INTERNATIONAL JOURNAL OF COMMUNICATIONS
Issue 2, Volume 4, 2010

Modeling VoIP Traffic on Converged IP Networks with Dynamic Resource Allocation

Imad Al Ajarmeh, James Yu, and Mohamed Amezziane

***Abstract*—** The exponential growth of reliable IP networks provides a suitable and cost-efficient replacement for the legacy TDM based voice networks. In this paper we propose a new framework for modeling Voice over IP (VoIP) traffic based on a non-homogeneous Poisson process (NHPP). We show that the NHPP can provide an exact fit for the call arrival data, and can also be approximated to a normal model under heavy traffic condition. The overall goal of traffic engineering is to minimize call blocking and maximize system resource utilization. Our study which is based on hundreds of millions of call arrival information shows that the Poisson process fails to model the traffic behavior of modern IP-based telecommunication systems. This failure is due to using fixed call arrival rate and static resource allocation scheme. Our proposed framework solves the two problems by modeling call arrivals as a function of time. This time-dependent function supports a dynamic resource allocation mechanism that can be easily applied to converged IP networks. The proposed model is validated by real traffic data, and is also applied to predict the behavior of future data. We conducted statistical tests which demonstrate the validity of our model and the goodness-of-fit of predicted data and actual data. Our statistical results also show that the NHPP can safely be approximated by a normal process under heavy traffic conditions.

***Key-Words*--** Traffic engineering, VoIP, Erlang-B, Call arrival modeling, NHPP.

I. INTRODUCTION

THE wide deployment of broadband, reliable, and cost-efficient IP networks is pushing towards a major paradigm shift in the telecommunication world. The wired as well as the wireless telecom industries are both heading towards an all-IP networks. According to [1], 42% of US businesses at the end of 2009 had a VoIP solution in at least one business location.

Traffic engineering is a fundamental requirement for designing and maintaining voice and data networks. It provides the tradeoff between service and cost. Traffic engineering of the Public Switched Telephone Network (PSTN) passed through many phases and matured enough to model the behavior of the legacy telecommunication networks. IP networks are packet-switched rather than circuit-switched. The resources of IP networks are different from those of circuit-switched networks. For example the major resource in a circuit-switched network is the number of circuits (trunks), but the concept of trunks is not applicable to IP networks. The non-blocking nature of packet networks requires adding a Call Admission Control (CAC) component [2][3].

The introduction of IP telephony and the wide spread of wireless technology have significantly affected the phone usage pattern, and the traffic behavior. This effect results in inadequacy of the traditional traffic engineering approaches. The paper provides an in-depth analysis for call arrival patterns on an IP Tandem network. This study is based on hundreds of millions of calls. The data shows the deficiency in the traditional approach which is based on the Poisson process, and this deficiency cause significant underutilization of network resources. In such traditional approaches, resources are statically allocated. When system resources are depleted, more resources are added to the network. Adding more resources may resolve congestion conditions temporarily, but it is not a cost-efficient solution on the long run. Static resource allocation problem is more severe on converged networks. For example some voice calls might be blocked due to the lack of resources while some portion of the resources statically allocated for data are idle.

We propose a new approach to traffic engineering by applying a Non-Homogeneous Poisson Process (NHPP) for call arrival rate. We then apply a generalized linear function to model call arrivals as a function of time. The proposed model supports dynamic allocation of network bandwidth based on predicted traffic, and modern network management system can easily support this dynamic bandwidth allocation procedure. Furthermore, a dynamic resource allocation system can adopt a de-allocation scheme which can significantly minimize call blocking probability and maximize the bandwidth utilization.

II. CALL ARRIVAL PROCESS

The study of the call arrival process is to identify the key parameters to model the behavior of incoming call traffic into the system. The call arrival function, *p(k,t)*, is the probability of *k* calls arriving during the next *t* seconds

A. Traditional Call Arrival Models

Traditionally, call arrival rate has been modeled using a *Poisson* distribution with a constant rate (λ)

$$p(k,t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad for\ k = 0,1,2,.. \qquad (1)$$

λ is the key parameter for the distribution. It indicates the average number of call arrivals in the given time interval (rate).

Under the Poisson assumption calls arrive independently from one another with a constant mean arrival rate. Therefore, the observed call arrival process consists of the sum of a large number of independent call arrivals [4]. The Poisson assumption provides relative simplicity in the corresponding

mathematical and analytical models. In practice, it is impossible to have a phone system with a constant call arrival rate over a long interval such as one day. Therefore, the model is designed for some maximum arrival rate, or the time of the day is divided into blocks and the arrival rate is assumed constant within each block. A separate queuing model is provided for each block.

The Erlang-B model is used to estimate the telecom network resource requirements. This model is based on the Poisson arrival distribution where the rate (λ) is constant and is measured based on the Busy Season Busy Hour (BSBH) which is the busiest hour in the busiest week during the year. Networks are designed to handle the traffic offered during this hour. Using a constant call arrival rate fails to adapt to the variation of traffic with respect to time, such as time of day, day of week, and day of year.

Under the BSBH approach, a considerable portion of network resources will remain idle for the majority of the year which results in poor resource utilization. Considering the traffic data that we collected over eight months, we noticed that during the BSBH we receive 300 calls per second, however for many days the call arrival rate does not exceed 200 calls per second. If we apply the BSBH engineering approach the system should be designed to handle 300 calls per second, therefore 33% of the resources will be idle for the majority of the year.

Such problems can be justified in the PSTN because of the difficulties associated with allocating and revoking network resources which are the number of trunks connecting central offices. Increasing or decreasing this number is a complicated and expensive process that involves the interaction of multiple parties. In the IP world resource allocation is flexible. Allocating more or less bandwidth for voice applications is a relatively simple process. Dynamic resource allocation for VoIP traffic can be useful especially for converged networks where voice and data share the same physical facilities. More bandwidth can be allocated to voice traffic during busy time, while providing non-used bandwidth for data applications during other time.

