Wireless Networks 7, 211–219, 2001

Hyper-Erlang Distribution Model and its Application in Wireless Mobile Networks *

YUGUANG FANG
Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA

Abstract. This paper presents the study of the hyper-Erlang distribution model and its applications in wireless networks and mobile computing systems. We demonstrate that the hyper-Erlang model provides a very general model for users' mobility and may provide a viable approximation to fat-tailed distribution which leads to the self-similar traffic. The significant difference from the traditional approach in the self-similarity study is that we want to provide an approximation model which preserves the Markovian property of the resulting queueing systems. We also illustrate that the hyper-Erlang distribution is a natural model for the characterization of the systems with mixed types of traffics. As an application, we apply the hyper-Erlang distribution to model the cell residence time (for users' mobility) and demonstrate the effect on channel holding time. This research may open a new avenue for traffic modeling and performance evaluation for future wireless networks and mobile computing systems, over which multiple types of services (voice, data or multimedia) will be supported.

Keywords: teletraffic, traffic modeling, PCS, wireless networks, mobile computing, channel holding times

1. Introduction

The future telecommunications networks (such as the third generation wireless networks) target to provide integrated services such as the voice, data and multimedia via inexpensive low-powered mobile computing devices over the wireless infrastructures, teletraffic modeling and resource dimensioning for such networks are challenging and important, which rely heavily on good traffic models to characterize the network dynamics with appropriate accuracy. The criteria for finding an appropriate traffic model are as follows: (1) it must be general enough to provide a good approximation to the field data; (2) it must also be simple enough to enable us to obtain analytically tractable results for performance evaluation. Current practices usually emphasize one and neglect the other, it will be worthwhile to find a good traffic model to satisfy these two criteria. Traffic models can be determined by probability distributions of time variables, such as interarrival times and service times in queueing systems, which characterize the traffic and service utilities in the considered networks.

In order to get tractable analytical results, researchers have used the exponential distribution to model time variables in the past for many years. Recent study on similar traffic [19] and intensive research followed strongly show the limitation of exponential distribution models. In the wireless and mobile computing arena, a plenty of evidences showed that channel holding times and interarrival times of cell traffic are no longer exponentially distributed [1,3,13,15,16].

Future wireless networks target to provide various bit-rate multimedia services of different types with various QoS (quality of service) requirement via inexpensive low powered portables. The session time for each service will vary significantly from application to application. In each cell, the cell traffic will consist of different traffic flows with distinct characteristics, similar to the wired environments, the cell traffic most probably shows self-similarity property, hence, the traditional traffic theory (relying on exponential model) will most likely be invalid. More general distribution models are needed to capture the essence of the network dynamics. It has been shown [19] that fat-tailed distribution is better suitable for the characterization of the self-similar traffic, unfortunately, this model complicates the resulting queueing model, which is not computationally tractable.

Recently, in wireless network and mobile computing research, two general models have been proposed. Rappaport and his colleagues [23,24] proposed a model called the SOHYP (the Sum of the Hyper-Exponential) distribution to model the channel holding time and cell dwell time (i.e., the cell residence time), by showing that the coefficient of variation (the ratio of square root of variance to mean) can be adjusted to be less than, equal and greater than unity, they showed the generality of the SOHYP models. In a series of research works on PCS networks, the current author and his colleagues [6,8] proposed another general while simpler distribution model called *the hyper-Erlang* distribution to model directly the cell residence time which leads to better characterization of channel holding time. We have qualitatively shown that the hyper-Erlang model provides a universal approximator to any general distribution of nonnegative random variable and also satisfies the two criteria we mentioned earlier. More importantly, this distribution model can be easily implemented in simulation software and provide more general traffic modeling platform for verification and testing.

* This research was sponsored in part by the New Jersey Institute of Technology under Grant SBR421980.

There are still some unanswered questions left for the hyper-Erlang distribution model: what moment properties does it have (in terms of first moment, variance and even high moments)? How can we apply it to solve the traffic modeling problems, particularly in wireless mobile networks? How do we do with the data fitting for it? This paper is an attempt to address some of the issues. We first present some results for the hyper-Erlang model, we show why the hyper-Erlang model is better suited for the characterization of traffic model for networks with different types of traffic flows, how this model can be potentially used to approximately characterize the network traffic with self-similarity property, and why this model can provide a more realistic approximation to general distribution. We discover that the hyper-Erlang distribution is a natural choice for network modeling with integrated services supported in the future telecommunications networks. It is observed that in wireless mobile networks, the important teletraffic parameters such as channel holding time can be characterized by the cell residence time and call holding time, we then apply the hyper-Erlang distribution to model the cell residence time for the study of channel holding time in wireless networks and mobile computing systems, easy-to-use analytical results have been derived. The sensitivity of the mobility (i.e., cell residence time) on the distribution of channel holding time has been discussed. Rather than studying the effect of variance of cell residence time on the channel holding time distribution as we did in the past [6,8], we investigate the effect of the coefficient of variation (CoV) of cell residence time on the channel holding time. It seems that the CoV is a better parameter for capturing the users' mobile behavior.

2. Hyper-Erlang distribution and its properties

In this section we present some new results on the hyper-Erlang distribution model.

We first demonstrate that the hyper-Erlang model is the natural choice for teletraffic modeling in communications networks with integrated services. As we mentioned, the future telecommunications networks (such as the third generation wireless networks) target to provide integrated services such as the voice, data and multimedia via inexpensive low-powered mobile computing devices over the wireless infrastructures, teletraffic modeling and resource dimensioning for such networks are challenging and important. Imagining that we have a wireless network with cellular infrastructure, we want to support M types of (real-time or non-realtime, prioritized or nonprioritized) services. Let λ_i denote the arrival rate of the call requests of type i in a cell, let $s_i(t)$ denote the cell session time (channel holding time) probability density function of service type i in the cell. If we assume some independence of services of all types, we can model the cell as a queue system with total arrival rate $\lambda = \sum_{i=1}^{M} \lambda_i$ and with service time probability density function:

$$s(t) = \sum_{i=1}^{M} \frac{\lambda_i}{\lambda} s_i(t). \tag{1}$$

This is in the mixed form of different (maybe simple) distributions in the spirit of the hyper-Erlang modeling. Another scenario comes from the research on self-similarity of traffic modeling, the ON/OFF model of Willinger et al. [26] where the traffic process is modeled by the superposition of a large number of independent $0/1$ reward processes whose ON/OFF durations are heavy-tailed. This scenario is very similar to the previous one for cellular networks, the interarrival time in a cell is also the mixed type of simple distributions.

