# ANN Based Forecasting of VBR Video Traffic for Dynamic Bandwidth Allocation in ATM Networks

*Song Chong, San-qi Li and Joydeep Ghosh*
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, Texas 78712

April 1, 1995

**Abstract**

Two time delay neural network (TDNN) based forecasting systems are proposed to perform dynamic bandwidth reservation for real-time, variable bit rate (VBR) video service in ATM networks. Both multilayered perceptron (MLP) and pi-sigma network (PSN) based systems are found to give highly reliable predictions even in a nonstationary environment. Their performance is quantified through simulation experiments on clippings from the movie, "Star Wars". In particular, the PSN-TDNN based dynamic bandwidth reservation achieves high network utilization, loss-free transmission and reasonable delay, and its lower computational requirements make it suitable for on-line operation.

**keywords:** ATM network, real-time VBR video, dynamic bandwidth allocation, time delay neural network, pi-sigma network, forecasting.

# 1 Introduction

Future broadband integrated services digital networks (B-ISDN) will support a wide range of services including voice, data, video and other modalities in an integrated and unified fashion. Asynchronous transfer mode (ATM) technology has evolved as a viable candidate for implementation of B-ISDN. An ATM network is a high-speed packet-switched network in which transmission is conducted over optical fibers. All information is segmented into a series of packets called "cells". Each cell consists of 44 bytes of data and 9 bytes of control information. Cells coming from several sources are statistically multiplexed through a buffer called "statistical multiplexer" and share common switching and transmission resources. Both packetization of information and statistical multiplexing of cells make ATM network inherently flexible and service independent; a wide range of bit rate and variable bit rate communications can be supported in an unified fashion.

B-ISDN is required to meet diverse service and performance requirements of multimedia traffic. Real-time voice, for example, requires rapid transfer, but the loss of small amounts of voice information is tolerable. In many data applications, real-time delivery is not of primary importance, but high throughput and strict error control are required. Real-time video communications, which is of primary concern in this paper, require loss-free transmission as well as rapid transfer. In video services, the bandwidth flexibility offered by the ATM technology makes it possible to use variable bit rate (VBR) video coding algorithms that offer consistent picture quality. On the other hand, unlike in circuit-switched networks, packetization of coded information and statistical multiplexing drive cell loss and delay at network buffers and consequent picture quality degradation. Managing the available transmission bandwidth to avoid congestion and provide guaranteed levels of GOS (Grade of Service) for connections is one of the most important tasks in ATM network research.

The link capacity allocation problem determines the minimum transmission bandwidth of a link to support the connections on the link with guaranteed levels of GOS. The central objective of the allocation is two-fold: to take advantage of statistical multiplexing of connections for transmission efficiency, and to prevent nodal congestion caused by bursty arrival of traffic. As in circuit-switched networks, the most conservative allocation scheme is to remove the statistical multiplexing by allocating a fixed link capacity to each connection according to its peak rate during the whole connection. This is certainly not cost-effective. On the other hand, on ATM there has been a great concern about potential gains in transmission efficiency through the statistical regularity that results when connections are combined through multiplexing. Many researchers have been

tried to characterize the bursty behavior of traffic with a set of statistical metrics called "traffic descriptors", such as peak rate, mean rate and average duration of a burst period of traffic. By using these metric as descriptors of connections, they have developed approximate algorithms for link capacity allocation to provide a certain GOS on a link where many connections are multiplexed. In reality, since much of multimedia traffic will be very high-volume (in connection time, range of hours and in bit rate, several hundred megabits per second) and highly bursty, the potential gains in transmission efficiency through statistical multiplexing via finite buffering should not be overestimated and the simple statistical traffic descriptors are less likely to be valid. In order to overcome these difficulties, new directions for link capacity allocation have been indicated in [2]. Instead of using predetermined static traffic descriptors during the entire connection and expecting a big role of statistical multiplexing in absorbing traffic fluctuation, the current traffic flow is measured in real-time and link capacity is dynamically allocated based on this measurement. This approach is strongly motivated by stringent performance requirements of multimedia service because the only way to guarantee negligible loss and delay in service transfer may be based on real-time traffic measurement and resource management. Especially, for real-time video service which requires loss-free transmission as well as rapid transfer, this approach may have significant impact.

