Mobile Netw Appl (2007) 12:31–41
DOI 10.1007/s11036-006-0004-y

# Queue Analysis and Multiplexing of Heavy-tailed Traffic in Wireless Packet Data Networks

**Shahram Teymori · Weihua Zhuang**

Published online: 14 December 2006

**Abstract** Recent research based on traffic measurements shows that Internet traffic flows have a fractal nature (i.e., self-similarity property), which causes an underestimation of network engineering parameters when using the conventional Poisson model. Preliminary field measurements demonstrate that packet data traffic in wireless communications also exhibits self-similarity. In this paper, we investigate the queuing behavior of self-similar traffic flows for data applications in a packet-switching single-server wireless network. The traffic is generated by an on–off source with heavy-tailed on periods and exponentially distributed off periods. We extend previous analysis of a relation among the asymptotic distribution of loss probability, traffic specifications, and transmission rate for a wireline system to a wireless system, taking into account wireless propagation channel characteristics. We also investigate the multiplexing of heavy-tailed traffic flows with a finite buffer for the downlink transmission of a wireless network. Computer simulation results demonstrate that assumptions made in the theoretical analysis are reasonable and the derived relationships are accurate.

**Keywords** delay · heavy-tailed traffic · loss probability · quality-of-service (QoS) · queueing analysis · resource allocation · self-similarity · statistical multiplexing · wireless packet transmission

## 1 Introduction

Wireless communication systems have been revolutionized by technological advances in the last decade. The third generation (and beyond) wireless networks use and will continue to use packet-switching technologies to provide high-speed data services such as File Transportation Protocol (FTP) and web browsing. These applications have stringent performance requirements in terms of throughput and transmission accuracy. Data traffic flows can tolerate a certain degree of transmission delay, depending on the application; however, they are sensitive to transmission errors and can normally tolerate a bit error rate (BER) up to $10^{-6}$. Selection of a proper model to analyze the queueing behaviors of network traffic flows plays an important role in network engineering. Recently, it has been shown that the conventional Poisson traffic model is not proper for packet data traffic in the Internet [1]. It is also observed that the aggregated Internet traffic flows have similar patterns in different time scales, referred to as self-similarity [2]. Similarly, the existence of self-similarity in wireless network traffic flows has been observed [3] and has attracted attention in the research community [4]. The self-similarity characteristics result in that

S. Teymori (✉)
Automation Systems Division, ATS Automation Tooling Systems Inc., 250 Royal Oak Rd., Cambridge, Ontario, Canada N3H 4R6
e-mail: steymori@atsautomation.com

W. Zhuang
Centre for Wireless Communications (CWC), Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

W. Zhuang
e-mail: wzhuang@uwaterloo.ca

traffic engineering techniques, based on the traditional Poisson model, underestimate the required resources (such as transmission rate) to achieve quality-of-service (QoS) satisfaction.

Willinger et al. prove that the superposition of a large number of on–off traffic sources, with heavy-tailed on or off periods, results in self-similar aggregated traffic [5]. The finding is important, as it indicates that heavy-tailed on–off periods can be the reason for self-similarity in TCP/IP[1] traffic. The result is confirmed in [6], which shows that individual data traffic source can be modelled by an on–off source with heavy-tailed on or off periods.

Unfortunately the network analysis with heavy-tailed traffic sources is very complicated. For example, in an M/G/1 queueing system, the average packet delay is proportional to the variance of the service time [7]. In the presence of heavy-tailed on–off sources, the service time variance is infinite, which implies an infinite value for the average delay. Furthermore, it can be shown that, for on–off sources with heavy-tailed on periods, the moment generation function of the service time is infinite [7]. This means that the Chernoff bound can not be used to analyze the queue delay performance. In [8, 9], a relation among queue length distribution, transmission rate, and traffic characteristics is derived for a single server system with a single on–off source in a wireline network for an infinite buffer size and a finite buffer size, respectively. The analysis is carried out under the assumption that the network transmission rate (server capacity) is a constant. The assumption is not valid in wireless communications. Due to the time-variant fading dispersive propagation medium, the link capacity in a wireless system changes with time, and the server capacity is not constant.

In this research, we extend previous analysis [8–10] for queueing behaviors and scheduling of self-similar packet data traffic in a wireline single-server network to a wireless environment. We derive relations among the packet loss probability, traffic specifications, channel characteristics, and capacity for both infinite buffer and finite buffer cases. In addition, we apply the relation for a finite buffer to packet scheduling in the downlink transmission of a wireless network, based on the general process sharing (GPS) principle [11]. This study provides insights of the impact of the self-similar property on network resource allocation for QoS provisioning. The remainder of this paper is organized as follows. Section 2 presents mathematical definitions and describes the system model. Details of the queue analysis with a self-similar traffic input for an infinite buffer size is given in Section 3. We study the queueing behavior for a single server and single input with a finite buffer size in Section 4. Section 5 investigates the queueing behavior of heavy-tailed sources under the GPS scheduling principle in the wireless network. Conclusions are given in Section 6.

## 2 System model

We first present some mathematical definitions and concepts, which are used in this work.

**Definition 1** [8] A cumulative distribution function (CDF) $F$ on $[0, \infty]$ is called heavy-tailed ($F \in \L$) if

$$\lim_{x\to\infty} \frac{1 - F(x-y)}{1 - F(x)} = 1, \quad y \in \Re. \tag{1}$$

**Definition 2** [8] A CDF function $F$ on $[0, \infty]$ is called subexponential ($F \in S$) if

$$\lim_{x\to\infty} \frac{1 - F^{*2}(x)}{1 - F(x)} = 2, \quad y \in \Re \tag{2}$$

where $F^{*2}(x)$ denotes the 2nd convolution of $F$ with itself, i.e, $F^{*2}(x) = \int_0^{+\infty} F(x-y)F(y)dy$.

A well known example of subexponentially distributed functions is functions of regular variation $R_\alpha$ (in particular Pareto family); $F \in R_\alpha$ if it is given by

$$F(x) = 1 - \frac{l(x)}{x^\alpha}, \qquad \alpha \geq 0 \tag{3}$$

where $l(x)$: $\Re_+ \longrightarrow \Re_+$ is a function of slow variation, i.e., $\lim_{x\to\infty} \frac{l(\delta x)}{l(x)} = 1$.

