ITC19/ Performance Challenges for Efficient Next Generation Networks 959-968
LIANG X.J. and XIN Z.H.(Editors)
V.B. IVERSEN and KUO G.S.(Editors)
Beijing University of Posts and Telecommunications Press

Performance of Prediction-Based Dynamic Bandwidth Provisioning

Wang Hao Huang Changcheng James Yan

Department of Systems and Computer Engineering, Carleton University
1125 Colonel By Drive, Ottawa, Ontario, Canada K1S 5B6
Email: nhwang@sce.carleton.ca

Department of Systems and Computer Engineering, Carleton University
1125 Colonel By Drive, Ottawa, Ontario, Canada K1S 5B6
Email: huang@sce.carleton.ca

Department of Systems and Computer Engineering, Carleton University
1125 Colonel By Drive, Ottawa, Ontario, Canada K1S 5B6
Email: jim.yan@sympatico.ca

Abstract: Future data networks are required to support different types of traffic with distinct QoS requirements. An attractive approach to meet this requirement is to use virtual networks to provide traffic segregation and resource partitioning in the network. Because of the bursty nature of network traffic, in order to reduce the cost of operating the virtual links of a virtual network, as well as to improve the efficiency of the underlying network, the bandwidth of the virtual links needs to be dynamically adjusted. In this paper, the performance of Autoregressive Moving Average (ARMA) and Fractional Autoregressive Integrated Moving Average (fARIMA) predictors in prediction-based dynamic bandwidth provisioning is studied. The analysis and simulation results show that the performance of prediction-based dynamic bandwidth provisioning depends not only on the accuracy of the predictor, but more importantly on the autocorrelation structure of the prediction error. This dependence has not been noticed in previous research. In addition, our results show that error compensation is a powerful technique to decorrelate the predictor error, and simple predictors, such as ARMA predictor, with error compensation are more suitable in prediction-based dynamic bandwidth provisioning themes than computationally complex fARIMA predictors.
Keywords: dynamic bandwidth provisioning, traffic prediction, ARMA predictor, fARIMA predictor, virtual network, overlay network, QoS.

1. INTRODUCTION

Future data networks are required to support different types of traffic with distinct QoS

This work was supported in part by NSERC Grant RGP 261769 - 2003 and by NSERC CRDPJ 282925 - 03.

requirements, such as VoIP, video conference, video-on-demand, online gaming and other emerging services. . One of the proposed solutions is to use overlay virtual network structure, which can be setup within a single network domain [3], or across multiple network domains [4]. Such an approach implies traffic segregation and resource partitioning in the network

Under the overlay virtual network structure, virtual links are purchased from the network operators to form a connected virtual network. The bandwidth capacity of the virtual link can be statically or dynamically set. However, because of the bursty nature of data traffic, statically partitioning the network resource will decrease the network efficiency. Also from the overlay virtual network operator's point of view, purchasing a virtual link with fixed bandwidth means paying money for the bandwidth that is not used most of the time. Therefore, to make the partitioned network more efficient and the virtual networks more profitable, dynamic bandwidth provisioning is needed [3] [4] [12].

In dynamic bandwidth provisioning, the traffic loads on the virtual links are monitored, and the link sizes, i.e. the bandwidth, are dynamically adjusted. At the time of the bandwidth resizing, only the information of the traffic load at the current and previous measurement times are available, but the bandwidth adjusting decision will affect the QoS of the traffic after the current time and until the time of the next resizing. Therefore, for the dynamic bandwidth allocation to work, the traffic load on the virtual link in the next time interval needs to be predicted based on the current and previous measurements history, and the bandwidth is adjusted according to this prediction. In prediction-based dynamic network resource partitioning, the size of the virtual links is adjusted before the actual traffic changes. In other words, if the network foresees a certain amount of traffic increase during the next measurement interval, it would be able to reserve that amount of additional bandwidth beforehand, therefore the QoS of the traffic, the efficiency of the network, as well as the profitability of the overlay virtual network will be improved.

The rest of the paper is organized as followed. Section 2 gives a briefly introduction to prediction and predictors. In Section 3, the effect of the autocorrelation of the prediction error on the system performance is analyzed. Simulation results are provided in Section4. Finally, Section 5 provides the conclusions.

