A Measurement-Based Modeling Approach for Network-Induced Packet Delay[1]

Daniel A. Vivanco and Anura P. Jayasumana
Department of Electrical and Computer Engineering,
Colorado State University, Fort Collins, CO 80523

***Abstract*— An approach is presented to capture and model Internet end-to-end packet delay behavior using *ARMA* and *ARIMA* models. Autocorrelation (*ACF*) and Partial Autocorrelation (*PACF*) functions are used to identify the most appropriate model and the model order. Impact due to sending rate and packet size of the probe, and the available link capacity on these two metrics are investigated. Results indicate that the models presented reflect accurately the effect of packet correlation induced by the network. Modeling Inter-Packet Gap (*IPG*) is an alternative for capturing the effect of the network on a packet stream. A methodology for fitting *ARMA* and *ARIMA* models to end-to-end packet delay and *IPG* series is presented.**

Keywords; Internet Measurments, Delay, Inter-packet gaps.

1. INTRODUCTION

Understanding the nature of end-to-end packet delay is important for several areas of application development, routing and transport protocols design, and congestion and flow control algorithm development. Due to random delays associated with queuing and processing, interaction of packet flows, and variation of load on different links, end-to-end packet delay is a complex random process. Network conditions change with time due to link congestion, link failure, routing table updates, among others. Thus first, second and higher order moments of end-to-end packet delay may also vary with time. Modeling efforts have to consider this to decide the time scales at which statistics should be collected and the type of statistics to collect. Traffic behavior, when described at different time scales, can be classified as microdynamics and macrodynamics [10]. The former characterizes the local behavior of the traffic flows in a short time scale, however the latter captures the global behavior of the traffic over longer time scales. Macrodynamics behavior of traffic is of significant importance for network performance analysis

Capturing accurately the end-to-end packet delay characteristics is crucial for understanding the dynamics of the end-to-end delivery process and for the development of accurate models. Such models can also have a great impact on operation and management of networks.. For instance, with both real-time[2] and non-real time congestion control mechanisms[6][9], delay based alternatives to packet-loss based approaches aim at preventing network congestion at early stages. Furthermore, end-to-end packet delay models can be applied to estimate the behavior of live streams under particular network conditions. A direct application for accurate delay models can be found on network emulators[8][22], in which incoming real packet data streams are altered based on traces capturing end-to-end behavior or analytical/simulation models representing such behavior. Such emulators are useful for investigating the performance and behavior of distributed applications and application software under different network conditions. An accurate measurement based model for end-to-end delay can replace large traces that would be required to reproduce network conditions. Such a model provides a solution to the problem, addressed in [17], of synthetically generating traffic characteristics that obey temporal patterns as observed in realistic networks.

It has been demonstrated that the distribution of the end-to-end packet delay by itself does not always result in a complete model, due to the fact that the correlation of packets belonging to the same stream is not considered[19]. Moreover, such an analysis will reveal only the overall behavior of the system and not its dynamic behavior. Also, reproducing the network state, either using simulation or emulation techniques, just by employing the observed end-to-end packet delay distribution will not yield to accurate representation of this metric, since its internal dynamics are not captured. An alternative for accurate representation could be using traces of packet delay. Such an approach also has many limitations including the large trace lengths needed, the fact that it only captures one specific packet delay flow under certain network conditions, and its inability to provide insights into network behavior.

End-to-end packet delay modeling has been attempted using several techniques. Queuing theory has been broadly used as a powerful tool for this; however, accurate queuing analysis requires that inter-arrival and inter-departure traffic distributions at each individual link are known, or follow a tractable distribution, which is rarely the case in Internet [24]. Even if the distribution of each link is available, the computational cost will grow dramatically as the network size increases. Compared to analytical models such as queuing models, time series models offer a methodology with smaller computational cost, that is easier to utilize and update, and in general, less complex to use. Time series offers a passive *black box* approach, in which model fitting does not require any assumption about the internal structure of the observed system. Time series methods have been used in the past for modeling the macrodynamic behavior of arrival processes and system responses. Correlation among Inter Packet Gaps *(IPG)* and network induced delays was investigated in [19], in which the effect of the network induced over a flow of probes was analyzed. This was accomplished by sending a continuous constant bit rate (*CBR)* stream of packet. [19] also pointed out

[1] This work is supported in part by the Engineering Research Center program of the National Science Foundation under NSF award number 0313747, Ixia University Partners Program and a grant from Agilent Technologies.

0742-1303/07 $25.00 © 2007 IEEE
DOI 10.1109/LCN.2007.136

IEEE computer society

the difficulty of modeling packet delay process when high sending rate flows are utilized. In this paper, we present a comprehensive approach for end-to-end packet delay modeling for a range of *CBR* flows with different sending rates and packets sizes. Our findings show that the behavior of end-to-end packet delay can be captured effectively by *ARMA* (Auto-Regressive Moving Average) and *ARIMA* (Auto-Regressive Integrated Moving Average) models. It is observed that autocorrelation of packet delay series is weak when sending rates are low compared to available network link capacity. As the sending bit rate increases, network bottleneck links tend to get more congested. Hence, end-to-end packet delay sequences get more auto-correlated. Our references below to the intensity of the sending rate as a fraction of the available link capacity refers to the amount of link congestion the flow produces on the network, and not its actual load. Choosing of the appropriate *ARMA* and *ARIMA* order is done by analyzing the observed *ACF* (Auto Correlation Function) and *PACF* (Partial Auto correlation Function) distribution of the observed packet delay series. Our findings indicate that *ACF* and *PACF* distributions are appropriate for capturing the stream behavior according to the network condition. Modeling *IPG* is also proposed as an alternative for capturing the effect of the network on a data stream. A methodology for fitting *ARMA* and *ARIMA* models to end-to-end packet delay and *IPG* series is presented.