In this paper we address a dynamic bandwidth allocation mechanism for real-time VBR video services with guaranteed GOS, which combines the new concepts of real-time traffic measurement and bandwidth management. This mechanism achieves high network utilization, loss-free transmission and rapid transfer at the same time. The idea of this scheme is as follows: Bandwidth is reserved based on current variations of activity from scene to scene, and buffering is used to absorb the other high frequency variations of activities from frame to frame, from line to line and from pixel to pixel. In other words, after separating the signal into high and low frequency components by appropriate low-pass filtering, the lower frequency component is transmitted like in "transparent pipe" by assigning the corresponding bandwidth. Precise forecasting of traffic is crucial in dynamic link capacity allocation. Since in practice the link capacity reservation of each trunk will be updated at every fixed time interval, the traffic behavior during the next time interval should be estimated. Certainly, the larger the update interval, the larger the forecasting error. The importance of and difficulties in accurate and robust forecasting of VBR video are summarized as follows:

- Video and video-multiplexed traffic will dominate the performance of entire network because of their high volume; the accuracy of forecasting such traffic is crucial to avoid possible nodal

congestion and information loss due to buffer overflow.

- In terms of feasibility of forecasting, there is a contradiction. While video traffic is predictable in the sense that it possesses strong correlation from scene to scene, from frame to frame and so on, it can also said to be unpredictable because it is in nature rapidly time-varying and highly bursty.

- On the other hand, in practice forecasting with longer lead-time is desirable for higher layer implementation of bandwidth reservation; so the key question faced is that in what time scale can the prediction operation meet both prediction accuracy and implementation requirements from higher layer at the same time.

In this paper, we extensively study the issues on forecasting feasibility of VBR video and its implementations based on several methods. As a testbed of our study, the DCT coded VBR full-motion movie "Star Wars" is selected [1]. Such real-time full-motion VBR video may be one of the most popular and performance-crucial service in an ATM environment. Techniques for different forecasting problems have been developed in the several areas, and can be mostly classified in two general categories. One approach treats the values to be forecast as a time series, and predicts the future using time series analysis techniques such as Kalman filtering, ARMA models and spectral expansion techniques. In general, time series approaches can be effective unless there is abrupt changes in the variables affecting the signal of interest. However they are computationally intensive and are prone to numerical instabilities. The other approach expresses the signal of interest as a function in a predetermined form, of other variables. Techniques such as regression are then used to estimate the function parameters. Such techniques require good apriori knowledge of the functional form, and cannot easily handle situations where the functional relationship is time varying.

Artificial neural networks (ANNs) offer a promising alternative for the automatic forecasting problem as they can combine the positive features of both general approaches mentioned above [6]-[12]. They are particularly suitable when a simple and accurace mathematical characterization of the signal of interest is not forthcoming, and when the functional relationship is time-varying, thus demanding adaptive techniques for accurate forecasting. In this paper, two different ANN based forecasting systems are proposed. One is a time delay neural network (TDNN) based on the multilayered perceptron (MLP) architecture and the other is a TDNN based on pi-sigma (PSN) architecture [9],[12]. When the problem is not well characterized by *a priori* mathematical model or statistics, a powerful enough ANN family of networks often out-performs more conventional

parametric methods. ANNs have adaptation ability which can accommodate nonstationarity/time-varying characteristic. ANNs have generalization capability which make them flexible and robust when faced with new and/or noisy data patterns. Once the training is completed, an ANN can be computationally inexpensive, even if it continues to adapt on-line.

