulates the long-term transition of the wireless channel states.

For fixed wireless channels as those in an LMDS system, a line-of-sight (LOS) propagation path from the transmitter antenna to the receiver antenna is usually required. Therefore, the received signal consists of the waves coming directly from the transmitter along the LOS path (the direct component, E_d) and a number of small waves scattered from adjacent houses, leaves or other scattering objects (the scattered component, E_s). Therefore, the amplitude of the received signal follows the Nakagami-Rician distribution,¹⁴ (i.e., Rician fading). When using the K-factor Rician fading model, the channel state is described by a tuple (K, SNR) , in which $K = \frac{E_d}{E_s}$, and SNR is the signal to noise ratio at the receiver.

The combined Markov model simulates a TDMA LMDS fixed wireless channel in the 28GHz frequency band. The modulation scheme used is $\pi/4$ -shifted DQPSK. The BER of the resultant 4 states are 2×10^{-2} , 4×10^{-3} , 10^{-4} , and 10^{-11} respectively. To study the performance of different reservation algorithms, the average channel BER is chosen to be approximately 3×10^{-3} . The retransmission limit is chosen to be 3. A clip (about 3001 frames) of the well-known action movie *Indiana Jones*, which contains many car chases and frequent scene changes, was used in the simulations.

3.2. Summary of simulation results

Table 1 summarizes the simulation results on the excess bandwidth reserved when using the different algorithms discussed above. The performance results, including residual pdf (rsd pdf), queue length (ql), and wasted channel bandwidth (wb) due to underflow, are then summarized in Table 2. For now, we assume that no cells are ever dropped by the base station, i.e., the queue has an infinite capacity.

Some studies have shown that VBR traffic is correlated.¹⁵⁻¹⁷ Figure 3 shows the auto-correlation functions (ACFs) of the normal traffic and the excess traffic of the MPEG-2 encoded VBR video clip, for I-frames only (upper) and at the aggregation of the Group-of-Picture (GoP) level (lower). Figure 3 confirms that the MPEG-2

Table 1. Excess bandwidth Reserved.

Algorithm	excess bandwidth reserved*	
	total (Mb)	mean (Kb)
fix($\alpha=1$)	35.02	11.67
sta($\alpha=1$)	35.05	11.68
dyn($d=2, \alpha=1$)	35.01	11.67
adp($d=2, \mu=0.5$)	36.04	12.01
adp($d=2, \mu=1$)	35.71	11.90
dyn _{best-effort}	37.81	12.60

*: the total normal bandwidth required is 482.05Mbits.

Table 2. A comparison of the performance for the different excess bandwidth reservation algorithms.

Algorithm	rsd pdf (cell)		ql (Kb)		wb (Kb)
	mean	std	mean	std	mean
fix($\alpha=1$)	0.000	12.2	2059	863	0.01
sta($\alpha=1$)	-0.002	5.3	56	32	0.02
dyn($d=2, \alpha=1$)	-0.005	6.6	12	4.4	0.01
adp($d=2, \mu=0.5$)	-0.78	8.2	5.8	4.5	0.35
adp($d=2, \mu=1$)	-0.54	9.6	6.0	4.2	0.24
dyn _{best-effort}	-2.1	7.9	1.3	1.8	0.94

encoded VBR video sequence is correlated and has a slow-decaying tail. We also observe that the excess bandwidth exhibit a similar slow-decaying auto-correlation, because the excess bandwidth is proportional to the original bandwidth statistically.

Studies have also shown that serving a highly correlated VBR traffic with a slow-decaying tail at a fixed rate which is not close to the peak value, can cause large queue build-ups and thus result in large delays¹⁸⁻²⁰ and potential cell losses due to finite buffer capability in practice. The simulation results as shown in Table 2 confirm that reserving bandwidth at a fixed rate close to the average for the highly-correlated excess traffic results in huge traffic build-ups at the queue due to the accumulation of the reservation residuals.

