ARCH-based Traffic Forecasting and Dynamic Bandwidth Provisioning for Periodically Measured Nonstationary Traffic

Balaji Krithikaivasan, Yong Zeng, Kaushik Deka, and Deep Medhi

Abstract— **Network providers are often interested in providing dynamically provisioned bandwidth to customers based on periodically measured nonstationary traffic while meeting service level agreements (SLAs). In this paper, we propose a dynamic bandwidth provisioning framework for such a situation. In order to have a good sense of nonstationary periodically measured traffic data, measurements were first collected over a period of three weeks excluding the weekends in three different months from an Internet access link. To characterize the traffic data rate dynamics of these data sets, we develop a seasonal AutoRegressive Conditional Heteroskedasticity (ARCH) based model with the innovation process (disturbances) generalized to the class of heavy-tailed distributions. We observed a strong empirical evidence for the proposed model. Based on the ARCH-model, we present a probability-hop forecasting algorithm, an augmented forecast mechanism using the confidence-bounds of the mean forecast value from the conditional forecast distribution. For bandwidth estimation, we present different bandwidth provisioning schemes that allocate or deallocate the bandwidth based on the traffic forecast generated by our forecasting algorithm. These provisioning schemes are developed to allow trade off between the underprovisioning and the utilization, while addressing the overhead cost of updating bandwidth. Based on extensive studies with three different data sets, we have found that our approach provides a robust dynamic bandwidth provisioning framework for real-world periodically measured nonstationary traffic.**

Index Terms— **Nonstationary traffic, Autoregressive conditional heteroskedasticity, Probability-hop forecasting, Heavy-tailedness, Bandwidth provisioning**

I. Introduction

In this paper, we attempt to address the following question: can we determine an effective dynamic bandwidth provisioning mechanism in a nonstationary traffic environment where measurements are collected periodically, say, every few minutes? Such an adaptive scheme is of importance to network providers in providing virtual services to customers while meeting service level agreements (SLAs) so that network providers can effectively allocate their resources to different customers. Such virtual services can be either based on the pseudo-wire concept or the virtual path concept in a packet-based network environment. There are two key issues to consider in answering this question: 1) any specific stochastic property of the traffic may not be available, 2) predicted bandwidth should not over-estimate (waste) too much and should avoid under-estimation (starvation) as much as possible; the starvation would certainly be based on service level agreements. An additional issue to consider is that the bandwidth update should not be frequent enough so that high *signaling* cost due to the bandwidth updates can be avoided. In general, a bandwidth update request in a network involves exchange of control messages as well as making appropriate entries in the provisioning and/or the billing system—this overhead is required to be minimized, if possible, and is referred to as the signaling cost here. We focus our work for a single customer's nonstationary, yet periodically measured traffic date rate on an end-to-end basis without any routing involved; that is, we are addressing dynamic provisioning at the time scale of minutes on a single virtual link basis.

It is well-known that network traffic has a cyclical behavior with a 24-hour cycle (while other seasonal variation from week to months are possible). In terms of bandwidth provisioning, an extreme case will be to consider the maximum traffic in a 24-hour period and determine the bandwidth needed based on this maximum traffic estimate; certainly, this will cause bandwidth update only once every 24 hours and the signaling cost is minimized to once-a-day update; on the other hand, the unused bandwidth other than during high traffic windows is under-utilized which could have been allocated, for example, to other customers supported in the same underlying network by the network provider. Thus, the end goal of this work is to develop a more accurate, dynamic bandwidth provisioning framework that can adapt to short-term traffic fluctuations (on the order of minutes) while adhering to the data loss (minimize starvation), utilization (reduce overprovisioning), and the signaling cost constraints.

There is limited work on understanding network dynamism in defining control strategies for dynamic bandwidth provisioning. Most such schemes in the literature are rooted on models that assumes that the nature/dynamics of the traffic is known. For example, there have been works [18], [6], [19] based on the Markovian assumption on the traffic flow arrival process for dynamic virtual path management in ATM networks. Groskinsky et. al [13] have investigated adaptive bandwidth control schemes for time-dependent Poisson traffic using a point-wise stationary fluid-flow approximation technique. In addition, for time varying traffic within an ATM virtual circuit, a hidden Markov model-based bandwidth estimation for future time interval has been proposed [1]. However,

Manuscript received June 21, 2004; revised October 8, 2005.

B. Krithikaivasan and D. Medhi are with the Department of Computer Science and Electrical Engineering, University of Missouri–Kansas City, MO 64110 USA (e-mail: bkrith@ieee.org, dmedhi@umkc.edu).

Y. Zeng is with the Department of Mathematics and Statistics, University of Missouri–Kansas City, MO 64110 USA (e-mail: zengy@umkc.edu).

K. Deka was with UMKC and is now with Demos Solutions, Norwell, MA 02061 USA.

none of these works consider a generically measured periodic traffic data and how to do predictive bandwidth estimate. The interest for generically measured data is further necessitated by the work on network traffic measurements documenting that Poisson models may not be appropriate for Internet traffic [21], and that the traffic over varied time scales exhibits self-similar properties and shows long range dependencies; for example, see [17]. Thus, it is important to consider the *lack* of the characterization of traffic dynamics/behavior as a factor in determining a *desirable* and a *more accurate* predictive bandwidth provisioning scheme which can be useful to network providers in meeting service level agreements with their customers.

To develop an adaptive bandwidth provisioning scheme and check its effectiveness, it is imperative that we use actual, measured data from a real-world environment. Toward this end, the collected measurements from the Internet link that connects the University of Missouri–Kansas City to MOREnet for connectivity to the rest of the Internet ("UMKC measured data") was found to be a good source of nonstationary, periodically measured data, especially due to our proximity to work with the campus network administrators. It may be noted that the packet level measurements of Internet traffic on a link collected on the granularity of minutes (say, averaged every 5 minutes) exhibit nonstationary, sometimes chaotic behavior; this observation is consistent with the findings reported in [17]. In such environments, statistical time series based techniques seem useful in practice. In particular, AutoRegressive Integrated Moving Average (ARIMA) time series models have been found to be appropriate in modeling nonstationary data traffic [12], [20] as well as in modeling time varying telephone traffic [8]. In our earlier work [16], we have considered an ARIMA model based on eight hours of data to characterize the traffic data rate and have proposed an augmented forecasting mechanism over the conventional Minimum Mean Square Error (MMSE) forecast for traffic prediction, in order to do dynamic bandwidth provisioning.

One of the issues that has not been addressed in [16] is to identify the ARIMA parameter re-estimation window. This is rather important since it is quite possible that the estimates of ARIMA parameters may change over a longer time duration. On further investigation, this was found to be true when a longer time window than eight hours was considered; see Appendix I. Thus, over time, such inadequacies can induce undesirable correlation in the innovations (noise terms) of the model which in turn affects the forecasts generated. Furthermore, the form of the model may change at times, due to the highly changing nature of the underlying traffic dynamics over, say 24 hours. Since the bandwidth prediction for the next time window depends heavily on the forecast generated from the ARIMA model, it is essential to address this issue in developing a more accurate bandwidth provisioning scheme. In essence, on further detailed investigation of the UMKC measured data (including additional data sets), it was observed that the measured data rate is *highly nonstationary of the second order*, having non-constant conditional variance.

Thus, our goal has been to develop a more robust model of the measured, periodic data overcoming these inadequacies that arise over time. In this regard, we sought out to characterize the traffic data rate behavior over a 24-hour window based on a history of 10-day measurements (only the week days). The intuition here is to take into account the *time-of-the-day* effect that is evidential from the collected measurements (see Fig. 1). This periodic effect can be easily incorporated in to ARIMA models by including multiplicative or additive seasonal autoregressive components. However, we made two interesting observations from the measurements that require *augmenting* the ARIMA model with additional components.

In the process of transforming the raw measurements into a stationary time series, we have noticed a clustering phenomenon in the resulting series (see Fig. 3). In other words, a large (small) noise term of either sign tends to follow by a large (small) noise term of either sign; Such an observation indicates the presence of non-zero correlation of second moments (square) of the innovation process. Consequently, the innovations of the underlying ARIMA model are no longer independent, but they are merely uncorrelated. To accommodate this behavior, a more accurate model than ARIMA became necessary. In this respect, a class of models called *AutoRegressive Conditional Heteroskedasticity (ARCH) models* used in Econometric modeling was found to be appropriate; ARCH models have been first introduced by Engle [11] to model time series with conditional non-constant variances, however with finite unconditional variance (in the asymptotic region). Since its introduction, considerable work has been done in the past two decades to integrate these models in the traditional ARIMA framework, especially for Econometric modeling. For excellent surveys on ARCH models, see [3], [4]. The observed non-constant conditional variance of traffic data rate can be explained by the inherent dependence of the amount of Internet activity (file uploads/downloads, number of websites accessed) on the time-of-the-day. Our other observation is with regard to the distribution of the noise/innovation process of the underlying ARIMA model. Based on the residuals, we found the distribution of the innovation process to depart significantly from normality, i.e., we found them to be relatively heavy-tailed rather than normally distributed. This behavior can be due to the heavy-tailed nature of the file size distribution being downloaded over Internet as reported in [9]. To accommodate this phenomenon into the innovation process of the model, we have introduced the choice of the Student-t distribution where the degree of heavy-tailedness can be controlled by the number of degrees of freedom. To summarize, in this paper, we present a new model of the periodically measured data which is an additive seasonal ARCH-based model with the innovation process generalized to the class of heavy-tailed distributions. Our conception is that such a model can, in general, be applicable in other network traffic environments.

Since our goal is to do dynamic bandwidth provisioning based on periodically measured nonstationary traffic, our next step is to generate forecasts using the proposed data model. Toward this end, we present an improved version of the probability-hop forecasting algorithm, first presented in [16]. We have developed this algorithm in an attempt to reduce the number of forecast misses as compared to the MMSE forecast. To accomplish our end goal of bandwidth provisioning, we map the forecast generated from the algorithm into a bandwidth requirement using various bandwidth quantization schemes. It can be argued that in the case of unpredictable, sudden high peaks that can arise occasionally due to non-regular traffic such as denial-of-service attacks, our approach tends to predict higher than the expected bandwidth for the next time window; this may not be desirable to do so in order to limit/restrict the bandwidth access to the non-regular traffic. Our scheme has a provision to limit the maximum allowable change in the bandwidth requirement from one time window to the next so that a network provider can impose such restrictions, if needed.

The rest of the paper is organized as follows. An overview of ARCH-based models is presented in Section II. In Section III, we describe the measured data sets used in our study followed by the statistical model identification. The MMSE forecast scheme is summarized in Section IV. We then present our modified probability-hop forecasting algorithm in Section V, showing effectiveness of the probability-hop forecast scheme as compared to the MMSE forecast scheme. Next, we describe several quantized bandwidth provisioning schemes in Section VI. Finally, in Section VII, we present evaluation results of our bandwidth provisioning framework followed by summary.