Recently, a very computationally efficient high-order network has been developed [11], that can capture underlying high-order correlation of input components while avoiding an exponentially increasing computational and memory cost that affects ordinary higher-order nets [12]. This architecture, called the Pi-Sigma Network (PSN) is the second basis for our forecasting system. For comparison purposes, one of the most sophisticated prediction algorithms in the field of adaptive filtering, called the recursive least square method (RLS), is utilized [14]-[16]. Both MLP-TDNN and PSN-TDNN give very reliable prediction even in testing sets collected from different statistics. Although the performances of both networks in terms of accuracy and robustness are almost the same, PSN-TDNN is much more computationally inexpensive thus is preferable in practice. With around 0.6 second of prediction lead-time, we can meet both prediction accuracy and demands from higher layers of the ATM network that ask for longer lead-time in order to make more global decisions. The role of a "maximum unit" in the forecasting system is found to be important in the sense that it can compensate the prediction risk caused by variance and underestimation. Previously, ANNs have been applied to the call admission problem in ATM [4][5], but to the authors' best knowledge, no application of ANNs to traffic prediction in ATM network have been reported so far.

The paper is organized as follows. In Section 2 we describe the dynamic bandwidth allocation scheme and contrast it with static schemes. MLP-TDNN, PSN-TDNN and RLS forecasting systems are introduced in Section 3. The performance of suggested ANN based forecasting systems and their effectiveness on dynamic bandwidth allocation problem is evaluated in Section 4. The forecasting systems are applied in Section 5 to a simple example network. The paper is concluded in Section 6.

## 2 Dynamic Bandwidth Allocation

Static bandwidth allocation assigns some predetermined value to the required bandwidth. For example, the most conservative static allocation will reserve bandwidth using peak rate of traffic. It guarantees no loss but is very inefficient in the sense that it wastes bandwidth. Another static

allocation method is to reserve bandwidth with a statistically determined "effective bandwidth" [3]. This effective capacity is determined somewhere between mean rate and peak rate of traffic, and is aimed at expecting the gain of statistical multiplexing. However, it is also considered to be inefficient. Furthermore, it takes risk of cell loss in the case of highly varing traffic. A third method called filtered peak [2] reserves bandwidth with peak rate of variation of activity from scene to scene.

The concept of dynamic bandwidth allocation is simple and clear [2]. In recent works in queueing theory, it is found that the low frequency component of the traffic, i.e., the component with slow time variation, dominates queueing performance. Therefore, in VBR video, the variation of activity from scene to scene will dominate queueing performance compared to the other activities such as frame by frame, line by line and pixel by pixel. Besides, this low frequency variation from scene to scene is hardly absorbed through buffering so that it should be taken care of by assigning the bandwidth accordingly. Let us look at the power spectrum of the 2-minute video segment in Fig. 1a. The lowest frequency power represents activity from scene to scene. This power will be taken care of by bandwidth reservation while the higher frequency powers can be absorbed through buffering. The implementation of this concept is shown schematically in Fig. 1b where the bandwidth reservation is controlled by the low-pass filtered component $\check{x}(t)$. All the bandwidth allocation schemes mentioned here are summarized in Fig. 2. In Fig. 3, page 56 (2-minute long) of "Star Wars"(a), and its filtered signal with cutoff frequency $\omega_c$=6.3 [rad](b) are shown. As we can observe in Fig. 3, the variation from scene to scene is essentially within 1.0 second. This implies that if we do prediction with lead-time larger than 1.0 second, we will inevitably lose the accuracy. In our study, it is found that lead-times of up to 0.6 second can yield highly accurate prediction. In this paper, we set lead-time to 0.56 seconds. Another comment here is that this 0.56-second based bandwidth reservation can be well supported by higher layer of networks. Usually, in existing data networks such as ARPANET, the routing operation is performed on about a 0.6 second basis.