From the above observation we know that the mixed types of distributions are the appropriate modeling approach. Rappaport and his colleagues [23,24] proposed the SOHYP to model the channel holding time in the cellular networks. We observe that the SOHYP is one of the mixed types. Recently, we [6] proposed a simpler model of mixed types, what we called the *hyper-Erlang* distribution model.

The hyper-Erlang distribution has the following density function and Laplace transform:

$$\begin{aligned} f_{\mathrm{he}}(t) &= \sum_{i=1}^{M} \alpha_i \frac{(m_i\eta_i)^{m_i} t^{m_i-1}}{(m_i-1)!} \mathrm{e}^{-m_i\eta_i t} \quad (t \geqslant 0), \\ f_{\mathrm{he}}^{*}(s) &= \sum_{i=1}^{M} \alpha_i \left(\frac{m_i\eta_i}{s + m_i\eta_i} \right)^{m_i}, \end{aligned} \tag{2}$$

where

$$\alpha_i \geqslant 0, \qquad \sum_{i=1}^{M} \alpha_i = 1,$$

and $M, m_1, m_2, \ldots, m_M$ are nonnegative integers, $\eta_1, \eta_2, \ldots, \eta_M$ are positive numbers. Define

$$\mathcal{H} = \left\{ f(t)\colon\ f(t) = \sum_{i=1}^{M} \alpha_i \frac{(m_i\eta_i)^{m_i} t^{m_i-1}}{(m_i-1)!} \mathrm{e}^{-m_i\eta_i t}, \right. \\ \left. M > 0,\ m_i \geqslant 0,\ \eta_i > 0,\ \alpha_i \geqslant 0,\ \sum_{i=1}^{M} \alpha_i = 1 \right\}.$$

This set is basically the set of all hyper-Erlang distribution models, it contains the exponential distribution, Erlang distribution, the hyper-exponential distribution. For this class of distribution models, we have

Theorem 1. The set $\mathcal{H}$ has the following properties:

(1) $\mathcal{H}$ is a convex set, i.e., any convex combination of hyper-Erlang distributions is also in this set. Therefore, any convex supposition of hyper-Erlang distributions is also a hyper-Erlang distribution.

(2) Let $\mathcal{F}$ denote the set of all probability density functions of nonnegative random variables, then $\mathcal{H}$ is a dense set

in $\mathcal{F}$, i.e., any probability density function of a nonnegative random variable can be approximated by hyper-Erlang distribution models.

(3) Hyper-Erlang distributions can be tuned to have the coefficients of variation (CoV) less than, equal to and greater than unity; they can be also tuned to have CoV as small as desired and as large as desired.

Proof. (1) can be easily verified, (2) has been shown in [6] (or referred to [17]). We focus on the proof of (3). For a hyper-Erlang distribution as in (2), the first moment and the second moment are given by (ξ is a random variable with the above hyper-Erlang distribution)

$$E[\xi] = \sum_{i=1}^{M} \alpha_i \frac{1}{\eta_i}, \qquad E\left[\xi^2\right] = \sum_{i=1}^{M} \alpha_i \frac{m_i+1}{m_i} \cdot \frac{1}{\eta_i}, \tag{3}$$

from which we obtain the coefficient of variation is given by

$$\mathrm{CoV}^2 = \frac{E[\xi^2] - E[\xi]^2}{E[\xi]^2} = \frac{\sum_{i=1}^{M} \alpha_i (1+1/m_i) x_i^2}{[\sum_{i=1}^{M} \alpha_i x_i]^2} - 1, \tag{4}$$

where $x_i = 1/\eta_i$. Let us choose $M = 2$, define

$$g(\alpha) = \frac{\alpha(1+1/m_1)x_1^2 + (1-\alpha)(1+1/m_2)x_2}{[\alpha x_1 + (1-\alpha)x_2]^2} - 1. \tag{5}$$

It is obvious that $g(0) = 1/m_2$ and $g(1) = 1/m_1$. Since $g(\alpha)$ is continuous in $[0,1]$, hence $g(\alpha)$ can assume any value in $[1/m_2, 1]$. If we choose $m_1 = 1$ and vary m_2 from 1 to ∞, then we can show that $g(\alpha)$ can assume any value in $(0,1]$. This implies that the CoV for hyper-Erlang distribution can be tuned to have any value in $(0,1]$.

Next, we show that $g(\alpha)$ can also assume any value in $[1,+\infty)$. Choose $m_1 = m_2 = 1$, then

$$h(\alpha, x_1, x_2) = \frac{2\alpha + 2(1-\alpha)(x_2/x_1)^2}{[\alpha + 2(1-\alpha)(x_2/x_1)]^2} - 1 \rightarrow \frac{2}{\alpha} - 1 \quad \left(\frac{x_2}{x_1} \rightarrow 0\right),$$

from which we conclude that when we choose α and x_2/x_1 sufficiently small (for example, choose $x_2/x_1 = \alpha^2$), $h(\alpha, x_1, x_2)$ can be as large as desired. Moreover, $h(0.5, x, x) = 1$. Since $h(\alpha, x_1, x_2)$ are continuous in $[0,1] \times (0,\infty) \times (0,\infty)$, hence $h(\alpha, x_1, x_2)$ can assume any value in $[1,\infty)$. Hence, the CoV can also be tuned to have value in $[1,\infty)$. In summary, this completes the proof of (3). □

Next, we want to show that the hyper-Erlang distributions can also provide good approximations to the fat-tailed distributions which lead to the self-similar traffic [25]. Let $F(t)$ denote the cumulative distribution of a hyper-Erlang distribution given in (2), let $\overline{F}(t) = 1 - F(t)$. It can be easily verified that

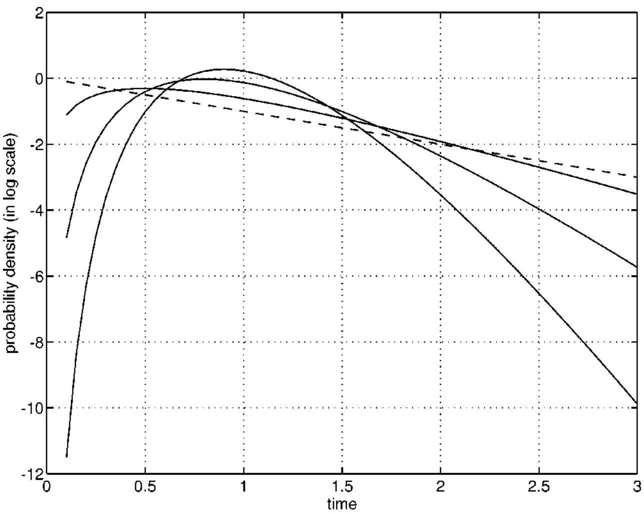

Figure 1. Erlang and exponential (dashed line) density functions in log-linear graph.

