Real-Time Network Traffic Prediction based on a Multiscale Decomposition

Guoqiang Mao
School of Electrical and Information Engineering
The University of Sydney
Email: g.mao@ieee.org

Abstract—**The presence of the complex scaling behavior in network traffic makes accurate forecasting of the traffic a challenging task. Some conventional prediction tools such as recursive least square method do not apply to network traffic prediction. In this paper we propose a multiscale decomposition approach to real time traffic prediction. The raw traffic data is first decomposed into multiple timescales using the $\grave{a}\ trous$ Haar wavelet transform. The prediction of the wavelet coefficients and scaling coefficients are performed independently at each timescale using ARIMA model. The predicted wavelet coefficients and scaling coefficient are then combined to give the predicted traffic value. This multiscale decomposition approach can better capture the correlation structure of traffic caused by different network mechanisms, which may not be obvious when examining the raw data directly. The proposed prediction algorithm is applied to real network traffic. It is shown that the proposed algorithm generally outperforms traffic prediction using neural network approach and gives more accurate result. The complexity of the prediction algorithm is also significantly lower than that using neural network.**

I. Introduction

It is well known that some characteristics of Internet traffic fall beyond the conventional framework of Markov traffic modelling. Leland et al. demonstrated self-similarity in a LAN environment (Ethernet) [1]. Paxson et al. showed self-similar burstiness manifesting itself in pre-World Wide Web WAN IP traffic [2]. Beran et al. demonstrated the self-similairty in variable-bit-rate (VBR) video traffic [3] and Crovella et al. showed self-similarity for WWW traffic [4]. Recent measurements and simulation studies further revealed that wide area network traffic has complex multifractal characteristics on small timescales, and is self-similar on large timescales [5], [6], [7]. The presence of the scaling behavior in network traffic is striking not only in its ubiquity, appearing in almost every kind of traffic, but also in the wide range of timescales over which the scaling holds. As network bandwidth increases over time, this scaling behavior may progressively extends over larger timescales [8].

Accurate forecasting of the traffic is important in the planning, design, control and management of networks. Traffic prediction at different timescales has been used in various fields of networks, such as long-term traffic prediction for network planning, design and routing; and short-term traffic prediction for dynamic bandwidth allocation, and predictive and reactive traffic and congestion control. The presence of the complex scaling behavior makes the accurate forecasting of Internet traffic a challenging task. An implication of the self-similarity (or equivalently long range dependence) in network traffic is that the autocorrelation function $r(k)$ of the traffic decays hyperbolically rather than exponentially fast:

$$r(k) \sim C_r k^{-\beta}\ ,\ 0 < \beta < 1 \tag{1}$$

where C_r is a positive constant and β is related to the Hurst parameter by $H = 1 - \beta/2$. Hurst parameter is a measure of the self-similarity. As a result the autocorrelation function is non-summable, i.e. $\sum_k r(k) = \infty$. This implies that the traffic process has an infinite variance. Therefore some conventional prediction tools such as recursive least square method do not apply to network traffic prediction.

Some algorithms have been proposed in the literature for real-time traffic prediction, which include FARIMA (fractional autoregressive integrated moving average) models [9], neural network approach [10], [11] and method based on α-stable models [12], [13], etc. Traffic prediction using FARIMA models relies on accurate estimation of the Hurst parameter. Despite a number of estimators reported in the literature, accurate estimation of the Hurst parameter remains a difficult problem even in off-line conditions. The presence of non-stationarity and complex scaling behavior in network traffic makes the situation even worse. Therefore the real applications of traffic prediction based on FARIMA models are not optimistic. Neural network approach can be quite complicated to implement in reality. The accuracy and applicability of neural network approach to traffic prediction is limited [11]. Finally, α-stable model is based on a generalized central limit theorem and its application is limited by that. It might achieve a good performance in heavy traffic or when there is a high level of traffic aggregations. However when traffic conditions deviate from that, the performance may be poor. Moreover, α-stable model is a parsimonious model, which may not be able to capture the complex scaling behavior of the traffic. In this paper we propose a traffic prediction algorithm based on a multiscale decomposition approach. Using the $\grave{a} - trous$ Haar wavelet transform, the traffic is decomposed into components at multiple timescales. Traffic component at each timescale is predicted independently with an ARIMA (autoregressive integrated moving average) model. Then they are combined to form the predicted traffic.

The rest of the paper is organized as follows: in section II, we shall introduce the use of the $\grave{a}\ trous$ Haar wavelet transform in decomposing the traffic into different timescales; in section III the prediction algorithm will be introduced; some simulation results using real traffic trace are given in IV and finally some conclusions and further work are summarized in section V.

II. Multiscale Traffic Decomposition

Wavelet tools have been widely used in the area of traffic analysis and they have many advantages when used for traffic analysis. Fundamentally, this is due to the non-trivial fact that the analyzing wavelet family itself possesses a scale invariant feature, a property not shared by other analysis methods. Quite different kinds of scaling features can be analyzed by the same technique.