Section 2 reviews the related work. In Section 3, we present an overview of time series and modeling techniques associated with the Internet end-to-end packet delay processes. Experiment setup, results and goodness-of-fit for packet delay modeling are presented in Section 4. A methodology for fitting *ARMA* and *ARIMA* models into packet delay and *IPG* processes is also presented in Section 4. Section 5 concludes the analysis and provides direction for future work.

2. Related Work

There have been several previous measurement based studies related to end-to-end packet delay in recent literature. The distribution of packets delay has been studied using queuing theory [3] and system identification techniques [18]. In [4], the end-to-end packet delay and loss behavior in the Internet have been examined using small UDP probe packets. In [16], the correlation between packet delays and packet losses experienced by a continuous-media traffic source is considered based on a per-packet measurement of these metrics.

Time series approaches have also been proposed for measurement-based traffic modeling and prediction, mainly motivated by the *black box* solution promised. In [2], predictive models for video packet delay using auto-regressive models are presented. In [18] a variable bit rate (*VBR*) flow of probes and *ARX* (Auto-Regressive eXogenous) are applied for modeling the end-to-end delay, in which the packet delay process is captured based not only on the observed end-to-end packet delays, but also on the probe's inter-departure packet gap distribution. A measurement-based tool for traffic modeling and queuing analysis is developed in [12], which uses *CMPP* (Circulant Modulated Poisson Process) for a traffic model. A comparative analysis of network traffic prediction based on both *ARMA* and *MMPP* (Markov-Modulated Poison Process) models is presented in [20]. Studies focused on the impact of packet autocorrelation on traffic modeling can also be found. For instance, in [12], time series approaches are used to study the impact of packet autocorrelation on the queue response. In[1][11]methodologies for modeling autocorrelation functions for Long-Range-Dependent (*LRD*) and Short-Range-Dependent (*SRD*) traffic are presented. Tools for replicating actual network conditions in controllable environments, such as network emulators [8][22], are impaired by the necessity of capturing and injecting packet autocorrelation from measurements into the emulated network traffic. This is crucial to reproduce the actual observed network conditions for experimental purposes. For instance, NistNet[8], uses correlation coefficients, to a limited extend, to generated delay values for data streams.

The work presented here uses *ARMA* and *ARIMA* modeling techniques, which in fact reflect accurately the effect of packet autocorrelation. Autocorrelation of end-to-end packet delay itself is analyzed and modeled to determine the effects that network and stream characteristics have on it. A methodology for fitting *ARMA* and *ARIMA* models into end-to-end packet delay processes is also presented.

3. Using Time Series For End-to-End Packet Delay Modeling

A time series $\{X(t)\}$ is defined as a set of observations ordered sequentially in time [15]. A series of n observations can be viewed as a random process of the variables $X_1, X_2, \ldots, X_n$, sampled at, often equidistant, time intervals $t_1, t_2, \ldots, t_n$. Time series can be considered as the output of a dynamic system of which external input can not be observed [14]. There are two main goals of time series analysis; prediction and modeling. The former aims at forecasting future system output values. However, we are interested in latter, in which the properties of the series are summarized and its salient features characterized.

In general, time series modeling focuses on series that are not deterministic but contains a random component. If this random component is stationary, powerful techniques for modeling can be developed. *ARMA* models are widely used for this purpose. However, most time series data on the Internet are non-stationary or weakly stationary. For such cases there are methods which transform a non-stationary series into a stationary one. In most cases first and second-order differencing are sufficient to remove any kind of trend existing in a time series [15]. *ARIMA* methodology is based on such an idea [5].

An *ARMA* model can be viewed as a special case of *ARIMA* models. *ARMA* and *ARIMA* are considered passive *black box* approaches in which model identification relies solely on the data, without prior information of the system that generated the data. *ARMA* and *ARIMA* models fit very well into the study of Internet data packets, since very little information is known to build up the state of the system, due to the complexity of the networks [24]. Identification and developing of *ARMA* and *ARIMA* models rely on *ACF* and *PACF* distribution and coefficients. These concepts are presented and associated with the Internet end-to-end packet delay processes next.

3.1. ACF and PACF for end-to-end packet delay.

ACF and *PACF* functions play an important role on time series modeling and prediction, since they provide useful measures of the degree of dependence between the sampled values at different times. *ACF* of a random process describes the correlation between the process at different points in time. Informally, *ACF* is a measure of how well a series matches a time-shifted version of itself, as a function of the amount of time shift. Sample *ACF*, $\hat{\rho}_h$, is defined as $\hat{\rho}_h = \frac{\hat{\gamma}_h}{\hat{\gamma}_0}$. Where $\hat{\gamma}_h$ is the sample auto-covariance function at lag *h*, which is presented on equation (1).