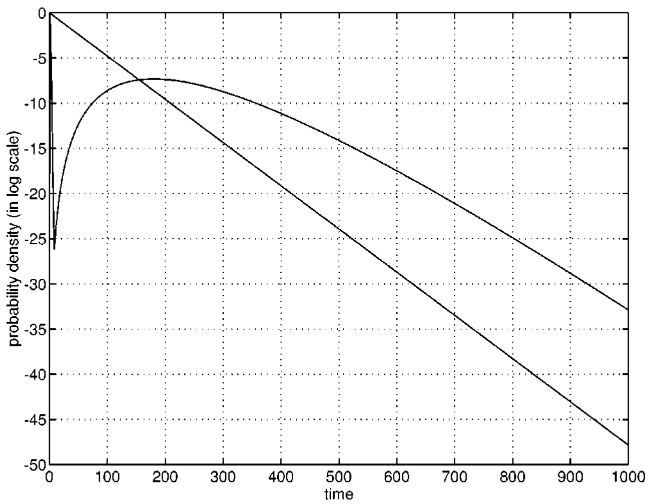

Figure 2. Hyper-Erlang distribution which shows some fatness property in a time interval of interest.

$$\overline{F}(t) = \sum_{i=1}^{M} \alpha_i \left(\sum_{k=0}^{m_i-1} \frac{(m_i \eta_i t)^k}{k!} \mathrm{e}^{-m_i \eta_i t} \right). \tag{6}$$

The distribution of a random variable X is said to be fat-tailed if $\Pr[X > t]$ is on the order of $1/t^r$ when t is sufficiently large for a $r > 0$. It is enough to study the property of $\overline{F}(t)$ for the fatness property. Intuitively, a fat-tailed distribution has a "fat tail" comparing to the exponential distribution, so we can also observe the probability density functions on the log-linear graphs to determine whether a distribution is a fat-tailed distribution. Figure 1 shows that the Erlang distributions are not fat-tailed distribution. However, as we observe from the figure, there are pieces where Erlang distributions do show "fatness" property: the Erlang distributions are higher than the straight (dashed) line. Suitable supposition of multiple Erlang distributions may lead to a fat-tailed distribution in the time interval of interest. Indeed, as shown in figure 2, the hyper-Erlang distribution does provide the fat-tailed property in the time range of interest by tuning the parameters in the hyper-Erlang model appropriately.

Another characteristic of the fat-tailed distribution is the infinite variance. We can demonstrate that the hyper-Erlang distributions can be tuned to have a finite bounded mean while a sufficiently large variance. From (3), we observe that we could choose α_i and η_i appropriately so that $E[\xi]$ is finite while $E[\xi^2]$ is sufficiently large. For example, choose $M = 2$, $n > 1$, $\alpha_1 = 1/\sqrt{n}$, $\eta_1 = \sqrt{n}$ and $\eta_2 = 1$, then $E[\xi] = 2 - 1/\sqrt{n} < 2$ and

$$E\left[\xi^2\right] = \left(1 + \frac{1}{m_1}\right)\sqrt{n} + \left(1 - \frac{1}{\sqrt{n}}\right)\left(1 + \frac{1}{m_2}\right) \geqslant \sqrt{n} \to \infty \quad (n \to \infty).$$

Thus, when n is sufficiently large, then the variance of the corresponding hyper-Erlang distribution can be sufficiently large, which also leads to the fat tail property.

There exists a nice structure about the hyper-Erlang distribution model. From theorem 1(1), we notice that a hyper-Erlang distribution has a layered structure: the Erlang distributions are the first layer, the convex combination of a set of Erlang distributions forms the second layer, another convex combination of a set of hyper-Erlang distributions forms another layer, and so on. The more the number of layers (related to the number of the Erlang distribution terms in the overall hyper-Erlang distribution), the better approximation to the real situation we achieve. This modeling bears a similarity to the self-similar traffic study: no matter how small the time scale is, the traffic pattern looks almost the same. This is the reason why hyper-Erlang distributions can be used for the modeling of self-similar traffic. Another observation is that our modeling is also in the similar vein to the ON/OFF modeling for self-similarity [26]. If we use the following hyper-Erlang distribution ($f_{\mathrm{hp}}(t;k)$ is a hyper-Erlang distribution),

$$h(t) = \sum_{k=1}^{N} \beta_i f_{\mathrm{hp}}(t;k), \quad \beta_i \geqslant 0, \quad \sum_{k=1}^{N} \beta_i = 1,$$

to model the interarrival times of call arrivals to a cell in a cellular network, then the cell traffic is in fact the switched traffic merge from N sources: a new call arrival belongs to the kth source with probability β_k ($k = 1, 2, \ldots, N$). We will investigate self-similar traffic using the hyper-Erlang models in more detail elsewhere.

3. Analytical results for channel holding times

In this section, we concentrate on the study of the channel holding time for the wireless networks and mobile computing systems (WINMOC) under hyper-Erlang cell residence time. In a WINMOC, service areas are equipped with cellular structure, where mobile users get their service via the base station in the cell they are traveling. The time the mobile users spend in the cell (not necessarily engaging in a service) is called the *cell residence time*, which characterizes the users' mobility. The time a mobile user locks on a channel in the cell is called the *channel holding time*, which corresponds to the service time in queueing systems study. In [8], we have studied the channel holding time under general cell residence time distributions. We have obtained the following result. As in the self-similar traffic study [11], variations of service time at certain node may result in the self-similar traffic downstream. In this section, we apply the hyper-Erlang distributions to model the cell residence time and investigate how integrated services will affect the channel holding time distribution.

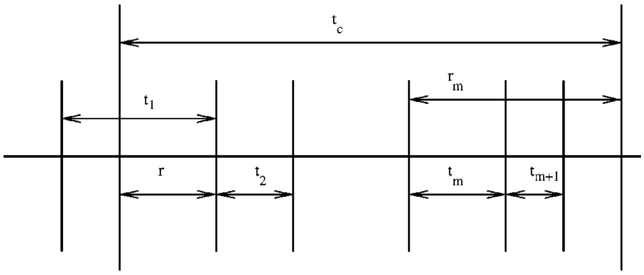

Figure 3. The time diagram for call holding time and cell residence time.