In this work, data traffic flows are modelled by on-off sources. An on–off source can be considered as a renewal process, with the $n$th renewal period $T_n = \tau_n^{on} + \tau_n^{off}$, $n \geq 0$, where $\tau_n^{on}$ and $\tau_n^{off}$ are the durations of on and off periods, respectively. The on periods are i.i.d. and follow a Pareto distribution with

$$P[\tau_n^{on} > x] = \frac{b}{x^\alpha}, \quad x \geq 0, \alpha > 0. \tag{4}$$

In this paper, we are interested in the case where $\alpha \in (1,\ 2)$, corresponding to an infinite variance of $\tau_n^{on}$ and a finite mean of $\tau_n^{on}$, given by $\bar{\tau}_n^{on} = E[\tau_n^{on}] = b\alpha/(\alpha - 1)$. The off periods are i.i.d. and follow an exponential distribution with

$$P[\tau_n^{off} > x] = e^{-\lambda x}, \quad x \geq 0,\ \lambda > 0$$

with $\bar{\tau}^{off} = E[\tau^{off}] = 1/\lambda$. During each on period, packets are generated at a constant rate, $r$. The probability

[1] Transmission Control Protocol/Internet Protocol.

that the source is generating packets at any time is given by

$$P_{on} = \frac{\bar{\tau}^{on}}{\bar{\tau}^{on} + \bar{\tau}^{off}}.$$

Let $A_n$ and $R_n$, $n \geq 0$, be two sequences of i.i.d. random variables, representing the (equivalent) numbers of packet arrivals and departures in the $n_{th}$ recursion, respectively. Let $X_n$ $(= A_n - R_n)$ denote the net increment of the packet number over the recursion.

Consider the transmission of data packets of equal length over a wireless channel. With a constant transmission rate, the transmission time for a packet is constant, referred to as packet time. An important QoS parameter in data service is the transmission accuracy, which can be described by the transmission BER at the physical layer. Consider the transmission over a wireless channel. Given the modulation and coding scheme, the channel model and parameters, and the receiver structure, the required BER can be mapped one-to-one to the required received signal to interference-plus-noise ratio (SINR)[12]. For a given maximum transmit power, the required SINR may not be achieved and the data transmission stops if the channel gain is below a pre-determined threshold (i.e., the channel experiences deep fading). When the channel condition improves over time (to a non deep fading condition), the required SINR can be ensured and the transmission over the channel can resume. As a result, we use a two-state i.i.d. channel model. The channel is said to be in a nonworking state in the former case and in a working state in the latter case. If the channel fades slowly with respect to the packet time, the channel state very likely remains the same over a packet time. Given the channel fading statistics, the working-state probability $P_w$ and nonworking-state probability $P_{nw}$ $(= 1 - P_w)$ can be calculated. The transmission system for each packet traffic flow can be described by a simple queueing model, where the packets are generated by a single on–off source and are transmitted through a two-state fading channel, as illustrated in Fig. 1. In the working state of the channel, the server transmits packets at a constant rate of $C$, and in the nonworking state the server stops the transmission to ensure the transmission accuracy and to save the resources. The scheduling principle based on the channel condition achieves a high utilization efficiency of the limited radio resources and has been used in third-generation wireless systems, e.g., in 1× evolution (1×EV) for code-division multiple access 2000 (cdma2000) and in high speed downlink packet access (HSPDA) for wideband CDMA. Let $c(t)$ denote the time varying channel capacity. Then, $c(t) = C$ if the channel is in the working state and $c(t) = 0$ if the channel is in the nonworking state, at time $t$. The QoS requirements of the data traffic are specified in terms of transmission accuracy and delay: For the case of an infinite buffer size, the probability that the queue length $Q$ exceeds the threshold $q_{th}$ should be less than a small value $\epsilon_1$, i.e., $P(Q > q_{th}) < \epsilon_1$; and for the case of a finite buffer size $B$, the packet loss probability should be less than a small value $\epsilon_2$, i.e., $P(Q = B) < \epsilon_2$. The scheduler calculates and allocates the bandwidth of $c(t)$ by considering the channel status and traffic specifications, which are assumed to be known to the scheduler. Given traffic source and wireless channel statistics, our objective is to determine the minimum capacity $C$ in order to guarantee the QoS requirements.

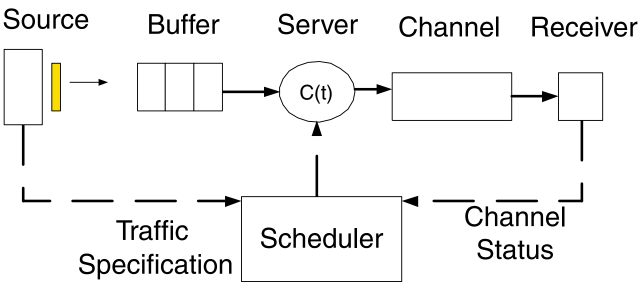

**Fig. 1** Single server single input wireless system

## 3 Queueing analysis for infinite buffer size

We start the analysis by using a classical result given in [13]. The queueing process in Fig. 1 for an infinite buffer size can be described by (Lindley recursion)

$$Q_{n+1} = (Q_n + X_n)^+, \qquad n \geq 0 \tag{5}$$

where $q^+ = \max(q, 0)$, and $Q_n$ is the queue size at the beginning of the $n$th recursion. According to [13], this recursion admits a unique stationary solution and, for all initial conditions, the probability $P[Q_n > x]$ converges to the stationary probability of $P[Q > x]$ under stability condition $E[X_n] < 0$. In the following derivations, we assume that the queueing system under consideration is in its stationary regime.

The residual life of a renewal process is defined as the duration between some fixed time $t$ and the starting point of the following renewal. It is one of the random variables that describe the local behavior of a renewal process. Another variable is the age at time $t$, defined as the time already elapsed in the current renewal. When we look at a renewal process in the reverse (backward) direction, we again observe a renewal process having the same probability structure, but the residual life time is called age [14]. Letting $F(x)$ denote the CDF of $X_n$, the CDF of the residual life, $F1(x)$, is given by

$$F1(x) = \frac{\int_0^x [1 - F(u)]du}{\bar{X}_n}. \tag{6}$$

$F1(x)$ is also called the integral tail distribution of $F(x)$.

**Theorem 1** [8] *If $F_X(x)$ is heavy-tailed with $E[X_n] < 0$ and $F1_X(x)$ is subexponential, for (5) it can be shown that*

$$P[Q_n > x] \sim \frac{\int_x^\infty P[X_n > u]du}{-E[X_n]}, \quad x \to \infty$$

*where $l(x) \sim j(x)$ denotes $l(x)/j(x) = 1$ as $x \to \infty$.*

**Lemma 1** [8] *Let $X$ and $Y$ be two independent random variables distributed as $F(x)$ and $G(x)$, respectively. If $F(x)$ is heavy tailed, $E[X]$ and $E[Y]$ are finite, and $Y$ is a positive non-heavy-tailed random variable, then*

$$P[X - Y > x] \sim P[X > x] = 1 - F(x), \quad x \to \infty.$$

Consider the system illustrated in Fig. 1, during the interval of $\Gamma_n = \tau_n^{on} + \tau_n^{off}$, the state of the channel may change several times, so does the capacity. We define $\Omega_n$ (in second) as the summation of all the time intervals (during the $\Gamma_n$ period) over which the channel is in the working state. In the working state with capacity $C$, if there are backlogged packets in the queue, the number of the transmitted packets is $C\Omega_n$. We define the effective capacity of the time varying channel as $c_n = \frac{C\Omega_n}{\Gamma_n} < r$. Instead of the time-varying capacity $c(t)$, the effective capacity that is constant (even though random) during $\Gamma_n$ will be considered.