$$\hat{\gamma}_h = \frac{\sum_{t=1}^{n-|h|} \left(X_{t+|h|} - \bar{X}_n \right)\left(X_t - \bar{X}_n \right)}{n} \quad (1)$$

where $-n < h < n$, $\bar{X}_n$ is the mean of the observed time series $\{X(t)\}$, and *n* is the number of data samples. *ACF* for end-to-end packet delay process denotes the amount of dependency the delay of the current packet has to previous packets. Contrary to *ACF*, *PACF* is used to measure the degree of association between the current sample of the series, X_t, and a previous sample, X_{t-k} ,when the effect of the other *k-1* time lags is removed [21]. *PACF* can be considered as the amount of correlation between a variable and a lag of itself that is not explained by correlations at all *lower-order*-lags. In practice, *ACF* of a end-to-end packet delay time series, $\{D(t)\}$, at lag 1 is the coefficient of correlation between D_t and D_{t-1}, which is also most likely to be the correlation between D_{t-1} and D_{t-2}. However, if D_t is correlated with D_{t-1}, and D_{t-1} is equally correlated with D_{t-2}, it has to result in a correlation between D_t and D_{t-2}. Thus, the correlation at lag 1 propagates to lag 2 and most probably to higher-order lags. The *PACF* at lag 2 is therefore the difference between the actual correlation at lag 2 and the expected correlation due to the propagation of correlation at lag 1[7].

Previous analysis have observed the relationship between the probe's sending rate, as a fraction of the available link capacity, and the end-to-end packet delay *ACF* [19] and *PACF* [2] distributions. Degree of link congestion itself depends on the sending bit rate and the available link capacity, and only sending bit rate can be controlled in an experiment. So when it is increased up to a point at which link congestion is perceived, packets get closer to each other and thus their correlation can be expected to become stronger. This effect manifests as a slower decay of the *ACF* function as the sending bit rate increases. Contrary to *ACF*, the *PACF* function decays towards zero faster as the sending bit rate increases[2]. The relationship between *PACF* and *ACF* is presented in equation (2):

$$\Phi = R_p^{-1} \Gamma_p \quad (2)$$

where Φ is the vector of the *PACF* coefficients, and R_p and Γ_p are presented in equations (3) and (4) respectively [7]. Note that the number of *PACF* coefficients, *p*, obtained through equation (2) depends on the number of *ACF* coefficients used.

$$R_p = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \cdot & \cdot & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \rho_2 & \cdot & \cdot & \rho_{p-2} \\ \rho_2 & \rho_1 & 1 & \rho_1 & \cdot & \cdot & \rho_{p-3} \\ \cdot & & & & & & \\ & & & & & & \\ & & & & & & \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \rho_{p-4} & & & 1 \end{bmatrix} \quad (3)$$

$$\Gamma_p = \left(\rho_1 \quad \rho_2 \quad \rho_3 \quad . \quad . \quad . \quad \rho_p \right)^T \quad (4)$$

For a *CBR* flow of probes, the scenario considered in this paper, end-to-end packet delay has a direct relationship to *IPG*, G_i, as shown in equation (5).

$$G_i = D_{i+1} - D_i \quad (5)$$

Thus a relationship between *ACF* and *PACF* distributions of *IPG* and the end-to-end packet delay can also be expected. This fact plays an important role on end-to-end packet delay modeling, since measuring *IPG* is less complex and more accurate. This alternative will be explained in detail in the Section 4.1.

3.2. ARMA and ARIMA model selection for end-to-end packet delay processes.

The main goal of Internet traffic modeling is to develop powerful models that represent closely the behavior and characteristics of the observed data values. The *black-box* modeling approach obtained through *ARMA* and *ARIMA* models is very appropriate for characterizing the impact of network on Internet traffic streams. Selection of the best model that accurately represents the observed process depends intrinsically on *ACF* and *PACF* characteristics. On one hand, a stationary series can be identified straightforwardly from *ACF* function, as their autocorrelation coefficients die out quickly. If this is not the case, the observed series has to be considered to be in the range of weakly stationary to non-stationary, depending on its degree of autocorrelation. In the former case, as well as for the stationary series, *ARMA* is suitable for representing the observed process. However, for non-stationary series, *ARIMA* models are better alternatives. Order of *ARIMA(p,q,d)* models are represented by their indexes. Where *p* and *q* indicate the order of the embedded *AR(p)* and *MA(q)* models, respectively. However *d* indicates the number of times the process has to be differentiated before it becomes a stationary one. *ARMA(p,q)* for modeling a $\{X(t)\}$ process is presented in equation (6).

$$\hat{X}_t - \sum_{i=1}^{p} \phi_i X_{t-i} = Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j} \tag{6}$$

where $\hat{X}_t$ is the best linear mean-square predictor of X_t based on the data up to time *t-1,* Z_t is assumed to be a sequence of independent and normal distributed random variables with zero and variance of σ^2 $\left(i.d. \sim \mathrm{N}(0,\sigma^2)\right)$, and ϕ_i and θ_j are the *AR* and *MA* coefficients, respectively. Note that *AR(p), MA(q)* models can obtained from equation (6) when taking *q*=0 and *p*=0, respectively. Equation (6) can also be written as $\phi(B)X_t = \theta(B)Z_t$, where $\phi(\bullet)$ and $\theta(\bullet)$ are the p^{th} and q^{th} degree polynomials,

$$\theta(B) = 1 + \theta_1 B + ... + \theta_q B^q \tag{7}$$

$$\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p \tag{8}$$

and **B** is the backward shift operator $\left(\mathbf{B}^j X_t = X_{t-j}\right)$ [7]. Following this methodology, *ARIMA(p,q,d)* models for non-stationary process are shown in equation (9).

$$\phi^*(B)\hat{X}_t \equiv \phi(B)(1-B)^d \hat{X}_t = \theta(B)Z_t \tag{9}$$

Note that *AR(p)* may be considered as a way of differentiation if model coefficients are close to unity [7], see equation (9); thus it is possible that *ARMA(p,q)* or *AR(p)* can represent non-stationary processes when coefficients are accurately selected. Note also that since *IPG* at the receiver side denotes the first differentiation, *d*=1, of the packet delay time series, $\{D(t)\}$, *ARMA(p,q)* of the *IPG* sequence is equivalent to *ARIMA(p,q,1)* of $\{D(t)\}$.