Wavelet analysis is based on the decomposition of a signal using orthogonal bases[1]. Discrete wavelet transform (DWT) consists of the collection of coefficients

$$c_J(k) =< X, \varphi_{Jk}(t) >, \quad d_j(k) =< X, \psi_{jk}(t) >, \;\; j,\; k \in Z, \tag{2}$$

where $< *, * >$ denotes inner product, $\{d_j(k)\}$ are the wavelet coefficients and $\{c_J(k)\}$ are the scaling coefficients. Equation (2) compares the signal X to be analyzed with a set of analysis functions

$$\psi_{jk}(t) = 2^{-j/2}\psi(2^{-j}t - k). \tag{3}$$

This set of analysis functions is constructed from a reference pattern $\psi(t)$ called the mother-wavelet by a time-shift operation and a dilation operation. The mother wavelet is required to satisfy the admissibility condition, whose weak form is

$$\int \psi(t)dt = 0, \tag{4}$$

which shows it is a band-pass or oscillating function, hence the name “wavelet”. Function $\varphi_{Jk}(t)$ is a time shifted function of the scaling function $\varphi_J(t)$: $\varphi_{Jk}(t) = \varphi_J(t-k)$. $\varphi_J(t)$ is a low-pass function which can separate large timescale (low frequency) component of the signal. Thus wavelet transform decomposes a signal into a large timescale approximation (coarse approximation) and a collection of details at different smaller timescales (finer details). The original signal can be recovered from the wavelet coefficients and the scaling coefficients using

$$X(t) = \sum_k c_J(k)\varphi_{Jk}(t) + \sum_{j=1}^{J}\sum_k d_j(k)\psi_{jk}(t). \tag{5}$$

Theoretically the scale j can span from $-\infty$ to ∞. For practical signals, i.e. network traffic, we limit the scale to $0 \sim J$, where scale J is the largest timescale and scale 0 is the smallest timescale.

Define a dilated and shifted function $\varphi_{jk}(t)$ of $\varphi(t)$ as

$$\varphi_{jk}(t) = 2^{-j/2}\varphi_(2^{-j}t - k). \tag{6}$$

Denote the subspace spanned by the basis functions $\{\varphi_{jk}, k \in Z\}$ as V_j and the subspace spanned by the basis functions $\{\psi_{jk}, k \in Z\}$ as W_j. Multiresolution analysis (MRA) requires the subspaces satisfy

$$V_J \subset V_{J-1} \subset \cdots \subset V_0 \quad and \quad V_{j+1} \bigoplus W_{j+1} = V_j. \tag{7}$$

Equation (7) means a signal can also be expressed as the combination of a small timescale (smaller than the timescale corresponding to scale J) approximation (finer approximation) and the details at even smaller timescales. In fact, we can zoom into any timescale that we are interested in and use the coefficients of a wavelet transform to directly study the scale dependent properties of the data. For example, if we fix a scale j and investigate certain statistics about the wavelet coefficients at that scale across time we can obtain information about the scaling behavior of the signal as a function of j (the global-scaling behavior). Alternatively, if we fix a point in time t and examine how the wavelet coefficients within the cone of influence of t change across scales as we examine finer and finer scales, we can determine the local irregularity (the local scaling behavior) of the signal about the point t. Moreover the analysis of each scale is largely decoupled from that at other scales [8]. Refer to [14], [15] for details of wavelet theory.

In addition to the characteristics of applications generating the traffic, traffic variations at different timescales are caused by different network mechanisms. Traffic variations at small timescales (i.e. in the order of ms or smaller timescale) are caused by buffering effect and scheduling algorithms etc. Traffic variations at larger timescales (i.e. in the order of 100ms) are caused by traffic and congestion control protocols, e.g. TCP protocols. Traffic variations at even larger timescales are caused by routing changes, daily and weekly cyclic shift in user populations. Finally long-term traffic changes are caused by long-term increases in user population as well as increases in bandwidth requirement of users due to the emergence of new network applications. This fact motivates us to decompose traffic into different timescales and predict traffic independently at each timescale. The proposed multiscale decomposition approach to traffic prediction allows us to explore the correlation structure of network traffic at different timescales caused by different network mechanisms, which may not be easy to investigate when examining the raw data directly.

The roles of the mother scaling and wavelet functions $\varphi(t)$ and $\psi(t)$ can also be represented by a low-pass filter h and a high pass filter g. Consequently, the multiresolution analysis and synthesis of a signal $x(t)$ can be implemented efficiently as a filter bank [14]. The approximation at scale j, $c_j(k)$ is passed through the low-pass filter h and the high pass filter g to produce the approximation $c_{j+1}(k)$ and the detail $d_{j+1}(k)$ at scale $j+1$. At each stage, the number of coefficients at scale j is decimated into half of that at scale $j+1$, due to downsampling. This decimation reduces the number of data points to be processed at coarser time scales and removes the redundancy information in the wavelet and the scaling coefficients at the coarser time scales. Decimation allows us to represent a signal X by its wavelet and scaling coefficients whose total length is the same as the original signal. However decimation has the undesirable effect that we cannot relate information at a given time point at the different scales in a simple manner. Moreover, while it is desirable in some applications (e.g. image compression) to remove the redundancy information, in time series prediction the redundancy information can be used to improve the accuracy of the prediction.