Consider the evolution of the queue length at the initial moment of the on period, denoted by $Q_n^p$, where the superscript $p$ stands for Palm probability (meaning that the queue length is observed at the beginning of the on period). According to the definition of $X_n$, it can be shown that $X_n = r\tau_n^{on} - C\Omega_n$, and we have

$$\begin{aligned}
Q_{n+1}^p &= (Q_n^p + X_n)^+ \\
&= (Q_n^p + r\tau_n^{on} - C\Omega_n)^+ \\
&= [Q_n^p + r\tau_n^{on} - c_n(\frac{\Gamma_n}{\Omega_n})\Omega_n]^+ \\
&= [Q_n^p + r(\tau_n^{on}) - c_n(\tau_n^{on} + \tau_n^{off})]^+ \\
&= [Q_n^p + \tau_n^{on}(r - c_n) - c_n\tau_n^{off}]^+.
\end{aligned}$$

By Lemma 1, for $x \to \infty$, we have

$$\begin{aligned}
P[X_n > x] &= P[\tau_n^{on}(r - c_n) - c_n\tau_n^{off} > x] \\
&\sim P[\tau_n^{on}(r - c_n) > x].
\end{aligned}$$

Since

$$P[X_n > x] = E_{c_n}[P(X_n > x|c_n)]$$

we have

$$\begin{aligned}
P[X_n > x] &\sim E_{c_n}\{P[\tau_n^{on}(r - c_n) > x|c_n]\} \\
&= E_{c_n}\left[\frac{b\,(r - c_n)^\alpha}{x^\alpha}\right] \\
&= \frac{bk}{x^\alpha}
\end{aligned} \tag{7}$$

where $k = E_{c_n}[(r - c_n)^\alpha]$. Equation (7) shows that $X_n$ also follows a Pareto distribution. Under the assumption that the on and off periods are stationary, it can be shown that

$$\begin{aligned}
E[X_n] &= E_{c_n}[\bar{\tau}_n^{on}(r - c_n) - c_n\bar{\tau}_n^{off}] \\
&= r\bar{\tau}^{on} - (\bar{\tau}^{on} + \bar{\tau}^{off})\bar{c}_n.
\end{aligned}$$

Assuming that $c_n$ is an ergodic process, we have $\bar{c}_n = CP_w$ and

$$\bar{X}_n = r\bar{\tau}^{on} - (\bar{\tau}^{on} + \bar{\tau}^{off})CP_w. \tag{8}$$

Considering the stability condition $\bar{X}_n < 0$, it is required that

$$rP_{on} < CP_w.$$

As the probability density function (PDF) of $c_n$ is difficult to obtain in our system model, we need to estimate the value of $c_n$ to proceed further. As a large queue length is mainly the result of large $\Gamma_n$ values (large bursts), the probability of $P[Q > x]$ for a large $x$ depends mainly on large $\Gamma_n$ values. With a large $\Gamma_n$ value, we assume that $\frac{\Omega_n}{\Gamma_n} \simeq P_w$ and the effective capacity $c_n \simeq CP_w$. In the case of a large queue, $X_n$ can be approximately estimated by

$$X_n \simeq \tau_n^{on}(r - CP_w) - \tau_n^{off}CP_w.$$

Due to the fact that the distribution of $X_n$ is heavy tailed, by Theorem 1, it can be shown that, as $x \to \infty$,

$$P[Q^p > x] \sim \frac{bk}{[CP_w(\bar{\tau}^{on} + \bar{\tau}^{off}) - r\bar{\tau}^{on}](\alpha - 1)x^{\alpha-1}}. \tag{9}$$

Substituting $k = E_{c_n}[(r - c_n)^\alpha] \simeq (r - CP_w)^\alpha$, we have, as $x \to \infty$,

$$P[Q^p > x] \approx \frac{b\,(r - CP_w)^\alpha}{[CP_w(\bar{\tau}^{on} + \bar{\tau}^{off}) - r\bar{\tau}^{on}](\alpha - 1)x^{\alpha-1}} \tag{10}$$

where $l(x) \approx j(x)$ denotes $l(x)/j(x) \simeq 1$ as $x \to \infty$.

Next, we extend the distributed for the Palm queue in (10) to the stationary probability $P[Q > x]$. Let $\Lambda_n^{on}$,

$-\infty < n < \infty$, be a random process that represents the beginning of the on periods in the stationary on–off process with $\Lambda_0^{on} < 0 \leq \Lambda_1^{on}$, and $B$ be a Bernoulli random variable with $P(B = 0) = 1 - P(B = 1) = P_{on}$. We define the residual on period, $\tau_n^{on,r}$, as the residual life time of $\tau_n^{on}$ with respect to time 0. Similarly, the residual off period, $\tau_n^{off,r}$, can be defined. $\Lambda_0^{on}$ can be represented by

$$-\Lambda_0^{on} = B(\tau_0^{on} + \tau_0^{off,r}) + (1 - B)\tau_0^{on,r}$$

where $\tau_0^{on,r}$ and $\tau_0^{off,r}$ follow integrated tail distributions of $\tau_n^{on}$ and $\tau_n^{off}$, respectively. The preceding equation means that, if the source is in the on state ($B = 0$) at time $t = 0$ , $-\Lambda_0^{on}$ is the age (residual life) of the on period, $\tau_0^{on,r}$; otherwise, $-\Lambda_0^{on}$ is the duration of the on period plus the age (residual life) of the off period. Furthermore, the net increment $X_0$ of the load that arrives to the queue in the interval $[\Lambda_0^{on}, 0]$ is equal to

$$X_0 = B[\tau_0^{\rm on}(r - c_0) - c_0\tau_0^{off,r}] + (1 - B)[(r - c_0)\tau_0^{on,r}]. \tag{11}$$

Let $Q_{T_0}$ denote the backlogged traffic observed at the beginning of the zeroth renewal period ($t = \Lambda_0^{\rm on}$). The asymptotic probability of the queue length in the zeroth renewal period, $P[Q_0 > x]$, is

$$\begin{aligned} P[Q_0 > x] &= P[Q_0 > x|B = 1]P[B = 1] \\ &\quad + P[Q_0 > x|B = 0]P[B = 0] \\ &= P[Q_{T_0} + (r - c_0)\tau_0^{on} - c_0\tau_0^{off,r} \\ &\quad > x|B = 1]P[B = 1] \\ &\quad + P[Q_{T_0} + (r - c_0)\tau_0^{on,r} \\ &\quad > x|B = 0]P[B = 0]. \end{aligned} \tag{12}$$

Note that $Q_{T_0}$ is equal to the Palm queue $Q_0^p$. Since $Q_{T_0}$ and $\tau_0^{on,r}$ are independent and subexponentially distributed, as $x \to \infty$, we have [15]

$$\begin{aligned} &P[Q_{T_0} + (r - c_0)\tau_0^{on,r} > x|B = 0] \\ &\quad \sim P[Q_{T_0} > x] + P[(r - c_0)\tau_0^{on,r} > x]. \end{aligned} \tag{13}$$