B. Call arrivals as Non-Homogeneous Poisson Process (NHPP)

Recently, there has been a growing interest in using more flexible arrival models whose capabilities are not restricted by the traditional assumptions, for example: Irina et al [5] modeled the SS7 signaling traffic as exponential BCMP[1] queuing network, Brown et al [6] models the call arrivals at a call center as a time-inhomogeneous Poisson process with piecewise constant rates. This interest is driven by the need to solve the problems associated with the inadequacy of Poisson assumption to satisfy the modern engineering requirements. Finding explicit analytical equations for systems with complex arrival flows might be very difficult [7]. Research in this field tends towards simulations [8] or towards analyzing the system under the condition of heavy traffic (many calls in the system) [9] and low traffic (the system is mostly idle) [10].

In the Poisson constant rate approach, the accuracy of the engineering process depends on the validity of the assumption that the arrival rate is constant within the given time blocks. In large systems with heavy traffic loads, it requires to have very small time blocks so that the rate can be assumed constant. Providing a separate queuing model for a large number of small time blocks is not a practical solution. A better approach is to model the call arrival process as a Non-Homogeneous Poisson Process (NHPP). The NHPP has independent call arrivals, the same as the Poisson Process; in addition it models non constant arrival rates [14]. The arrival rate is a function of time and can be captured using an appropriate time-dependent function. We adopt NHPP approach for modeling call arrivals of the IP tandem network, and we research on finding time-dependent function that models the variation of call arrivals.

III. CALL ARRIVAL MODELING FRAMEWORK

This study is based on real VoIP traffic data obtained from one of the major Tandem carriers in the United States. This carrier recently migrated their TDM-based circuit switched networks to IP networks.

A. IP tandem network

Tandem networks play critical roles in the telecommunications hierarchy. They interconnect different central offices together by means of toll switches. Central offices might belong to the same carrier or to different carriers. In the later case the tandem service provides interconnectivity and switching between different carriers (inter-carrier switching). As a result, tandem networks are expected to carry large amount of traffic and should be designed for high capacity, high availability, high scalability, and cost effective operations

An IP-Based Tandem service utilizes IP core network instead of the legacy TDM as a transport for the voice traffic. Using a converged IP network for data and voice provides substantial cost saving for network design and management. It is important to develop a traffic engineering scheme suitable for these converged networks. The goal is to increase resource utilization by supporting more subscribers without additional infrastructure cost.

B. Data collection and processing

During this study we have collected several hundreds of millions of call detail records (CDR's) from an IP tandem network. We developed a library to collect raw data from the different sources and then to filter, aggregate, process and visualize the data according to our study needs. Fig. 1 illustrates a typical IP tandem network. The legacy PSTN is connected through TDM trunks.

[1] BCMP network is a heterogeneous queuing network with multiple classes of customers having different distributions, it is considered as an extension to a Jackson [15]

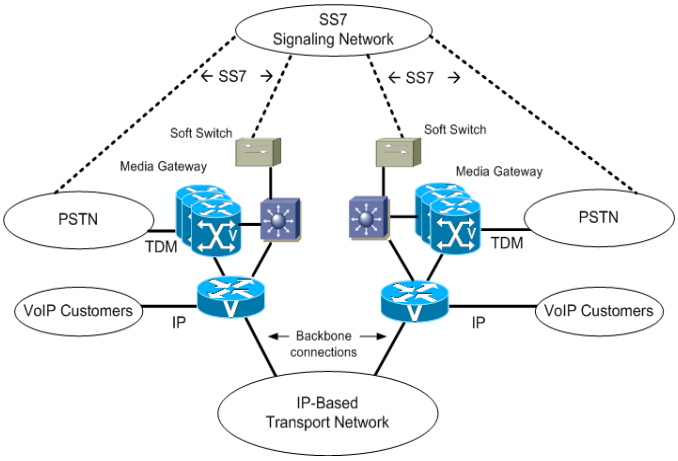

Fig. 1 Typical IP-based Tandem network

VoIP customers are connected via IP links. The network has an IP core which is used to interconnect different sites. The limiting resources on the network can be the IP backbone connections between different tandem offices, the IP connections to the VoIP customers, or the TDM connection to the legacy PSTN. The scope of our research is to optimize the first 2 IP-based resources.

A Call Detail Record (CDR) is kept for every call on a local Billing Server located at each tandem office. Our scripts poll the distributed billing servers for CDR's every day over a period of eight months. Once we have our copy of the CDR we divide the traffic into three categories:

- Wireless traffic
- landline traffic
- VoIP traffic.

The wireless traffic is usually delivered by wireless carriers over TDM links. This categorization is based on the type of incoming call transport, however all calls leaving the tandem office to another tandem office are converted into VoIP so they can be transported on the IP backbone. The calls leaving the office to carrier networks (customers) will be converted to VoIP only if the carrier is connected to the office by means of IP circuits.

Fig. 2 shows a comparison between the three traffic categories in a typical tandem office. It is clear that traffic coming to the office from the carrier networks over IP links is only 15% of the overall traffic; however, it is important to notice that all other traffic (wireless and landline) will be converted to VoIP to be transferred to another office. In addition, all the traffic coming over the backbone from other offices is VoIP. In other words, all the incoming traffic [over the backbone as well as over the carrier links] is either VoIP or "potential" VoIP.

After traffic is categorized, we extract traffic information of our research interest. In order to satisfy the requirements of this research, we keep the raw time of arrival (TOA) for each call. We also generate aggregated forms of the data by dividing the day into time blocks and finding the mean of the call arrival rate over each time block. We generate 10, 100, 1200, 3600 seconds aggregated data files.

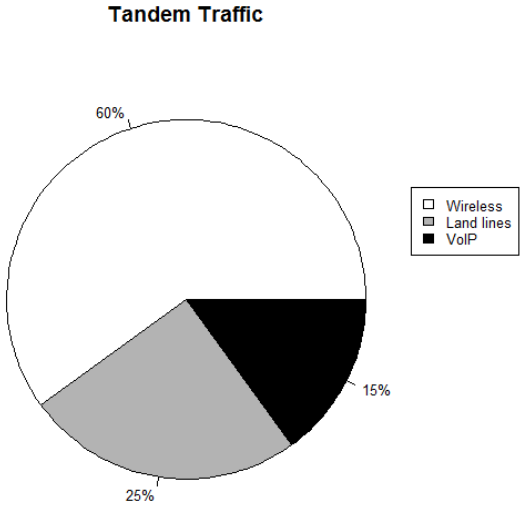

Fig. 2 Tandem traffic categories

Our modeling results are identical for all traffic aggregations which indicate the goodness and significance of the proposed model and hence the correctness and robustness of the proposed engineering framework.