In this paper, we use a redundant wavelet transform, i.e. the $\grave{a} - trous$ wavelet transform, to decompose the signal [16]. Using the redundant information from the original signal, the $\grave{a} - trous$ wavelet transform produces smoother approximations by filling the “gap” caused by decimation. Using the $\grave{a} - trous$ wavelet transform, the scaling coefficients and the wavelet coef-

[1] Some other wavelet bases exist, such as semi-orthogonal or bi-orthogonal wavelet bases. However in this research, we only consider orthogonal bases.

ficients of $x(t)$ at different scales can be obtained as:

$$c_0(t) = x(t) \tag{8}$$

$$c_j(t) = \sum_{l=-\infty}^{\infty} h(l) c_{j-1}(t + 2^{j-1} l). \tag{9}$$

where $1 \leq j \leq J$, and h is a low-pass filter with compact support. The detail of $x(t)$ at scale j is given by

$$d_j(t) = c_{j-1}(t) - c_j(t). \tag{10}$$

The set $d_1, d_2, ..., d_J, c_J$ represents the wavelet transform of the signal up to the scale J, and the signal can be expressed as a sum of the wavelet coefficients and the scaling coefficients:

$$x(t) = c_J(t) + \sum_{j=1}^{J} d_j(t) \tag{11}$$

Many wavelet filters are available, such as Daubechies' family of wavelet filters, $B3$ spline filter, etc. Here we choose Haar wavelet filter to implement the $\grave{a} - trous$ wavelet transform. A major reason for choosing the Haar wavelet filter is the calculation of the scaling coefficients and wavelet coefficients at time t uses information before time t only. This is a very desirable feature in time series prediction. The Haar wavelet uses a simple filter $h = (1/2, 1/2)$. The scaling coefficients at higher scale can be easily obtained from the scaling coefficients at lower scale:

$$c_{j+1,t} = \frac{1}{2}(c_{j,t-2^j} + c_{j,t}). \tag{12}$$

The wavelet coefficients can then be obtained from Equation (10).

III. The prediction algorithm

In this section, we use the aforementioned $\grave{a} - trous$ Haar wavelet decomposition for traffic prediction. Instead of predicting the original signal $X(k), X(k-1), ..., X(k-N)$ directly, we predict the wavelet coefficients and the scaling coefficients independently at each scale and use the wavelet coefficients and the scaling coefficients to construct the prediction of the original signal.

Fig. 1 shows the architecture of the prediction algorithm. Coefficient prediction can be represented mathematically as

$$\begin{aligned} \widehat{c}_J(k+p) &= \widehat{F}_J(c_J(k), c_J(k-1), ..., c_J(k-m)), &(13)\\ \widehat{d}_j(k+p) &= \widehat{f}_j(d_j(k), d_j(k-1), ..., d_j(k-n_j)), &(14)\end{aligned}$$

where m and n_j is the number of coefficients taken for prediction and p is the prediction depth. In this paper, we only use one-step prediction, i.e. p=1. Multistep prediction can be achieved by using the predicted value as the real value or by aggregating the traffic into larger time interval.

$ARIMA(p,d,q)$ model is used for prediction. An ARMA(p,q) (autoregressive moving average) model can be represented as:

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \tag{15}$$

where Z_t is a Gaussian distributed random variable with zero mean and variance σ^2, , i.e. $Z_t \sim WN(0, \sigma^2)$ and the polynomials $(1 - \phi_1 z - \cdots - \phi_p z^p)$ and $(1 + \theta_1 z + \cdots + \theta_q z_{t-q})$ have no common factors [17]. If $p = 0$, then the model reduces to a pure MA process and if $q = 0$, then the process reduces to a pure AR process. Equation (15) can also be written in a more concise form as:

$$\phi(B) X_t = \theta(B) Z_t, \tag{16}$$

where ϕ and θ are polynomials of degree p and q respectively and B is the backward shift operator:

$$B^j x_t = X_{t-j}, j = 0, 1, ... \tag{17}$$

ARMA model assumes the time series are stationary. If the time series exhibits variations that violate the stationarity assumption, differencing operation can be used to remove the non-stationary trend in the time series. We define the lag-1 difference operator ∇ by

$$\nabla X_t = X_t - X_{t-1} = (1 - B) X_t. \tag{18}$$

An ARIMA(p,d,q) model is an ARMA(p,q) model that has been differenced d times. Therefore it can be represented as:

$$\phi(B)(1 - B)^d X_t = \theta(B) Z_t \tag{19}$$

If the time series has a non-zero average value through time, then Equation (19) also features a constant term δ on its right hand side.

Fig. 2 and Fig. 3 shows the wavelet coefficients and the scaling coefficients of an hour-long LAN traffic trace. The time series being analyzed is the data rate of the LAN trace measured in terms of byte/s during 1s measurement interval. The details of the traffic trace will be introduced later. A visual inspection of the scaling coefficients and wavelet coefficients indicates that the wavelet coefficients can be reasonably treated as a stationary time series with zero mean. Therefore wavelet coefficients can be modelled using ARMA(p,q) model, or equivalently ARIMA(p,0,q) model. However there is significant non-stationarity in the scaling coefficients. This non-stationarity becomes more obvious when examining the scaling coefficients over longer time period as shown in Fig. 4. Therefore for scaling coefficients it is more appropriate to use ARIMA(p,d,q) model.