# 3 ANN Based Forecasting Systems

A. TDNN based forecasting systems

Define $\Delta$ to be the 0.56-second time slot. In practice, in order to perform dynamic bandwidth reservation every $\Delta$ seconds, we need to predict the traffic behavior during the next $\Delta$ in advance and then reserve bandwidth according to the peak rate in this next time period. How can we get

an accurate prediction of peak value during the next $\Delta$? We propose a configuration of ANN based forecasting system as in Fig. 4. The ANN is designed by a TDNN consisting of 14 inputs. The network has 4 outputs representing four different and equally distributed prediction points within next $\Delta$ respectively. At a given time $t$, 14 inputs are collected from the time interval [t-3.5$\Delta$, t]. The reason we predict four points within next $\Delta$ is that our objective is not to predict value at predetermined future point but to find peak value of next $\Delta$ period. In this sense, our problem is unlike ordinary regression type prediction problems. By increasing the number of prediction points within $\Delta$ (here 4 points), we can search for the peak value with higher resolution. The maximum unit selects the maximum value among the 4 TDNN outputs. The quantizer performs discretization of analog output of the max. unit because in practice bandwidth reservation will be performed in units of of cells/slice. An important thing we note here is that the cooperation between multiple outputs of TDNN and the following max unit can compensate for the effect of estimation variance and underestimation, so that we can get very reliable bandwidth reservation. In bandwidth reservation, underestimation is a severe problem but overestimation is acceptable. This is another interesting feature of our prediction problem.

As shown in Fig. 5, the TDNNs used are implemented using two different neural network architectures, with an MLP (Fig. 5a) and a PSN (Fig. 5b) being respectively used as the feedforward network. Prime motivation for using a PSN is to employ its capability to capture high-order correlation of inputs and its efficient computation. For example, in the case of k-th order PSN with N inputs and M outputs using asynchronous learning rule, the number of multiplications needed is (N+1)$MK^2$+NMK. In an MLP with N inputs, H hidden units and M outputs, number of multiplications is (5N+6M)H. Detailed learning algorithms and implementations can be found in [9][10][12].

B. RLS based forecasting system

For comparison purposes, we also construct an on-line adaptive transversal filter using recursive least squares[13]-[16]. The RLS algorithm may be viewed as the deterministic counterpart of Kalman filer theory. The rate of convergence is shown to be typically an order of magnitude faster than that of LMS (least mean square) algorithm. Unlike the LMS algorithm, the rate of convergence of the RLS algorithm is essentially insensitive to variations in the eigenvalue spread of the correlation matrix of the input vector. By setting the forgetting factor to less than unity i.e. by giving the algorithm only a finite memory, RLS attains the capability to track slow statistical

variations. The detailed RLS algorithm can be found in [13]-[16]. An important thing we should note here is that RLS algorithm is based on a linear model, it has transient behavior and it requires much more computation during operation. Considering these, TDNN based forecasting will be the better choice if they are able to provide comparable accuracy. Since in our prediction problem we need four predictions within $\Delta$, the so called indirect approach is used in which underlying RLS algorithm performs parameter estimation and then based on this estimate four consecutive predictions are performed.

# 4 Performance of TDNN Forecasting and Dynamic Bandwidth Allocation

A. Training and testing

For training, we select a 2-minute segment (page 56) of the movie, as shown in Fig. 3. Training is carried on the filtered signal in Fig. 3(a). We use 40 hidden units in the case of MLP and 2 hidden units (i.e. we make a second order approximation) in the case of PSN. As mentioned in the previous section, 14 inputs are sampled during the time interval from [t-3.5 $\Delta$, t], and the TDNN then performs predictions of 4 points within next $\Delta$ seconds. We scan the 2-minute segment along the time axis to generate 207 examples. After adding a 0.28-second offset, we scan again the same 2-minute segment. This rescaning will generate another 207 examples. By repeating this procedure seven times, 1449 (207x7) examples are prepared. In training, the most important thing is normalization of data. Due to the use of the sigmoid function by the output units, it is recommended that the dynamic range of desired output be normalized to [0,1]. In the case of PSN, asynchronous update rule is employed because randomized updating rule cannot be applied when momentum is used. The learning was carried out on a SPARC-10. The number of epoch is set to 5,000. The training time was 5 hours and 49 minutes for MLP and 1 hour and 7 minutes for PSN including all I/O executions in our code. Roughly speaking, in this problem we observe 5 times efficient computation in the case of PSN-TDNN. The training results for both MLP and PSN are given in Fig. 6. The plot in Fig. 6a shows the prediction output from MLP-TDNN. Fig. 6b shows the output from the maximum unit in the case of MLP-TDNN. Fig. 6c shows the prediction output from PSN-TDNN. Fig. 6d shows the output from the maximum unit in the case of PSN-TDNN. As we can observe in Fig. 6a and Fig. 6c, MLP-TDNN has a somewhat