II. Overview of ARCH-based Models

The general form of AutoRegressive Integrated Moving Average model, ARIMA(p, d, q), for an observable random variable z_t can be given as follows:

$$z_t = \sum_{i=1}^{p+d} \varphi_i z_{t-i} + \sum_{k=1}^{q} \theta_k \varepsilon_{t-k} + \varepsilon_t \tag{1}$$

where $\{\varphi_i\}_{i=1}^{p+d}$ and $\{\theta_k\}_{k=1}^{q}$ are the autoregressive and the moving average parameters, respectively. Here, p stands for the autoregressive order, d for the order of differencing, and q for the moving average order. The innovation (disturbance) variable, ε_t, is assumed to be an independent and identically distributed normal random variable with mean 0 and variance σ^2. Thus, $E[\varepsilon_t^2|\mathcal{F}_{t-1}] = \sigma^2$ where $\mathcal{F}_{t-1}$ includes all the past information up to and including time $t-1$, i.e., the innovation variance is time independent.

The ARIMA model with conditionally heteroskedastic disturbances can be given by extending model (1) to allow the conditional variance of ε_t to change over time. In addition to $p + d + q$ parameters from the ARIMA model, conditionally heteroskedastic extension of order m introduces additional $m + 1$ parameters. An ARIMA(p, d, q)-ARCH(m) model can be expressed as follows:

$$z_t = \sum_{i=1}^{p+d} \varphi_i z_{t-i} + \sum_{k=1}^{q} \theta_k \varepsilon_{t-k} + \varepsilon_t \tag{2a}$$

$$\varepsilon_t = \eta_t \sqrt{h_t} \tag{2b}$$

$$h_t = \alpha_0 + \sum_{k=1}^{m} \alpha_k \varepsilon_{t-k}^2 \tag{2c}$$

where η_t is assumed to be an independent and identically distributed normal random variable with zero mean and unit variance. The additional $m + 1$ parameters are $\{\alpha_i\}_{i=0}^{m}$. Note that disturbances, ε_t, are assumed to be uncorrelated but not independent (higher moments may be correlated) unlike model (1), i.e., $E[\varepsilon_t \varepsilon_{t-1}] = 0$ and $E[\varepsilon_t^2 \varepsilon_{t-1}^2] \neq 0$. Under the given assumptions, it follows then that the conditional distribution of ε_t, given the past information up to and including time $t - 1$, is normal with mean 0 and variance h_t (time dependent).

These models can be further extended with additive or multiplicative seasonal components to take into account the periodic effect that may arise at times in a stochastic time series. The multiplicative seasonal models require a special structure on the underlying time series while the additive seasonal models are direct extensions. In particular, additive seasonal models can be treated as special cases of *subset* models where, in addition to the lag-dependency at neighboring points on the time axis, it includes parameters at those lags that are integer multiples of some period T. Thus, if there exists a periodic effect with period T in the underlying time series, (2a) can be replaced with

$$z_t = \sum_{i=1}^{p+d} \varphi_i z_{t-i} + \sum_{j=1}^{s} \phi_j z_{t-jT} + \sum_{k=1}^{q} \theta_k \varepsilon_{t-k} + \varepsilon_t \tag{3a}$$

where $\{\phi_j\}_{j=1}^{s}$ are additive seasonal autoregressive parameters with s being the seasonal autoregressive order; all other components of the model are the same as model (2).

For some observable time series, it may be the case that the conditional distribution of $(\varepsilon_t|\mathcal{F}_{t-1})$ follows a more general distribution (non-normal) from the class of symmetric distributions. For instance, to accommodate heavy-tailedness observed in network traffic, it deemed suitable to consider $(\varepsilon_t|\mathcal{F}_{t-1})$ to be Student-t distributed with scale $\mathcal{M}_t$ and ν degrees of freedom; then, model (2) becomes:

$$z_t = \sum_{i=1}^{p+d} \varphi_i z_{t-i} + \sum_{j=1}^{s} \phi_j z_{t-jT} + \sum_{k=1}^{q} \theta_k \varepsilon_{t-k} + \varepsilon_t \tag{4a}$$

$$\varepsilon_t = \eta_t \sqrt{\mathcal{M}_t} \tag{4b}$$

$$\mathcal{M}_t = h_t \Big(\frac{\nu - 2}{\nu}\Big) \tag{4c}$$

$$h_t = \alpha_0 + \sum_{k=1}^{m} \alpha_k \varepsilon_{t-k}^2 \tag{4d}$$

where η_t is Student-t distributed with unit scale, ν degrees of freedom, and m is the ARCH order. It may be noted that the conditional variance of ε_t remains at h_t. It is well known that as $\nu \rightarrow \infty$, the Student-t distribution approaches the standard normal distribution; in fact, the scale parameter, $\mathcal{M}_t$, in (4b) approaches h_t as $\nu \rightarrow \infty$. Thus, model (2) can be considered as a special case of model (4).

III. ARCH Model Fitting for Measured Data

A. Data sets

The measurements are collected on the Internet link connecting the University of Missouri-Kansas City to MOREnet (an Internet Service Provider) based on Multi Router Traffic Grapher (MRTG), a tool for monitoring traffic load on network

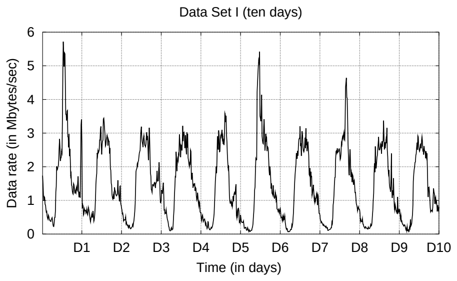

Fig. 1. Data set I: 10-day view

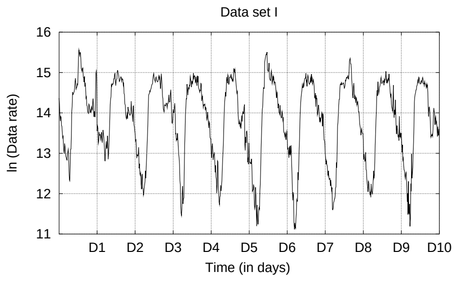

Fig. 2. Log Transformed Data

links [22]; these measurements represent the data rate of the traffic from outside world to UMKC averaged over every 5-minute interval (granularity). We have grouped the measurements into three data sets referred to as data set I, data set II and data set III. Each data set spans measurements over 24-hours for 15 days, excluding weekends and any holidays as our interest is to model for the weekday behavior. The measurements were collected in the months of September, November and December of year 2003. In Table I, we present the average and the standard deviation of the maximum data rate and the minimum data rate observed for each day over the first ten days in each data set (see also Fig. 1); the first ten days' data were used in determining the parameters' estimates and then, test our model for the next five days.

TABLE I
DATA SETS (UNITS: BYTES/SEC)

	Min. Data rate		Max. Data rate	
	Avg.	Dev.	Avg.	Dev
Data set I	143,543.9	91,470.93	3,869,781.7	1,008,676.05
Data set II	155,606.2	52,854.91	3,888,929.5	784,254.87
Data set III	150,372.5	53,542.56	3,375,274.0	409,973.24

B. Model Identification

First, we have averaged three successive sample points (of 5 minutes each) of the collected measurements to obtain the data rate over every 15-minute interval. Our intent behind this aggregation is under the assumption that a network provider may not want to do bandwidth updates more frequently then every fifteen minutes; this also gives us one-step prediction to look out for the next fifteen minutes. Note that this is done for the purpose of our study and the ARCH model presented is quite general. Over ten days, we thus consider 960 samples for model identification in each data set with 96 samples per day. In general, we denote the sample size per day using T; for this study, T happens to be 96. Fig. 1 shows a view of 15-minute aggregate data rate observed in data set I spanning ten days. A similar coarse-level behavior is observed in the other two data sets as well.

The first obvious observation from Fig. 1 is the evidence of the time-of-the-day effect. In a wider sense, the data rate behavior over any given weekday is similar to that of any other weekday. However, we have observed a few noticeable peaks on some days in the data sets. For example, relatively higher peaks can be observed on day 1, day 6 and day 8 than the rest in Fig 1. These peaks result in outliers (data points that are extreme relative to the rest of the sample) that can distort the innovation distribution as well as the statistical estimates of the parameters of the resulting model. Nevertheless, these outliers are genuine and essential in characterizing the traffic dynamics since they might arise due to, for example, some course submission deadlines to be met by students (which happens frequently) in an academic environment. Thus, we do not eliminate them in our analysis; rather, we transform the aggregated time series using the natural logarithmic transformation. Such a transformation is shown to be quite effective in stabilizing the data points (free of outliers). In fact, this stabilization effect can be clearly observed from Fig. 2 for data set I. Later, our goodness-of-fit results indeed confirm the transformation to be appropriate for the data sets. If z_t represent the aggregated time series (15-minute interval) then the log-transformed series is referred to as w_t, i.e., $w_t = \ln z_t$.

Following Box-Jenkins' homogeneous nonstationarity assumption [5], we obtain the first-difference of series w_t. On investigating the stationarity of the resulting series based on the auto-correlation function, we have observed a slow decay of the autocorrelation at integer multiples of seasonal lag T. Consequently, we do further difference of the first-differenced series at seasonal lags to obtain a new series y_t. That is, series y_t is given as follows:

$$y_t = (w_t - w_{t-1}) - (w_{t-T} - w_{t-T-1}). \qquad (5)$$

For illustration of series y_t for data set I, see Fig. 3. Evidently, the mean of series y_t appears to be stationary; however, on a closer look, we can observe the presence of clustering of large values and small values. In other words, the underlying large innovations and the small innovations of series y_t are clustered. This clustering effect arises from the strong dependence of the innovation variance at any given sample point over the innovation variance of past sample points, i.e., innovation variance is correlated. Such a dependency can be modeled by including a conditionally heteroskedastic component in an ARIMA model. In fact, (2b) - (2c) precisely captures such a clustering behavior.

In time series modeling, the distribution of forecasts generated relies critically on the distributional assumption of the underlying innovations (disturbances). It is a common practice to impose the normality assumption on the innovation distribution following the central limit theorem. However, the validity of the central limit theorem depends heavily on the

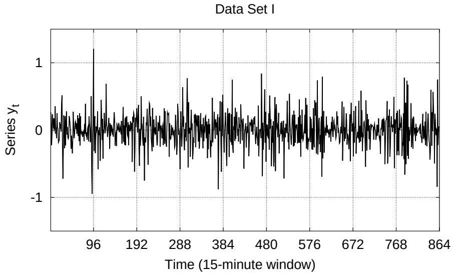

Fig. 3. Seasonal and First Differenced Log Transformed Data (Data Set I, nine days)

chosen asymptotic region. In fact, the innovation distribution, at times, follows a more general distribution from the class of symmetric distributions; this means that Gaussian assumption on the disturbances might lead to an inadequate model. We found that for our data sets, the normality test on the residuals proposed in [15] showed a strong disagreement on the Gaussian assumption. In addition, we found the sample kurtosis (a measure of tail behavior) to be around 5.0 revealing the heavy-tailed nature of the series. Thus, we consider the choice of the Student-t distribution to characterize the disturbances in our modeling exercise. Since, the degrees of freedom of the Student-t distribution is also estimated from the time series data through the maximum-likelihood estimation, the applicability of the Student-t distribution for the innovation process can be quite general here.