Box-Jenkins forecasting methodology is used to establish the ARIMA(p,d,q) model for prediction at each scale. Box-Jenkins methodology involves four steps [17]:

- The first step is the tentative identification of the model parameters. This is done by examining the sample autocorrelation function:

$$r_k = \frac{\frac{1}{n-k} \sum_{t=1}^{n-k} (X_t - \bar{X})(X_{t+k} - \bar{X})}{\frac{1}{n} \sum_{t=1}^{n} (X_t - \bar{X})^2}, \tag{20}$$

where

$$\bar{X} = \frac{\sum_{t=1}^{n} X_t}{n} \tag{21}$$

and the sample partial autocorrelation function:

$$r_{kk} = \begin{cases} r_1 & \text{if } k = 1 \\ \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} r_j} & \text{if } k = 2, 3, ... \end{cases}, \tag{22}$$

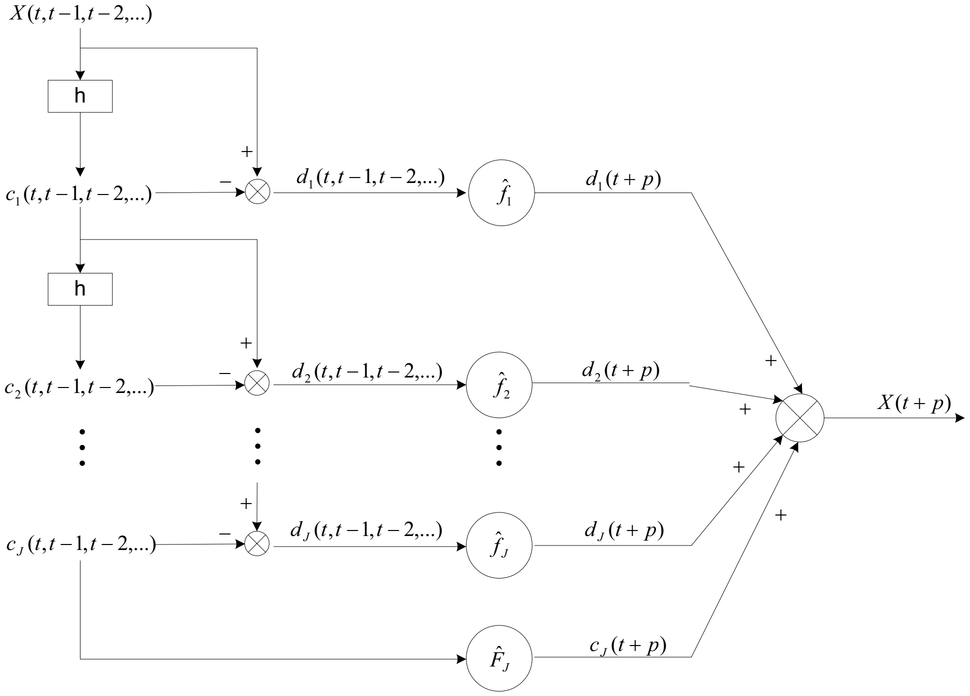

Fig. 1. Architecture of the prediction algorithm

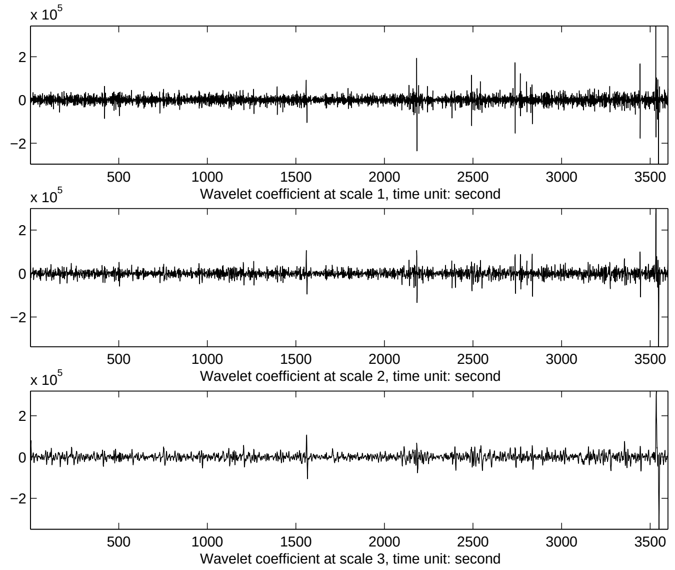

Fig. 2. Wavelet coefficients from scale 1 to 3

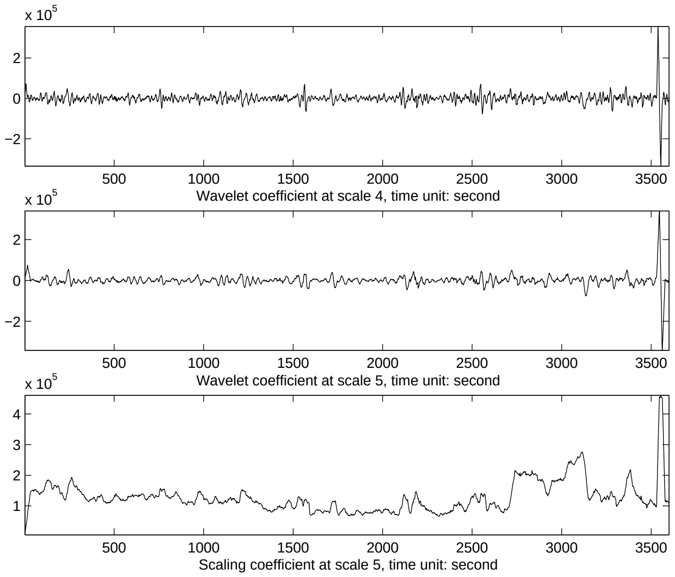

Fig. 3. Wavelet coefficients at scales 4 & 5 and Scaling coefficients at scale 5

where

$$r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j}, \text{for } \ j = 1, 2, ..., k-1 \qquad (23)$$

of the time series X.

- Estimation step. Once the model is established, the model parameters can be estimated using either a maximum likelihood approach or a least mean square approach. In this paper both the maximum likelihood approach and the least mean square approach were tried and their results are almost exactly the same. Thus we stick to the least mean square approach to estimate the model parameters for its simplicity.