C. Call arrival pattern

We study hundreds of millions of call arrivals collected over several months. We notice that the minimum call load occurs near 4 AM. Based on this finding we redefine the day from a traffic engineering point of view as the period between 4 AM and 4 AM of the next day. Furthermore, we notice that different days have different patterns. For example the difference between the call load on Fridays and that on Sundays is noticeable and should not be ignored. Fig 3 shows call arrival patterns for a typical week. We notice call arrival difference of 70% between Friday and Sunday. Our proposed model takes the daily effect into consideration.

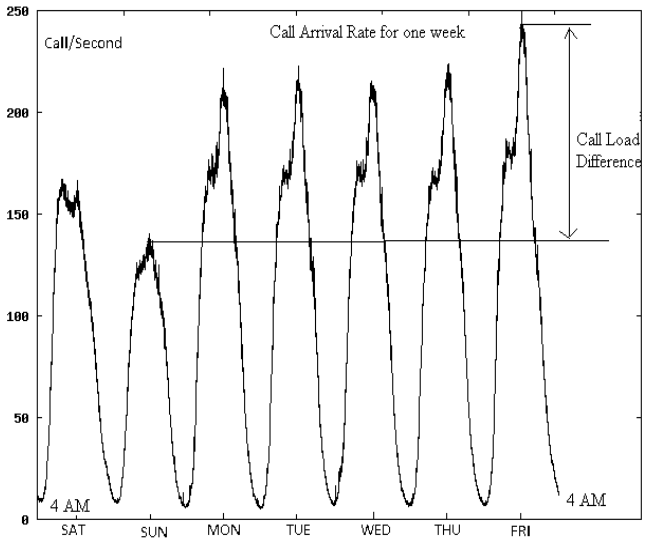

Fig. 3 Call arrival pattern for a typical week

D. Model formulation

After choosing NHPP to model the call arrival process, we need to find a time-dependent function that captures the variation of call arrivals. This step might vary from one system to the other depending on the nature of traffic (business, residential, enterprise, or mixed) and the scope and requirements of the engineering process. We show an example of using real data to derive a time function that takes the daily effect into consideration, we propose a model for the call arrival rate represented by a time-dependent intensity function $[\boldsymbol{\lambda(t)}]$

Given our data, we are inspired to construct a model that describes the variation of call arrival rates during a week. It is common in statistical analysis to model the logarithm of λ(t) instead of λ(t) itself for count data [11]. Such transformation would guarantee that the estimate of the intensity function is always non-negative. Our model takes into consideration the daily arrival patterns and has the time-dependent intensity function of:

$$log[\lambda(t)] = \mu + \sum_{i=1}^{k_o} [\alpha_i \sin(i\omega_o t) + \beta_i \cos(i\omega_o t)] + \sum_{j=1}^{6} \gamma_j I_j(t) \quad (2)$$

Where: $\boldsymbol{\lambda(t)}$ is a function of time (t).

$\boldsymbol{I_j(t)}$ is day Indicator function where $\boldsymbol{j}$ is the day of the week. The value of $\boldsymbol{I_j(t)}$ is 1 if the time $\boldsymbol{t \in j}$ and 0 otherwise. $\boldsymbol{K_o}$ is the number of harmonics in the model. $\boldsymbol{\mu}$ represents the model central tendency without daily effects. $\boldsymbol{\gamma_j}$ is the effect of day *j* and represents the difference between $\boldsymbol{\mu}$ and the mean number of calls for day *j*. $\boldsymbol{\alpha_i}$, and $\boldsymbol{\beta_i}$ are the contribution of the *ith* harmonic to the model.

Throughout the different days in a week, a similar pattern occurs regularly. In general, the rate of arrival during a given time of the day does not change significantly after a period of 24 hours, unless that day falls during the week-end or a special day. This pattern looks like a sinusoidal wave, and since the period is $\boldsymbol{T = 24 * 3600}$ seconds, we consider the frequency $\boldsymbol{\omega_o = \frac{2\pi}{T}}$. The model for $\boldsymbol{\lambda(t)}$ belongs to a family of generalized linear models, known as Poisson regression models, which have been extensively studied and applied to analyze data from scientific sources [12].

E. Parameters estimation

We use Maximum likelihood estimation to fit our proposed λ(t) to the actual call arrivals. As explained in section III.B the processed call arrival data is aggregated into non-overlapping time intervals ($\boldsymbol{\delta}$) of 10, 100, 1200, and 3600 seconds. Thus we will use the total number of calls within time intervals rather than the exact call time of arrival. Let $\boldsymbol{n_1, n_2, .., n_{m-1}, n_m}$ denote the number of calls arrived at the system in non-overlapping intervals $\boldsymbol{(a_1, a_2], (a_2, a_3], \dots, (a_{m-2}, a_{m-1}], (a_{m-1}, a_m]}$

Therefore, the likelihood function $\boldsymbol{L}$ is given as

$$L = \exp\left\{-\sum_{i=1}^{m}\int_{a_i}^{a_{i+1}} \lambda(t)dt\right\} \prod_{i=1}^{m} \frac{\left(\int_{a_i}^{a_{i+1}} \lambda(t)dt\right)^{n_i}}{n_i!} \quad (3)$$

And the log-likelihood, apart from a given constant, is given as

$$\boldsymbol{l = \sum_{i=1}^{m} n_i log \int_{a_i}^{a_{i+1}} \lambda(t)dt \quad - \sum_{i=1}^{m} \int_{a_i}^{a_{i+1}} \lambda(t)dt} \quad (4)$$

Where: $\boldsymbol{m}$ is the number of intervals within each day. Given that $\boldsymbol{\delta}$ is the aggregation time interval, we can say that $\boldsymbol{a_{i+1} = a_i + \delta}$ and $\boldsymbol{a_m - a_0 = m.\delta}$
The value of $\boldsymbol{\delta}$ is very small compared to the whole study duration. So practically, the integrals in (3) and (4) can be evaluated using the following approximation:

$$\int_{a_i}^{a_{i+1}} \lambda(t)dt \approx \delta\lambda(t_i)$$

Where $t_i = (a_{i+1} + a_i)/2$.
The approximation error is of order $\boldsymbol{o(\delta^2)}$. Hence (4) becomes:

$$l = \sum_{i=1}^{m} n_i log[\delta\lambda(t_i)] \quad - \sum_{i=1}^{m} \delta\lambda(t_i) \quad (5)$$