smaller estimation variance. However, after passing through the max unit, the effect of difference in estimation variance is basically removed (See Fig. 6b and Fig. 6d).

Six 2-minute segments of the movie i.e. page 39-41 (6 minutes) and page 55-57 (6 minutes) are selected as testing sets. Note that the training set was page 56 (2 minutes). Page 39-41 was selected to represent different statistics. Fig. 7 shows the filtered video traffic of testing sets. In order to demonstrate the statistical difference between p56 (training set) and the testing set pages, quantile-quantile plots (Q-Q plots) are given in Fig. 8. If the Q-Q plot represents a straight line in Fig. 8, it implies that the probability density function (PDF) of the segment is same as that of page 56. As we can see, pages 39-41 have much different PDF from the training set as compared to pages 55 and 57. In Figs. 9 and 10, the test results of page 40 and 41 are shown for PSN-TDNN. As evidenced by Fig. 9c and 10c, the reservation covers original traffic very well and it is difficult to find underestimated portions. In fact, about the same performance was achieved for all the testing sets for both MLP-TDNN and PSN-TDNN, and we have just shown Figs. 9 and 10 for demonstration purposes.

B. Performance of dynamic bandwidth allocation based on ANN forecasting

Here, we evaluate the performance of dynamic bandwidth allocation scheme based on PSN-TDNN forecasting. This evaluation was made through cell-level queueing simulations with buffer size of 250 cells. The testing was carried on page 39-41 and 55-57. The results are summarized in Table 1. In the case of static allocation based on peak rate, we can achieve 46% network utilization with an average queue length of 5.18 cells. The ideal dynamic allocation achieves 80% utilization with only 6 more cells queued on average. Thus for both MLP and PSN we can acheive performance comparable to the ideal case. The small degradation of utilization as compared to ideal case arises because, as seen in Figs. 9 and 10, both PSN and MLP forecasting systems over-reserve the bandwidth a little. In the case of RLS forecasting system, the performance is also pretty good. Therefore, in this specific study, we could not demonstrate the superiority of ANN based forecasting systems, except in terms of computational efficiency. Since both ANN based and RLS based approaches performed prediction so well, it was difficult to distinguish their performances. However, if we increase prediction lead-time it is expected that ANN based prediction will have better performance because RLS will involve tracking delay.

## 5 Application to an ATM Network Multiplexer

In this section we demonstrate an application of the dynamic bandwidth reservation scheme based on TDNN-PSN. Consider an ATM network multiplexer shown in Fig. 11. Five 2-minute video segments are fed to the system, and the network manager divides the given link capacity based on the following rule:

$$\mu_j(t) = \frac{\check{x}_j(t)\mu}{\sum_{i=1}^{5} \check{x}_i(t)},$$

where $\check{x}_j(t)$ is the filtered traffic of $x_j(t)$.

The results obtained with $\mu = 158.0$ cells per unit time, are summarized in Table 2. The access to network by the five sources is seen to be quite fair.

## 6 Conclusions

Two TDNN based forecasting systems are proposed for performing dynamic bandwidth reservation for real-time video service in ATM networks. In terms of prediction performance, both MLP and PSN give highly reliable predictions over the entire periods of the video test data taken from "Star Wars". Interestingly, even for test data that exhibited quite different statistics (e.g. different PDF and different autocorrelation) both TDNN forecasting systems works very well, thus indicating their generalization capability. In this context it should be noted that the use of multiple prediction points followed by a maximum unit plays an important role in overcoming the effect of prediction variance and underestimation.