As in [8], figure 3 shows the time diagram for our study. Let t_{c} be the call holding time (the time of the requested connection to a wireless network) for a typical new call, t_m be the cell residence time in the mth cell the mobile user travels, r be the time between the instant the new call is initiated at and the instant the new call moves out of the cell if the new call is not completed (we call it the residual cell residence time), let r_m ($m > 1$) be the residual call holding time when the call finishes mth handoff successfully. Assume that the call holding times are exponentially distributed with parameter μ and the cell residence time is generally distributed with mean $1/\eta$. Let t_{nh} and t_{hh} denote the new call channel holding time and the handoff call channel holding time, respectively (i.e., the channel holding times for new calls and handoff calls, respectively). Then, from figure 3, the new call channel holding time is

$$t_{\mathrm{nh}} = \min\{t_{\mathrm{c}}, r\}, \tag{7}$$

and the handoff call channel holding time is

$$t_{\mathrm{hh}} = \min\{r_m, t_m\}. \tag{8}$$

Let λ and λ_{h} denote the arrival rates for new calls and handoff calls, respectively. Let t_{ch} denote the channel holding time (i.e., the channel holding time no matter whether the call is new call or handoff call), thus, $t_{\mathrm{ch}} = t_{\mathrm{nh}}$ with probability $\lambda/(\lambda + \lambda_{\mathrm{h}})$ and $t_{\mathrm{ch}} = t_{\mathrm{hh}}$ with probability $\lambda_{\mathrm{h}}/(\lambda + \lambda_{\mathrm{h}})$.

Let $f_{\mathrm{c}}(t)$, $f(t)$, $f_r(t)$, $f_{\mathrm{nh}}(t)$, $f_{\mathrm{hh}}(t)$ and $f_{\mathrm{ch}}(t)$ denote, respectively, the probability density functions of t_{c}, t_m, r, t_{nh}, t_{hh} and t_{ch} with their corresponding Laplace transforms $f_{\mathrm{c}}^*(s)$, $f^*(s)$, $f_r^*(s)$, $f_{\mathrm{nh}}^*(s)$, $f_{\mathrm{hh}}^*(s)$ and $f_{\mathrm{ch}}^*(s)$, respectively. In [8], we obtain the following result.

Theorem 2. For a wireless network with exponential call holding times and Poisson new call arrivals with arrival rate λ, we have the following statements:

(i) The Laplace transform of the probability density function of the new call channel holding time is given by

$$f_{\mathrm{nh}}^*(s) = \frac{\mu}{s+\mu} + \frac{\eta s}{(s+\mu)^2}\big[1 - f^*(s+\mu)\big], \quad (9)$$

and the expected new call channel holding time is

$$E[t_{\mathrm{nh}}] = \frac{1}{\mu} - \frac{\eta}{\mu^2}\big[1 - f^*(\mu)\big]. \quad (10)$$

(ii) The Laplace transform of the probability density function of the handoff call channel holding time is given by

$$f_{\mathrm{hh}}^*(s) = \frac{\mu}{s+\mu} + \frac{s}{s+\mu} f^*(s+\mu), \quad (11)$$

and the expected handoff call channel holding time is

$$E[t_{\mathrm{hh}}] = \frac{1}{\mu}\big(1 - f^*(\mu)\big). \quad (12)$$

(iii) Let λ_{h} denote the handoff call arrival rate to a cell, then the Laplace transform of the probability density function of channel holding time is given by

$$f_{\mathrm{ch}}^*(s) = \frac{\lambda}{\lambda+\lambda_{\mathrm{h}}} f_{\mathrm{nh}}^*(s) + \frac{\lambda_{\mathrm{h}}}{\lambda+\lambda_{\mathrm{h}}} f_{\mathrm{hh}}^*(s), \quad (13)$$

and the expected channel holding time is given by

$$E[t_{\mathrm{ch}}] = \frac{1}{\mu} - \frac{\lambda\eta}{(\lambda+\lambda_{\mathrm{h}})\mu^2}\left[1 - \left(1 - \frac{\lambda_{\mathrm{h}}\mu}{\lambda\eta}\right) f^*(\mu)\right]. \quad (14)$$

(iv) The handoff call arrival rate λ_{h} is given by

$$\lambda_{\mathrm{h}} = -\eta(1 - p_{\mathrm{o}})\lambda \times \sum_{p\in\sigma_{\mathrm{c}}} \operatorname*{Res}_{s=p} \frac{1 - f^*(s)}{s^2[1 - (1 - p_f) f^*(s)]} f_{\mathrm{c}}^*(-s), \quad (15)$$

where σ_{c} is the set of poles of $f_{\mathrm{c}}^*(-s)$ on the right complex plane, $\mathrm{Res}_{s=p}$ is the residue at a pole $s = p$, p_{o} and p_f are the blocking probabilities for the new calls and handoff calls, respectively.

In the current literature, we observe that most performance analyses were carried out under the assumption that the channel holding times for new calls and handoff calls are identically distributed (some with exponential distribution), i.e., any calls, either new calls and handoff calls, were assumed to have the same identically distributed channel holding time with the same parameter, in which case the one-dimensional Markov chain can be used to obtain the blocking probabilities for new calls and handoff calls. From theorem 2 we could show that the new call channel holding time and the handoff call channel holding time are having different distributions, even having different average values. Figure 4 shows that the average channel holding times for new calls and handoff calls are different. In this figure, we assume that the cell residence time is Gamma-distributed with shape parameter varying at $\gamma = 0.1$, $\gamma = 0.5$, $\gamma = 2$, and $\gamma = 10$. The average channel holding times for new calls and handoff calls can be computed using the formulae in theorem 2. We observe in the figure that the difference is sometimes very significant. Thus, the one-dimensional Markov chain model for call blocking performance assuming that the new calls and handoff calls are identically distributed may not be appropriate, the multidimensional Markov chain (basically two-dimensional Markov chain) may be needed. This observation calls for the necessity of characterizing new call channel holding time and handoff call holding time under more general mobility assumption.

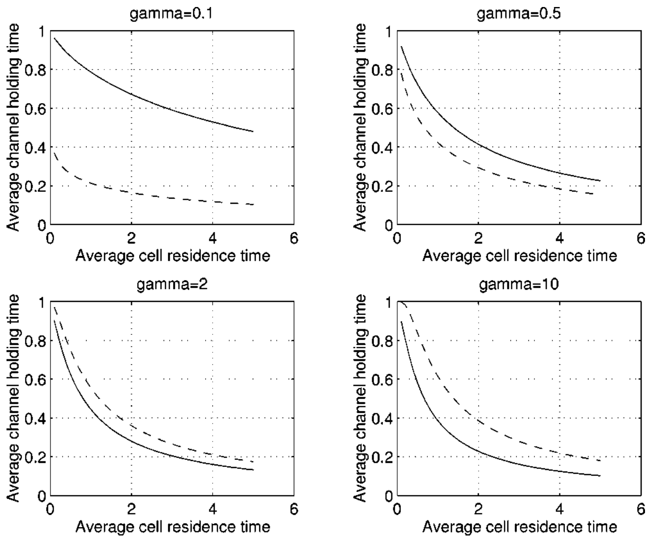

Figure 4. Average channel holding times for new calls and handoff calls: solid line for the new calls and the dashed line for the handoff calls.