2. TRAFFIC PREDICTION

2.1. Introduction

Suppose that the traffic load on a virtual link is monitored, and the mean of traffic data rate (bits/s) for every time interval $[t_1,t_2)$, $[t_2,t_3)$, $[t_3,t_4)$, … is calculated, where $t_{i+1}-t_i=T_c$ is the size of the observation window. In this way, we get a time series of arriving mean traffic rate $X_1, X_2, \ldots, X_t, \ldots$. In traditional queuing theory models, it is assumed that the arrivals are Poisson, which means the arrival process is memoryless and the X_i and X_j from the arriving traffic rate series are independent, therefore uncorrelated, if $i \neq j$. Under this assumption, the best prediction one can make about the future traffic, if we assume that the random process $\{X_t\}$ is stationary, is $\mathrm{E}\{X_t\}$. But in the study of real network traffic, both short-range dependence (SRD) and long-range dependence (LRD) are observed [8]. This correlation structure makes it possible to predict the future traffic based on the traffic history. [3] [5] [9] [10] [11]

If the task of the predictor is to forecast the value of X_{t+1} based on the values of $X_t, X_{t-1}, \ldots, X_1$, it is called a one-step predictor, which is the predictor studied in this paper. h-step predictors, with $h > 1$, do exist but are not discussed in the paper. For predictions of different time length into the future, the task can be achieved by changing the value of T_c.

Time series prediction has been a research area for a long time and has generated many prediction models, among which the Autoregressive Moving Average (ARMA) and Fractional Autoregressive Integrated Moving Average (fARIMA) models are the most frequently used.

2.2. ARMA Predictor

The general model for an ARMA(p , q) process is defined as

$$\phi(B)X(t) = \theta(B)\varepsilon_t \tag{1}$$

in which $\phi(B)$ and $\theta(B)$ are polynomials of B:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{2}$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \tag{3}$$

where B is the backward shift operator defined as $BX_t = X_{t-1}$, $B^2 X_t = X_{t-2}$, etc; ϕ_i and θ_i are real valued coefficients of the polynomials, and ε_t is a white noise process.

In prediction practice, ARMA(p , q) model with $q = 0$, also called AR(p) model, is often used, because it does not require to deduce ε_t from the history, although in general an ARMA(p , q) model requires fewer parameters than an AR(p) model. By using an AR(p) model, the predicted value of X_{t+1} is

$$\hat{X}_{t+1} = \phi_1 X_t + \phi_2 X_{t-1} + \cdots + \phi_p X_{t-p+1} \tag{4}$$

The ARMA model is simple yet powerful, and is widely used in predictions. However it can only model the SRD of the process. Since the early 90's, numerous studies have shown that most network traffic can be better modeled as self-similar, which has the LRD characteristic [8] [11]. For an SRD process, such as the ones depicted by ARMA models, the autocorrelation function decays exponentially. The LRD process, on the other hand, has an autocorrelation function which decays hyperbolically. In order to be able to describe the LRD of actual traffic, models with LRD capability were introduced. The fARIMA model is one of the new LRD models and is frequently used in self-similar traffic prediction.

2.3. fARIMA predictor

A fARIMA(p , d , q) process is defined as

$$\phi(B)(1-B)^d X(t) = \theta(B)\varepsilon_t \tag{5}$$

where B, $\phi(B)$ and $\theta(B)$ are the same as defined in the ARMA model. $(1-B)^d$ is the fractional differencing operator, with $-1/2 < d < 1/2$, and $d = H - 1/2$ where H is the Hurst parameter which is the defining factor of LRD [1][8]. Similar to the AR(p) model case, fARIMA(p , d , q) model with $q = 0$ is called fAR(p , d) model.

Let Δ^d denote the fractional differencing operator, and we can expand it as

$$\Delta^d = (1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k \tag{6}$$

where $\binom{d}{k} = \dfrac{\Gamma(d+1)}{\Gamma(K+1)\Gamma(d-K+1)}$, and Γ denotes the Gamma function.

Because of the non-integer values of d, the summation in Δ^d is genuinely over an infinite number of indices [6], which means that fARIMA predictor needs infinite number of traffic observations to produce the one-step-ahead prediction. But in reality, storing infinite observations is impossible. Even storing a very large number of historical data may not be practical either. When applied in practice, the fARIMA predictors are usually truncated at some desired level.

Since the fARIMA models have the fractional differencing operator, which can describe the LRD [1][6], they are believed to describe real traffic traces more accurately than ARMA models, and fARIMA predictors are often regarded as much more accurate than ARMA predictors [11].