In practice *MA(q)* models are called *q-correlated* processes, as their *ACF* is reduce to zero for all lags greater than *q*. Hence, the *ACF* is a good indication of the *MA(q)* order. However for a strongly correlated series, *ACF* tails off but never approaches zero for any *q* values. In such cases, it is difficult to characterize the process based on *ACF* only. For these cases, *AR(p)* models are better alternatives. *AR(p)* models apply similar methodology as *MA(q)* models for identifying the model order but, on the contrary, use the *PACF* function[7]. However, in general *ARMA(p,q)* and *ARIMA(p,q,d)* models are often used in time series modeling, combining the benefits of the two previous models and often providing lower order models. Note that in general *ACF* and *PACF* functions are assumed to reach zero, or cut off, when lying within the 95% Confidence Interval [7].

3.3. Optimization criteria and fitting procedures for ARMA models.

Although *ARMA* and *ARIMA* model order can go has high as the number of available data samples, *n*, over-specified models may fail to distinguish the systematic effects of the data from its random effects[7][14]. Thus, it is our goal to find a model that fits accurately the observed sampled data values with the smallest number of parameters. Scoring methods have been developed to quantify the relative goodness-of-fit of statistical models for a given data. These methods add a penalty factor to the negative log-likelihood for each parameter of the fitted model. One of the widely used methods is the Akaike's Information Corrected Criterion (*AICC*). For an *ARMA(p,q)* process, the *AICC* score is computed by [7];

$$AICC = -2\ln\left(L(\tilde{s})\right) + \frac{2n \times (p+q+1)}{n-p-q-2} \tag{10}$$

where n is the sample size and $L(\tilde{s})$ is the Gaussian Likelihood of an *ARMA* process with *n* observations.

$$L(\tilde{s}) = \frac{1}{\sqrt{\left(2\pi\sigma^2\right)^n r_0 \ldots r_{n-1}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} \frac{\left(X_j - \hat{X}_j\right)^2}{r_{j-1}} \right\} \tag{11}$$

where $r_{t-1} = E\left(X_t - \hat{X}_t\right)^2 / \sigma^2$ and σ^2 is the white noise variance of the fitted model. $\hat{X}_t$ and X_t are the modeled and actual data samples at time *t,* respectively. The values *p, q* and σ^2 that maximize equation (11), $L(\tilde{s})_{\max}$,are called the Maximum Likelihood Estimators, which are interpreted as the *ARMA* parameter values most likely to be responsible for the observed data values. As a result, the model that has a lower *AICC* score is a better representation of the process than those with higher scores. After selecting the right order of the *ARMA* or *ARIMA* process, estimation of its parameters has to be done. Several techniques exist for this. *Yule-Walker* and *Burg* procedures apply to the fitting of pure autoregressive models, although the former can be adapted to models with $q > 0$ its performance is less efficient that when *q*=0. On the other hand, *Innovation* and *Hannan-Rissanen* algorithms are used also to provide preliminary estimates of *ARMA* parameters when $q > 0$. For pure autoregressive models *Burg's* algorithm usually gives higher likelihoods than the *Yule-Walker* equations. For pure moving average models the *Innovation* algorithm often gives slightly higher likelihoods than the *Hannan-Rissanen* algorithm. For mixed models the *Hannan-Rissanen* algorithm usually gives better fitting. Detailed information on the above mentioned techniques can be found in[7].

3.4. Diagnostic checking for ARMA and ARIMA models.

Prediction and modeling analysis using *ARMA* and *ARIMA* models typically judges the goodness of fit of a statistical model to a set of data by comparing the observed values with the corresponding predicted values obtained from the fitted model. It is known that if the fitted model is appropriate, then the residual should have properties consistent with those of a white noise sequence[7][14]. Residuals, $\hat{W}_t$, are defined to be the rescaled one-step predictor errors[7];

$$\hat{W}_t = \left(X_t - \hat{X}_t\right) / \sqrt{r_{t-1}} \tag{12}$$

To check the appropriateness of the model we can therefore examine the residual series and check that it resembles a $WN(0,1/n)$ sequence[7]. *ACF/PACF* distribution, histogram and data plot generated from the model residual can be compared to the expected generated by a $WN(0,1)$ sequence when evaluating the correctness of the model[18].

4. RESULTS

In this section, experiment setup, methodology and results of end-to-end packet delay modeling are presented. The data used for this analysis is from a previous study described in detail in [19]. Here *CBR UDP* traffic streams of 20,000 packets each, corresponding to 64 and 256 bytes packet size, were sent from California Polytechnic to Colorado State University, using Ixia 1600T chassis [25] at both sides. Average one-way delay was found to be 22 milliseconds. Experiments were run on consecutive days at the same time to maintain consistency. Non-peak times of the days were chosen for running the experiments to keep cross traffic within a narrow range. End-to-end packet delay and *IPG* values were collected for a variety of sending rates and packet sizes.