To summarize, we propose a model of the form (4) to characterize the dynamics of series y_t. For the statistical model identification, we used SAS, a well-known software package [23]. Based on the auto-correlation and partial auto-correlation function, we investigated various candidate models. Using Akaike Information Criterion (AIC) [2], Bayesian Information Criterion (BIC) [5] and the likelihood ratio tests, we have found appropriate choices of p, s, q and m to be 4, 2, 0 and 1, respectively, for all the data sets leading to a consistent fit with reference to model (4). Note that $s = 2$ implies a strong seasonal dependence of y_t on y_{t-T} and y_{t-2T}. Thus, our seasonal ARCH model for y_t takes the final form:

$$y_t = \sum_{i=1}^{4} \varphi_i y_{t-i} + \sum_{j=1}^{2} \phi_j y_{t-jT} + \varepsilon_t \tag{6a}$$

$$\varepsilon_t = \eta_t \sqrt{\mathcal{M}_t} \tag{6b}$$

$$\mathcal{M}_t = h_t \Big(\frac{\nu - 2}{\nu} \Big) \tag{6c}$$

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2. \tag{6d}$$

Since y_t is already a differenced series, d is not considered explicitly in (6). The unknown parameters of model (6) are: $\{\varphi_i\}_{i=1}^4$, $\{\phi_j\}_{j=1}^2$, α_0, α_1 and ν. These parameters are estimated based on the sample maximum likelihood function conditioned on the first $10T$ observations; see [14, Chapter 21] for more details. The estimates of all the parameters along with their t-values are listed in Appendix III. It may be noted that the estimates of α_1 and ν are *highly statistically significant* from zero (i.e., at the 99% confidence level to reject the null hypothesis that they are zero). Thus, the ARCH effect and the heavy-tailedness are justified. To evaluate the goodness of fit of the chosen statistical model for the data sets, we conducted the generalized portmanteau tests as well as the non-parametric rank tests proposed in [10]. The major advantage of these test statistics against the conventional Ljung-Box test statistic [5] is that they can be applied to non-normal time series as well. The p-values of chi-square statistics were in the range of 35%-50% (for all the data sets) confirming the adequacy of the chosen model. Such a well-fit statistical evidence set a solid empirical foundation for the forecasting and the bandwidth provisioning.

IV. MMSE Forecast Computation

In this section, we discuss MMSE forecast computation before presenting the probability-hop forecasting method in the next section. We will follow the convention that we have information up to time $t - 1$ and our aim is to forecast for time t. That is, we only consider a one-step forecast here in order to make use of the observations as soon as they are available. For example, for the specific data sets studied, the one-step prediction from the model translates to the expected bandwidth requirement over the next 15 minutes.

It is known that for any class of distributions with finite second order moments, the minimum mean square error (MMSE) forecast of a future observation y_t can be given by the conditional expectation $E[y_t|\mathcal{F}_{t-1}]$. The distribution of ε_t is irrelevant for such a derivation. Recall that series y_t from (6) is a differenced and a transformed series of the actual series z_t. Using the relation (5) and $w_t = \ln z_t$, we can rewrite (6a) in terms of $\ln z_t$ as follows:

$$\begin{aligned} \ln z_t = {} & \sum_{i=1}^{5} \chi_i \ln z_{t-i} + \sum_{j=T}^{T+5} \chi_j \ln z_{t-j} + \sum_{k=2T}^{2T+1} \chi_k \ln z_{t-k} \\ & + \sum_{s=3T}^{3T+1} \chi_s \ln z_{t-s} + \varepsilon_t \end{aligned} \tag{7}$$

where χ's can be computed from the estimates of φ and ϕ. Since we are interested in forecasting z_t's to be used by our bandwidth provisioning subsystem (discussed later in Section VI), it is imperative to derive the conditional expected value of z_t from w_t. Since $w_t = \ln z_t$, the forecast of z_t is given by $E[e^{w_t}|\mathcal{F}_{t-1}]$, the moment generating function (MGF) of the conditional distribution of w_t. It is evident from (7) that the conditional distribution of $w_t|\mathcal{F}_{t-1}$ follows a Student-t distribution. However, it is not possible to derive an exact expression for $E[z_t|\mathcal{F}_{t-1}]$ from $E[w_t|\mathcal{F}_{t-1}]$ since the MGF of the Student-t distribution is not well-defined. Thus, we propose here, the following approximation for the expected value of one-step ahead forecasts

$$\begin{aligned} E[z_t|\mathcal{F}_{t-1}] \simeq {} & \Big(1 + \chi_1' \ln z_{t-1} + \sum_{i=2}^{5} \chi_i \ln z_{t-i} \\ & + \sum_{j=T}^{T+5} \chi_j \ln z_{t-j} + \sum_{k=2T}^{2T+1} \chi_k \ln z_{t-k} + \sum_{s=3T}^{3T+1} \chi_s \ln z_{t-s} \Big) z_{t-1} \end{aligned} \tag{8}$$

where $\chi_1' = \chi_1 - 1$; the derivation is shown in Appendix II. In most cases, we found the error due to the first order Taylor approximation to be negligible. Henceforth, we refer to $E[z_t|\mathcal{F}_{t-1}]$ from (8) as the MMSE forecast for z_t.

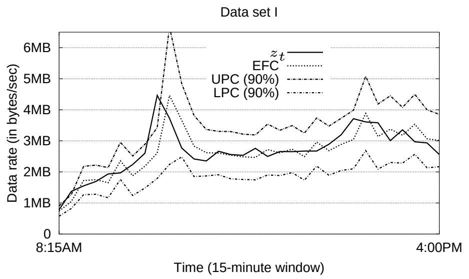

Fig. 4. Exact forecast curve along with 90% forecast limits, shown for part of Day-11 (from 8:15AM till 4:00PM)

V. Modified Probability-Hop Forecasting

The MMSE forecast, by definition, does not differentiate between the positive and the negative deviation from the unknown true series value of the future. In other words, the MMSE forecast is unbiased about the direction of the deviation. However, from the bandwidth provisioning perspective, forecast miss (under-forecast) is more undesirable than the over-forecast. Hence, there is a need to augment the MMSE forecast in *some* way to reduce the number of forecast misses. Our approach involves considering the probability limits (derived from the conditional forecast distribution) as well as the forecast errors from the last two periods. We consider the two recent forecast errors based on the assumption that in a dynamic traffic context, these errors reflect the transient nature of traffic dynamics well. Below, we present a modified version of an algorithm from our previous work [16].

Let $\hat{z}_{t-1}(1)$ denote the one-step forecast of z_t at time $t-1$. Then, the confidence limits (probability limits) of $\hat{z}_{t-1}(1)$ for model (6) can be given as follows (see Appendix II for details):

$$z_t(\pm) = \hat{z}_{t-1}(1) \pm t_{\nu,\frac{\beta}{2}} \sqrt{\mathcal{M}_t} z_{t-1} \qquad (9)$$

where $t_{\nu,\frac{\beta}{2}}$ is the deviate from the Student-t distribution with ν degrees of freedom corresponding to the $100(1-\beta/2)\%$ limits. We define three curves: Upper Probability Curve (UPC), Exact Forecast Curve (EFC) and Lower Probability Curve (LPC) with their values at time t given as $\hat{z}_{t-1}(1) + t_{\nu,\frac{\beta}{2}} \sqrt{\mathcal{M}_t} z_{t-1}$, $\hat{z}_{t-1}(1)$ and $\hat{z}_{t-1}(1) - t_{\nu,\frac{\beta}{2}} \sqrt{\mathcal{M}_t} z_{t-1}$, respectively. Note that EFC represent the MMSE forecast curve. Fig. 4 shows EFC, UPC and the LPC for 90% probability limits along with the actual measurements for part of Day-11 for data set I. The observed value z_t has been found to stay with in the 90% probability limits most of the time.

A. Algorithm

The fundamental idea behind the probability-hop forecast algorithm is to increase the *sensitivity of the forecast* toward sudden rise/drops by hopping among the UPC, the EFC and the LPC values at each forecast instant in an attempt to reduce the number of forecast misses as well as to reduce the extent of over-forecast. The curve on which the effective one-step forecast value (probability-hop forecast) falls at each time instant is decided based on the gradient of the weighted sum of the two most recent forecast errors with respect to *sensitivity quantum*. We define the sensitivity quantum in terms of bandwidth units. In particular, it represents a fraction of the total available bandwidth in the system. By taking into account the weighted sum of two recent forecast errors in deciding the effective forecast, we can conservatively address the sudden spikes and declines that cannot be effectively dealt otherwise. In order to be more flexible with the degree of sensitivity toward spikes and declines, we define *sensitivity-up-quantum* and *sensitivity-down-quantum* respectively rather than a single parameter. A positive weighted forecast error is measured against sensitivity-up-quantum whereas a negative weighted forecast error is measured against sensitivity-down-quantum. We refer to the forecast curve generated by the probability-hop forecast algorithm as the Probability-hop Forecast Curve (PFC). We have observed that the PFC value conform to the EFC value in regions that are not punctuated by high system perturbations, thus giving a good, close fit to the time series. However, if a sudden spike (decline) occurs, PFC will hop from EFC to UPC (LPC) in the next forecast instant. Without loss of generality, we set the significance level of gradient to be 1.0 in order to initiate a shift in the corresponding direction. A higher (lower) significance level can be achieved by increasing (decreasing) sensitivity-up(down)-quantum.

For clarity, notations used in the probability-hop forecast algorithm are summarized in Table II. In addition, we include UB and LB as upper and lower bounds on the system bandwidth, respectively. The formal description of the probability-hop forecast mechanism is presented in Algorithm 1. The version presented here differs from the one presented in [16] in two aspects. First, if PF$_{t-1}$ was at the UPC value, it will remain at the UPC value at time t if the gradient evaluated at time t is still greater than zero; if the gradient at time t is less than zero, PF$_t$ will shift to the EFC value. This adjustment in the algorithm further reduced the number of forecast misses without increasing the overall minimum mean square error significantly. Second, an adaptive mechanism is introduced here to determine convex weights, γ_{t-1} and γ_{t-2}, associated with forecast errors as compared to fixed values, γ_{t-1} = γ_{t-2} = 0.5, used in [16].