When compared with on-line RLS forecasting systems, it was difficult to show significant superiority of TDNN based forecasting. However, when we increase the prediction lead-time, it is expected that TDNN based systems will have comparatively better performance as RLS based system will inevitably involve tracking delay. In addition, the RLS approach suffers from initial transient behavior, and is relatively more computationally expensive during operation. As we demonstrated through queueing simulations, the proposed dynamic bandwidth reservation scheme based on PSN-TDNN provides high utilization of network, loss-free transmission and reasonable delay. In terms of computation, PSN-TDNN appears to be the best choice. Finally, when we consider the nature of VBR video traffic and demand from higher ATM network layers for longer lead-times, lead-times of around 0.6-second will be a practical choice.

## References


[1] M. Garrett and M. Vetterli, "Congestion Control Strategies for Packet Video," presented in the*Fourth International Workshop on Packet Video*, Kyoto, Japan, Aug. 1991.

[2] S.Q. Li, S. Chong, C. Hwang and X. Zhao, "Link Capacity Allocation and Network Control by Filtered Input Rate in High Speed Networks," submitted to *IEEE Globecom '93*

[3] G. Guerin, H. Ahmadi and M. Naghshineh, "Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks," *IEEE J. Select. Areas in Communications,* Vol. SAC-9, pp.968-981, Sept. 1991.

[4] A. Hiramatsu, "ATM Communications Network Control by Neural Networks," *IEEE Trans. on Neural Networks*, Vol.1, No.1, pp.122-130, Mar. 1990.

[5] A. Hiramatsu, "Integration of ATM Call Admission Control and Link Capacity Control by Distributed Neural Networks," *IEEE J. Select. Areas in Communications*, Vol.9, No.7, pp.1131-1138, Sept. 1991.

[6] D. Park et al., "Electric Load Forecasting Using An Artificial Neural Network", *IEEE Trans. on Power Systems*, Vol.6, No.2, pp.442-449, May 1991.

[7] H. Yang, T. Akiyama and T. Sasaki, "A Neural Network Approach to the Identification of Real-time Origin-destination Flows from Traffic Counts", *Proc., International Conference on Artificial Intelligence Applications in Transportation Engineering '92*, pp.253-269, June 1992.

[8] E. Hartman and J. Keeler, "Predicting the Future: Advantages of Semi-local Units", unidentified literature.

[9] K. Hornik, M. Stinchcombe and H. White, "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks*, Vol.2, No.23, pp.359-366, 1989.

[10] J. Ghosh and K. Tumer, "Structural adaptation and Generalization in Neural Networks," Technical Report 92-14-89, The Computer and Vision Research Center, The University of Texas at Austin, Nov. 1992.

[11] J. Ghosh and Y. Shin, "Efficient High-order Neural Networks for Classification and Function Approximation," *Intl. Journal of Neural Systems*, Vol. 3, No. 4, 1992.

[12] C. Giles and T. Maxwell, "Learning, Invariance, and Generalization in a High-order Neural Network," *Applied Optics*, Vol.26, No.23, pp. 4972-4978, 1987.

[13] G.C. Goodwin and K.S. Sin, "*Adaptive Filtering, Prediction and Control*," Prentice-Hall, Englewood Cliffs, N.J., 1984.

[14] S. Haykin, "*Adaptive Filter Theory*," Prentice-Hall, Englewood Cliffs, N.J., 1986.

[15] R. Hastings-James and M. Sage, "Recursive Generalized-least-squares Procedure for Online Identification of Process Parameters," *Proc. IEE*, Vol.116, pp.2057-2062, 1969.

[16] E. Eleftheriou and D. Falconer, "Tracking Properties and Steady State Performance of RLS Adaptive Filter Algorithms," Report SCE-84-14, Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada, 1984.