4.1. Methodology for fitting ARMA and ARIMA models into end-to-end packet delay processes .

Fitting an end-to-end packet delay series into the appropriate *ARMA* or *ARIMA* model depends on several factors, such as probe's packet sizes and probe's sending bit rate, as a fraction of the available link capacity. Moreover, collecting end-to-end packet delays requires clock synchronization techniques on both sides to prevent clock skew issues. Collecting *IGP* samples is presented here as an alternative. *IPG* not only avoids the synchronization dilemma between the sender and the receiver, but also represents an alternative for modeling non-stationary end-to-end packet delay processes, as was explained in Section 3.2. End-to-end packet delay model can be obtained afterwards, integrating the *IPG* model. Note that higher order of differentiation may be needed in some cases.

For both end-to-end packet delay and *IPG* modeling, *AR* and *MA* model order can be estimated by observing the *ACF* and *PACF* distributions, respectively. These in turn depend on probe's sending rate, as a fraction of the available link capacity, and probe's packet size, as will be seen in Section 4.2. However, *ARMA* and *ARIMA* represent a mixture of *AR* and *MA* models, and their orders are calculated through scoring methods. Thus a clear relationship of their orders to the probe's sending rate and packet size can not be expected, as it is on *AR* and *MA* models. In this paper the effect of the network induced on a *CBR* flow of probes, is captured by finding an optimum *ARMA/ARIMA* model that represents the observed probe's packet delay or *IPG* series. A comparison of packet delay and *IPG* modeling approaches is given for varied bit rate and packet size scenarios in the following subsection.

Note that for very low sending bit rate scenarios, as a fraction of the available link capacity, packet delay autocorrelation is very weak and thus its distribution may be enough to represent the process. However, this conclusion can only be reached after examining the *ACF*/*PACF* distributions of the series.

4.2. Modeling results.

Figure 1 shows the sample *ACF* of the end-to-end packet delay series, for different sending bit rates, using 64 bytes for packet size. The *ACF* distributions for low sending rates decay faster than ones coming from medium or high sending rates. *PACF* distributions are shown in Figure 2 for four packet delay traces, generated with four different sending bit rates; 0.25,1,30 and 70Mbps, also using 64 bytes for packet size. It can be seen from here that, as the lag increases all *PACF* coefficients diminish to zero faster than their corresponding *ACF,* as was anticipated on Section 3.1. However, we note that it takes a larger number of lags for the *PACF* to die off for the one generated by the smallest sending bit rate (0.25Mbps).

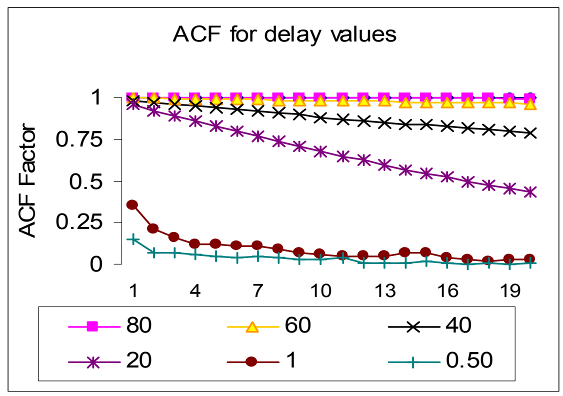

Figure 1. ACF function for lag 1-20 of packet delay values for varied sending bit rates (0.5 – 80Mbps) using 64 bytes packet size.

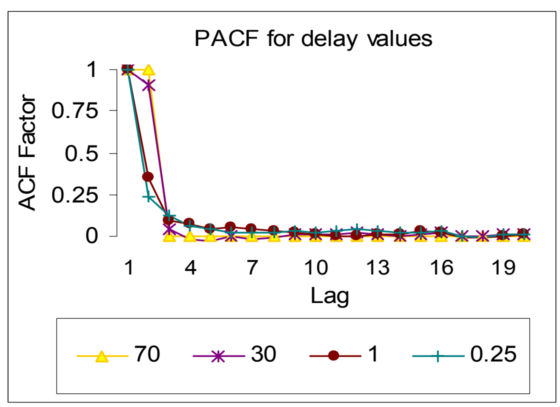

Figure 2. PACF function for lag 1-20 of packet delay values for varied sending bit rates(0.25,1,30 and 70Mbps) using 64 bytes packet size.

This can be explained since as sending bit rate increases, *IPG's* tend to decrease [19] and thus adjacent packets get closer to each other and more likely to be aligned together on the same buffer [2]. As a result correlation between adjacent delay samples becomes stronger. When applying *PACF,* this chain of dependency is broken (strong influence of intermediate samples is removed), consequently *PACF* coefficients will decays faster than their corresponding *ACF* ones. Since packet delay chain of dependency for high sending rates streams is stronger than for low ones, it can be expected that *PACF* coefficients tend to decay faster, as was anticipated in Section 3.1, and previously observed on [2].The inverse relationship between *ACF* and *PACF* coefficients (Φ and Γ_p) can also be seen mathematically from equations (2),(3)and (4).

End-to-end Packet Delay and IPG Modeling Fitting						
Sending Bit Rate	***ARMA* model for packet delay series**			***ARMA* model for *IPG* series**		
	p	***q***	$-2\ln(L(\tilde{s})_{\max})$	***p***	***q***	$-2\ln(L(\tilde{s})_{\max})$
0.5 Mbps	3	4	6.06E+03	4	5	6.12E+03
1 Mbps	6	6	4.35E+03	3	3	4.47E+03
10 Mbps	2	7	-2.33E+04	7	1	-2.32E+04
20 Mbps	1	7	-3.96E+04	7	5	-3.95E+04
30 Mbps	1	6	-4.27E+04	1	7	-4.26E+04
40 Mbps	7	5	-.486477E+05	7	7	-.485815E+05
50 Mbps	1	6	-.535692E+05	4	4	-.535711E+05
70 Mbps	1	0	-.799783E+04	1	0	-.800136E+04
80 Mbps	1	0	-.125418E+05	1	0	-.125485E+05

Table 1. End-to-end packet delay model fitting for different sending bit rate scenarios using 64 bytes packet size.