TABLE II
Notations

$\Delta\mathcal{S}_u$	Sensitivity-up-quantum (input parameter)
$\Delta\mathcal{S}_d$	Sensitivity-down-quantum (input parameter)
PF$_t$	Probability-hop forecast for time t made at time $t-1$
e_t	One-step probability-hop forecast error at time t (z_t − PF$_t$)
γ_t	Weight associated with the probability-hop forecast error at time t ($0 \leq \gamma_t \leq 1$)
ϑ	Weighted sum of the two recent probability-hop forecast errors
β	Critical value to derive $100(1-\frac{\beta}{2})\%$ limits of one-step forecast distribution (input parameter)

The role of convex weights, γ_{t-1} and γ_{t-2}, that satisfy $\gamma_{t-1} + \gamma_{t-2} = 1$ is to achieve smoothing of probability-hop forecast errors. It is evident from Algorithm 1 that the choice of these weights play a crucial role in controlling ϑ and thereby, creating the shift among the three forecast curves. Our adaptive approach to determine γ_{t-1} and γ_{t-2} is described in

Algorithm 1 *MoPHForecast*($\Delta\mathcal{S}_u$, $\Delta\mathcal{S}_d$, β, UB, LB)

```
PF$_{t-2} \leftarrow \hat{z}_{t-3}(1)$; PF$_{t-1} \leftarrow \hat{z}_{t-2}(1)$
while true do
  $e_{t-2} \leftarrow z_{t-2} - \text{PF}_{t-2}$; $e_{t-1} \leftarrow z_{t-1} - \text{PF}_{t-1}$
  $\gamma_{t-1} \leftarrow ComputeWeights(t-1, e_{t-1}, e_{t-2})$
  $\gamma_{t-2} \leftarrow 1.0 - \gamma_{t-1}$
  $\vartheta \leftarrow e_{t-1}\gamma_{t-1} + e_{t-2}\gamma_{t-2}$
  EFC$_{t-1} \leftarrow \hat{z}_{t-2}(1)$; EFC$_t \leftarrow \hat{z}_{t-1}(1)$
  UPC$_{t-1} \leftarrow \hat{z}_{t-2}(1) + t_{\nu,\frac{\beta}{2}}\sqrt{\mathcal{M}_{t-1}}z_{t-2}$
  UPC$_t \leftarrow \hat{z}_{t-1}(1) + t_{\nu,\frac{\beta}{2}}\sqrt{\mathcal{M}_t}z_{t-1}$
  LPC$_{t-1} \leftarrow \hat{z}_{t-2}(1) - t_{\nu,\frac{\beta}{2}}\sqrt{\mathcal{M}_{t-1}}z_{t-2}$
  LPC$_t \leftarrow \hat{z}_{t-1}(1) - t_{\nu,\frac{\beta}{2}}\sqrt{\mathcal{M}_t}z_{t-1}$
  if $\vartheta \geq 0$ then
    if $\frac{\vartheta}{\Delta\mathcal{S}_u} \geq 1.0$ then
      if ((PF$_{t-1}$=EFC$_{t-1}$ or PF$_{t-1}$=UPC$_{t-1}$) and UPC$_t \leq$ UB) then
        PF$_t \leftarrow$ UPC$_t$
      else
        PF$_t \leftarrow$ EFC$_t$
    else
      if (PF$_{t-1}$=LPC$_{t-1}$ or PF$_{t-1}$=EFC$_{t-1}$ or (PF$_{t-1}$ = UPC$_{t-1}$ and UPC$_t >$ UB)) then
        PF$_t \leftarrow$ EFC$_t$
      else
        PF$_t \leftarrow$ UPC$_t$
  else
    if $\frac{\vartheta}{\Delta\mathcal{S}_d} \leq -1.0$ then
      if (PF$_{t-1}$=EFC$_{t-1}$ and LPC$_t >$ LB) then
        PF$_t \leftarrow$ LPC$_t$
      else
        PF$_t \leftarrow$ EFC$_t$
    else
      PF$_t \leftarrow$ EFC$_t$
  $t \leftarrow t + 1$
end while
```

the next section.

B. Adaptive Determination of Weights: γ_{t-1} and γ_{t-2}

From (23) in Appendix II, the conditional distribution of one-step forecast error at time t is given by

$$z_t - E[z_t|\mathcal{F}_{t-1}] \sim \text{Student-}t(0, \mathcal{M}_t z_{t-1}^2, \nu) \quad (10)$$

It follows that the conditional distribution of one-step *probability-hop* forecast error at time t is also Student-t distributed with scale $\mathcal{M}_t z_{t-1}^2$ and ν degrees of freedom. However, the mean of such a distribution is zero only if PF$_t$ falls on the EFC value. Otherwise, the mean will be less than or greater than zero depending on whether PF$_t$ falls on the UPC value or the LPC value, respectively. Using the knowledge of the conditional error distribution, we obtain the weights as follows:

1) If e_{t-1} is the one-step probability-hop forecast error observed at time $t-1$, we compute the area, to be denoted by a_{t-1}, that is enclosed between $-|e_{t-1}|$ and $|e_{t-1}|$ from the conditional distribution. Similarly, we compute area a_{t-2} for e_{t-2}.
2) Set $\gamma_{t-1} = \dfrac{a_{t-1}}{(a_{t-1} + a_{t-2})}$, and thus, $\gamma_{t-2} = 1 - \gamma_{t-1}$.

This step is $ComputeWeights(\cdot)$ in Algorithm 1.

C. Algorithm Parameters: $\Delta\mathcal{S}_u$, $\Delta\mathcal{S}_d$ and β

We now discuss three crucial parameters, $\Delta\mathcal{S}_u$, $\Delta\mathcal{S}_d$ and β, that are input to Algorithm 1.

It may be recalled that both $\Delta\mathcal{S}_u$ and $\Delta\mathcal{S}_d$ represent a fraction of the maximum available bandwidth in the system except that we associate a direction with it. It follows from the definition that as we increase $\Delta\mathcal{S}_u$ ($\Delta\mathcal{S}_d$), the sensitivity of the PFC toward the UPC (LPC) decreases. Consequently, the PFC will follow the EFC closely. On the other hand, as we decrease $\Delta\mathcal{S}_u$ ($\Delta\mathcal{S}_d$), the PFC becomes highly sensitive to the transient network dynamics, thereby, capturing most of the sudden spikes (declines). As a result, the PFC will oscillate among the EFC, UPC and the LPC much more frequently leading to an increase in the mean square error of forecasts. Thus, a match between a low mean square error and a desired level of sensitivity is sought in the selection of the sensitivity quantum in each direction.

It is evident from Algorithm 1 that parameter β determines both the UPC and the LPC. The range of values for β could be set to any value from 0.1 to 0.99; the smaller the value, the larger will be the difference between the EFC and the UPC/LPC. Thus, a larger mean square error (MSE) will be observed than that of using EFC. At the same time, smaller values of β reduces the number of forecast misses. As a result, the choice of β can be guided by a desired balance between the fractional increase in the MSE and the fractional decrease in the number of forecast misses.

D. Algorithm Evaluation

In this section, we evaluate the modified probability-hop forecast against the MMSE forecast for the Day-11 data. We consider three metrics: (i) Fractional Increase in error due to Over-Forecast (FIOF), (ii) Fractional Decrease in error due to Forecast Misses (FDFM), and (iii) Difference in the number of Forecast Misses (DFM). Let the cumulative error observed due to the over-forecast and the under-forecast for the forecasting scheme S be denoted by CE_S^{of} and CE_S^{uf}, respectively. Furthermore, let NFM_S be the number of forecast misses observed for the forecasting scheme S. Then, the metrics are defined as follows:

$$\text{FIOF} = \frac{(CE_{PF}^{of} - CE_{MMSEF}^{of})}{CE_{MMSEF}^{of}} \quad (11)$$

$$\text{FDFM} = \frac{(CE_{PF}^{uf} - CE_{MMSEF}^{uf})}{CE_{MMSEF}^{uf}} \quad (12)$$

$$\text{DFM} = NFM_{MMSEF} - NFM_{PF}. \quad (13)$$

For our evaluation, we varied values of input parameters, $\Delta\mathcal{S}_u$ and $\Delta\mathcal{S}_d$, from 2%, 4%, 5%, 8% to 10%. In particular, we are primarily interested in the pairs, ($\Delta\mathcal{S}_u$,$\Delta\mathcal{S}_d$), with $\Delta\mathcal{S}_d \geq \Delta\mathcal{S}_u$. The consequence of having $\Delta\mathcal{S}_d$ lower than

TABLE III
PROBABILITY HOP FORECAST RESULTS

	Data Set I						Data Set III					
	50% Limits			90% Limits			50% Limits			90% Limits		
$\Delta\mathcal{S}_u$,$\Delta\mathcal{S}_d$	DFM	FDFM	FIOF	DFM	FDFM	FIOF	DFM	FDFM	FIOF	DFM	FDFM	FIOF
(2%,2%)	1	-14.23%	46.19%	13	-15.48%	191.17%	6	-24.89%	34.58%	14	-10.28%	202.00%
(2%,4%)	1	-14.23%	47.65%	14	-16.60%	193.96%	6	-26.19%	42.50%	16	-19.31%	208.44%
(2%,5%)	2	-18.14%	46.81%	14	-16.60%	193.96%	8	-27.81%	45.62%	18	-33.12%	214.65%
(2%,8%)	3	-20.29%	46.25%	8	-28.44%	168.64%	10	-32.73%	46.31%	19	-43.73%	203.22%
(2%,10%)	3	-20.29%	49.88%	9	-32.53%	165.31%	11	-34.09%	47.55%	20	-49.50%	197.54%
(4%,4%)	1	-14.23%	43.71%	14	-16.60%	183.05%	2	-16.42%	27.94%	12	-7.50%	187.70%
(4%,8%)	3	-20.29%	42.32%	8	-28.44%	157.73%	6	-22.66%	33.83%	13	-26.33%	154.73%
(4%,10%)	3	-20.29%	45.95%	9	-32.53%	154.39%	7	-24.02%	35.06%	14	-32.10%	155.97%
(5%,5%)	0	-13.22%	27.98%	6	-10.87%	134.69%	0	-12.95%	20.37%	12	-11.20%	160.45%
(5%,8%)	2	-16.28%	29.74%	7	-21.39%	136.70%	3	-19.11%	28.04%	10	-22.79%	129.77%
(5%,10%)	2	-16.28%	33.37%	8	-25.48%	133.36%	4	-20.47%	29.28%	11	-28.56%	131.00%

$\Delta\mathcal{S}_u$ is that it forces the probability-hop forecast to be inclined more toward the LPC than the UPC. Such a behavior has the direct implication of liberal bandwidth deallocation from the provisioning module leading to increased data loss. Then, the choices of β considered are 0.5 and 0.1 which translate into 50% and 90% confidence limits, respectively. Due to space limitations, we present our evaluation results for data set I and data set III in Table III.