As we mentioned earlier, the hyper-Erlang distribution model is general enough for field data and simple enough for tractable analysis. We can apply this hyper-Erlang distribution to model the cell residence time (mobility). Assume that the cell residence time is hyper-Erlang distributed as in (2) with parameter $1/\eta = \sum_{i=1}^{M} \alpha_i/\eta_i$ and the call holding time is exponentially distributed, applying theorem 2 with some mathematical manipulations, we can obtain

Theorem 3. For a wireless network or mobile computing system with exponential call holding time, Poisson new call arrivals and hyper-Erlang distributed cell residence time, we have

(i) The Laplace transform of the density function of the new call channel holding time is given by

$$f_{\mathrm{nh}}^*(s) = \sum_{i=1}^{M} \alpha_i \left[\frac{\mu}{s+\mu} + \frac{\eta s}{(s+\mu)^2} - \frac{\eta s}{(s+\mu)^2}\left(\frac{m_i\eta_i}{s+m_i\eta_i}\right)^{m_i}\right], \quad (16)$$

where $\eta = [\sum_{i=1}^{M} \alpha_i/\eta_i]^{-1}$, and the expected new call channel holding time is

$$E[t_{\rm nh}] = \frac{1}{\mu} - \frac{\eta}{\mu^2} + \frac{\eta}{\mu^2}\left[\sum_{i=1}^{M} \alpha_i \left(\frac{m_i\eta_i}{\mu + m_i\eta_i}\right)^{m_i}\right]. \tag{17}$$

(ii) The Laplace transform of the density function of the handoff call channel holding time is given by

$$f_{\rm hh}^*(s) = \sum_{i=1}^{M} \alpha_i \left[\frac{\mu}{s+\mu} + \frac{s}{s+\mu}\left(\frac{m_i\eta_i}{s+\mu+m_i\eta_i}\right)^{m_i}\right], \tag{18}$$

and the expected handoff call channel holding time is

$$E[t_{\rm hh}] = \frac{1}{\mu}\left[1 - \sum_{i=1}^{M} \alpha_i \left(\frac{m_i\eta_i}{\mu + m_i\eta_i}\right)^{m_i}\right]. \tag{19}$$

(iii) The Laplace transform of the density function of channel holding time is given by

$$f_{\rm ch}^*(s) = \frac{\lambda}{\lambda + \lambda_{\rm h}} f_{\rm nh}^*(s) + \frac{\lambda_{\rm h}}{\lambda + \lambda_{\rm h}} f_{\rm hh}^*(s), \tag{20}$$

and the expected channel occupancy time is given by

$$E[t_{\rm ch}] = \frac{\lambda}{\lambda + \lambda_{\rm h}} E[t_{\rm nh}] + \frac{\lambda_{\rm h}}{\lambda + \lambda_{\rm h}} E[t_{\rm hh}]. \tag{21}$$

(iv) The handoff call arrival rate $\lambda_{\rm h}$ can be computed by the following formula:

$$\lambda_{\rm h} = \frac{\eta(1-p_{\rm o})[1 - f^*(\mu)]\lambda}{\mu[1 - (1-p_f) f^*(\mu)]}, \tag{22}$$

where

$$f^*(\mu) = \sum_{i=1}^{M} \alpha_i \left(\frac{m_i\eta_i}{\mu + m_i\eta_i}\right)^{m_i}.$$

In current PCS (Personal Communications Services) networks and some mobile computing systems with high transmission rates, services are mainly accomplished via the circuit switching mode (session switching in mobile computing). In this scenario, call holding time (session time) can still be appropriately modeled by exponential distributions. Thus, the above result can be applied to study the channel holding time for performance evaluation and design. We observe two important features about the above result. The first is the simplicity. All computations involved with the result are the manipulations of the rational functions, hence the partial fractional expansion technique can be used to find the probability density functions of the channel holding time (the inverse Laplace transform). The second feature is the generality. Due to the universal approximation capability of the hyper-Erlang models, we can use the hyper-Erlang distributions to approximate any distribution function of cell residence time. Since the cell residence time captures the users' mobility, we can use the hyper-Erlang distribution models to characterize the users' mobility. Thus, if field data is available, then we can apply statistical method to fit the field data by the hyper-Erlang distribution, as such the channel holding time distribution is determined.

If the call holding time is not exponentially distributed, then results for handoff call channel holding time both in theorem 3 and in [8] cannot be applied, where the memoryless property had been used. In the reminder of this section, we give some results for this case.

Let $r_{\rm c}$ denote the residual life of the call holding time (it is r_m if the call has been handed off m times). The statistics for $r_{\rm c}$ can be obtained in a similar way as we do for the call holding time $t_{\rm c}$ and the cell residence time t_i, which do not involve the network operation. Let $f_{r{\rm c}}(t)$ denote the probability density function for $r_{\rm c}$ with the Laplace transform $f_{r{\rm c}}^*(s)$ and let $F_{r{\rm c}}(t)$ denote the cumulative distribution function of $t_{\rm hh}$. For handoff call channel holding time, we have

$$t_{\rm hh} = \min\{r_{\rm c}, t_m\}. \tag{23}$$

From equation (23), we obtain

$$\begin{aligned} F_{\rm hh}(t) &= \Pr(t_{\rm hh} \leqslant t) \\ &= \Pr(t_{r{\rm c}} \leqslant t \text{ or } t_m \leqslant t) \\ &= \Pr(t_{r{\rm c}} \leqslant t) + \Pr(t_m \leqslant t) - \Pr(t_{r{\rm c}} \leqslant t, t_m \leqslant t) \\ &= \Pr(t_{r{\rm c}} \leqslant t) + \Pr(t_m \leqslant t) - \Pr(t_{r{\rm c}} \leqslant t)\Pr(t_m \leqslant t). \end{aligned} \tag{24}$$