2.4. Comparing ARMA and fARIMA predictors

The most obvious criterion when comparing predictors is the accuracy, and it has been treated as the most decisive factor affecting the performance of prediction-based bandwidth provisioning in most research. Many research papers have shown that fARIMA predictors have obvious advantage in accuracy over ARMA predictors, because fARIMA model can describe the LRD that exists in real network traffic [11]. But this is only true in the ideal case, where infinite traffic history is available. In practice, because of limited available traffic history or limited computation power, the fARIMA predictors have to be truncated. Some papers have shown that in such cases the advantage of fARIMA predictors over ARMA predictors is not as significant as researchers originally thought [3] [5] [9].

3. PREDICTION ERROR

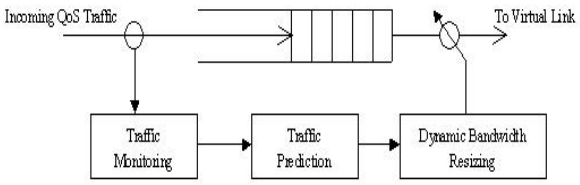

Fig. 1 Predictive Bandwidth Resizing in Edge Routers

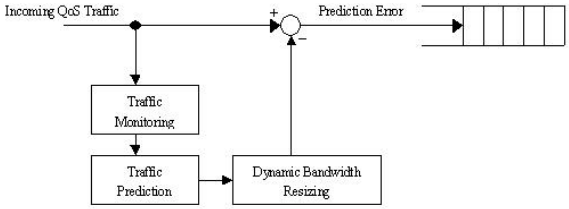

Fig. 2 Equivalent diagram of Predictive Bandwidth Resizing.

In this paper, it is discovered that beside accuracy, which has been treated as the only decisive factor in the predictor's performance, the autocorrelation structure of the prediction error also plays a very important role in determining the performance of prediction-based bandwidth provisioning. To the authors' best knowledge, the effect of the autocorrelation of prediction error on the performance of prediction-based bandwidth provisioning has not been studied before, and our analysis is given in this section.

In an overlay virtual network structure, the traffic transported on the virtual links is monitored at the edge routers, where traffic prediction is performed. Based on the predicted traffic load, the bandwidth of the virtual links is dynamically adjusted. Fig. 1 shows the diagram at the edge router. The performance of the system is determined by the performance

of the packet queue in the edge router executing the prediction algorithm. Equivalently, Fig. 1 can be redrawn as shown in Fig. 2, in which the prediction error can be of either positive or negative value. If it is positive, the resizing module adds packets into the queue; if the error is negative, the module draws packets out of the queue when it is not empty.

Let $e(t)$ denote the prediction error at time t as

$$e(t) = X(t) - \hat{X}(t) \tag{7}$$

where $X(t)$ is the incoming traffic, and $\hat{X}(t)$ is the prediction of $X(t)$ based on $X(t-1)$, $X(t-2)$, …, $X(t-m)$.

The size of the queue, denoted as $Q(t)$, can be represented as a function of $e(t)$.

$$Q(t) = \max(\,[Q(t-1) + e(t)],\, 0\,) \tag{8}$$

If the prediction is unbiased, which is the case in both ARMA and fARIMA predictors, we have $E\{e(t)\} = 0$. The more accurate a predictor is, the smaller the variance of $e(t)$. If $e(t)$ is not correlated, it will be safe to say that the smaller the variance is, the better the queue performs. But is $e(t)$ uncorrelated in reality? No previous research has been done to investigate this question.

It is well known that the autocorrelation structure of the traffic feeding into a queue has considerable impact on the queue performance. If the input has LRD, the performance of the queue will degrade considerably, compared to the uncorrelated Poisson arrival case, depending on how significant the LRD is. Therefore, as Fig. 2 illustrate, it is also very important to investigate the autocorrelation structure of the prediction errors of the predictors.

In real networks, $X(t)$ is long-range dependent [5] [8] [9] [11]. Since ARMA models cannot capture LRD, it is reasonable to expect that the prediction error of ARMA predictor has LRD. The fARIMA predictor is recognized for its ability of describing LRD. This may lead to a perception that the prediction error of fARIMA predictor does not have LRD. However, this may not be true in reality. As shown earlier in equation (6), $(1-B)^d$ is the component in the model that expresses LRD, but it is a polynomial of infinite order, which is impossible to implement in reality where only limited history is available. Because of this, the prediction error of fARIMA predictor may still have LRD.

In the simulations followed, the autocorrelation of the prediction errors of both ARMA and fARIMA predictors are inspected, and the effect of LRD of the prediction error on the performance of the queue is also analyzed.