Table 1 shows results of the model fitting done for a set of sending bit rates scenarios for the experiment set up described above. Order of the *ARMA* models, as well as model parameters were obtained using the methodologies presented in Section 3 and equation (10). ITSM package is used for the model fitting [26]. Table 1 shows also the negative log-maximum likelihood estimator, $-2\ln(L(\tilde{s})_{\max})$, which represents the goodness-of-fit for the modeled series [7]. Model that has lower $-2\ln(L(\tilde{s})_{\max})$ is considered better fit for the analyzed series, as explained in Section 3.3. From Table 1, it can be seen that *ARMA* order selection for packet delay and *IPG* series show no clear connection to the sending bit rate, as was anticipated on Section 4.1. Although from Table 1, it can be seen that *IPG ARMA* models offer smaller negative log-maximum likelihood estimator than packet delay for scenarios with sending bit rate greater than 50 Mbps. Also as the sending bit rate increases, greater than 70Mbps for this experiment set up, model order for both series tend to decrease. On one hand, as sending bit rate increases, packet delay becomes a pure *AR* process since its *PACF* decays faster. For the same conditions *IPG* becomes less correlated [19]. In Table 1 the *ARMA* model with smaller $-2\ln(L(\tilde{s}))$, among the ones obtained through packet delay and *IPG* is highlighted.

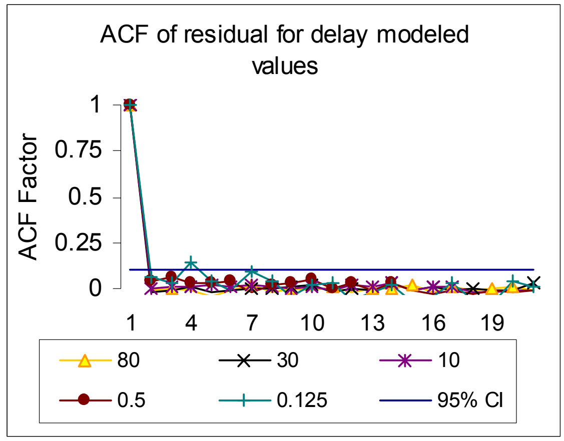

Figure 3.ACF function for residual of packet delay modeled values for sending bit rates(0.125,0.5,10,30 and 80Mbps) using 64 bytes packet size.

Modeling efforts presented in this paper aim to generate a modeled process which characterizes the overall behavior of an end-to-end packet delay and/or *IPG* process, and not to obtain a perfect match of this one at any given time. Thus modeling performance evaluation has been done by the applying the diagnostic checking methods presented in Section 3.4. Figure 3 shows the *ACF* distribution of the residual series, $\hat{W}_t$, for the packet delay models, see equation (12), together with the 95% Confidence Interval (CI)[7]. It is known that an *ARMA* model is considered an suitable fit to the data if all or almost all residual *ACF* coefficients for positive lags are inside the 95% CI [7][14]. Figure 3 shows that almost all residual are inside the confidence interval, indicating that end-to-end packet delay dynamics is well modeled by the fitted *ARMA* model. Also from Figure 3, it can be seen that *ACF* distribution of the residual series for the packet delay models die faster for high sending bit rates scenarios. This implies that residual resembles better a white noise sequence, and thus fitted model is more accurate. Table 1 confirms this, since $-2\ln(L(\tilde{s}))$ tends to get smaller as the sending bit rate increases.

Figure 4 and 5 show the *ACF* and *PACF* distributions for packet delays for varied sending bit rate scenarios for 256 bytes packet size. From here, it can be seen that autocorrelation of the process at different points in time is weaker than the one observed using 64 bytes packet size for the same probe sending rate. This is expected since sender *IPG* for these scenarios are 4 times bigger than with corresponding 64 packet size cases for the same probe stream bit rate. From Figure 5 it can be seen that *PACF* coefficients die off abruptly after a small number of lags, in fact, for sending bit rates higher than 10 Mbps *PACF* distribution cuts off after the second lag. This phenomenon tells us that the chain of dependency of intermediate samples is easily broken, see Section 3.1. Comparing Figure 5 to Figure 2, it can be seen that the chain of dependency of intermediate packet samples does not only depend on probe's sending bit rate but also on probe's packet size. Thus, it can be expected that auto-regressive models of packet delay and *IPG* series will vary according these two parameters, and thus both have to be considered when extracting the effect of the network on the captured packet stream.

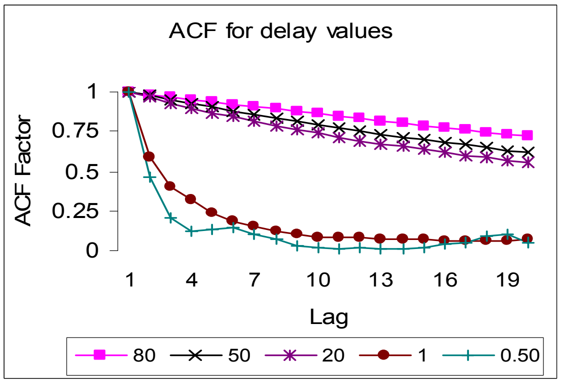

Figure 4.ACF function for lag 1-20 of packet delay values for varied sending bit rates (0.5 - 80Mbps) using 256 bytes packet size.