Also by Lemma 1 [8],

$$\begin{aligned} &P[Q_{T_0} + (r - c_0)\tau_0^{on} - c_0\tau_0^{off,r} > x|B = 1] \\ &\quad \sim P[Q_{T_0} > x], \quad x \to \infty \end{aligned} \tag{14}$$

as both $\tau_0^{on}$ and $\tau_0^{off,r}$ under the condition of $B = 1$ are non heavy-tailed. Equation (12) can then be simplified by using (13) and (14), as $x \to \infty$,

$$\begin{aligned} P[Q_0 > x] &\sim P[Q_{T_0} > x] \\ &\quad + P[(r - c_0)\tau_0^{on,r} > x]P(B = 0). \end{aligned} \tag{15}$$

From (4) and (6), we have, for $x \geq 0$,

$$\begin{aligned} P[(r - c_0)\tau_0^{on,r} > x] &= E_{c_0}\left[\frac{\int_{\frac{x}{r-c_0}}^{\infty} \frac{b}{u^\alpha}du}{\bar{\tau}_0^{on}}\right] \\ &= E_{c_0}\left[\frac{b(r - c_0)^{\alpha-1}}{\bar{\tau}_0^{\rm on}(\alpha - 1)x^{\alpha-1}}\right]. \end{aligned} \tag{16}$$

Using $c_0 \simeq CP_w$, we have

$$P[(r - c_0)\tau_0^{on,r} > x] \simeq \frac{b(r - CP_w)^{\alpha-1}}{\bar{\tau}_0^{\rm on}(\alpha - 1)x^{\alpha-1}}. \tag{17}$$

With $Q_{T_0} = Q_0^p$ and $Q$ being stationary (i.e., $P[Q_0 > x] = P[Q_n > x] = P[Q > x]$), applying (9) and (15) to (17), we have

$$\begin{aligned} P[Q > x] \approx &\frac{b(r - CP_w)^\alpha}{[CP_w(\bar{\tau}^{\rm on} + \bar{\tau}^{off}) - r\bar{\tau}^{\rm on}](\alpha - 1)x^{\alpha-1}} \\ &+ \frac{b(r - CP_w)^{\alpha-1}}{(\bar{\tau}^{\rm on} + \bar{\tau}^{off})(\alpha - 1)x^{\alpha-1}} \end{aligned} \tag{18}$$

as $x \to \infty$. In (18), we estimate the asymptotic behavior of $Q$ for a large $x$. By using this relationship, we are able to calculate the required link capacity (transmission rate) $C$ based on traffic parameters ($r$, $\bar{\tau}_n^{on}$, $\bar{\tau}_n^{off}$, $b$, $\alpha$) and channel characteristics ($P_w$) to reach a pre-defined QoS level (in terms of delay specified by $q_{th}$ and $\epsilon_1$).

Due to the complexity of the queue analysis with the heavy-tailed input traffic flow, we have to make several simplified assumptions for tractability in deriving (18). To validate the derivation process, computer simulations were carried out and the results are presented in the following.

The single server single input system illustrated in Fig. 1 is simulated. The information bits of the input traffic flow are generated during each on period at a rate of $r$. The on periods are determined from the sample values of the Pareto random variable, generated by using the random generator as specified in [16]. The off periods are determined from the sample values of the exponential random variable. As the data packets have the same length and are transmitted at a constant rate, without loss of generality, we simulate the generation and transmission of bit flows instead of packet flows.

The simulation parameters are chosen based on the previous measurements for FTP application in wireless networks [17] to be $\bar{\tau}^{on} = 0.256$ s, $\bar{\tau}^{off} = 15$ s, $r = 64$ KBps, and frame duration of 10 ms. Although $\alpha = 1.1$ is suggested in [17], we choose $\alpha = 1.4$ in the simulation, because the accuracy of the Pareto random generator is relatively poor for a small $\alpha$ value. The two-state channel status is characterized by a Bernoulli random process with working probability ($P_w$), changes

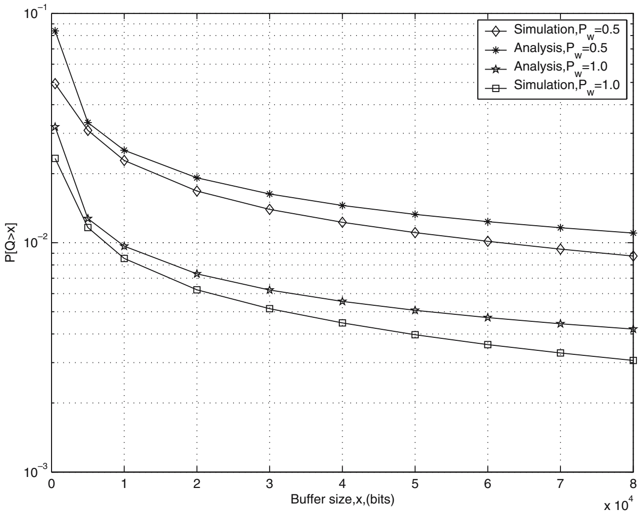

**Fig. 2** Effect of the wireless channel quality on the queue length distribution

independently from frame to frame. Each run in the simulations consists of at least 10e8 frames.

Figure 2 shows the tail distribution of the queue length for the buffer, with $P_w$ being 0.5 and 1.0 respectively and capacity $C$ of 40 KBps. The analytical results are obtained from (18). It is observed that the simulation results agree very well with the analytical results. Figure 3 plots the probability, $P[Q > 8,000\ (\text{bits})]$, as a function of the system capacity $C$ with $P_w$ being 0.5. Again, we observe a close agreement between the analytical results obtained from (18) and the simulation results. Note that, in both Figs. 2 and 3, the simulation results are slightly smaller than the corresponding analytical results, due to the infinite variance of the traffic on periods in the analysis but a limited variance in the simulations.