Differentiating (24), we obtain

$$\begin{aligned} f_{\rm hh}(t) &= f_{r{\rm c}}(t) + f(t) - f_{r{\rm c}}(t)\Pr(t_m \leqslant t) \\ &\quad - \Pr(t_{r{\rm c}} \leqslant t) f(t) \\ &= f_{r{\rm c}}(t)\int_t^\infty f(\tau)\,{\rm d}\tau + f(t)\int_t^\infty f_{r{\rm c}}(\tau)\,{\rm d}\tau. \end{aligned} \tag{25}$$

Notice that the Laplace transform of $\int_t^\infty g(\tau)\,{\rm d}\tau$ is $(1 - g^*(s))/s$ where $g(t)$ can be $f_{r{\rm c}}(t)$ or $f(t)$, applying Laplace transform to both sides of equation (25) and the inverse Laplace transform theorem, we obtain

$$\begin{aligned} f_{\rm hh}^*(s) &= \int_0^\infty f_{r{\rm c}}(t)\left[\int_t^\infty f(\tau)\,{\rm d}\tau\right]{\rm e}^{-st}\,{\rm d}t \\ &\quad + \int_0^\infty f(t)\left[\int_t^\infty f_{r{\rm c}}(\tau)\,{\rm d}\tau\right]{\rm e}^{-st}\,{\rm d}t \\ &= \int_0^\infty \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} f_{r{\rm c}}^*(z){\rm e}^{zt}\,{\rm d}z \\ &\quad \times \left[\int_t^\infty f(\tau)\,{\rm d}\tau\right]{\rm e}^{-st}\,{\rm d}t \\ &\quad + \int_0^\infty f(t)\frac{1}{2\pi j} \\ &\quad \times \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{1 - frc^*(z)}{z}{\rm e}^{zt}\,{\rm d}z\,{\rm e}^{-st}\,{\rm d}t \\ &= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} f_{r{\rm c}}^*(z) \\ &\quad \times \int_0^\infty \left[\int_t^\infty f(\tau)\,{\rm d}\tau\right]{\rm e}^{-(s-z)t}\,{\rm d}t\,{\rm d}z \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{1-f_{rc}^*(z)}{z}\\
&\times\int_0^\infty f(t)\mathrm{e}^{-(s-z)t}\,\mathrm{d}t\,\mathrm{d}z\\
&=\frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} f_{rc}^*(z)\frac{1-f^*(s-z)}{s-z}\,\mathrm{d}z\\
&+\frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{1-f_{rc}^*(z)}{z}f^*(s-z)\,\mathrm{d}z\\
&=\frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\left[f_{rc}^*(z)\frac{1-f^*(s-z)}{s-z}\right.\\
&\qquad\left.+\frac{1-f_{rc}^*(z)}{z}f^*(s-z)\right]\mathrm{d}z, \quad (26)
\end{aligned}
$$

where σ is the real number appropriately chosen for the use of the inverse Laplace transforms. Let $\sigma_{r\mathrm{c}}$ and σ_{p} denote the sets of poles of $f_{rc}^*(-s)$ and $f^*(-s)$ in the right half of the complex plane, respectively. It can be verified that $z=s$ is a removable singular point of the integrand of the last equation of (26). From the residue theorem [20], we obtain (using a contour in the right half of the complex plan)

$$
\begin{aligned}
f_{\mathrm{hh}}^*(s) &= -\sum_{p\in\sigma_{\mathrm{p}}}\operatorname*{Res}_{z=s+p}\left[f_{rc}^*(z)\frac{1-f^*(s-z)}{s-z}\right.\\
&\qquad\left.+\frac{1-f_{rc}^*(z)}{z}f^*(s-z)\right], \quad (27)\\
&= -\sum_{p\in\sigma_{rc}}\operatorname*{Res}_{z=s+p}\left[f_{rc}^*(z)\frac{1-f^*(s-z)}{s-z}\right.\\
&\qquad\left.+\frac{1-f_{rc}^*(z)}{z}f^*(s-z)\right]. \quad (28)
\end{aligned}
$$

In particular, if the call holding time t_{c} is exponentially distributed, from the strong memoryless property [6] $t_{r\mathrm{c}}$ is also exponentially distributed with the same distribution as t_{c}, so $f_{rc}^*(s)=\mu/(s+\mu)$. From (28), we obtain

$$
\begin{aligned}
f_{\mathrm{hh}}^*(s) &= -\sum_{p\in\sigma_{rc}}\operatorname*{Res}_{z=s+p}\left\{f^*(z)\left[1-\frac{\mu}{s-z+\mu}\right]\right.\\
&\qquad\left.+\frac{1-f^*(z)}{z}\frac{\mu}{s-z+\mu}\right\}\\
&= \operatorname*{Res}_{z=s+\mu}\left\{f^*(z)\frac{1}{z-(s+\mu)}\right.\\
&\qquad\left.+\frac{1-f^*(z)}{z}\frac{\mu}{z-(s+\mu)}\right\}\\
&= \frac{\mu}{s+\mu}+\frac{s}{s+\mu}f_{\mathrm{c}}^*(s+\mu),
\end{aligned}
$$

which is the same as in [8].

3.1. Remarks

(1) We witness the powerful approximation of the hyper-Erlang distribution models, hence, we can use hyper-Erlang distributions to approximately model the residual life of the call holding time and cell residence time. In this case, the complex functions under the Res operator in (27) and (28) will be rational functions, hence, the partial fractional expansion techniques can be used to find the inverse Laplace transform, i.e., the probability density function $f_{\mathrm{hh}}(t)$.

(2) For the new call channel holding time, we can obtain the similar result by substituting the r_{c} with t_{c} and t_m with r, respectively. This is left to the readers.

(3) As a final remark, the probability distribution for r_{c} can be approximately modeled by the Residual Life Theorem [18], so the Laplace transform of the probability density function of r_{c} can be given by

$$
f_{\mathrm{rc}}^*(s)=\frac{\mu[1-f_{\mathrm{c}}^*(s)]}{s}, \quad (29)
$$

which also gives the exponential distribution when the call holding time is exponentially distributed.

4. Performance studies

In this section we present our findings on how the distribution of cell residence time affects the distribution of channel holding time. We use the hyper-Erlang distribution to model the cell residence time.

We use the hyper-Erlang distribution model of two Erlang terms for our numerical study. Our idea here is as follows: if the users' mobility could be described by the hyper-Erlang distribution, what would happen to the resulting channel holding time distribution? Could we still use the exponential model to approximate it? Can we use the first and second order statistics of the cell residence time to characterize the channel holding time distribution?