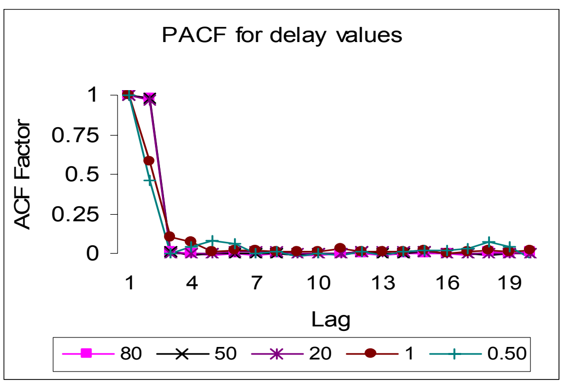

Figure 5. PACF function for lag 1-20 of packet delay values for varied sending bit rates (0.5 - 80Mbps) using 256 bytes packet size.

Table 2 shows the results of the model fitting done for a set of sending bit rates scenarios for 256 bytes packet size. Results were obtained in a similar manner to those in Table 2. Form Table 2 it can be seen that *ARMA* model orders, for both packet delay and *IPG* series, decreases rapidly for sending bit rates higher than 10Mbps. This can be expected from the *PACF* distribution observed on Figure 5. Also it can be seen that *IPG ARMA* models offer smaller negative log-maximum likelihood estimator than packet delay for scenarios with sending bit rate greater than 10 Mbps. By comparing results obtained from Table 2 to the ones observed on Table 2, it can be concluded that the packet delay model becomes an *AR* process at smaller sending data rates for the 256 bytes packet size experiment than for the 64 bytes packet size one. Also, in general *ARMA/ARIMA* packet delay and *IPG* models show lower orders for the 256 bytes packet size cases than their corresponding 64 bytes packet size cases. This can be explained since both *ACF* and *PACF* distributions decay faster for 256 bytes packet size than for 64 bytes packet size. From here it can be concluded that *ACF* and *PACF* distributions and therefore *ARMA/ARIMA* models depends not only on the probe's sending data rate but also on the probe's packet size. Also *IPG* has proven to be a feasible way of capturing the effect of the network on a data stream, and has demonstrated to offer better goodness-of-fit than packet delay models for high sending rate probes scenarios. This was anticipated earlier, since *IPG* becomes less correlated as sending bit rate increases [19]. Note that when both packet delay and *IPG* processes show very similar $-2\ln\left(L(\tilde{s})_{max}\right)$ values, process selection can be done based on the model order, *p*+*q*.

End-to-end Packet Delay and IPG Modeling Fitting

Sending Bit Rate	*ARMA* model for packet delay series			*ARMA* model for *IPG* series		
	p	***q***	$-2\ln\left(L(\tilde{s})_{max}\right)$	***p***	***q***	$-2\ln\left(L(\tilde{s})_{max}\right)$
0.5 Mbps	5	7	1.99E+05	5	7	2.00E+05
1 Mbps	4	1	1.80292E+05	3	1	1.80492E+05
10 Mbps	6	4	1.43592E+05	1	2	1.43871E+05
20 Mbps	1	0	1.6921E+05	1	2	1.67212E+05
30 Mbps	1	0	1.2275E+05	1	1	1.22264E+05
40 Mbps	1	0	1.4591E+05	1	1	1.4517E+05
50 Mbps	1	0	1.41872E+05	1	1	1.41851E+05
70 Mbps	1	0	1.37721E+05	2	1	1.37531E+05
80 Mbps	1	0	1.3579E+05	2	1	1.3521E+05
100 Mbps	1	0	1.3991E+05	2	1	1.3972E+05

Table 2. End-to-end packet delay model fitting for different sending bit rate scenarios using 256 bytes packet size.

Diagnostic checking for the 256 bytes packet size also shows packet delay dynamics are well modeled by the fitted *ARMA* model. Due to the space limitation these results are not presented in this paper. Although experiments with other packet sizes are not presented in this paper, relationship between packet correlation, packet size and process modeling can be inferred. For further information refer to [23].

Note that since the above experiments were conducted at non peak times of the day, network cross traffic remained within a narrow range. Thus non-stationarity of the observed packet delay samples was modeled successfully using *ARIMA(p,q,d)*, where *d={0-1}* , and not higher orders of *d* were needed. *d=0* represent the *ARMA* model of packet delay and *d=1* the *ARMA* process of *IPG*.

5. Conclusions and Projected Future work

This paper investigated the effects of correlation among delays of packets in a stream of Internet traffic and presented a system for modeling end-to-end packet delay and *IPG* by means of time series techniques. Packet delay autocorrelation plays an important role in capturing the effects of the network induced delay on the packets of an end-to-end flow, and it has to be included when packet traffic modeling efforts are intended.

Our finding shows that the behavior of end-to-end packet delay and *IPG* sequences can be captured effectively by *ARMA* and *ARIMA* models, when *CBR* probe flows are used. Autocorrelation and partial autocorrelation functions have been used to identify the most appropriate model. Effects of sending bit rate, packet size, and available link capacity on these two metrics have also been analyzed. Methodology for fitting *ARMA* and *ARIMA* models into packet delay and *IPG* series were presented. This methodology is based on the probe's sending bit rate, as a fraction of the available link, and the probe's packet size. Modeling *IPG* as an alternative for modeling end-to-end packet delay has also been considered. Results prove that *IPG* shows better goodness-of-fit than packet delay, for the same probe stream bit rate, as the combination of data rates and packet size increases.