Substituting the function $\boldsymbol{\lambda(t)}$ given in (2) into the log-likelihood function, and excluding the constant that does not depend on the parameters, (5) becomes:

$$\boldsymbol{l = \sum_{k=1}^{m} n_k \left(\mu + \sum_{i=1}^{k_o} [\alpha_i \sin(i\omega_o t_k) + \beta_i \cos(i\omega_o t_k)] + \sum_{j=1}^{6} \gamma_j I_j(t_k) \right) - \delta \sum_{k=1}^{m} exp\left(\mu + \sum_{i=1}^{k_o} [\alpha_i \sin(i\omega_o t_k) + \beta_i \cos(i\omega_o t_k)] + \sum_{j=1}^{6} \gamma_j I_j(t_k) \right)} \quad (6)$$

Equation 6 can be rewritten as

$$l = n\mu + \sum_{i=1}^{k_o} [\alpha_i S_i + \beta_i C_i] + \sum_{j=1}^{6} \gamma_j F_j - \delta \sum_{k=1}^{m} G_k \quad (7)$$

Where

$$S_i = \sum_{k=1}^{m} n_k sin\,(i\omega_o t_k)$$

$$C_i = \sum_{k=1}^{m} n_k cos\,(i\omega_o t_k)$$

$$F_j = \sum_{k=1}^{m} n_k I_j(t_k)$$

And

$$G_k = exp\left(\mu + \sum_{i=1}^{k_o} [\alpha_i \sin(i\omega_o t_k) + \beta_i \cos(i\omega_o t_k)] + \sum_{j=1}^{6} \gamma_j I_j(t_k)\right) \quad (8)$$

Notice that F_j is the total number of calls on day j. The terms S_i and C_i do not depend on the parameters while the terms G_k are exponential terms.

The ML estimators are obtained by taking the partial derivatives of the log-likelihood, with respect to the model parameters: $\mu, \alpha_i, \beta_i,$ and each of γ_j. Hence, we obtain the following score equations:

$$\frac{\partial l}{\partial \mu} = n - \delta \sum_{k=1}^{m} G_k \quad (9)$$

$$\frac{\partial l}{\partial \alpha_i} = S_i - \delta \sum_{k=1}^{m} G_k sin(i\omega_o t_k) \quad (10)$$

$$\frac{\partial l}{\partial \beta_i} = C_i - \delta \sum_{k=1}^{m} G_k cos(i\omega_o t_k) \quad (11)$$

$$\frac{\partial l}{\partial \gamma_j} = F_j - \delta \sum_{k=1}^{m} G_k I_j(t_k) \quad (12)$$

The ML estimators are obtained by solving:

$$\frac{\partial l}{\partial \mu} = \mathbf{0}, \frac{\partial l}{\partial \alpha_i} = \mathbf{0}, \frac{\partial l}{\partial \beta_i} = \mathbf{0}, \frac{\partial l}{\partial \gamma_j} = \mathbf{0} \; for \; j = 1, 2, .., 6 \; .$$

An implicit form of the solution to the first equation can be obtained easily as follows:

$$\hat{\mu} = log\left(\frac{\delta}{n} \sum_{k=1}^{m} G_k{}'\right) \quad (13)$$

Where

$$G_k{}' = exp\left(\sum_{i=1}^{k_o} [\alpha_i \sin(i\omega_o t_k) + \beta_i \cos(i\omega_o t_k)] + \sum_{j=1}^{6} \gamma_j I_j(t_k)\right) \quad (14)$$

The other score equations cannot be solved analytically; therefore we resort to Newton type nonlinear optimization methods. We use R language to code our model and estimate its parameters. Our estimations are based on Fisher scoring method which is a simplification of the famous Newton-Raphson algorithm. The results are presented in the next section.

F. Inference about the model: significance, validation and prediction

To assess the performance of the estimators and hence of the model, we need to study the variability of these estimators and test their significance. This task requires computing the covariance matrix of the estimators. ML estimation theory tells us that when the sample size is sufficiently large, as is the case of our call arrival data, the covariance matrix is equal to I^{-1} [13], where I is the information matrix, obtained by evaluating the negative expectation of the Hessian matrix of the log-likelihood function. The diagonal elements of the information matrix are,

$$-E\left(\frac{\partial^2 l}{\partial \mu^2}\right) = \delta \sum_{k=1}^{m} G_k \quad (15)$$

$$-E\left(\frac{\partial^2 l}{\partial {\alpha_i}^2}\right) = \delta \sum_{k=1}^{m} G_k sin^2(i\omega_o t_k) \quad (16)$$

$$-E\left(\frac{\partial^2 l}{\partial {\beta_i}^2}\right) = \delta \sum_{k=1}^{m} G_k cos^2(i\omega_o t_k) \quad (17)$$

$$-E\left(\frac{\partial^2 l}{\partial {\gamma_j}^2}\right) = \delta \sum_{k=1}^{m} G_k I_j(t_k) \quad (18)$$

Where the operator *E(.)* denotes the expectation of a random variable. The off-diagonal terms are computed by taking mixed partial derivatives of order 2. We evaluated the variance terms for each parameter, then we used them to conduct Wald's significance test: H_o: $\theta = 0$ against H_1: $\theta \neq 0$, where θ is any parameter of interest (α_i, β_i, γ_j and μ). Table 1 shows the values of the estimated parameters, their standard errors and the corresponding p-values of Wald's test.