From results shown in Table III, we have the following major observations: (i) in general, the difference in the number of forecast misses (DFM) is greater for the probability-hop forecast with 90% limits than with 50% limits, (ii) the difference in the FIOF between 50% limits and 90% limits is much greater than the difference in FDFM, and (iii) for a given β, increasing $\Delta\mathcal{S}_d$ with fixed $\Delta\mathcal{S}_u$ improves FDFM while increasing $\Delta\mathcal{S}_u$ with fixed $\Delta\mathcal{S}_d$ decreases FDFM. The first and second observations follow directly from the relatively larger enclosed area associated with 90% probability limits as compared to 50% probability limits. The third observation can be explained by the relative decrease in the sensitivity of the probability-hop forecast toward the LPC than toward the UPC and vice versa. Moreover, it may be noted from Table III that for some choices of ($\Delta\mathcal{S}_u$, $\Delta\mathcal{S}_d$), even if DFM is zero, there is a gain in FDFM, complemented by an increase in FIOF. Since the gain in FDFM translates to reduced loss, this effect is desirable. As a final note, it can be inferred from Table III that choosing 50% probability limits against 90% probability limits i.e., choosing β = 0.5 achieves a good balance between maintaining the mean square error closer to the MMSE and in reducing the number of forecast misses.

VI. QUANTIZED BANDWIDTH PROVISIONING SCHEMES

A predictive bandwidth provisioning scheme provisions bandwidth for a future time instant based on a forecasted bandwidth value from the forecasting subsystem as well as on the desired performance metrics reflecting the service level requirements of the traffic. In this section, we present bandwidth provisioning schemes that take as input the probability-hop forecast computed using *MoPHForecast*(.) presented in Algorithm 1. The underlying assumption of these provisioning schemes is that the bandwidth can be allocated or deallocated only in discrete units referred to as *bandwidth quantum*. We further assume that bandwidth quantum is expressed as a fraction of maximum available bandwidth on a link (similar to $\Delta\mathcal{S}_u$ and $\Delta\mathcal{S}_d$).

Suppose that PF_t represents the probability-hop forecast for time t, $\Delta\mathcal{B}$ represents the bandwidth quantum, and $\mathcal{C}_{max}$ represents the maximum available bandwidth in the system. Then, the *bandwidth requirement* at time t is determined as

$$\mathcal{B}_t = \min\left\{ \left\lceil \frac{\text{PF}_t}{\Delta\mathcal{B}} \right\rceil \times \Delta\mathcal{B}, \mathcal{C}_{max} \right\}. \tag{14}$$

That is, we choose either the upper bound of the interval $[(k-1)\Delta\mathcal{B}, k\Delta\mathcal{B}]$ ($k \geq 1$) in which PF_t falls, or, the maximum available bandwidth if PF_t falls outside the available bandwidth region.

In an attempt to reduce the short-term data loss and to increase the bandwidth utilization, an utopian provisioning scheme would provision according to $\mathcal{B}_t$ at each instant. However, such a scheme triggers frequent bandwidth updates and can lead to oscillatory/unstable behavior; such oscillatory behavior was reported earlier in [13] for time-varying Poisson traffic. In our case, frequent update also results in high signaling overhead. Thus, our provisioning schemes are also aimed at reducing the signaling overhead in addition to meeting the loss and the utilization constraints. Henceforth, we refer to the utopian scheme as the Instantaneous Bandwidth Provisioning ($\mathcal{IBP}$) scheme.

A. Stabilized Bandwidth Provisioning ($\mathcal{SBP}$)

The principle of this scheme is to allocate immediately when there is a need while following a conservative approach in deallocation. We achieve this conservative deallocation through maintaining a Hold Down Timer (HDT). The idea of using a timer to slow down the message exchanges has been previously used in computer operating systems to prevent thrashing, in routing protocols to reduce the communication overhead and oscillation, and so on. Let $\mathcal{SBPL}_t$ denote the bandwidth provisioned at time t using this scheme and t_l denote the last time instant at which the bandwidth was updated. Then,

$$\mathcal{SBP}_t = \begin{cases} \max\{\mathcal{B}_t, \mathcal{SBP}_{t-1}\} & (t-1) - t_l < \hat{t} \\ \mathcal{B}_t & (t-1) - t_l = \hat{t}. \end{cases} \tag{15}$$

Here $\hat{t}$ is the hold-down timer. As long as the timer is running, if an allocation is required, $\mathcal{B}_t$ as defined in $\mathcal{IBP}$ is considered. However, if deallocation is required, the value remains at $\mathcal{SBP}_{t-1}$ until the timer expires. Upon expiry of the timer, we use again $\mathcal{B}_t$ to update the bandwidth requirement.

B. Stabilized Bandwidth Provisioning with Local Maxima ($\mathcal{SBPL}$)

This scheme is a variant of the $\mathcal{SBP}$ scheme. The difference is only whether the bandwidth is to be released upon the expiration of the hold-down timer. If $\mathcal{BW}^{max}$ represent the local maximum of the bandwidth requirements (recorded in the last HDT period), then

$$\mathcal{SBPL}_t = \begin{cases} \max\{\mathcal{B}_t, \mathcal{SBPL}_{t-1}\} & (t-1) - t_l < \hat{t} \\ \max\{\mathcal{B}_t, \mathcal{BW}^{max}\} & (t-1) - t_l = \hat{t}. \end{cases} \tag{16}$$

Note that if $\mathcal{B}_t$ is less than $\mathcal{BW}^{max}$, then upon the expiry of the hold-down timer, we deallocate only up to $\mathcal{BW}^{max}$ rather than $\mathcal{B}_t$ as in the case of the $\mathcal{SBP}$ scheme.

VII. Results

In this section, we present results for our integrated provisioning framework. We first state the performance metrics used to evaluate the provisioning framework. Recall that z_t denotes the observed data rate; $\mathcal{BW}_t$ denotes the bandwidth provisioned at time t using one of the schemes.

A. Performance Metrics

1) *Average Utilization ($\mathcal{U}_{avg}$)*: The average utilization measures the fraction of bandwidth used to serve the data traffic. It is computed as follows:

$$\mathcal{U}_{avg} = \frac{1}{T}\sum_{t=1}^{T} \min\left\{\frac{z_t}{\mathcal{BW}_t}, 1.0\right\}. \tag{17}$$

2) *Average Under-Provisioning Ratio ($\mathcal{AUPR}_{avg}$)*: The average under-provisioning ratio gives a measure of the bytes dropped at the interface due to bandwidth under-provisioning. It is computed as follows:

$$\mathcal{AUPR}_{avg} = \frac{1}{T}\sum_{t=1}^{T} \max\left\{\frac{z_t - \mathcal{BW}_t}{\mathcal{BW}_t}, 0\right\}. \tag{18}$$

3) *Signaling Frequency ($\mathcal{SF}$)*: The signaling frequency helps to evaluate a bandwidth provisioning scheme in terms of how often bandwidth allocation/deallocation is done. It directly affects the signaling cost that will be incurred.

4) *Gain Factor ($\mathcal{GF}$)*: In our provisioning framework, the MMSE forecast generated from the time series model undergoes the first level of smoothing through ($\Delta\mathcal{S}_u$, $\Delta\mathcal{S}_d$) and the next level through $\Delta\mathcal{B}$. This cumulative smoothing effect produces a significant gain in terms of reduced data loss. As an attempt to quantify this gain, we define a gain factor relating the total number of under-provisioned instances with the under-forecast instances.

$$\mathcal{GF} = \frac{\sum_{t=1}^{T}(I_{\{z_t - \mathrm{PF}_t > 0\}} - I_{\{z_t - \mathcal{BW}_t > 0\}})}{\sum_{t=1}^{T} I_{\{z_t - \mathrm{PF}_t > 0\}}} \tag{19}$$

where $I_{\{a>0\}}$ assumes 1.0 when $a > 0$ and 0, otherwise.

5) *Cost Function ($\mathcal{CF}$)*: It is a linear weighted function that takes into account both the total slack bandwidth and the total under-provisioned bandwidth over the period T. Precisely, it captures the area between the bandwidth envelope and the forecast envelope. As it is undesirable to have under-provisioned bandwidth at any instant, we associate a cost factor $\tau \geq 1$ with the under-provisioned bandwidth. Then, it follows that

$$\mathcal{CF} = \sum_{t=1}^{T}(\mathcal{BW}_t - z_t)_+ + \tau\sum_{t=1}^{T}(z_t - \mathcal{BW}_t)_+ \tag{20}$$

where $(x-a)_+$ is equivalent to $(x-a)$ if $x \geq a$ and assumes 0 otherwise.

B. Evaluation

In this section, we evaluate the provisioning results using the above metrics. We consider the probability-hop forecast (PFC) as well as the MMSE forecast (EFC) in our evaluation. The choices of $\Delta\mathcal{B}$ are 2%, 4%, 5%, 8% and 10% and the choices of ($\Delta\mathcal{S}_u$, $\Delta\mathcal{S}_d$) pairs for the probability-hop forecast are based on results from Section V-D. For each data set, we use the first 10 weekday data for developing a time series model and consider the next 5 weekday data for evaluating the effectiveness of our approach. Due to space limitation, we present the evaluation results of our bandwidth provisioning framework only for the 11th Day.

1) Average Utilization and Loss Metrics: We primarily discuss results for the $\mathcal{SBP}$ scheme and the $\mathcal{SBPL}$ scheme. In general, the $\mathcal{IBP}$ scheme displays markedly high utilization and relatively higher loss as compared to the other two schemes when all the other parameters are being the same. Such a behavior is due to the increased sensitivity of the $\mathcal{IBP}$ scheme toward traffic fluctuations. For example, if we compare $\mathcal{AUPR}_{avg}$ of data set I for $\Delta\mathcal{B}$ = 2%, we observed an average gain (over various values of the pair $\Delta\mathcal{S}_u, \Delta\mathcal{S}_d$) of nearly 40% from the $\mathcal{SBP}$ scheme with respect to the $\mathcal{IBP}$ scheme. However, $\mathcal{U}_{avg}$ falls down only by 4% on average.