Figure 4 shows the ACFs of the reservation residuals for the I-frames only (upper) and the aggregation at GoP level (lower), when the different reservation algorithms were applied. The ACFs indicate that the prediction errors are not auto-correlated at either the I-frame level or at the GoP aggregation level when the dynamic or the adaptive algorithms are used. In addition, the distribution of the pdfs of the residuals of the dynamic and the adaptive algorithms indicate that the residuals are Gaus-

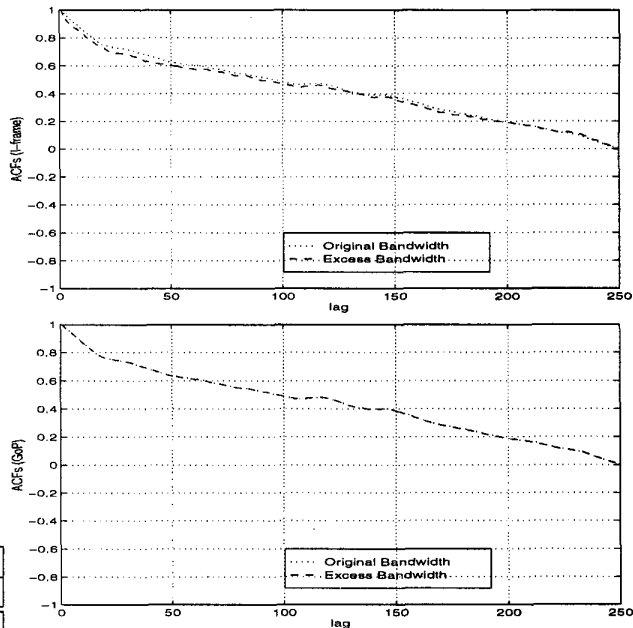


Figure 3. ACFs of normal traffic and the excess traffic of the video clip from *Indiana Jones*: (upper) I-frames only, (lower) at GoP level of aggregation.

sian distributed with a mean close to zero.

Statistically, the excess bandwidth is proportional to the original bandwidth, and it is a function of the parameters of the type-I hybrid ARQ scheme including the retransmission limits and error correction capability. In addition, the excess bandwidth is a stochastic process, due to the variations in the performance of the underlying wireless channel performance. Hence, we can observe from Table 2 that, by reserving the excess bandwidth based on non-fixed algorithms, a smaller queue capacity is needed at the base station, i.e., shorter delays are experienced by the packets, in particular, we notice this for the dynamic and the adaptive algorithms, when compared to the fixed reservation algorithm

For the dynamic reservation algorithm, the average retransmission probability of the preceding frames is used. While in the adaptive reservation algorithm, the retransmission ratio is updated based on the d -order linear predictor, which is a linear combination of the actual retransmission ratio of the preceding d frames. The updates of the coefficients of the linear predictor depend on the relative residuals. Therefore, the adaptive algorithm is more efficient in adjusting the prediction parameters to reflect the fluctuation of the channel conditions. Therefore, the average queue lengths for the adaptive schemes are approximately half of that needed by the dynamic scheme.

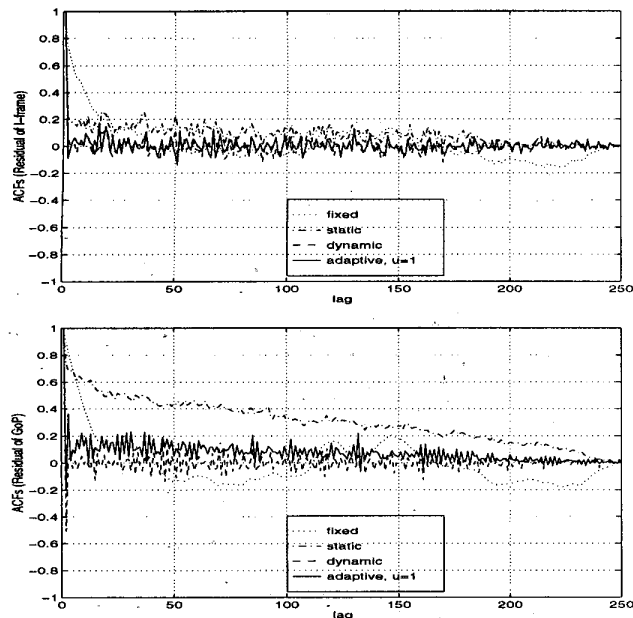


Figure 4. ACFs of the reservation residual of the video clip from *Indiana Jones* when using different prediction algorithms: (upper) I-frames only, (lower) at GoP level of aggregation.

The requirement on the queue capacity is further reduced when the best-effort dynamic algorithm is applied. As shown in Table 2, the mean of the pdf for the best-effort algorithm has a negative value, indicating that by serving an additional bandwidth which is equal to the residual of the previous frame, the probability of over-reservation is larger than that of under-reservation. Therefore, the maximum and the average queue length are reduced dramatically.