- Diagnostic check step. Diagnostic checks can be used to see whether or not the model that has been tentatively identified and estimated is adequate. This can be done by examining the sample autocorrelation function of the error signal, i.e. the difference between the predicted value and the real value. If the model is inadequate, it must be modified and improved.
- When a final model is determined, it can be used to forecast future time series values.

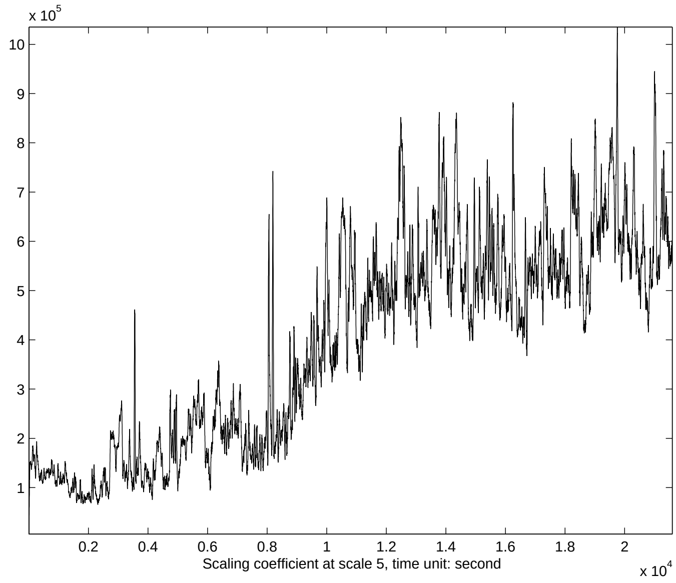

Fig. 4. Scaling coefficients at scale 5 over a 6-hour period

IV. Simulation

In this section, we apply the proposed model to the real network traffic for prediction. The traffic traces used were collected by WAND research group at the University of Waikato Computer Science Department. It is the LAN traffic at the University of Auckland on campus level. The traffic traces were collected between 6am and 12pm from on June 9, 2001 to June 13, 2001 on a 100Mbps Ethernet link. IP headers in the traffic trace are GPS synchronized and have an accuracy of $1\mu s$. More information on the traffic trace and the measurement infrastructure can be found on their webpage: http://atm.cs.waikato.ac.nz/wand/wits/auck/6/. Fig.5 shows the traffic rate of the traffic trace measured between 6am and 12am on June 12, 2001. The traffic rate is measured on 1 second intervals.

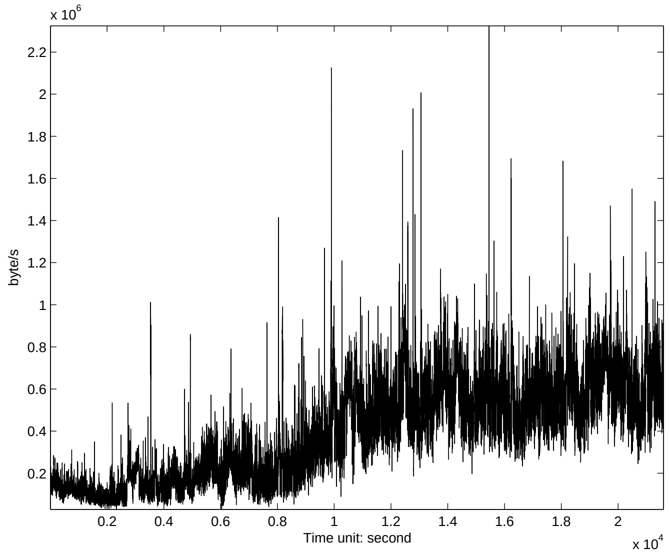

Fig. 5. Traffic rate of the LAN trace measured between 6am and 12am on June 12, 2001

Five traffic traces are available. Table II shows information of the traffic traces.