In Fig. 5 and Fig. 6, we present the average utilization ($\mathcal{U}_{avg}$) and the average under-provisioning ratio ($\mathcal{AUPR}_{avg}$), respectively, for the $\mathcal{SBP}$ scheme for all the three data sets. We have considered the hold-down timer to be 3 time units here. For any given $\Delta\mathcal{B}$ value, in all three data sets, $\mathcal{AUPR}_{avg}$ is relatively higher for symmetric ($\Delta\mathcal{S}_u$, $\Delta\mathcal{S}_d$) pairs i.e., (2%, 2%), (4%, 4%) and (5%, 5%) as compared to the other pairs. On the other hand, in general, $\mathcal{AUPR}_{avg}$ for pairs with $\Delta\mathcal{S}_d$ higher than $\Delta\mathcal{S}_u$ is relatively lower than that of the EFC value. The evidence for this behavior is more pronounced in data set III than the other two data sets. Furthermore, with all other

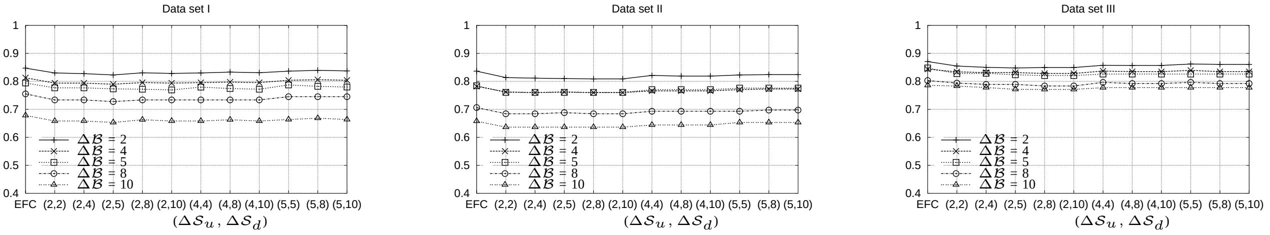

Fig. 5. $\mathcal{U}_{avg}$ for $\mathcal{SBP}$ scheme

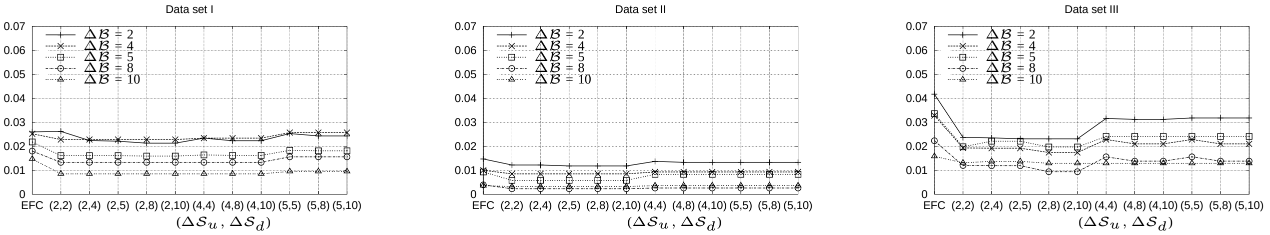

Fig. 6. $\mathcal{AUPR}_{avg}$ for $\mathcal{SBP}$ scheme

parameters being the same, the average under-provisioning ratio and the average utilization tend to be higher in data set III relative to the other data sets. This behavior can be explained as follows. The variance of the one-step forecast depends on estimates of α_0 and α_1 from (6); this, in turn, affects the bandwidth envelope. Since the estimate of α_1 for data set III (see Table VI) is almost twice that of the other data sets, the mean square error tends to be significantly higher relative to the other data sets, and hence, an increased chance of forecast misses (thereby under-provisioning).

In the case of the $\mathcal{SBPL}$ scheme, we observed a similar behavior as that of the $\mathcal{SBP}$ scheme with respect to all three data sets. However, in general, we noticed a decrease in $\mathcal{AUPR}_{avg}$ and a decrease in $\mathcal{U}_{avg}$ relative to the other two schemes due to the conservative deallocation policy.

2) Provisioned Bandwidth Envelope: In order to present a picture at a finer granularity (as the series evolves), we present the provisioned bandwidth envelope along with the observed data rate z_t on the eleventh day (for data sets I and III) for $\Delta\mathcal{B}$ = 2% and $\Delta\mathcal{B}$ = 10% in Fig. 7-10, respectively. The x-axis represents the time index based on the 15-minute time window with “D11” on the x-axis representing the 96th sample on the eleventh day. From the results above, we observed a greater similarity between data set I and data set II than data set III. Thus, we restrict ourselves to present the envelope for data sets I and III. Furthermore, the value of pair ($\Delta\mathcal{S}_u$, $\Delta\mathcal{S}_d$) is restricted to (2%, 10%) since our intent here is to compare the provisioning schemes.

In the regions of traffic rise, all schemes respond in the same manner in order to minimize the data loss. However, as it can be observed, these schemes respond very differently to the fall in the traffic data rate. For example, in Fig. 7, at t = 41, there is a big upward jump in the bandwidth envelope with all three schemes. Following this, at t = 42, the bandwidth envelope associated with the $\mathcal{IBP}$ scheme falls while it does not with the other two schemes. After three time periods (the duration of the hold-down timer), the bandwidth envelope of the other two schemes falls down. It may be noted that, with the $\mathcal{SBPL}$ scheme, the drop is only up to the local maximum during the previous HDT period. A similar trend can be observed in Fig. 9 for data set III in the neighborhood of t = 53. When $\Delta\mathcal{B}$ is increased from 2% to 10%, we observe a relatively lesser fluctuation in the bandwidth envelope, less frequent bandwidth updates (see Fig. 8 and Fig. 10). However, the trend among the schemes remains similar to that of $\Delta\mathcal{B}$ = 2%.

3) Results on Signaling Frequency: We present results for $\Delta\mathcal{B}$ = 2% and $\Delta\mathcal{B}$ = 10% in Table IV for data set I. Our intent here is to summarize the impact of increasing/decreasing $\Delta\mathcal{B}$ on the signaling overhead with respect to the provisioning schemes. Hence, we omit the results for data sets II and III. From Table IV, it can be observed that for $\Delta\mathcal{B}$ = 2%, the $\mathcal{SBP}$ scheme incurs less significant overhead than the $\mathcal{IBP}$ scheme. Furthermore, for the $\mathcal{SBPL}$ scheme, subsequent reduction on the overhead is evident here attributing to the more conservative deallocation policy. For $\Delta\mathcal{B}$ = 10%, a similar trend can be noticed, however, with less intensity. When we consider the impact of change in the $\Delta\mathcal{S}_u$ and the $\Delta\mathcal{S}_d$ on the signaling overhead, we observe a more oscillatory behavior with respect to all the schemes. In general, an increase in $\Delta\mathcal{S}_u$ brings out a gradual reduction in the overhead. Additionally, it can be pointed out that the signaling overhead due to using the EFC (MMSE forecast) is lesser than that of using the probability-hop forecast for a few of the ($\Delta\mathcal{S}_u$, $\Delta\mathcal{S}_d$) choices. It may be recalled that the data loss is maximum with the EFC across all provisioning schemes (see Fig. 6 for data set I). Moreover, the provisioning schemes are *not* designed to minimize the signaling overhead, rather to guarantee an acceptable under-provisioning ratio while maintaining a reasonably high utilization and hence, the observed behavior.

4) Results on Smoothing Effect: In order to illustrate the smoothing effect of the modified probability-hop forecast induced by the bandwidth provisioning module, we present $\mathcal{GF}$ for data set I in Table V for two extreme values of $\Delta\mathcal{B}$. For $\Delta\mathcal{B}$ = 2%, an average gain of around 50% is achieved in terms of avoiding under-provisioning for the $\mathcal{IBP}$ scheme. For the other two schemes, the gain further improves to around 80%. The implication here is that a forecast miss (under-forecast) need *not* lead to under-provisioning all the time. Such a gain is

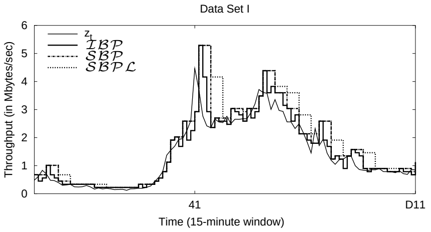

Fig. 7. Provisioning for Day-11, Data Set I ($\Delta\mathcal{B}$ = 2%)

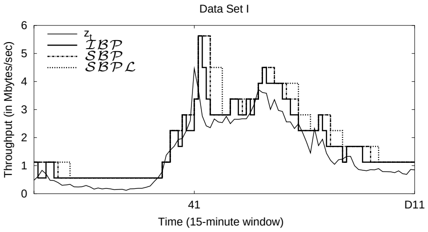

Fig. 8. Provisioning for Day-11, Data Set I ($\Delta\mathcal{B}$ = 10%)

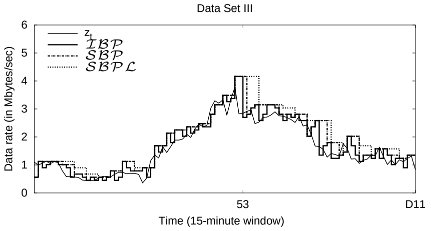

Fig. 9. Provisioning for Day-11, Data Set III ($\Delta\mathcal{B}$ = 2%)

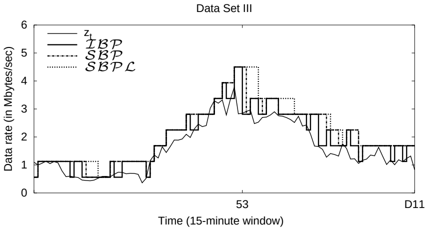

Fig. 10. Provisioning for Day-11, Data Set III ($\Delta\mathcal{B}$ = 10%)

TABLE IV
SIGNALING FREQUENCY (DATA SET I)

	$\Delta\mathcal{B}$ = 2%			$\Delta\mathcal{B}$ = 10%		
$\Delta\mathcal{S}_u$,$\Delta\mathcal{S}_d$	$\mathcal{IBP}$	$\mathcal{SBP}$	$\mathcal{SBPL}$	$\mathcal{IBP}$	$\mathcal{SBP}$	$\mathcal{SBPL}$
EFC	67.71%	37.50%	34.38%	32.29%	19.79%	14.58%
(2%,2%)	67.71%	39.58%	36.46%	34.38%	21.88%	19.79%
(2%,4%)	69.79%	38.54%	35.42%	34.38%	21.88%	19.79%
(2%,5%)	70.83%	39.58%	35.42%	35.42%	23.96%	16.67%
(2%,8%)	70.83%	37.50%	35.42%	32.29%	21.88%	19.79%
(2%,10%)	70.83%	37.50%	35.42%	32.29%	21.88%	19.79%
(4%,4%)	69.79%	38.54%	35.42%	34.38%	21.88%	19.79%
(4%,8%)	70.83%	37.50%	35.42%	32.29%	21.88%	19.79%
(4%,10%)	70.83%	37.50%	35.42%	32.29%	21.88%	19.79%
(5%,5%)	68.75%	37.50%	33.33%	30.21%	18.75%	18.75%
(5%,8%)	69.79%	36.46%	33.33%	29.17%	18.75%	18.75%
(5%,10%)	69.79%	36.46%	33.33%	29.17%	18.75%	18.75%

offset by some increase in the slack bandwidth. As we increase $\Delta\mathcal{B}$ from 2% to 10%, the effect is more pronounced.