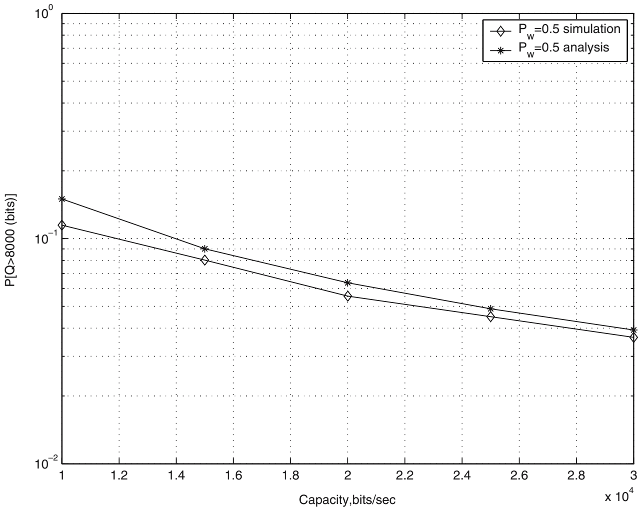

**Fig. 3** Effect of the capacity on the queue length distribution

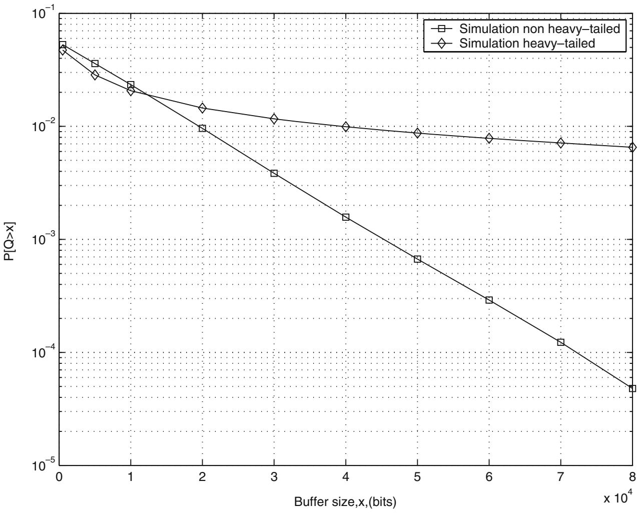

**Fig. 4** Impact of the traffic model on the queue length distribution ($P_w = 0.5$)

Figure 4 shows the impact of the heavy-tailed traffic on the asymptotic tail-distribution of queue length. Two traffic flows with the same average rate and equal expected on and off periods respectively are simulated. One is a non heavy-tailed on–off source, where both on and off periods are exponentially distributed. The other is a heavy-tailed traffic, whose on periods have a Pareto distribution with $\alpha = 1.4$. The system parameters are the same for both inputs with $C = 40$ KBps and $P_w = 0.5$. The distribution for the non heavy-tailed traffic decreases very fast (exponentially), but not in the case of the heavy-tailed source. This is the most important impact of the heavy-tailed traffic flows on network performance. It is observed that the distribution for the heavy-tailed traffic has a relatively large value even for a very large queue length, which results in the possibility of a very large delay and a large probability of packet loss.

## 4 Queueing analysis for finite buffer size

For wireless packet data services, the main performance parameters are transmission BER and packet loss rate due to buffer overflow. In addition, for interactive data applications (where traffic also exhibits the heavy-tailed property), transmission delay is another important performance parameter. The transmission BER requirement is to be met by ensuring that the received SINR is not below the required threshold, while the packet loss probability and transmission delay requirements are to be satisfied by properly choosing the link capacity $C$ and buffer size. In the following, we focus on queueing analysis for packet loss probability and queueing delay, given that the BER requirement can be satisfied by a proper selection of channel parameters (i.e., the channel gain threshold and, therefore, the channel working state probability $P_w$). We consider a queueing model with only one on–off source as shown in Fig. 1, where the buffer size is finite, denoted by $B$. In the following, we carry out the queueing analysis for the performance measures, under the assumption of a heavy-tailed traffic source. The analysis helps to determine the wireless link capacity in order to satisfy both transmission accuracy and delay requirements. Due to the complexity of the analysis, the approach of analyzing the corresponding fluid renewal process [8–10] is used to study the heavy-tailed traffic in the system here.

Consider the system model shown in Fig. 1, where the renewal process of the on–off traffic source has the $n$th renewal period $T_n = \tau_n^{on} + \tau_n^{off}$. At the beginning of the on periods, the queue length evolves according to[2]

$$Q_{n+1}^B = \{\min[(Q_n^B + (r - c_n)\tau_n^{on}, B] - c_n\tau_n^{off}\}^+, \quad n \geq 0 \tag{19}$$

where $Q_n^B$ is the queue length at the beginning of the $n$th recursion. Let $A_n = (r - c_n)\tau_n^{on}$ and $R_n = c_n\tau_n^{off}$. The effective capacity $c_n$ is a random variable bounded by $r$, depending on the channel condition, but remains unchanged over each renewal period. It has been shown that, for all initial conditions, $Q_n^B$ converges to a stationary distribution [9]. Here, we consider that the recursion (19) is in its stationary regime.

**Theorem 2** [9] *If $A$ is subexponential and $EA < ER$, the stationary number of lost packets $E(Q_n^B + A_n - B)^+$ satisfies $E(Q_n^B + A_n - B)^+ = E(A - B)^+(1 + o(1))$ as $B \to \infty$.*

Note that $o(1) \to 0$ as $B \to \infty$. The asymptotic number of lost packets, for a heavy-tailed source, is independent of $R_n$.

Similar to the derivation of (7), it can be derived that

$$\begin{aligned} P[A_n > x] &\sim E_{c_n}(P[\tau_n^{on}(r - c_n) > x | c_n]) \\ &= E_{c_n}[\frac{b(r - c_n)^\alpha}{x^\alpha}] \\ &= \frac{bk}{x^\alpha}, \quad x > 0 \end{aligned} \tag{20}$$

i.e., $A_n$ follows a Pareto distribution.

The condition ($EA < ER$) in Theorem 2 is to be satisfied by a properly chosen capacity $C$ value, which is a design constraint. Now that both the necessary conditions for Theorem 2 can be satisfied, in the following, we first derive the average loss rate $\lambda_{loss}^B$ in the fluid system based on Theorem 2 and then obtain the packet loss probability. The long-term time average loss rate for the fluid queue is defined as

$$\lambda_{loss}^B \equiv \lim_{t \to \infty} \frac{L(0, t)}{t} \tag{21}$$

where $L(0, t)$ is the amount of fluid lost in $(0, t)$. Let $L_n \equiv E[Q_n^B + (r - c_n)\tau_n^{on} - B]^+$, where $n \geq 0$ and the expectation is taken with respect to the random on period, be a sequence of random variables representing the number of lost packets in the $n$th renewal period. Letting $N_t = \sup[n : T_n < t]$, we have [9]

$$\sum_{n=1}^{N_t - 1} L_n \leq L(0, t) \leq \sum_{n=0}^{N_t} L_n \tag{22}$$

The strong law of large numbers for a renewal process yields

$$\lim_{t \to \infty} \frac{N_t}{t} = \frac{1}{\bar{\tau}^{on} + \bar{\tau}^{off}} \quad \text{almost surely.} \tag{23}$$

Similarly, (23) and Birkhoff's strong law of large numbers imply [9]

$$\lim_{N_t \to \infty} \frac{\sum_{n=0}^{N_t} L_n}{N_t} = E[L_1] \quad \text{almost surely.} \tag{24}$$

By dividing (22) with $t$ and letting $t \to \infty$ and using (23) and (24), we have

$$\lambda_{loss}^B = \frac{E[Q_n^B + (r - c_n)\tau_n^{on} - B]^+}{\bar{\tau}^{on} + \bar{\tau}^{off}}(1 + o(1)) \tag{25}$$

as $B \to \infty$. Considering that $c_n$ is a random variable, from (25), we have

$$\lambda_{loss}^B = \frac{E_{c_n}\{E[(Q_n^B + (r - c_n)\tau_n^{on} - B)^+ | c_n]\}}{\bar{\tau}^{on} + \bar{\tau}^{off}}. \tag{26}$$

[2] Another way of representing the queue length evolution is to use $\tilde{Q}_{n+1}^B = \min[(\tilde{Q}_n^B + A_n - R_n)^+, B]$. Similar to the analysis in [9], it can be shown that the loss rates for both queue length representations are asymptotically equivalent. Note that the $Q_{n+1}^B$ and $\tilde{Q}_{n+1}^B$ representations correspond to the upper and lower bounds of the packet loss number in each renewal period.