We use the traffic rate measured in the previous 1s time intervals to predict the traffic rate in the next second. Prediction over longer or shorter time intervals can be achieved by reducing the length of the time interval or by multistep prediction. To validate the performance of the proposed prediction model, one of the traffic traces (i.e. trace 4) was picked randomly to establish the prediction model and the prediction model is then applied to other traffic traces for prediction.

Table II shows the model parameters of the ARIMA(p,d,q) model at each scale. Three scales are chosen. The choice on the number of scales used for prediction is made based on the trade-off between model complexity and accuracy. Further increase in the number of scales significantly increases the complexity of the algorithm but there is only a modest increase in accuracy. As shown in the table, most noise in the model comes from wavelet coefficients at scale 1. In comparison with wavelet coefficients and scaling coefficients at other scales, wavelet coefficients at scale 1 has very weak autocorrelations and a white noise like power spectral density. It is almost like white noise. It is the wavelet coefficients at scale 1 that limit the overall performance that can be achieved by the prediction algorithm. Fig. 6 shows the autocorrelation function of the wavelet coefficients at scale 1.

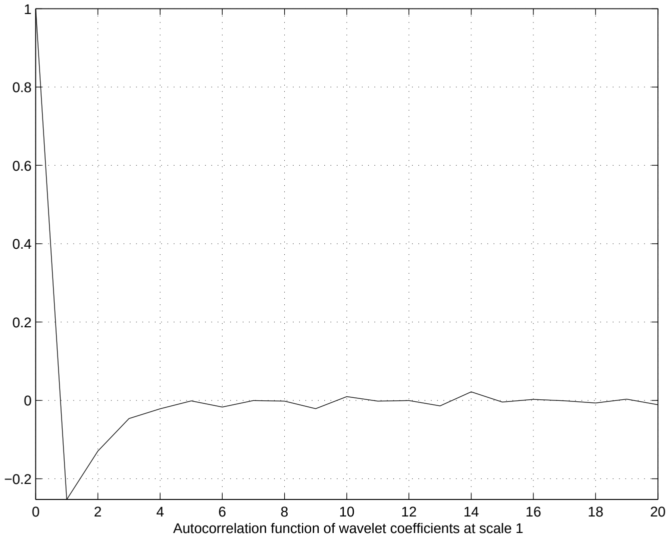

Fig. 6. Autocorrelation function of wavelet coefficients at scale 1

The ARIMA models developed from trace 4 are then applied to the other traffic traces to establish the performance of the prediction algorithm. To measure the performance of the prediction algorithm, two metrics are used. One is the normalized mean square error (NMS):

$$NMSE = \frac{\frac{1}{N}\sum_{n=1}^{N}(X(n)-\hat{X}(n))^2}{var(X(n))} \qquad (24)$$

where $\hat{X}(n)$ is the predicted value of $X(n)$ and $var(X(n)$ denotes the variance of $X(n)$. The other is the mean absolute rel-

TABLE I

TRACE TRACE USED IN THE SIMULATION

Trace ID	File name	Measurement time	Duration
1	20010609-060000-e0.gz	Saturday June 9, 2001	6am-12pm
2	20010610-060000-e0.gz	Sunday June 10, 2001	6am-12pm
3	20010611-060000-e0.gz	Monday June 11, 2001	6am-12pm
4	20010612-060000-e0.gz	Tuesday June 12, 2001	6am-12pm
5	20010613-060000-e0.gz	Wednesday June 13, 2001	6am-9am

TABLE II

MODEL PARAMETER OF THE PREDICTION MODEL

Scale	Model name	Model parameters ϕ	Model parameters θ	Noise σ^2
Wavelet coefficient 1	ARIMA(1,0,4)	$\phi_1 = 0.8842$	$\theta_1 = 1.311, \theta_2 = -0.2185,$ $\theta_3 = 0, \theta_4 = -0.1008$	2.147×10^9
Wavelet coefficient 2	ARIMA(4,0,4)	$\phi_1 = 1.443, \phi_2 = -0.4782,$ $\phi_3 = 0.04215, \phi_4 = -0.02682$	$\theta_1 = -0.04322, \theta_2 = 1.768$ $\theta_3 = 0.04953, \theta_4 = -0.7767$	5.847×10^8
Wavelet coefficient 3	ARIMA(4,0,8)	$\phi_1 = 1.384, \phi_2 = -0.435$ $\phi_3 = 0.02306, \phi_4 = -0.004911$	$\theta_1 = -0.1833, \theta_2 = -0.1531,$ $\theta_3 = -0.1824, \theta_4 = 1.751,$ $\theta_5 = 0.1789, \theta_6 = 0.1508,$ $\theta_7 = 0.1782, \theta_8 = -0.7583$	1.422×10^8
Scaling coefficient 3	ARIMA(2,1,8)	$\phi_1 = 0.508, \phi_2 = 0.02201$	$\theta_1 = -0.07853, \theta_2 = -0.08036$ $\theta_3 = -0.07985, \theta_4 = -0.08014,$ $\theta_5 = -0.07935, \theta_6 = -0.08083,$ $\theta_7 = -0.0796, \theta_8 = 0.9188$	1.348×10^8

ative error (MARE), which is defined as follows:

$$MRE = \frac{1}{N} \sum_{n=1}^{N} \left| \frac{X(n) - \hat{X}(n)}{X(n)} \right| \quad (25)$$