TABLE V
SMOOTHING EFFECT OF $\Delta\mathcal{B}$ - $\mathcal{GF}$ (DATA SET I)

	$\Delta\mathcal{B}$ = 2%			$\Delta\mathcal{B}$ = 10%		
$\Delta\mathcal{S}_u$,$\Delta\mathcal{S}_d$	$\mathcal{IBP}$	$\mathcal{SBP}$	$\mathcal{SBPL}$	$\mathcal{IBP}$	$\mathcal{SBP}$	$\mathcal{SBPL}$
(2%,2%)	54.00%	68.00%	80.00%	74.00%	88.00%	92.00%
(2%,4%)	54.00%	70.00%	80.00%	74.00%	88.00%	92.00%
(2%,5%)	57.14%	71.43%	81.63%	75.51%	87.76%	93.88%
(2%,8%)	58.33%	72.92%	79.17%	77.08%	87.50%	91.67%
(2%,10%)	58.33%	72.92%	79.17%	77.08%	87.50%	91.67%
(4%,4%)	54.00%	70.00%	80.00%	74.00%	88.00%	92.00%
(4%,8%)	58.33%	72.92%	79.17%	77.08%	87.50%	91.67%
(4%,10%)	58.33%	72.92%	79.17%	77.08%	87.50%	91.67%
(5%,5%)	56.86%	68.63%	78.43%	74.51%	86.27%	90.20%
(5%,8%)	59.18%	71.43%	77.55%	75.51%	85.71%	89.80%
(5%,10%)	59.18%	71.43%	77.55%	75.51%	85.71%	89.80%

5) Results on Cost Function: To evaluate the overall utility of our bandwidth provisioning framework, we present results of the cost function ($\mathcal{CF}$) for τ = 1, 5, 10, 25, 50 and 100. Our objectives are twofold here: 1) to show the impact of τ on the cost function, 2) to show the relative reduction in the cost (as we defined here) that can be achieved through the probability-hop forecast (PFC) over the MMSE forecast (EFC). Adhering to these objectives, for various τ, we present the fractional increase/decrease in the cost due to using the PFC, with respect to the cost incurred with using the EFC. Thus, in Fig. 11 and Fig. 12, we represent $\mathcal{CF}_{\text{EFC}}$ using the zero axis (reference axis). A shift on the negative (positive) direction for a given τ and ($\Delta\mathcal{S}_u$, $\Delta\mathcal{S}_d$), indicates a decrease (increase) in its cost relative to the EFC.

It can be inferred from Fig. 11 that for the $\mathcal{IBP}$ scheme, except for τ = 1, the cost incurred from the framework (using the probability-hop forecast) is still less than the cost due to using the MMSE forecast for provisioning. When we choose τ = 1, we associate equal penalty with both the under-provisioned bandwidth and the slack bandwidth and hence, the cost is dominated by the slack bandwidth. In fact, it can be recalled from Table III that the fractional increase in error due to over-forecast (FIOF) is much higher than the fractional decrease in error due to forecast misses (FDFM) almost all the times. Hence, an increase in the cost can be expected. With the $\mathcal{SBP}$ and the $\mathcal{SBPL}$ schemes, we have an increased cost (with respect to the EFC) for even τ = 5 in addition to τ = 1. Since, these two schemes follow a conservative deallocation policy, the total amount of slack bandwidth will be in general more than that of the $\mathcal{IBP}$ scheme thereby, contributing to an increase in the cost. With increasing $\Delta\mathcal{B}$ values, the average under-provisioning ratio decreases to a greater extent and hence, its contribution to the cost function decreases reducing the overall cost (see Fig. 12 for $\Delta\mathcal{B}$ = 10%). Among various choices of $(\Delta\mathcal{S}_u, \Delta\mathcal{S}_d)$ for a given τ, we notice a lesser cost for lower magnitudes of $\Delta\mathcal{S}_u$ and $\Delta\mathcal{S}_d$. This can be attributed to the decreased sensitivity of the probability-hop forecast envelope toward the UPC or the LPC on increasing $\Delta\mathcal{S}_u$ values.

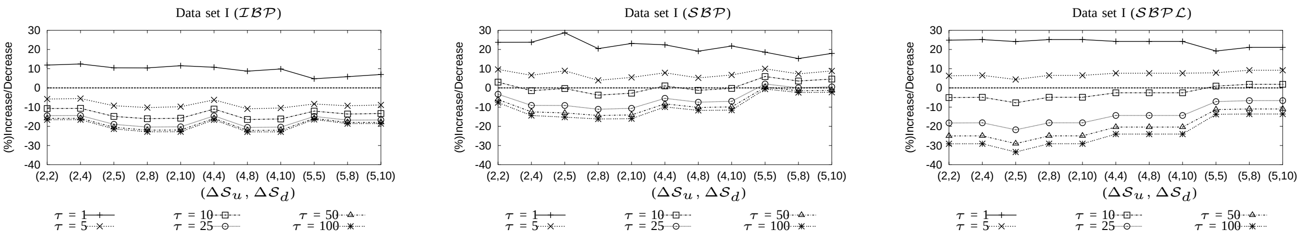

Fig. 11. Fractional Increase/Decrease in $\mathcal{CF}$ with respect to $\mathcal{CF}_{\text{EFC}}$ ($\Delta\mathcal{B}$ = 2%)

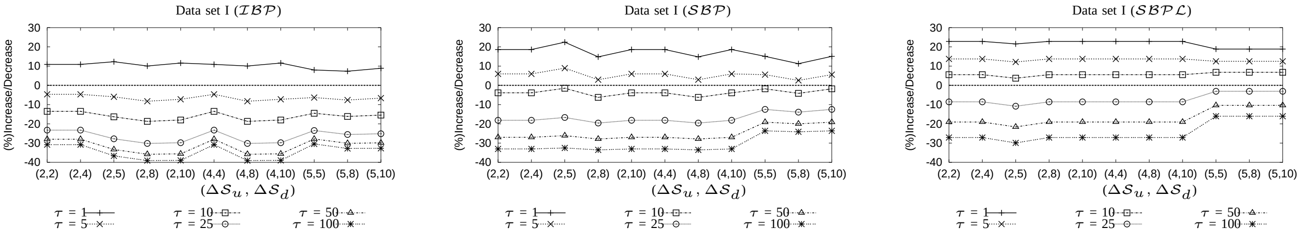

Fig. 12. Fractional Increase/Decrease in $\mathcal{CF}$ with respect to $\mathcal{CF}_{\text{EFC}}$ ($\Delta\mathcal{B}$ = 10%)

C. Controlling Bandwidth Increase

So far we have presented evaluation of our provisioning framework through various metrics. In this section, we discuss the impact of setting a maximum limit on the allowable increase in the bandwidth requirement from one time window to the next. This exercise is aimed to control any *liberal* bandwidth allocation that can result in the case of unpredictable high traffic spikes due to denial-of-service attacks.

Under the assumption that such events can be detected through an external anomaly detection system, our provisioning framework can react to it by setting a limit on the maximum allowable fractional increase from the current bandwidth provisioned. In Fig. 13, we illustrate this in the case of data set I for $\Delta\mathcal{S}_u$ = 2%, $\Delta\mathcal{S}_d$ = 10% and $\Delta\mathcal{B}$ = 2%. For illustration, we have set the bandwidth increase limit here to 60%. For comparison, we have included the bandwidth envelope in the unrestricted mode where there does not exist any limit on the allowable increase in the bandwidth. It may be noted that such a limit is irrelevant in the case of deallocation of the bandwidth (bandwidth decrease). We note that there are two time instants ($t = 35, 42$) at which the fractional increase in the bandwidth requirement from the previous time instant is more than 60%. For example, at t = 42, the bandwidth rise with the $\mathcal{SBP}$ scheme in the unrestricted mode is higher than that of the restricted mode. In spite of the limit on the maximum allowable change, the average under-provisioning ratio remains the same as that of the unrestricted case (without any limit). However, the average utilization increased slightly. If the maximum limit is further reduced, we may observe an increase in the average under-provisioning ratio. In general, our framework is flexible enough to incorporate restrictions on the maximum allowable increase of the bandwidth thereby, addressing the rare undesirable events. Furthermore, our provisioning schemes can be tuned to toggle between the restricted and the unrestricted mode in real-time depending on the input from the anomaly detection system.

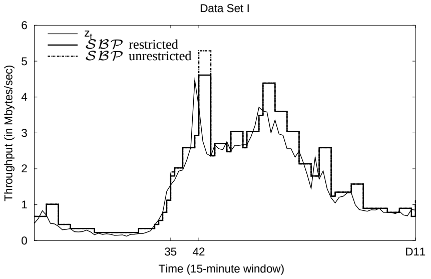

Fig. 13. Comparison between $\mathcal{SBP}$ unrestricted and $\mathcal{SBP}$ restricted

VIII. Summary

In our pursuit to answer a basic question on whether an effective dynamic provisioning scheme can be developed for nonstationary periodically measured data lacking information about the underlying stochastic process, we have made several key contributions:

- On investigating real-world nonstationary periodically measured data, we found that such data has nonstationarity of the second order. It may be noted that standard ARIMA based models are not adequate in capturing this behavior. More importantly, we have developed a seasonal AutoRegressive Conditional Heteroskedasticity (ARCH) based model with heavy-tailed innovations (disturbances) to characterize the traffic data rate dynamics. To our knowledge, this is the first attempt to characterize periodically measured nonstationary network traffic through an ARCH-based model that incorporates heavy-tail distribution; it may be noted that ARCH-models have been previously used primarily in Econometric modeling.
- We illustrated why MMSE-based forecast is not appropriate in the context of dynamic bandwidth provisioning since under-forecast should be avoided as much as possible in order to meet service level agreements between customers and network providers. Instead, we proposed a modified probability-hop forecast algorithm for forecasting one-step ahead forecast which extends the earlier work on [16]. We illustrated the intractability of deriving the one-step forecast of the observed series z_t

from the forecast of log transformed series (of which the heavy-tailed ARCH model is based) and proposed a reasonably good approximation to compute the same.

- Building on the probability-hop forecast, we present dynamic bandwidth provisioning schemes which incorporate quantized bandwidth steps as well as a hold-down timer to avoid oscillation.
- We have developed several performance metrics to assess the effectiveness of dynamic bandwidth provisioning schemes. Based on extensive computational studies, we have shown that indeed we have developed a robust framework for dynamic bandwidth provisioning. Specifically, our approach minimizes under-forecast while keeping the utilization high, yet the signaling overhead is also indirectly reduced.

While we have presented our studies using the measurement windows to be of 15 minutes interval, our dynamic bandwidth provisioning approach is general and can be applied to any time window of measurements. Still a question remains: no matter how small or large this interval window is, it is quite possible that there is a legitimate traffic spike *between* two successive measurements for which the provisioning is quite inadequate. Such in-between spikes can be labeled as *microcongestion*. An important question then is: how does our approach handle a *microcongestion*? Our approach does not directly address this question. Rather, our approach can be adapted *indirectly* to handle such a situation by considering a commonly used principle followed by large Internet service providers when they do traffic engineering of their networks; this principle limits the average link utilization to typically no more than, say, 60%, for traffic engineering—this then allows rooms for any short-term spikes being adequately handled by the provisioned network capacity. Using a similar idea, if we use another parameter, say λ (with $0 < \lambda \leq 1$), then the quantized bandwidth determined by our approach can be thought of as the *raw* bandwidth estimate which is then divided by λ to obtain the true provisioned bandwidth. We did not explicitly include this parameter in our presentation on bandwidth quantization; such a parameter will automatically increase the capacity to the point where no loss would be perceived at all. Since our goal was also to understand the loss trade-off, especially in our analysis, we did not include this as a built-in parameter. Secondly, depending on the terms of the service level agreements (SLAs) between a customer and its network provider, some amount of short-term loss might be acceptable as long as SLAs are met over a longer term, agreed upon window. Nevertheless, a parameter such as λ can be added to our schemes if provisioning is required to address any microcongestion issues. In any case, it is important to design a provisioning scheme in such a way that under-provisioning is minimized to avoid any consequential effect on the time series itself.