However, the downside to making over-reservations is that more wireless channel resource is going to be wasted due to underflows, which reduces the overall channel utilization. In addition, the cost for the reduction in the queue capacity consists of a slight increase in the excess bandwidth reserved by the adaptive and the best-effort dynamic algorithms,

In general, the choice of the proper non-stationary bandwidth reservation algorithm comes down to a trade-off among the availability of the channel resource, the requirement on the queue length (or the queuing delay) at the base station, and the potential wasted channel resource due to over-estimation and underflow. Overall, the dynamic or the adaptive based algorithms work effectively when properly applied as indicated by the simulation results.

REFERENCES

1. B. Khasnabish, "Broadband to the home (BTTH): Architectures, access methods, and the appetite for it," *IEEE Network Magazine*, pp. 58-69, January 1997.
2. W. Honcharenko, J. P. Kruys, D. Y. Lee, and N. J. Shah, "Broadband wireless access," *IEEE Communication Magazine*, pp. 20-26, January 1997.
3. L. Lugand, D. Costello, and R. Deng, "Parity retransmission hybrid ARQ using rate 1/2 convolutional codes on a non-stationary channel," *IEEE Transactions on Communications* 37(4), pp. 755-765, 1989.
4. M. Nakamura and T. Kodama, "Performance evaluation for ARQ schemes in power and/or bandwidth limited systems," *Transactions of IEICE* E72(5), pp. 491-501, 1989.
5. Y. Wang and S. Lin, "A modified selective-repeat type-II hybrid ARQ system and its performance analysis," *IEEE Transactions on Communications* 31(5), pp. 124-133, 1983.
6. S. B. Wicker, "Reed-solomon error control coding for Rayleigh fading channels with feedback," *IEEE Transactions on Vehicular Technology* 41(2), pp. 124-133, 1992.
7. J. B. Cain, G. C. Clark, and K. M. Geist, "Punctured convolutional codes of rate (n-1)/n and simplified maximum likelihood decoding," *IEEE Transactions on Information Theory* IT-25, pp. 97-100, January 1979.
8. H. Ma and M. E. Zarki, "MPEG-2 encoded VoD services over fixed wireless channels using ARQ schemes," *SPIE's symposium on voice, video and data communications*, Boston, 1998.
9. H. Ma and M. E. Zarki, "MPEG-2 video transmission over wireless access networks using type-I hybrid ARQ schemes for VoD services," in the *Proceedings of the ninth International Packet Video Workshop*, Columbia University, New York City, April, 1999.
10. H. Ma and M. E. Zarki, "Transmission MPEG-2 encoded VoD services over wireless access networks using type-II hybrid ARQ schemes with rpsc codes," to appear in *New Trend in Multimedia Systems, Maui, Hawaii, January 2000*.
11. S. Haykin, *Adaptive filter theory*, Prentice Hall, 1991.
12. P. Pancha and M. E. Zarki, "Bandwidth requirements of variable bit rate MPEG sources in ATM networks," in the *Proceedings of IEEE INFOCOM93*, pp. 902-909, San Francisco, March 1993.
13. P. Papazian, G. A. Hufford, R. J. Achatz, and R. Hoffman, "Study of the local multipoint distribution service radio channel," *IEEE Transactions on Broadcasting* 43, June 1997.
14. M. D. Yacoub, *Foundations of Mobile Radio Engineering*, CRC Press Inc., Boca Raton, FL., 1993.
15. J. Beran, R. Sherman, M. Taqqu, and W. Willinger, "Variable-bit-rate video traffic and long-range dependence," *IEEE Transaction on Networking*, 1993.
16. M. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," in the *Proceedings of ACM Sigcomm94*, pp. 269-280, London, 1994.
17. P. Pancha and M. E. Zarki, "Variable bit rate video transmission," *IEEE Communication Magazine* 32(5), pp. 54-66, May, 1994.
18. A. Adas and A. Mukherjee, "On resource management and QoS guarantees for long range dependent traffic," in the *Proceedings of IEEE INFOCOM95*, pp. 779-787, Boston, April 1995.
19. M. Livny, B. Melamed, and A. K. Tsolis, "The impact of auto-correlation on queuing systems," *Management Science*, pp. 322-339, March 1993.
20. P. Pancha and M. E. Zarki, "Leaky bucket access control for VBR MPEG video," in the *Proceedings of IEEE INFOCOM95*, pp. 796-803, Boston, April 1995.