Table 1 Estimated parameters for $\boldsymbol{\lambda(t)}$

Parameter	Estimated value	Std. Error	p-value
$\boldsymbol{\mu}$	12.4851183	0.0002360	< 2e-16
$\boldsymbol{\alpha_1}$	0.6244975	0.0002387	< 2e-16
$\boldsymbol{\alpha_2}$	0.3730669	0.0002675	< 2e-16
$\boldsymbol{\alpha_3}$	0.1122494	0.0002224	< 2e-16
$\boldsymbol{\beta_1}$	-1.2787258	0.0003443	< 2e-16
$\boldsymbol{\beta_2}$	-0.4221888	0.0002767	< 2e-16
$\boldsymbol{\beta_3}$	-0.1487193	0.0002205	< 2e-16
$\boldsymbol{\gamma_1}$	-0.2266414	0.0003971	< 2e-16
$\boldsymbol{\gamma_2}$	-0.5476155	0.0004534	< 2e-16
$\boldsymbol{\gamma_3}$	0.0833744	0.0003486	< 2e-16

The small magnitude of the parameters' p-values suggests that the considered parameters are significant. Parameters with p-values larger than 0.05 were removed since their presence would be a nuisance to the model and might contribute to variance inflation. The significance of the model's parameters reflect the significant of the model itself and that it explains the variability in the data as can be seen from the plot in Fig. 4 below.

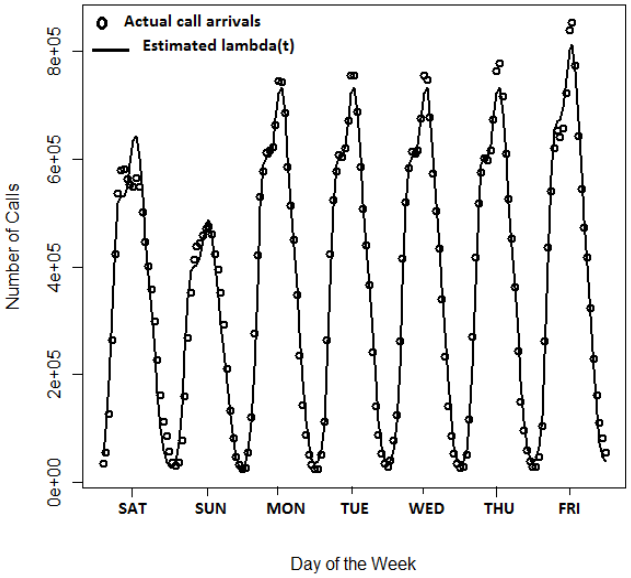

Fig. 4 fitting actual call arrivals to the suggested model

The model significance can also be evaluated by conducting the likelihood ratio test:

$$H_o: The\ process\ is\ HPP \Rightarrow \alpha_i = \beta_i = \gamma_j = 0 \Rightarrow (\lambda(t) = e^{\mu}\ (constant))$$
$$H_1: The\ process\ is\ a\ NHPP\ (\lambda(t) is\ time\ dependent)$$

If the null hypothesis is true, all the parameters in the model are non-significant and the process becomes a traditional HPP. If the null hypothesis is rejected then our model is significant and hence the call arrival process is generated by the NHPP with $\lambda(t)$ as described in (2). The likelihood ratio test statistic used here is evaluated as the ratio of the likelihood function for the restricted model (HPP under H_o), and of likelihood of the full model (NHPP with $\lambda(t)$) [11]. The null distribution of the test statistic is a chi square whose number of degrees of freedom equals the number of parameters minus 1. For our model and data, this value is equal to 34,676,131 with 8 degrees of freedom, corresponding to a p-value that is practically 0. Such very small p-value confirms our earlier results that the considered model is a very good fit to the data. Therefore, we can conclude that the observed call arrivals are definitely, not generated by the traditional homogeneous Poisson process and hence the need for modeling a time-dependent arrival is well justified.

In this example, the model is written in terms of 3 day indicators and 3 harmonics ($j=1,2,3$ and $K_o=3$), however in some cases we might consider including more indicators and harmonics and even relevant predictors (covariates) if necessary. The abundance of significant parameters in our model is due to the fact that the data set is very large, which results in many of the estimated terms being significant. Another way of looking at the proposed $\boldsymbol{\lambda(t)}$, is that it resembles a semi-parametric model in which the nonparametric component consists of a Fourier series expansion while the parametric component is a linear combination of indicator functions.

The traditional HPP models involve unjustifiable approximations that lead to improperly engineered systems. The traditional traffic engineering process involves collecting sample call information and estimating a constant call arrival rate which is fed to a system based on Erlang-B model in order to calculate the required resources. In our approach we follow the same sample collection process, and then we feed the collected sample (actual call information) to a system that will construct the model, estimate its parameters and compute the required resources accurately.

The importance of using a significant model lies in the capability of such model to predict future data. In this section we use our proposed framework to construct a model and estimate its parameters based on data collected in week 1 and then we use these parameters to predict data for week-2 and week-3. The model validation is based on the accuracy of the prediction results of week-2 and week-3.

We compare the predicted data to the actual data that we already have for these weeks. Fig. 5 shows a plot of the predicted data against the actual data of two random weeks. The figure shows clearly that the actual observations fall very close to the curve of the estimated model.

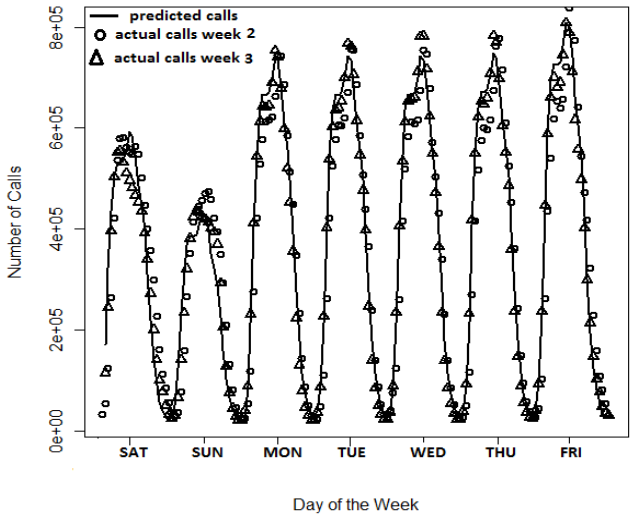

Fig. 5 Predicted against actual call arrivals for two random weeks

This approach of modeling call arrival rate allows us to build different time-dependent models based on the exact engineering requirements. For example, one might consider the variation of call arrivals from one week to another, or from one month to another or even from one hour to another and hence construct a different $\boldsymbol{\lambda(t)}$. Holidays, and special occasions can easily be dealt with by giving them indicator functions.