Finally, our forecasting step was to look out one-period ahead. As a part of future work, we plan to investigate forecasts for multiple periods ahead. We are also currently investigating how to directly incorporate the signaling overhead cost as part of the bandwidth provisioning schemes. These works, when completed, will be reported elsewhere.

Acknowledgment

We thank the anonymous reviewers for their constructive comments which have helped improve the presentation of the paper considerably. We also thank Justin Malyn of UMKC Information Services for his help in obtaining measurements data on the Internet link connecting the University of Missouri–Kansas City to MOREnet.

Appendix I

From data set I, we isolated the first 100 samples (little over a day) based on 15-minute time windows of measurements. We identified and fitted ARIMA(1,1,1) based on the first 32 samples, i.e. over 8 hours; the Ljung-Box diagnostic test confirmed the adequacy of the model. The parameters of the model are estimated using the maximum-likelihood function. In order to monitor the changes that might happen in the parameters of the identified model, we adopt the cumulative score approach proposed by Yuzhi et. al [7]. This method is applicable to general ARIMA time series models. Essentially, the cumulative score approach does not provide a measure of new estimate rather, indicates a possible direction of drift from the current estimates. The cumulative score as a function of time is shown in Fig. 14; here, the x-axis represent the time index of the isolated samples from data set I. The autoregressive coefficients and the moving average coefficients are denoted here by ϕ and θ, respectively. We use (ϕ_1^0, θ_1^0) to denote the current estimates and (ϕ_1^1, θ_1^1) to denote the other possible estimates. Until $t = 50$, there is a gradual decrease in the score for both parameters. However, between t = 50 and t = 54, the rate of decay becomes much higher for both ϕ_1 and θ_1 indicating inadequacy of their current estimates at that point in time. This observed behavior of cumulative scoring function strongly indicates an onset of inadequacy at around t = 50 in the current estimate of ϕ_1 and θ_1.

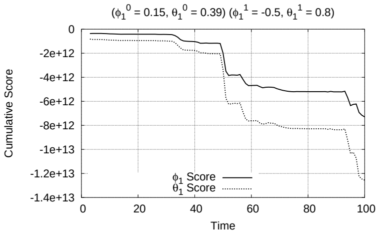

Fig. 14. Score of ARIMA(1,1,1) model

Appendix II

For a sequence $z_1, z_2, \ldots$, the rate of change between two successive values is given by $(z_k - z_{k-1})/z_{k-1}$ for $k \geq 2$. From the logarithmic series expansion, we have

$$\ln\left(\frac{z_k}{z_{k-1}}\right) = \ln\left(1 + \left(\frac{z_k}{z_{k-1}} - 1\right)\right) \simeq \left(\frac{z_k}{z_{k-1}} - 1\right) \quad (21)$$

for $(z_k/z_{k-1}-1) \ll 1$. Now, by subtracting $\ln z_{t-1}$ from both sides of (7) and using (21), we obtain an approximation for $(z_k/z_{k-1}-1)$ from which it follows that

$$\begin{aligned} z_t \simeq \Big(1 + (\chi_1 - 1)\ln z_{t-1} + \sum_{i=2}^{5} \chi_i \ln z_{t-i} \\ + \sum_{j=T}^{T+5} \chi_j \ln z_{t-j} + \sum_{k=2T}^{2T+1} \chi_k \ln z_{t-k} \\ + \sum_{s=3T}^{3T+1} \chi_s \ln z_{t-s}\Big) z_{t-1} + \varepsilon_t z_{t-1}. \qquad (22) \end{aligned}$$

By taking the conditional expectation of (22) with respect to $\mathcal{F}_{t-1}$, we have the desired result (8).

In order to derive the probability limits, we assume an equality relationship in (22). Then, we have,

$$z_t = E[z_t|\mathcal{F}_{t-1}] + \varepsilon_t z_{t-1}. \qquad (23)$$

From (6b) and (6c), it is evident that the conditional distribution of $\varepsilon_t|\mathcal{F}_{t-1}$ is Student-t distributed with ν degrees of freedom and scale $\mathcal{M}_t$. From (23), it follows that the distribution of statistic $(z_t - E[z_t|\mathcal{F}_{t-1}])/z_{t-1}$ is same as ε_t. Therefore, the confidence limits of $E[z_t|\mathcal{F}_{t-1}]$ can be given by

$$E[z_t|\mathcal{F}_{t-1}] \pm t_{\nu,\frac{\beta}{2}} \sqrt{\mathcal{M}_t} z_{t-1} \qquad (24)$$

where $t_{\nu,\frac{\beta}{2}}$ is the random deviate from the Student-t distribution with ν degrees of freedom and unit scale such that $\int_0^{t_{\nu,\frac{\beta}{2}}} f_{t_v}(x)\,dx = (1-\beta)/2$ and $f_{t_v}(x)$ is the corresponding probability density function.

APPENDIX III

Here, we present the estimates of parameters along with their t-statistics in parenthesis evaluated through the maximum likelihood function for all the three data sets in Table VI.

TABLE VI
PARAMETER ESTIMATES

	Data set I	Data set II	Data set III
φ_1	-0.123291844 (-3.87)	-0.053048458 (-1.64)	-0.096107493 (-3.11)
φ_2	-0.091455106 (-3.32)	-0.113222742 (-4.37)	-0.165460441 (-6.22)
φ_3	-0.075865242 (-2.63)	-0.05263004 (-1.99)	-0.110163493 (-4.31)
φ_4	-0.105211972 (-3.62)	-0.047591632 (-1.74)	-0.112527888 (-4.67)
ϕ_1	-0.636980162 (-20.52)	-0.611640344 (-20.31)	-0.545004245 (-18.71)
ϕ_2	-0.259639144 (-8.41)	-0.327395026 (-10.12)	-0.32442873 (-11.66)
α_0	0.033308668 (8.63)	0.026132488 (6.52)	0.0265452466 (7.54)
α_1	0.2364614211 (2.75)	0.2537706326 (2.64)	0.4786862913 (3.73)
ν	5 (4.55)	4 (4.75)	4 (5.10)

REFERENCES

[1] Y. Afek, M. Cohen, E. Haalman, Y. Mansour, "Dynamic Bandwidth Allocation Policies," *Proc. of IEEE INFOCOM'96*, pp. 880-887, 1996.

[2] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716-723, 1974.

[3] A. K. Bera, M. L. Higgins, "ARCH Models: Properties, Estimation and Testing," *Journal of Economic Surveys*, vol. 7, no. 4, pp. 305-362, 1993.

[4] T. Bollerslev, R. F. Engle, D. B. Nelson, "ARCH Models," *Handbook of Econometrics*, vol.4, pp. 2961-3038, 1994.

[5] G. E. P. Box, G. Jenkins, *Time Series Analysis: Forecasting and Control*, Prentice Hall, February 1994.

[6] C. Bruni, P. D. Andrea, U. Mocci, C. Scoglio, "Optimal Capacity Management of Virtual Paths in ATM Networks," *Proc. of GLOBECOM '94*, pp. 207-211, December 1994.

[7] Y. Chai, N. Davies, "Monitoring the Parameter Changes in General ARIMA Time Series Models." *Journal of Applied Statistics*, vol. 30, no. 9, pp. 983-1001, November 2003.

[8] P. Chemouil, J. Filipiak, "Modeling and Prediction of Traffic Fluctuations in Telephone Networks," *IEEE Trans. on Comm.*, vol. 35, pp. 931-941, 1987.

[9] M. E. Crovella, A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 835-846, 1997.

[10] J. Dufour, R. Roy, "Generalized Portmanteau Statistics and Tests of Randomness," *Communications in Statistics - Theory and Methods*, vol. 15, pp. 2953-2972, 1986.

[11] R.F. Engle, "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, vol. 50, no. 4, pp. 987-1007, July 1982.

[12] N. K. Groschwitz, G. C. Polyzos, "A Time Series Model of Long-Term NSFNET Backbone Traffic," *Proc. of ICC'94*, pp. 1400-1404, 1994.

[13] B. Groskinsky, D. Medhi, and D. Tipper, "An Investigation of Adaptive Capacity Control Schemes in a Dynamic Traffic Environment," *IEICE Trans. on Comm.*, vol. E84-B, no. 2, pp. 263-274, February 2001.

[14] J. D. Hamilton, *Time Series Analysis*, Princeton University Press, 1994.

[15] C. M. Jarque, A. K. Bera, "Efficient Tests for Normality, Homoskedasticity and Serial Independence of Regression Residuals," *Economic Letters*, vol. 6, no. 3, pp. 255-259, 1980.

[16] B. Krithikaivasan, K. Deka, D. Medhi, "Adaptive Bandwidth Provisioning based on Discrete Temporal Network Measurements," *Proc. of IEEE INFOCOM'04*, pp. 1786-1796, Hong kong, March 2004.

[17] W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic," *Proc. of ACM SIGCOMM '93*, pp. 183-193, September 1993.

[18] S. Ohta, K. Sato, "Dynamic Bandwidth Control of the Virtual Path in an Asynchronous Transfer Mode Network," *IEEE Trans. on Comm.*, vol. 40, no. 7, pp. 1239-1247, July 1992.

[19] A. Orda, G. Pacifici, D. E. Pendarakis, "An Adaptive Virtual Path Allocation Policy for Broadband Networks," *Proc. of IEEE INFOCOM '96*, pp. 329-336, March 1996.

[20] K. Papagiannaki, N. Taft, Z. Zhang, C. Diot, "Long-Term Forecasting of Internet Backbone Traffic: Observations and Initial Models," *Proc. of IEEE INFOCOM'03*, pp. 1178-1188, April 2003.

[21] V. Paxson, S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Trans. Networking*, vol. 3, pp. 226-244, 1995.

[22] *MRTG: Multi Router Traffic Grapher*, `http://www.mrtg.org/`

[23] *SAS Documentation*, `http://v9doc.sas.com/sasdoc/`

Balaji Krithikaivasan is completing PhD in Computer Networking at the University of Missouri–Kansas City.

Yong Zeng is Associate Professor of Statistics in the Department of Mathematics and Statistics at the University of Missouri–Kansas City. For additional information, go to `http://mendota.umkc.edu`

Kaushik Deka received his MS in Computer Science from UMKC; he is currently working at Demos Solutions, developing forecasting models for financial industry.

Deep Medhi is Professor of Computer Networking in the Computer Science & Electrical Engineering Department at the University of Missouri–Kansas City. For additional information, go to `http://www.sce.umkc.edu/~dmedhi`