From Theorem 2, given $c_n$,

$$E[Q_n^B + (r - c_n)\tau_n^{on} - B]^+$$

$$\sim E[(r - c_n)\tau_n^{on} - B]^+$$

$$= \int_B^\infty P[(r - c_n)\tau_n^{on} > x]dx$$

$$= \frac{b\,(r - c_n)^\alpha}{(\alpha - 1)\,B^{\alpha-1}} \qquad (27)$$

Substituting (27) to (26) gives

$$\lambda_{loss}^B \simeq \frac{b\,E[(r - c_n)^\alpha]}{(\bar{\tau}_n^{on} + \bar{\tau}_n^{off})(\alpha - 1)\,B^{\alpha-1}}. \qquad (28)$$

Letting $G(0, t)$ denote the amount of time that the buffer is full, we have

$$G(0, t) = \frac{L(0, t)}{E[r - c_n]}.$$

As it is very difficult to obtain the PDF of the effective capacity $c_n$ in our system model, we approximate $c_n$ by its mean value in the following, i.e., $c_n \simeq CP_w$. By ergodicity of $Q^B(t)$ and (28), as $B \to \infty$,

$$P[Q^B = B] = \lim_{t\to\infty} \frac{G(0, t)}{t} = \frac{\lambda_{loss}^B}{r - CP_w}$$

$$\simeq \frac{b\,(r - CP_w)^{\alpha-1}}{(\bar{\tau}^{on} + \bar{\tau}^{off})(\alpha - 1)\,B^{\alpha-1}}. \qquad (29)$$

Equation (29) establishes a relationship among the traffic parameters, buffer size, network capacity, channel quality, and loss probability. The transmission delay requirement specifies the maximum buffer size $B$, and the packet loss probability specifies the minimum link capacity $C$.

Computer simulations were carried out to validate (29) which is derived under several simplified assumptions for tractability. The single server single input system illustrated in Fig. 1 is simulated with the parameters the same as those given in Section 3, but with a finite buffer size. Figure 5 shows the loss probability for different buffer sizes, with $P_w$ being 0.5 and 1.0, respectively. The analytical results are obtained from (29). It is observed that the simulation results agree very well with the analytical results when the buffer size is relatively large (e.g., $B \geq 2,000$ bits). Note that the relation given by (29) is derived under the assumption of a large buffer size. When the buffer size is very small (e.g., $B < 1,000$ bits), the analysis is not accurate. For a large buffer size, the loss probability from the simulation is slightly smaller than that from the analysis, again due to the finite variance of the heavy-tailed on periods in the simulation. It is observed that, as the buffer size increases the loss probability decreases, as expected.

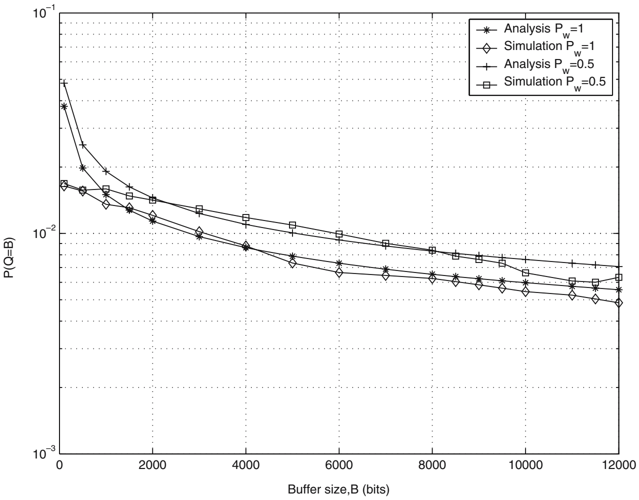

**Fig. 5** Asymptotic loss probability for $P_w$ equal to 0.5 and 1.0, respectively

However, the rate of the loss probability decrease versus the buffer size increase is low. Figure 6 plots the probability, $P[Q = 4,000\ \text{(bits)}]$, as a function of the system capacity $C$ with $P_w = 0.7$. Again, we observe a close agreement between the analytical results obtained from (29) and the simulation results. The small difference between the analytical and simulation results is due to the assumption of a large buffer size in the analysis and a limited variance of the traffic on period in the simulation. It is observed that the loss probability is more sensitive to a capacity change closer to the rate $r = 64$ KBps. Increasing the capacity to reduce the loss rate is effective only when the capacity is large, which is an impact of the heavy-tailed on periods. In order to guarantee the loss probability, we not only should provide sufficient capacity but also should reserve a large buffer space for heavy tailed traffic, if the delay requirement permits.

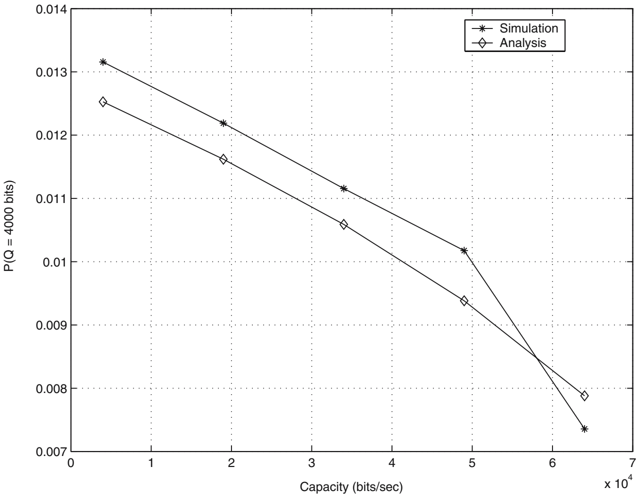

**Fig. 6** Effect of the capacity on loss probability ($P_w = 0.7$)

## 5 Multiplexing of heavy-tailed traffic sources

Consider a network model as illustrated in Fig. 7, which represents a typical downlink transmission system in centralized wireless communications where the buffer sharing is possible (e.g., at the base station of a cellular system or the access point of a wireless local area network). There are $N$ homogeneous heavy-tailed on–off sources, whose on periods are i.i.d. Pareto distributed and whose off periods are i.i.d. exponentially distributed. All flows are queued in a common buffer with a finite size of $B$. Buffer sharing is unrestricted as long as there is available space, i.e., the workloads evolve as if they were in an infinite buffer. When the buffer fills up, the flows with the maximum number of packets in the buffer are subject to penalty. They experience a minimum necessary loss probability in order to accommodate other flows with smaller workloads [10]. These flows share the network total capacity $C_T$ but have their own i.i.d. two-state radio channels. Without losing the generality, we assume that the capacity is fairly divided among flows with a good SINR value in periodic and tiny time intervals, referred to as frames. We assume that, during each frame, the status of every channel does not change.