IV. HEAVY TRAFFIC AND NORMAL APPROXIMATION

Although the proposed NHPP model was developed based on large scale tandem traffic data; however the model is designed as an exact and general purpose model so it can accurately capture wide variety of traffic patterns in terms of size and nature. Under certain circumstances some approximations can be made in order to simplify the model. For example if we apply this model to a sample of traffic data where the arrival rate is almost constant and the variation with respect to time is very small, in this case the estimated $\boldsymbol{\alpha, \beta, and\ \gamma}$ will be close to zero, and the model will fall back to a homogeneous Poisson process with fixed arrival rate of $\boldsymbol{\mu}$.

Likewise, in the presence of heavy traffic as is the case of our tandem data, the arrival rate $\lambda(t)$ tends to assume very large values. When this happens, the Poisson distribution of the number of calls can be approximated by a normal distribution $N(\lambda(t)\ ,\ \lambda(t))$ [17]. The rationale behind this approximation is due to the fact that Poisson random variables or processes are used to model rare events that occur at low arrival rates and when the rate of occurrence becomes large, the corresponding events cease being rare and become frequent, thus rendering the event "normal" instead of rare. More rigorous explanation of this approximation can be found in [18]. Therefore, our NHPP under heavy traffic can be approximated by a Gaussian process where at a given time t, the number of calls X_t follow the Gaussian (normal) distribution $N(\lambda(t)\ ,\ \lambda(t))$.

In section III.C, we considered Poisson distribution for the count data; therefore we used a log-linear model to fit the instantaneous Poisson rate of the NHPP. In this section we are considering a normal distribution, and because of its nature, we fit a linear model to the approximately normal data:

$$X_k = \lambda(t_k) + \lambda^{\frac{1}{2}}(t_k)\varepsilon_k, \qquad \varepsilon_k \text{ are } iid\ N(0,1)$$

$$\lambda(t) = \mu + \sum_{i=1}^{k_o}\left[\alpha_i \sin(i\omega_o t) + \beta_i \cos(i\omega_o t)\right] + \sum_{j=1}^{6}\gamma_j I_j(t)$$

If bias reduction or variance stabilization is needed, the data can be transformed before a linear model can be fit.

The above linear model can be written as:

$$\frac{X_k - \lambda(t_k)}{\lambda^{\frac{1}{2}}(t_k)} = \varepsilon_k, \qquad \varepsilon_k \text{ are } iid\ N(0,1)$$

So, the likelihood function is equal to:

$$L = \prod_{i=1}^{k_o}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left[\frac{X_k - \lambda(t_k)}{\lambda^{\frac{1}{2}}(t_k)}\right]^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^{k_o}\exp\left(-\frac{1}{2}\sum_{i=1}^{k_o}\left[\frac{X_k - \lambda(t_k)}{\lambda^{\frac{1}{2}}(t_k)}\right]^2\right)$$

Therefore, the log-likelihood function, apart from two data-independent constants, is given as:

$$l = -\sum_{i=1}^{k_o}\frac{1}{\lambda\ (t_i)}\left(X_i - \lambda(t_i)\right)^2$$

Maximizing the log-likelihood is equivalent to minimize a weighted least squares criterion with weights $w_i = \lambda^{-1}(t_i)$. Since the weights depend on the actual parameters and hence are unknown, we start by minimizing the ordinary least squares loss function to obtain estimates of these parameters and hence of the weights. Then we replace the weights by their estimates to calculate the weighted least squares and minimize it to obtain more accurate estimates of the parameters. This operation can be re-iterated if deemed necessary after the examination of the residuals. For a detailed list of methods and algorithms commonly used to obtain weighted least-squares estimators, the reader is referred to Gentle (2002).

In addition to the parameters estimates, this process delivers standard errors as well and therefore enables us to conduct Wald' test for each parameter in order to remove the non-significant ones. Table 2 shows the estimated, standard errors and the corresponding p-values for the significant parameters.

Table 2 Estimated parameters for the normalized model

Parameter	Estimated Value	Std. Error	p-value
$\boldsymbol{\mu}$	369305.06	8150.67	< 2e-16
$\boldsymbol{\alpha_1}$	166804.64	8502.85	< 2e-16
$\boldsymbol{\alpha_2}$	17381.67	8280.95	0.037639
$\boldsymbol{\alpha_3}$	-29661.11	8275.31	0.000468
$\boldsymbol{\alpha_4}$	-28387.62	8298.99	0.000822
$\boldsymbol{\beta_1}$	-291436.81	8126.73	< 2e-16
$\boldsymbol{\beta_2}$	-33645.36	8333.34	8.92e-05
$\boldsymbol{\beta_3}$	-20236.73	8326.22	0.016363
$\boldsymbol{\gamma_1}$	-25034.94	8621.29	0.004294
$\boldsymbol{\gamma_2}$	34989.04	13532.59	0.010759

Since we are fitting a linear model, the model significance is evaluated through the analysis of variance (ANOVA) and related Fisher test:

$H_o: \alpha_i = \beta_i = \gamma_j = 0 \ \Rightarrow\ (\lambda(t) = e^{\mu}\ (constant))$
$H_1: Model\ is\ significant$

The p-value of this test is less than 2.2e-16, meaning that the proposed model is highly significant and hence we reach the same conclusion as the one in section III.D This conclusion is reinforced by an $R^2 = 0.9487$, which means that our model explains 94.9% of the total variability in the data and hence we can ascertain that the normal approximation approach for heavy traffic is valid and can be safely used for prediction purposes.

Fig. 6 shows the normal (Gaussian) approximation graph plotted against real call arrivals and NHPP. The graph obtained based on the normal approximation model is very close to the one obtained by NHPP in Fig. 4 and its predicted curve is very close to the one in Fig. 5. These results support the theory that the approximate normality-based model performs very closely to the exact Poisson model when the arrival rate is very large.