There are two approaches for these problems. First, we compare the channel holding time distribution and its approximation when the cell residence time is approximated by exponential fit. In this case, the channel holding time distributions can be obtained from our analytical results by taking the hyper-Erlang and exponential cell residence times into theorem 3. Figures 5 and 6 show the comparisons. In figure 5, we use one Erlang term and one exponential term in our hyper-Erlang model, we vary the coefficient of variation in order to observe the change of channel holding time distribution and its approximation. Obviously, the exponential distribution model does not have good fit for the real distribution, even when the coefficient of variation is close to unity (a signature for exponential distribution). Figure 6 displays the cases when the hyper-Erlang distribution model has two Erlang terms. This figure exhibits more mismatches between the real channel holding time distribution and its approximations.

The second approach is to use the exponential approximation directly to the channel holding time. Figure 7 shows the comparison between the channel holding time distribution and its exponential approximation where the coefficient of variation (CoV) for cell residence time is very close to

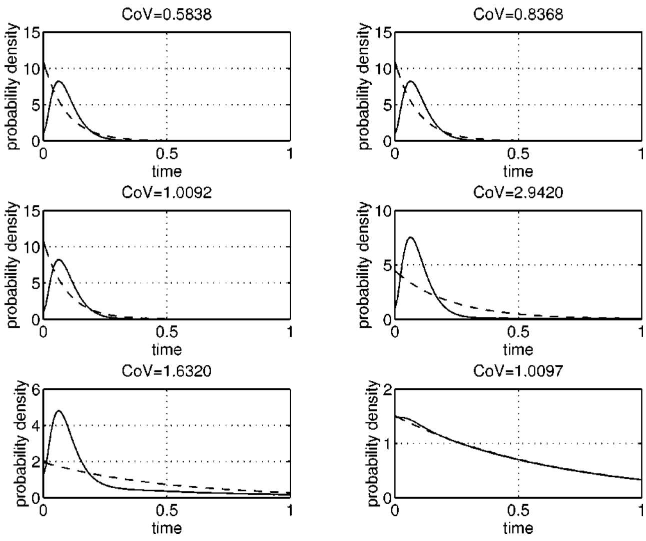

Figure 5. Handoff call channel holding time and its statistical fit (dashed line) when the exponential fit for cell residence time is used: one term in hyper-Erlang model is exponential.

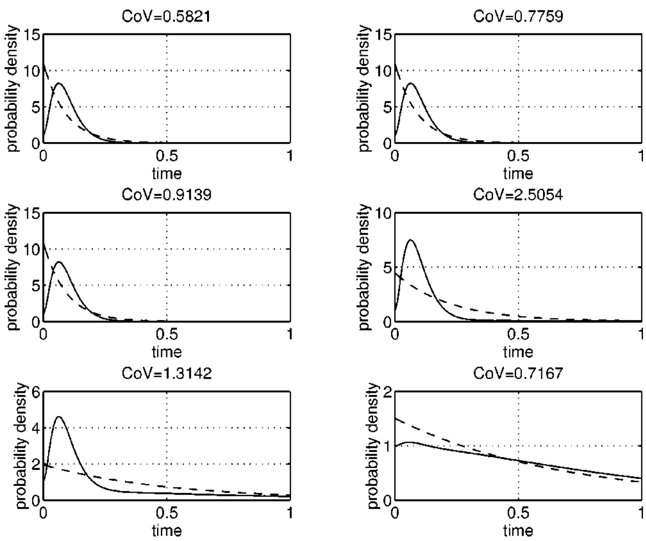

Figure 6. Handoff call channel holding time and its statistical fit (dashed line) when the exponential fit for cell residence time is used.

unity. The mismatch between the channel holding time distribution and its exponential approximation is obvious.

From preceding discussions, we observe that the exponential modeling is not appropriate, the first and second order statistics of cell residence time do not provide sufficient information to determine the channel holding time distribution. The statistical details (the distributions) of the cell residence time are needed to characterize the channel holding time. The hyper-Erlang distribution gives a good set of distribution models for approximation, and our analytical results provide the tools for the analytical evaluation of channel holding time. If we could draw a good approximation using hyper-Erlang model from the field data for users' mobility, then we could fully characterize the channel holding time (and other performance metrics such as blocking probabilities).

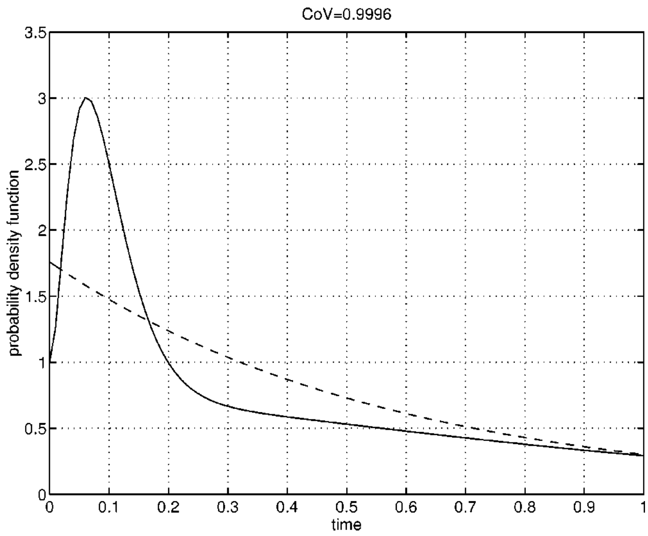

Figure 7. Handoff call channel holding time and the exponential fit (dashed line) for the channel holding time.

5. Conclusions

In this paper, we present some new properties of the hyper-Erlang model we proposed for mobility modeling. We show the generality of such model, which can be used to model not only cell residence time (users' mobility) but also other time variables in wireless networks and mobile computing systems. The future work is to study how the hyper-Erlang model can be used to approximate the network traffic. In the wireless and mobile systems, each cell can be modeled as a queueing system with two streams of arrivals (new calls and handoff calls) with distinct channel holding time distributions, it will be possible to use the multidimensional Markov chain to study the call blocking performance. This work in relation to call admission control is under way, which will be presented in a separate paper.

References

[1] F. Barcelo and J. Jordan, Channel holding time distribution in cellular telephony, in: *The 9th International Conference on Wireless Communications (Wireless'97)*, Vol. 1, Alberta, Canada (9–11 July 1997) pp. 125–134.

[2] V.A. Bolotin, Modeling call holding time distributions for CCS network design and performance analysis, IEEE Journal on Selected Areas in Communications 12(3) (1994) 433–438.

[3] E. Chlebus and W. Ludwin, Is handoff traffic really Poissonian?, in: *IEEE ICUPC'95*, Tokyo, Japan (6–10 November 1995) pp. 348–353.