Equation (29) can be used to determine the necessary buffer size and capacity to guarantee the loss probability and transmission (queueing) delay for the system with a single traffic source. For multiple heavy-tailed traffic flows with statistical multiplexing, because of the on and off nature of traffic sources and random alternations of the radio channel states, the problem gets more complex. The GPS principle, which is the core of many practical scheduling schemes, guarantees the minimum capacity for each flow [19]. It can be used for the scheduler to guarantee the minimum necessary capacity, as specified by (29). For example, with $N$ homogeneous heavy-tailed traffic flows, each of which requires at least $\xi$ KBps to guarantee the necessary QoS, the GPS scheduler can guarantee the QoS by setting the total capacity as $\xi N$ KBps. However, this is not the optimum way to use the limited radio resources, because it does not explore the multiplexing gain from the multiple sources sharing the capacity. In this section, we study the multiplexing of heavy-tailed sources in a wireless network, under the GPS scheduling principle. Based on computer simulation, we investigate the capacity sharing and propose a heuristic for calculating the necessary capacity, taking into account the QoS requirements, wireless channel characteristics, and traffic parameters.

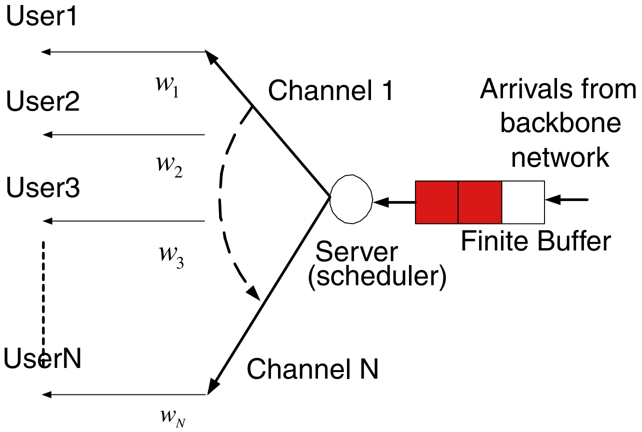

**Fig. 7** Scheduler for downlink communications

A similar scheduling problem has been studied in [10] for a wireline network having $N$ heavy-tailed sources with a total capacity $C_T$. Flow $i$ has a minimum service rate guarantee $\rho_i$ that exceeds its long-term average demand, $i = 1, 2, \ldots, N$. It has been shown that the loss rate of a particular flow $i$ is asymptotically equal to the loss rate in a reduced system with capacity of $(C_T - \sum_{j\neq i} \rho_j)$ and buffer of size $B$, where this flow $i$ is served in isolation [10]. In this case, (29) can still be used to study the queueing behavior.

Letting $w_i$ denote the weight of flow $i$ under GPS for resource sharing, a necessary system stability condition is $\rho_i \leq w_i P_w C_T$. As an example, consider homogeneous traffic sources with $N = 10$, $C_T = 64$ KBps, and $\rho_i = \rho$, $i = 1, 2, \ldots, N$, in the system illustrated in Fig. 7 for FTP applications. For $P_w = 0.7$ and 1.0, respectively, and for different values of buffer size $B$, we measure the percentage of time during which flow $i$ $(= 5)$ causes buffer overflow, $P[Q_5 = B]$. Also, we measure the percentage of time during which the total queue $(Q_T = \sum_{i=1}^{N} Q_i)$ experiences overflow, $P[Q_T = B]$. Taking into account the similar asymptotic behaviors of all the homogeneous flows in the system, each of them causes overflow independently with the same probability. Under the assumption that the probability of two or more sources simultaneously cause an overflow is low, the probability of any source causing an overflow can be estimated by $P[\text{buffer overflows}] \simeq N \cdot P[\text{flow } i \text{ overflows}]$, i.e., $P[Q_T = B] \simeq N \cdot P[Q_i = B]$ where $i \in \{1, 2, \ldots, N\}$.

Figure 8 shows $N \cdot P[Q_5 = B]$ and $P[Q_T = B]$ based on simulations. The results support the approximation $P[Q_T = B] \simeq N \cdot P[Q_i = B]$. This is consistent with the idea that, under a strict stability condition in GPS $(\rho_i \leq w_i P_w C_T)$, the workload build-up of traffic flow $i$ is unlikely to be caused by other flows [10]. Furthermore, considering the heavy tailed nature of the on periods of flow $i$, it can be said that the overflow scenario occurs most likely because of a single long

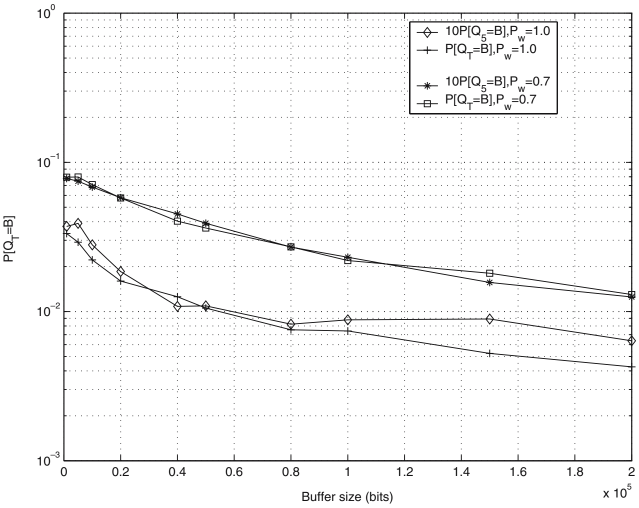

**Fig. 8** Asymptotic total loss probability based on simulations of ten traffic flows

on period of flow $i$. Hence, during an overflow, with a very high probability all other flows exhibit an average behavior with their average rate $\rho$. That is, during a long on period of flow $i$, the other flows asymptotically consume a total capacity of $(N-1)\rho$ on average. This translates to that, during its on periods, flow $i$ asymptotically uses the available capacity when its channel is in a good state with probability $P_w$. As a result, the average total capacity available to flow $i$ can be estimated by

$$C_i = [C_T - (N-1)\rho]P_w. \tag{30}$$

In addition, the average behavior of flow $j$ $(\neq i)$ yields that its build-up queue $(Q_j)$ behaves as $Q_j = O(1)$ and, thus, flow $i$ can potentially occupy up to $B - O(1)$ buffer space, where $O(1) \to 0$ when $B \to \infty$. For the system shown in Fig. 7, the asymptotic loss probability of flow $i$ can be approximately estimated by the loss probability of flow $i$ in a single input system shown in Fig. 1, with the capacity specified by (30) and buffer size $B$. The above argument is confirmed by preliminary simulation results given in Fig. 9. For different values of the buffer size, we measure the percentage of time that flow $i = 5$ causes an overflow. For the system shown in Fig. 1, we set the capacity by (30) and calculate the loss probability by using (29). It is observed that, for a large buffer size, the simulation results for the multiple-input system match well with the analysis results derived for the single-input system, for the FTP traffic. When the buffer size is large, the simulation results are slightly smaller than the corresponding analysis results because of a finite variance of the traffic on periods in the simulation.