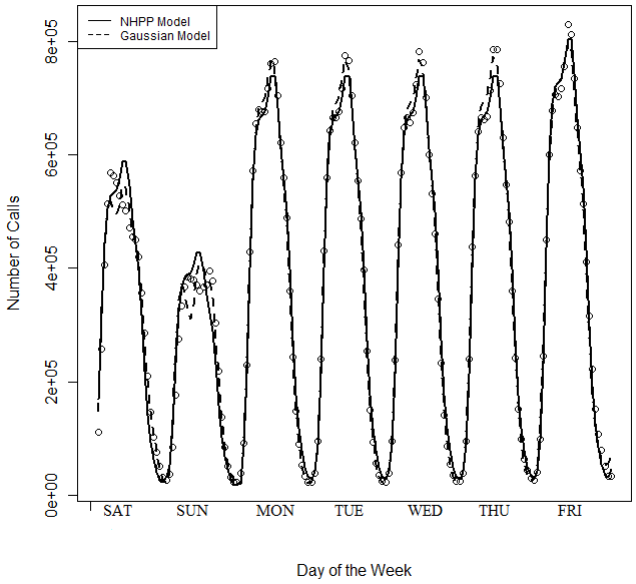

Fig. 6 Normal approximation against NHPP and real call arrivals

While taking the normality approach, it should be noted that Because of the time dependency of the data, a correlogram might be checked to test for autocorrelation among residuals. When such dependency exists, we need to fit a "model with autoregressive errors" to the data

V. DYNAMIC RESOURCE ALLOCATION

We illustrated in this paper how static resource allocation based on BSBH limits the utilization of the telecommunication system. In addition, under this paradigm it is difficult to response to events that involve high variability in traffic loads. This problem is becoming more severe in today's complex, integrated, and fast changing networks. The objectives of a dynamic resource allocation (DRA) system are:

- Increase system utilization
- minimize blocking among all services sharing the system resources
- Provide foundation for dynamic system behavior which can respond to real-time events instantly.

The dynamic resource allocation scheme can be implemented in a way that takes into consideration the QoS requirements. The benefit of this approach is to achieve the target grade of service and at the same time to assess voice traffic constrains. Resources are allocated based on the demand as well as the required QoS [16]. QoS parameters of our interest include end-to-end delay, packet loss, and delay jitter. In some advanced systems, the number of resources allocated to activities can be dynamically adjusted according to traffic conditions, such techniques are known as Adaptive Resource Allocation (ARA) or Flexible Resource Allocation (FLA) [19].

The decision of resource allocation is made on a higher level such that a global system goal will be maximized. The telecommunication system can be looked at as a set of *m* resources $(R_1,R_2,\ldots,R_m)$ where $m \geq 1$, and a set of *n* independent resource-requiring activities $(r_1,r_2,\ldots,r_n)$ where $n \geq 1$ and n activities occur over T time periods. We assume that the major resource for the network is bandwidth so we can view the available bandwidth as virtual channels each with pre-defined capacity. Therefore we can compute the maximum call load a system can handle [3] . In this scheme the bandwidth is a time-shared resource where a virtual channel is occupied for a period of time (duration of a call) and then the channel is released to be occupied by another call.

In converged networks, the allocation process allocates the shared resources between data and voice based on the traffic demand. Under this basic DRA scheme incoming calls might be blocked if no resources are available to carry the calls. Call blocking probability can be significantly decreased if we de-allocate some of the resources given to some low-priority data services and re-allocate them to the new voice calls. This approach is known as Dynamic Resource Allocation with De-allocation scheme DSA and is commonly used in wireless communications [19]. On one hand, the DSA approach has the advantage of decreasing call blocking probability, on the other hand it has the disadvantage of degrading the quality of some data services; in addition it requires coordination point between the dynamic resource allocation process, call admission control, and QoS authorization [20]. Under all dynamic resource allocation schemes, it is always important to protect ongoing calls from dropping. From the user's perspective it is more objectionable to terminate ongoing calls due to blocked-calls-dropped discipline than to block new incoming calls [21]

The NHPP model we proposed in the paper is best applied in converged networks with DRA schemes. The reason behind this is the fact that we treat the call arrivals as a varying function of time. At the times where high call loads are received more resources can be dynamically allocated to the voice traffic, and when the call load decreases, theses resources can be re-allocated to data traffic if needed. DRA scheme should be integrated with Traffic Engineering (TE) scheme; in this manner we can optimize the transportation of IP packets in the most efficient, reliable and rapid way while maximizing the overall system utilization. In IP networks, multiple paths might exist between a given sources and destination. TE scheme is needed in order to prevent the routing protocols from using the shortest path always which might cause some network resources (paths) to be over utilized while other resources are idle [22].

Resource allocation in converged IP networks is commonly achieved by providing virtual separation between data and voice traffic. Traffic policing and shaping techniques are implemented in order to limit the bandwidth for each type of traffic. In this case, the bursting nature of the data traffic will not affect the delay-sensitive voice traffic. At the same time resources can be allocated, deal located, and reallocated to traffic categories based on the demand and anticipated load.

This approach requires the use of Call Admission Control (CAC) and its implementation can be enforced by the VoIP signaling protocols such as SIP or H.323

VI. CONCLUSION

This research presents a new framework to model VoIP traffic based on real data collected from an IP tandem network. The data shows that the traditional Poisson process is not appropriate to model the VoIP traffic, while a non-homogeneous Poisson Process is able to capture the traffic behavior. A major contribution of this research is modeling the call arrival rate as a function of calendar time. We validate the model behavior with real traffic data over several months. The statistical analysis of predicted data and actual data shows strong model validity and goodness-of-fit. Furthermore, we prove that under heavy traffic conditions, the proposed NHPP model can be approximated by a normal model. This approximation simplifies the queuing system used to calculate the required resources based on specific traffic loads and target blocking probability. This traffic engineering model could support network management systems to develop a dynamic bandwidth allocation procedure. During the peak time of voice traffic, more bandwidth is allocated to the voice application. When the voice traffic is low, more bandwidth is allocated for data services. Our next step in this research is to study the call holding time and find a modern model that captures its variability. Once we have models for call arrival rate and call holding time, we will provide a queuing system to calculate the network resources

REFERENCES

[1] http://www.researchandmarkets.com/reportinfo.asp?report_id=1202379&t=d&cat_id=

[2] Solange R. Lima, Paulo Carvalho, and Vasco Freitas. "*Admission Control in Multiservice IP Networks: Architectural Issues and Trend*," IEEE Communications, Vol. 45 No. 4, April 2007, 114-121

[3] James Yu and Imad Al-Ajarmeh, "Design and Traffic Engineering of VoIP for Enterprise and Carrier Networks," IARIA On-Line Journal, Volume-1, January 2009