Acknowledgement: The authors thank Dr. Nischal Piratla for the traffic data and his suggestions to this work.

REFERENCES

[1] D.K. Arrowsmith, R.J.Mondragón, J.M. Pitts, and M. Woolf, "Internet packet congestion", *IEEE International Symposium on Circuits and Systems*, Bangkok, Thailand, May 2003.

[2] A. C. Begen, Mehmet A. Begen and Yucel Altunbasak, "Predictive modeling of video packet delay in IP networks", *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Atlanta, GA, Oct. 2006.

[3] J.-C Bolot, "End-to-end packet delay and loss behavior in the Internet", *Proc. ACM SIGCOMM '93*, pp. 289-298, Sept. 1993.

[4] J.-C.Bolot, "Characterizing end-to-end packet delay and loss in the Internet," *Journal of High-Speed Networks*, vol. 2, pp. 305-323, Dec. 1993.

[5] G. E. P. Box, G. M. Jenkins,, and, G. C. Reinsel, "Time Series Analysis, Forecasting and Control", *3rd ed. Prentice Hall*, Englewood Clifs, NJ, 1994.

[6] L. S. Brakmo, S. W. O'Malley, and L. L. Peterson, "TCP Vegas: New techniques for congestion detection and avoidance," *in Proceedings of ACM SIGCOMM '94*, pp. 24-35, Oct. 1994.

[7] P. Brockwell and R. Davis , "Introduction to Time Series and Forecasting (2^{nd} ed.)", *New York: Springer-Verlag, 2002. ISBN 0-387-95351-1.*

[8] M. Carson., and D. Santay, "NIST Net - A Linux based network emulator tool", *ACM SIGCOMM Computer Communications Review,* Vo1.33, Jan 2003, pages 111-126.

[9] R. Jain, "A delay-based approach for congestion avoidance in interconnected heterogeneous computer networks," *ACM Computer Communication Review, vol. 19*, pp. 56-71, Oct. 1989.

[10] L. A. Kulkarni and S. Q. Li, ``Measurement-Based Traffic Modeling: Capturing Important Statistics" *Journal of Stochastic Model, Vo. 14, No. 5, 1998.*

[11] M. Li, W. Jia, and W. Zhao, "A Method for Modeling Autocorrelation Functions of Asymptotically LRD Traffic and Its Verification," *Conf. Proc., ICCT2000, 16^{th} IFIP World Computer Congress*, *IEEE Press*, 21-25, Aug., 2000, Beijing, China, pp. 62-65.

[12] S.-Q. Li, S. Park, and D. Arifler, "SMAQ: Ameasurement-based tool for traffic modeling and queueing analysis Part I - design methodologies and software architecture", *IEEE Communications Magazine*, vol. 36, Aug. 1998.

[13] S. Q. Li & C. L. Hwang, "Queue Response to Input Correlation Functions: Discrete Spectral Analysis," *IEEE/ACM Transactions on Networking*, vol. 1, No. 5, Oct. 1994, pp. 522-533.

[14] L. Ljung. "System Identification - theory for the user". *N.J.: Prentice Hall, Englewood Cliffs, 1987.*

[15] S. Makridakis, "A survey of time series", *Int. Stat. Rev*., 44(1):29-70, 1976.

[16] S. B Moon., J.Kurose, P. Skelly, and D. Towsley, "Correlation of packet delay and loss in the Internet", *Report UM-CS-1998-0I1,* UM-CS-1998-011, Mar. 1998.

[17] A. Nussi, A. Sridharan and N. Taft, " The problem of synthetically generating IP traffic matrices: Initial recommendations," *ACM SIGCOMM Computer Communication Review*, 35(3), 19-32July 2005.

[18] H. Ohsaki, M. Murata, and H. Miyahara, " Modeling end-to-end packet delay dynamics of the Internet using system identification". *In Proceedings of the International Teletraffic Congress 17*, pages 1027-1038, Dec. 2001.

[19] N. Piratla, A. P. Jayasumana and H. Smith, "Overcoming Effects of Correlation in Packet Delay Measurements using Inter-packet Gaps," *Proc. IEEE International Conference on Networks*, Singapore, November 2004, pp. 233-238.

[20] Aimin Sang and San-qi Li, "A predictability analysis of networktraffic", *In Proceedings of the 2000 IEEE Computer and Communications Societies Conference on Computer Communications (INFOCOM-00)*, pages 342–351, 2000.

[21] Christian Spielvogel. "Design and Implementation of a Network Resource Service for QoS Aware Servers", *Master's thesis*, University Klagenfurt, Institute of Information Technology, November 2003.

[22] D. Vivanco and A. P. Jayasumana, "OTRENET: Overall Trend Replicating Network Emulator Tool", *Proc. International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'05),* Philadelphia, PA, July 2005, pp.863-872.

[23] D. Vivanco, "Network Emulator Tool for Large IP Networks; Methodology, Implementation and Analysis", *Ph.D. dissertation,* Dept. of Electrical and Computer Eng., Colorado State University (work in progress).

[24] M. Yang, X.R. Li, H. Chen, and N.S.V. Rao, "Predicting Internet end-to-end delay: An overview", in Proc. of 36th *IEEE Southeastern Symposium on Systems Theory*, pp. 210-214, Atlanta, Mar. 2004.

[25] Ixia Incorporated webpage, http://www.ixiacom.com

[26] ITSM 2000 Professional Version 6.0, developed by Peter J. Brockwell and Richard (http://www.stat.colostate.edu/~pjbrock/).