Since the relative error may be unduely affected by vary small values of $X(n)$, to make meaningful observations, we only count the MARE of $X(n)$ whose value is not small than the average value of $X(n)$. Table III shows the performance of the prediction algorithm. For comparison purpose, the performance of traffic prediction using neural network approach is also shown in the table. A number of neural network models with different number of input nodes, hidden nodes and transfer functions are evaluated, including those reported in [11], [18]. It is found that the 32-16-4-1 network architecture used in [18] gives the best performance. Hyperbolic tangent sigmoid transfer function is used in the hidden layer and linear transfer function is used in the output layer. The performance of the 32-16-4-1 neural network model is shown in Table III to represent the prediction performance using neural networks. To achieve a fair comparison, the same trace used for building ARIMA(p,d,q) models is used to train the neural network. The very large data size in the training trace ensures the convergence of the neural network, which is also confirmed by a visual inspection of the error signal.

As shown in Table III, the ARIMA model with multiscale decomposition (referred to as multiscale ARIMA model for simplicity) gives better performance than neural network in most cases except for trace 2, where the MARE metric of neural network approach is slightly better than that achieved by multiscale ARIMA approach. However, the NMS metric of neural network approach is much worse than that of multiscale ARIMA approach for trace 2. Therefore the exception on trace 2 cannot be used as an evidence that neural network performs better for trace 2. Fig. 7 and Fig. 8 show the autocorrelation function of the error signal for traffic trace 5 using multiscale ARIMA model and using neural network respectively. The autocorrelation function of the error signal for other traffic traces demonstrates similar characteristics. The autocorrelation function of the error signal using multiscale ARIMA model is much weaker than that using neural network prediction and it dies down faster. This also indicates that the performance of multiscale ARIMA model is better than neural network prediction as the error is closer to white noise. As such, it can be concluded that ARIMA model with multiscale decomposition generally achieves better performance than neural network. Moreover, only three scales are employed in the proposed prediction algorithm, which requires a memory length (here memory length refers to the number of past raw data samples required for prediction) of about 8. In comparison, neural network requires a memory length of 32. The computation using multiscale ARIMA model is also much easier than that using neural network.

V. CONCLUSION AND FURTHER WORK

In this paper we proposed a real-time network traffic prediction algorithm based on a multiscale decomposition. The raw traffic data is first decomposed into different timescales using the $\grave{a}\ trous$ Haar wavelet transform. The prediction of the

TABLE III
PERFORMANCE OF THE PREDICTION MODEL

Trace ID	Multiscale ARIMA		Neural network	
	NMS	MARE	NMS	MARE
1	0.1319	0.1633	0.1603	0.1667
2	0.2296	0.2165	0.3168	0.2053
3	0.1507	0.1403	0.1565	0.1493
4	0.1592	0.1313	0.1622	0.1386
5	0.21972	0.1731	0.2258	0.1823

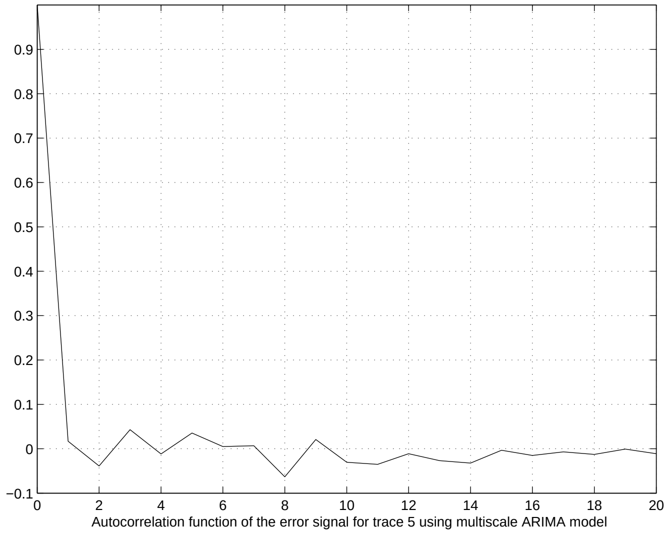

Fig. 7. Autocorrelation function of the error signal for trace 5 using multiscale ARIMA model