[4] P. Abry, P. Borgnat, F. Ricciato, A. Scherrer, and D. Veitch ,"*Revisiting an old friend: On the observability of the relation between Long Range Dependence and Heavy Tail*," Telecommunication Systems Journal, Springer Vol. 43, Numbers 3-4, April 2010

[5] Irina Buzyukova, Yulia Gaidamaka, Gennady G. Yanovsky , "*Estimation of GoS Parameters in Intelligent Network*", 9th International Conference, NEW2AN 2009 and Second Conference on Smart Spaces, ruSMART 2009, St. Petersburg, Russia, September 2009

[6] Lawrence Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, and Linda Zhao, "*Statistical Analysis of a Telephone Call Center A Queueing-Science Perspective*", Journal of the American Statistical Association, March 2005, Vol. 100, No. 469, Applications and Case Studies

[7] Kyungsup Kim, and Hoon Choi, "*Mobility Model and Performance Analysis in Wireless Cellular Network with General Distribution and Multi-Cell Mode*," Wireless Personal Communications Journal, Springer Netherland, Volume 53, Number 2 / April, 2010

[8] Jarosław Bylina, Beata Bylina, Andrzej Zoła, Tomasz Skaraczyński, "*A Markovian Model of a Call Center with Time Varying Arrival Rate and Skill Based Routing*", Computer Networks: 16th Conference, CN 2009, Wisla, Poland, June 2009

[9] L. G. Afanas'eva and E. E. Bashtova , "*Limit theorems for queueing systems with doubly stochastic poisson arrivals (Heavy traffic conditions)*", Problems of Information Transmission, Vol. 44, No. 4. pp. 352-369, December 2008

[10] Bashtova, E.E., "*The Small Workload Mode for a Queueing System with Random Unsteady Intensity*", Mat. Zametki, 2006, vol. 80, no. 3, pp. 339–349 [Math. Notes , vol. 80, no. 3–4,pp. 329–338], 2006

[11] Dobson, Barnett, "*An Introduction to Generalized Linear Models*," Third Edition, Chapman & Hall/CRC Textx in statistical Science. 2008

[12] S. L. Rathbun, and S. Fei, "*A Spatial Zero-Inflated Poisson Regression Model for Oak Regeneration***,**" Environmental and Ecological Statistics Volume 13, Number 4 / December, 2006

[13] Deng, X. and Yuan, M, "*Large Gaussian covariance matrix estimation with Markov structures*," Journal of Computational and Graphical Statistics, 18(3), 640-657, (2009)

[14] Hsien-Lun Wong, Shang-Hsing Hsieh, Yi-Hsien Tu, "*Application of Non-Homogeneous Poisson Process Modeling to Containership Arrival Rate*," icicic, pp.849-854, 2009 Fourth International Conference on Innovative Computing, Information and Control, 2009

[15] Baskett, F., Chandy, K.M., Muntz, R.R., Palacios, F.G.: "*Open, Closed, and Mixed Networks of Queues with Different Classes of Customers*", Journal of the ACM 22(2), 248–260 (1975)

[16] Zoran Perisic, Zoran Bojkovic, "*Model for quality of service management on Internet based on resource allocation*", WSEAS NNA-FSFS-EC 2002, February 11-15, 2002, Interlaken, Switzerland.

[17] Casella, G. and R.L Berger, "*Statistical Inference*", Second Edition. Pacific Grove, California: Duxbury Press, 2002

[18] Gentle, James, "*Elements of computational statistics*", Springer, 2002, ISBN 0387954899

[19] Sandoval-Arechiga, R., Cruz-Perez, F.A., Ortigoza-Guerrero, L., "Dynamic Resource Allocation in Integrated Voice/Data Wireless Networks With Link Adaptation", Global Telecommunications Conference (GLOBECOM), St. Louis, MO, 2005

[20] Fausto Andreotti, Nicola Ciulli, Marilia Curado, Giada Landi, and Cristina Nardini, "*Signalling Extensions for QoS Support in a IEEE 802.16d/e Domain*", 7th WESEAS International Conference on Electronics. Hardware, Wireless and Optical Communications, Cambridge, UK, Feb 2008

[21] Roxana Zoican, and Dan Galatchi, "*Analzyes of a MAC protocol for GPRS network***",** the 4th WSEAS International Conference on Telecommunications and Informatics, Prague, Czech Republic, 2005

[22] Francesco Palmieri, and Ugo Fiore, "*Genetic-based Traffic Engineering in GMPLS networks*", Proceedings of the 5th WSEAS International Conference on Simulation, Modeling AND Optimization, Corfu, Greece, August 2005

Imad. H. Al Ajarmeh is a PhD student at DePaul University, Chicago, USA. his main research is in Voice over IP traffic engineering and modeling. He received a MS degree in Network security from DePaul University in 2007, and a Bachelor in electrical Engineering/Telecommunications from Mu'tah University, Jordan in 1992.

Imad currently holds a position of Network Architect in Neutral Tandem (Chicago, USA). He is an IEEE member and his other research interests include XML-Based network management protocols, and wireless routing protocols.

Dr. James T. Yu is an associate professor at DePaul University where he is teaching and researching in the areas of Wireless LAN security, Fault Tolerant Networks, Voice over IP, and XML-based Network Management. He is the Program Chair of the Telecommunications and Data Networking Committee, responsible for the BS/MS curriculum development. He was the director of Network Technology at ARBROS Communications, responsible for network design and management. He had 15 years of experience at Bell Laboratories and was appointed to Distinguished Member of Technical Staff. He received the MS and Ph.D degrees in Computer Sciences from Purdue University

Dr. Mohamed Amezziane received his Diplome d'Ingénieur d'Etat in Industrial Engineering from L'Ecole Mohammedia des Ingénieurs (Rabat, Morocco) in 1996. He obtained a M.S. in Simulation Modeling and Analysis and a M.S. in Statistical Computing, both from the University of Central Florida (Orlando, Florida, USA) in 2000 and a Ph.D. in Mathematics from the same institution in 2004.

Mohamed currently holds a position of Assistant Professor at the Department of Mathematical Sciences at DePaul University. His main research interests are stochastic modeling and nonparametric and semiparametric estimation.