4. SIMULATIONS

4.1. Simulation Setup

A single server queue, as shown in Fig. 1, is implemented. The queue size is unlimited. The input traffic trace is a real traffic trace, the Bellcore traffic trace BC-pAug89.TL [8], which is openly available and is often used in network research papers as standard test data. The traffic trace records the time-stamp and the packet size of 1 million Ethernet packets over the time interval of about 3142.82 seconds.

The trace is processed using window size of 200 millisecond to get a time series of traffic rate X_t. Each X_t represents the mean traffic rate during an observing window. To simulate realistic situations, the traffic history available to the predictors is limited to 20. Four

predictors are simulated, AR(5), AR(20), fAR(5) and fAR(20), respectively. An ideal predictor, whose prediction of the traffic rate for the next time interval is the actual traffic rate, is also included in the simulations. It is impossible to implement this ideal predictor in real world, but we use it as the benchmark in evaluating the performance of other predictors.

Each predictor is fitted on the first 200 seconds of data, and then the predictors are used to predict the traffic rate of the rest of the trace. For the fAR predictors, the value of d in equation (6) is set to be 0.3, based on the Hurst parameter estimation in [11].

Two scenarios - with and without error compensation - are run on each predictor. In the case of no error compensation, the predicted traffic rate of the next time interval is simply the output from the predictor; in the error compensation case, as suggested in [3], the prediction error of last time interval is added to the current output of the predictor:

$$\tilde{X}(t+1) = \hat{X}(t+1) + E(t) \tag{9}$$

$$E(t) = X(t) - \hat{X}(t) \tag{10}$$

in which, $\hat{X}(t)$ is the direct output from the predictor; $X(t)$ is the actual traffic load; $E(t)$ is the error of the predictor; and $\tilde{X}(t)$ is the final prediction value.

To deal with traffic burstiness and prediction error, the actual bandwidth was resized according to the traffic prediction multiplied by a factor $R > 1$. By changing the value of R, the link utilization can be controlled, as $u = 1/R$.

4.2. Simulation Results

The Mean Square Error (MSE) is used as the metric measuring the accuracy of the predictors. Table 1 shows the MSEs of different predictors, normalized by the variance of the traffic. From Table 1, it can be seen that fAR predictors are more accurate than AR predictors, but the advantage of fAR predictors is not very significant, as expected due to truncation. It is also noticed that the MSEs of predictors with error compensation are substantially larger than the corresponding predictors without error compensation. If accuracy is the only factor affecting the performance of the predictor, as it was treated in previous research, one may jump to the conclusion that the performance of predictors with error compensation are much worse than those without error compensation. But the simulation results shown later prove that it is not true.

Predictor	**MSE**	
	Without Error Compensation	With Error Compensation
AR(5)	0.7288	1.3438
AR(20)	0.7509	1.3549
fAR(5)	0.6764	1.3684
fAR(20)	0.7156	1.3716

Table 1 MSEs of different predictors (normalized by the variance of the traffic).

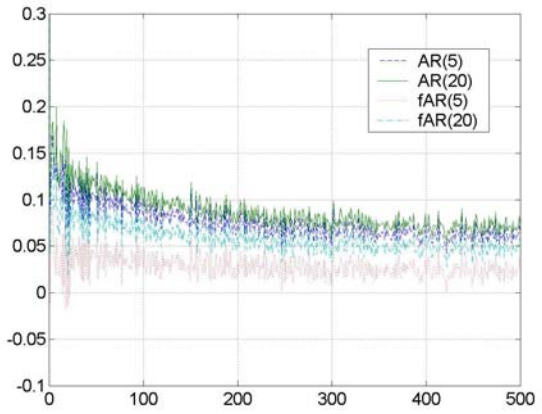

Fig. 3 Autocorrelation of prediction errors without error compensation.

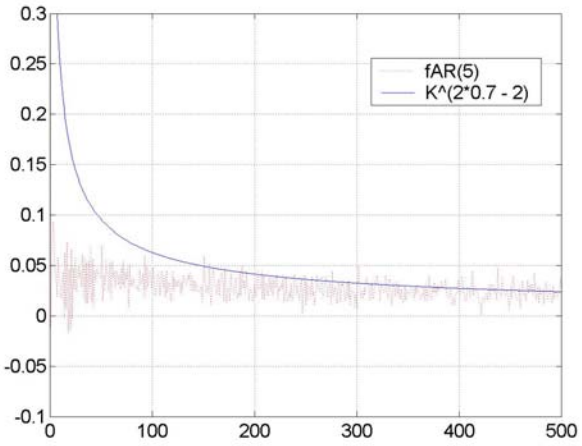

Fig.4 Autocorrelation of prediction error using fAR(5) predictor.