[4] D.C. Cox, Wireless personal communications: what is it?, IEEE Personal Communications Magazine (April 1995) 20–35.

[5] E. Del Re, R. Fantacci and G. Giambene, Handover and dynamic channel allocation techniques in mobile cellular networks, IEEE Transactions on Vehicular Technology 44(2) (1995) 229–237.

[6] Y. Fang and I. Chlamtac, Teletraffic analysis and mobility modeling for PCS networks, IEEE Transactions on Communications 47(7) (July 1999) 1062–1072.

[7] Y. Fang and I. Chlamtac, A new mobility model and its application in the channel holding time characterization in PCS networks, in: *Proceedings of INFOCOM'99*, New York, NY (March 1999).

[8] Y. Fang, I. Chlamtac and Y.B. Lin, Channel occupancy times and handoff rate for mobile computing and PCS networks, IEEE Transactions on Computers (June 1998).

[9] Y. Fang, I. Chlamtac and Y.B. Lin, Modeling PCS networks under general call holding times and cell residence time distributions, IEEE Transactions on Networking 5(6) (December 1997) 893–906.

[10] Y. Fang, I. Chlamtac and Y.B. Lin, Call performance for a PCS network, IEEE Journal on Selected Areas in Communications 15(8) (October 1997) 1568–1581.

[11] A. Feldmann, A.C. Gilbert and W. Willinger, Data networks as cascades: investigating the multifractal nature of Internet WAN traffic, in: *Proceedings of the ACM/SIGCOMM'98*, Vancouver, Canada (September 2–4, 1998).

[12] G.J. Foschini, B. Gopinath and Z. Miljanic, Channel cost of mobility, IEEE Transactions on Vehicular Technology 42(4) (1993) 414–424.

[13] R.A. Guerin, Channel occupancy time distribution in a cellular radio system, IEEE Transactions on Vehicular Technology 35(3) (1987) 89–99.

[14] D. Hong and S.S. Rappaport, Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and non-prioritized handoff procedures, IEEE Transactions on Vehicular Technology 35(3) (1986) 77–92.

[15] C. Jedrzycki and V.C.M. Leung, Probability distribution of channel holding time in cellular telephony systems, in: *IEEE Vehicular Technology Conference (VTC'96)* (1996) pp. 247–251.

[16] J. Jordan and F. Barcelo, Statistical modelling of channel occupancy in trunked PAMR systems, in: *The 15th International Teletraffic Conference (ITC'15)*, eds. V. Ramaswami and P.E. Wirth (Elsevier, 1997) pp. 1169–1178.

[17] F.P. Kelly, *Reversibility and Stochastic Networks* (Wiley, New York, 1979).

[18] L. Kleinrock, *Queueing Systems: Theory*, Vol. I (Wiley, New York, 1975).

[19] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, On the self-similar nature of Ethernet traffic (extended version), IEEE/ACM Transactions on Networking 2 (1994) 1–15.

[20] W.R. LePage, *Complex Variables and the Laplace Transform for Engineers* (Dover, New York, 1980).

[21] Y.B. Lin, S. Mohan and A. Noerpel, Queueing priority channel assignment strategies for handoff and initial access for a PCS network, IEEE Transactions on Vehicular Technology 43(3) (1994) 704–712.

[22] A.R. Noerpel, Y.B. Lin and H. Sherry, PACS: Personal access communications system – a tutorial, IEEE Personal Communications 3(3) (June 1996) 32–43.

[23] P.V. Orlik and S.S. Rappaport, Traffic performance and mobility modeling of cellular communications with mixed platforms and highly variable mobilities, Proceedings of the IEEE (1998) (to appear).

[24] P.V. Orlik and S.S. Rappaport, A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions, IEEE Journal on Selected Areas in Communications 16(5) (1998) 788–803.

[25] W. Stallings, *High-Speed Networks: TCP/IP and ATM Design Principles* (Prentice-Hall, 1998).

[26] W. Willinger, M. Taqqu, R. Sherman and D. Wilson, Self-similarity through high-variability: statistical analysis of Ethernet traffic at the source level, in: *Proc. ACM SIGCOMM'95* (1995) pp. 100–113.

Yuguang Fang received the BS and MS degrees in mathematics from Qufu Normal University, Qufu, Shandong, China, in 1984 and 1987, respectively, a Ph.D. degree in systems and control engineering from Department of Systems, Control and Industrial Engineering at Case Western Reserve University, Cleveland, OH, in January 1994, and a Ph.D. degree in electrical engineering from Department of Electrical and Computer Engineering at Boston University, MA, in May 1997. From 1987 to 1988, he held research and teaching positions in both Department of Mathematics and the Institute of Automation at Qufu Normal University. From September 1989 to December 1993, he was a teaching/research assistant in Department of Systems, Control and Industrial Engineering at Case Western Reserve University, where he held a research associate position from January 1994 to May 1994. He held a post-doctoral position in Department of Electrical and Computer Engineering at Boston University from June 1994 to August 1995. From September 1995 to May 1997, he was a research assistant in Department of Electrical and Computer Engineering at Boston University. From June 1997 to July 1998, he was a Visiting Assistant Professor in Department of Electrical Engineering at the University of Texas at Dallas. From July 1998 to May 2000, he was an Assistant Professor in the Department of Electrical and Computer Engineering at New Jersey Institute of Technology, Newark, NJ. Since May 2000, he has been an Assistant Professor in Department of Electrical and Computer Engineering at the University of Florida, Gainesville, FL. He has published more than 40 papers in the professional journals and refereed conferences. He is an Editor for IEEE Journal on Selected Areas in Communications: Wireless Communications Series, an Editor for IEEE Transactions on Communications, an Editor for Wiley International Journal on Wireless Communications and Mobile Computing, an Area Editor for ACM Mobile Computing and Communications Review. His recent research interests include wireless networks and mobile communications, personal communication services (PCS), call admission control and resource allocations. Dr. Fang is the Program Vice-Chair for the 2000 IEEE Wireless Communications and Networking Conference (WCNC'2000), the Program Vice-Chair for the Seventh International Conference on Computer Communications and Networking (IC^3N'98), a member of Technical Program Committee for the International Conference on Computer Communications (INFOCOM'98 and INFOCOM'00), a member of the Technical Program Committee for the First International Workshop for Wireless and Mobile Multimedia (WOW-MoM'98), and a member of the Technical Program Committee for the 1999 IEEE Wireless Communication and Networking Conference (WCNC'99).
E-mail: fang@ece.ufl.edu