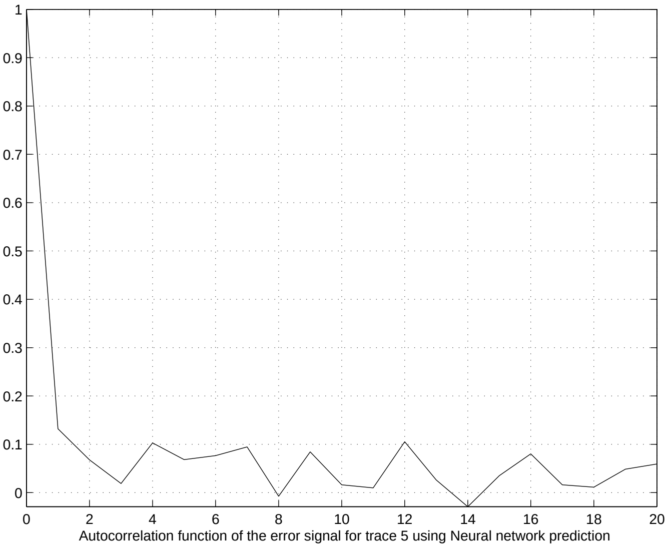

Fig. 8. Autocorrelation function of the error signal for trace 5 using neural network prediction

wavelet coefficients and scaling coefficients are performed independently at each timescale using ARIMA model. The predicted wavelet coefficients and scaling coefficient are then combined to give the predicted traffic value. As traffic variations at different timescales are caused by different network mechanisms, the proposed multiscale decomposition approach to traffic prediction can better capture the correlation structure of traffic caused by different network mechanisms, which may not be obvious when examining the raw data directly.

The prediction algorithm was applied to real network traffic. The autocorrelation of the error signal of the prediction algorithm is very weak, which is an indication of the adequacy of the model. The performance of the prediction algorithm was compared with that using neural network. It is shown that the proposed algorithm generally outperforms traffic prediction algorithm using neural network approach and gives more accurate prediction. The complexity of the prediction algorithm is also significantly lower than that using neural network.

As pointed out in the paper, traffic variations at large timescales are caused by routing changes, daily and weekly cyclic shift in user populations. However the length of the traffic trace (6-hour maximum) used in our analysis prohibits us to do large timescale traffic analysis and prediction. It is excepted that some future work can be carried out in this area, which allows us to build prediction models for large time scale traffic variations and incorporate them into our prediction algorithm. This will improve both the accuracy and the generalization capability of the prediction algorithm.

REFERENCES

[1] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, 1994.

[2] V. Paxson and S. Floyd, "Wide area traffic: The failure of poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, 1995.

[3] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Transactions on Communications*, vol. 43, no. 2/3/4, pp. 1566–1579, 1995.

[4] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, 1997.

[5] A. Feldmann, A. Gilbert, W. Willinger, and T. Kurtz, "The changing nature of network traffic: Scaling phenomena," *Computer Communication Review*, vol. 28, no. 2, pp. 5–29, 1998.

[6] A. Feldmann, A. C. Gilbert, and W. Willinger, "Data networks as cascades : Investigating the multifractal nature of internet wan traffic," 1998, Proceedings of SIGCOMM'98, pp. 42–55.

[7] A. Erramilli, O. Narayan, A. Neidhardt, and I. Saniee, "Performance impacts of multi-scaling in wide area tcp/ip traffic," in *INFOCOM 2000*, Tel Aviv , Israel, 2000, vol. 1, pp. 352–359.

[8] P. Abry, P. Flandrin, M. S. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation, and synthesis of scaling data," in *Self-Similar Network*

Traffic and Performance Evaluation, K. Park and W. Willinger, Eds., pp. 39–88. John Wiley and Sons, Inc., 2000.

[9] Y. Shu, Z. Jin, L. Zhang, and L. Wang, "Traffic prediction using farima models," in *IEEE International Conference on Communications*, 1999, vol. 2, pp. 891–895.

[10] Y. Liang, "Real-time vbr video traffic prediction for dynamic bandwidth allocation," *IEEE Transactions on Systems, Man and Cybernetics, Part C*, vol. 34, no. 1, pp. 32–47, 2004.

[11] J. Hall and P. Mars, "Limitations of artificial neural networks for traffic prediction in broadband networks," *Communications, IEE Proceedings-*, vol. 147, no. 2, pp. 114–118, 2000, TY - JOUR.

[12] M. Lopez-Guerrero, J.R. Gallardo, D. Makrakis, and L. Orozco-Barbosa, "Optimizing linear prediction of network traffic using modeling based on fractional stable noise," in *2001 International Conferences on Info-tech and Info-net*, 2001, vol. 2, pp. 587–592 vol.2.

[13] A. Karasaridis and D. Hatzinakos, "Network heavy traffic modeling using $alpha$-stable self-similar processes," *IEEE Transactions on Communications*, vol. 49, no. 7, pp. 1203–1214, 2001.

[14] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, 1996.

[15] I. Daubechies, *Ten Lectures on Wavelets*, Capital City Press, Montpelier, Vermont, 1992.

[16] M.J. Shensa, "The discrete wavelet transform: wedding the a trous and mallat algorithms," *IEEE Transactions on Signal Processing*, vol. 40, no. 10, pp. 2464–2482, 1992.

[17] B. L. Bowerman and R. T. O'Connell, *Time Series Forecasting - Unified Concepts and Computer Implementation*, PWS publishers, 2 edition, 1987.

[18] Y. Liang and E.W. Page, "Multiresolution learning paradigm and signal prediction," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2858–2864, 1997.