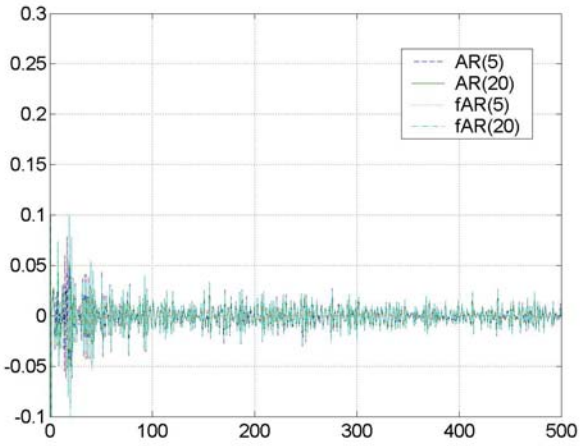

Fig.5 Autocorrelation of prediction errors with error compensation.

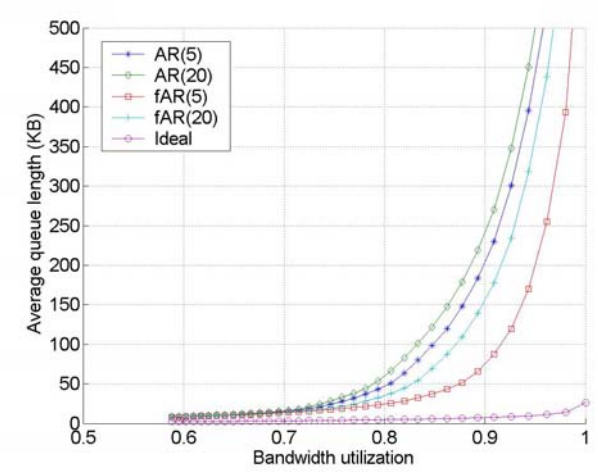

Fig.6 Average queue length using predictors without error compensation.

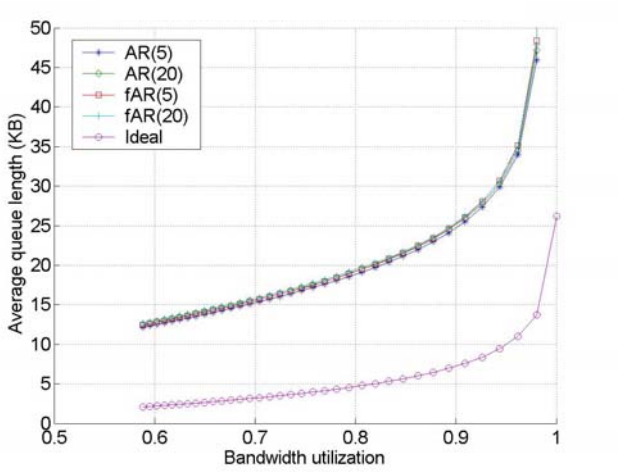

Fig.7 Average queue length using predictors with error compensation.

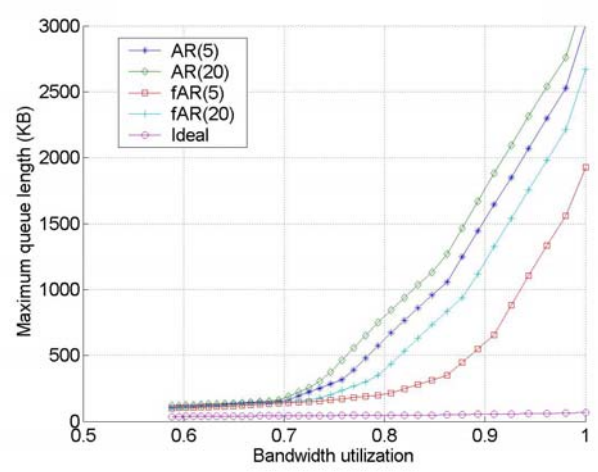

Fig.8 Maximum queue length using predictors without error compensation.