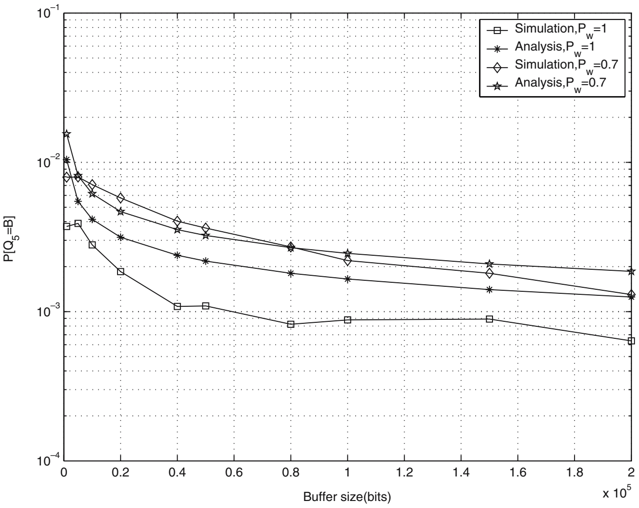

**Fig. 9** Asymptotic loss probability of flow 5 based on simulations of ten traffic flows

## 6 Conclusions

This paper investigates the impact of the self-similarity property of the source traffic flow(s) on the queue length distribution and packet loss probability in a wireless system having a single server. The wireless channel is characterized by an i.i.d. two-state model. In the case of a single input with an infinite buffer size and a finite buffer size, respectively, the close-form expressions are derived for the relation among the traffic parameters, the channel working state probability, the server capacity, and the queue distribution for an infinite buffer size or the packet loss probability for a finite buffer size. The simulation results demonstrate that the assumptions made in the analysis are reasonable and the derived closed-form expressions are accurate. For a multiple-input system with a finite buffer size, we study the queueing behavior under the GPS scheduling principle, with buffer sharing. The preliminary investigation demonstrates that asymptotically each traffic flow is served by the total capacity minus the average rate of other traffic sources in an isolated system with an infinite buffer. The studies extend the research presented in [8–10] for wireline networks to a wireless communication environment.

**Acknowledgements** This research was supported by research grants from the Bell University Laboratories (BUL) at the University of Waterloo and the Communications Information Technology Ontario (CITO).

## References

1. Paxson V, Floyd S (1995) Wide area traffic: the failure of Poisson modeling. IEEE/ACM Trans Netw 3:226–244

2. Crovella M, Bestavros A (1997) Self-similarity in world wide web traffic: evidence and possible causes. IEEE/ACM Trans Netw 5:835–846
3. Jiang M, Nikolic M, Hardy S, Trajkovic L (2001) Impact of self-similarity on wireless data network performance. In: Proc. IEEE ICC 2001, Finland, vol 2, pp 477–481
4. Cheng M, Chang LF (1999) Wireless dynamic channel assignment performance under packet data traffic. IEEE J Sel Areas Commun 17(7):1257–1269.
5. Willinger W, Taqqu M, Sherman R, Wilson DV (1997) Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at source level. IEEE/ACM Trans Netw 5:71–86
6. Park K, Kimy G, Crovellaz M (1996) On the relationship between file sizes, transport protocols, and self-similar network traffic. In: Proceedings of the 4th international conference on network protocols (ICNP'96), Columbus, Ohio, USA, pp 171–180
7. Jelenkovic P EE E6762 Broadband networks lecture notes chapter 3: heavy tails and fluid queues. http://www.ee.columbia.edu/~predrag/
8. Jelenkovic PR, Lazar AA (1997) Multiplexing on–off sources with subexponential on periods: part 1. In: Proc. IEEE Infocom'97, Kobe, Japan, pp 187–195
9. Jelenkovic P (1999) Subexponential loss rates in a GI/GI/1 queue with applications. Queueing Syst 33:91–123
10. Jelenkovic P, Momcilovic P (2002) Finite buffer queue with generalized processor sharing and heavy tailed input process. J Comput Netw 40(3):433–443
11. Parekh AK, Gallager, RG (1993) A generalized processor sharing approach to flow control in integrated services networks: the single-node case. IEEE/ACM Trans Netw 1:344–356
12. Collins B, Cruz, R (1999) Transmission policies for time varying channels with average delay constraints. In: Proc. 1999 Allerton conf. communication, control, and computer. Urbana, Illinois, USA
13. Loynes RM (1968) The stability of a queue with non independent inter arrival and service time. In: Proceedings of the Cambridge philosophical society, vol 58, pp 496–520
14. Ross S (2002) Introduction to probability models, 8th edn. Academic, New York
15. Jelenkovic PR, Lazar AA (1999) Asymptotic results for multiplexing subexponential on–off processes. Adv App Prob 31(2):392–421
16. Kramer G (2001) Self-similar network traffic. Dept. of Computer Science, University of California. http://www.csif.cs.ucdavis.edu/~kramer/research.html.
17. Iacovoni G Wg2 interaction. Technical report, Ericson Lab Italy
18. Jelenkovic PP (2001) Capacity region for network multiplexing with heavy-tailed fluid on–off sources. In: Proc. IEEE INFOCOM 2001, Anchorage, Alaska, USA, vol 1, pp 289–298
19. Zhang H (1995) Service discipline for guaranteed performance service in packet-switching networks. In: Proc. IEEE vol 83, pp 1374–1396

**Shahram Teymori** received the B.S. degree in computer engineering from Sharif University of Technology, Iran, in 1992 and the M.S. degree in electrical engineering from the University of Waterloo, Canada, in 2004. He was a research assistant at the University of Waterloo, Canada, where he carried out research on call packet scheduling for heavy-tailed traffic in wireless network. Since 2004, he has been with ATS Automation Tooling Systems Inc., Cambridge, Ontario, Canada.

**Weihua Zhuang** received the B.S. and M.S. degrees from Dalian Maritime University, China, and the Ph.D. degree from the University of New Brunswick, Canada, all in electrical engineering. Since October 1993, she has been with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, where she is a Professor. She is a co-author of the textbook *Wireless Communications and Networking* (Prentice Hall, 2003). Her current research interests include multimedia wireless communications, wireless networks, and radio positioning. Dr. Zhuang received the Distinguished Performance Award in 2006 from the Faculty of Engineering, University of Waterloo, the Outstanding Performance Award in 2005 from the University of Waterloo for outstanding achievements in teaching, research, and service, and the Premier's Research Excellence Award (PREA) in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions. She is an Editor/Associate Editor of *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, *EURASIP Journal on Wireless Communications and Networking*, and *International Journal of Sensor Networks.*