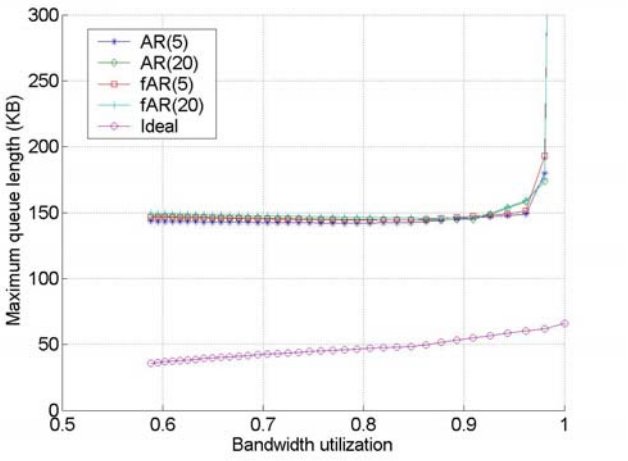

Fig. 9 Maximum queue length using predictors with error compensation.

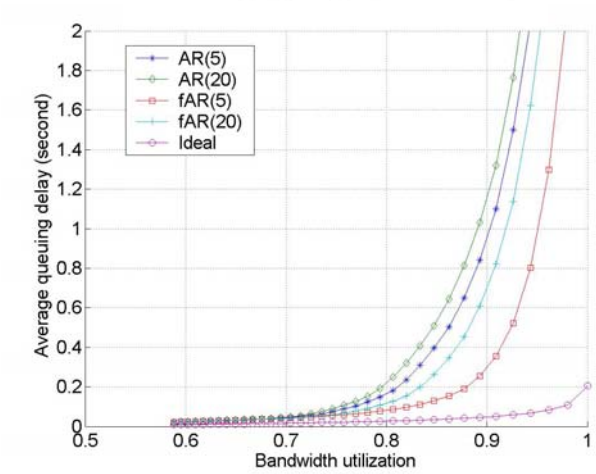

Fig. 10 Average queuing delay using predictors without error compensation

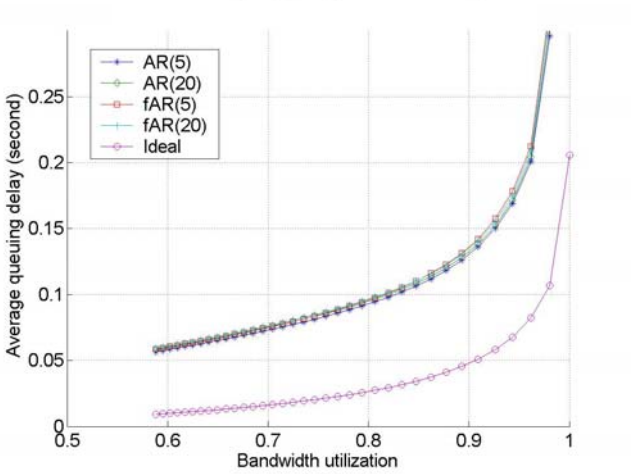

Fig. 11 Average queuing delay using predictors with error compensation.

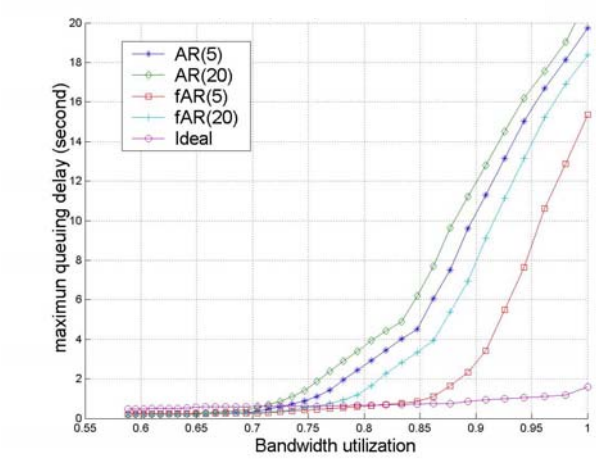

Fig. 12 Maximum queuing delay using predictors without error compensation.

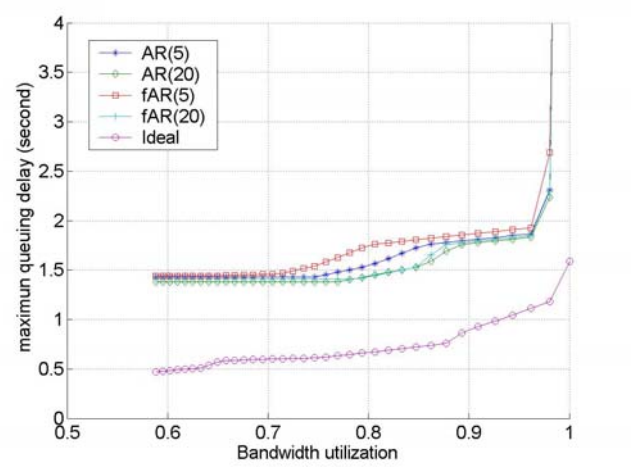

Fig. 13 Maximum queuing delay using predictors with error compensation.

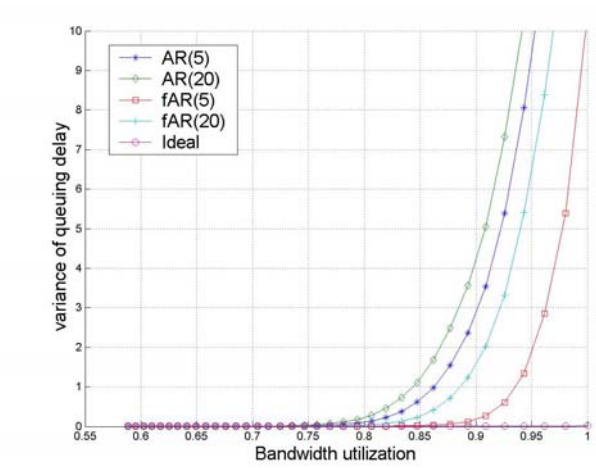

Fig. 14 Variance of queuing delay using predictors without error compensation.

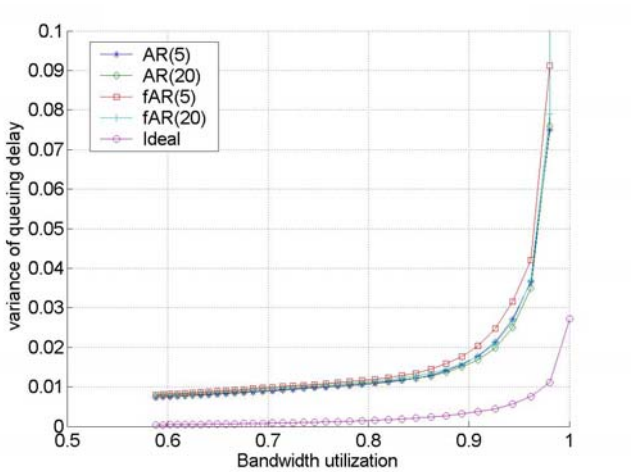

Fig. 15 Variance of queuing delay using predictors with error compensation.

Fig. 3 shows the autocorrelation functions of prediction errors using predictors without error compensation. It can be seen that, as we expected in previous analysis, the prediction errors of both the AR predictors and the fAR predictors have clearly visible LRD structure. The autocorrelation function of the prediction error with the least LRD, which is the one of the fAR (5) predictor, is selected and plotted in Fig. 4. It shows that the Hurst parameter for this prediction error is around 0.7. In Fig. 5, the autocorrelation functions of prediction errors using predictors with error compensation are shown. For all the four predictors, there is no observable LRD, and the autocorrelation functions are almost the same.

Fig. 6 and Fig. 8 show the average and maximum queue length at the edge router using predictors without error compensation. The curve at the bottom is the queue performance of the ideal predictor. It can be seen that compared to the ideal case the queue lengths using AR and fAR predictors without error compensation are much longer, especially when the link utilization is high. It should also be noticed that the queue performance curves in Fig. 6 and Fig. 8 have the same order as the autocorrelation functions in Fig. 3. The average and maximum queue lengths using predictors with error compensation are shown in Fig. 7 and Fig. 9 respectively. Compared with Fig. 6 and Fig. 8, the queue lengths are significantly reduced. The performances of the different predictors are almost identical, and are close to the curve of the ideal predictor, even when the link utilization is high.

Fig. 10 and Fig. 12 show the average and maximum queuing delays using predictors without error compensation. Similar to the performance of queue length, when predictors without error compensation are used, the queuing delays are much bigger than the ideal case. The order of the queuing delay curves shown in Fig. 10 and Fig. 12 are also the same as the order of autocorrelation functions in Fig. 3. Compared with Fig. 10 and Fig. 12, the average and maximum queuing delays using predictors with error compensation, as shown in Fig. 11 and Fig. 13, are much smaller, especially when the link utilization is high. The performances using different predictors with error compensation are very close.

Similar performance of queuing delay variance using predictors with and without error compensation can also be observed through Fig. 14 and Fig. 15.

It can be observed that the queue performance curves in Fig. 6, Fig. 8, Fig. 10, Fig. 12 and Fig. 14 all have the same order as the autocorrelation functions shown in Fig. 3. The less LRD the prediction error has, the better the queue performs. This suggests that the LRD in the prediction error has substantial influence on performance of the queue, which agrees with our analysis earlier.

In previous research, prediction accuracy was the only criterion used in evaluating the performance of different predictors. And it is widely accepted that the more accurate a predictor is, the better its performance. Such belief can in some way explain the results shown in Fig. 6 Fig. 8, Fig. 10, Fig. 12 and Fig. 14, because the order of the curves shown in those graphs are also the same as the order of MSEs of the predictors shown in Table 1. However, prediction accuracy cannot explain the results shown in Fig. 7, Fig. 9, Fig. 11, Fig. 13 and Fig. 15. The queue performance of predictors with error compensation is significantly better than the same predictors without error compensation, despite the fact that the accuracy of predictors with error compensation is much worse than the ones without error compensation (the MSEs nearly doubled as listed in Table 1).

The analysis in Section3, together with the simulation results in this section, show that besides accuracy, the autocorrelation structure of the prediction error, which has been neglected in previous research, is a very important factor in determining the performance of a

predictor in prediction-based bandwidth provisioning.

5. CONCLUSION

The analysis and simulations in this paper show that the performance of prediction-based bandwidth provisioning for overlay virtual networks depends not only on the accuracy of the predictor, as thought in previous research, but more importantly on the autocorrelation structure of the prediction error. Predictors with error compensation can remove LRD from the prediction error, and by removing the LRD of the prediction error the performance of the predictor can be improved significantly.

The ability of fARIMA predictor in describing LRD is limited in real application where the traffic history information is limited, and the prediction error of fARIMA predictor in real implementation still has LRD. Without error compensation, there is no noticeable performance advantage of fARIMA predictor over ARMA predictor. When the error compensation technique is applied, the performances of ARMA and fARIMA predictors are almost identical. This suggests that error compensation combined with the computationally less complex ARMA prediction is more suitable in prediction-based dynamic bandwidth provisioning applications.

REFERENCES:

[1] J. Beran, *Statistics for Long-memory Processes*, Chapman & Hall, London, 1994.

[2] P.J. Brockwell and R.A. Davis, *Time Series: Theory and Methods*, Springer-Verlag, New York, 1987.

[3] W. Cui and M. Bassiouni, "Virtual Private Network Bandwidth Management with Traffic Prediction", *Computer Networks*, V. 42, No. 6, P. 765-778, August 2003.

[4] Z. Duan, Z. Zhang, and Y. Hou, "Service Overlay Networks: SLAs, QoS, and Bandwidth Provisioning", *IEEE/ACM Trans. on Networking*, V. 11, No. 6, December 2003.

[5] M. Ghaderi, "On the Relevance of Self-Similarity in Network Traffic Prediction", *Technical report, CS-2003-28*, School of Computer Science, University of Waterloo, October 2003. (http://www.cs.uwaterloo.ca/cs-archive/ CS-2003/28/TR-CS-2003-28.pdf)

[6] J. R. M. Hosking, "Fractional Differencing", *Biometrika*, V. 68, No. 1, P.165-176, 1981.

[7] F. Kelly, "Note on Effective Bandwidths", *Stochastic Networks: Theory and Applications*, P. 141-168, Oxford Science Publications, 1996.

[8] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)", *IEEE/ACM Transactions on Networking (TON)*, V.2, No.1, p.1-15, Feb. 1994.

[9] S. Östring and H. Sirisena, "The Influence of Long-range Dependence on Traffic Prediction", *IEEE International Conference on Communications (ICC)* 2001.

[10] N. Sadek, A. Khotanzad and T. Chen, "ATM Dynamic Bandwidth Allocation Using F-ARIMA Prediction Model", *12th International Conf. on Computer Communications and Networks (ICCCN 2003)*, Oct. 2003.

[11] Y. Shu, Z. Jin, L. Zhang, and L. Wang, "Traffic Prediction Using FARIMA Models", *IEEE International Conference on Communications (*ICC*)* 1999.

[12] J. Yan, "Adaptive Configuration of Elastic High-Speed Multiclass Networks", *IEEE Communication Magazine*, May 1998.