Effect of Wireless Channel Process on Queueing Delay - Approximate Analysis Using Peakedness Function

Amal Ekbal and John M. Cioffi
Department of Electrical Engineering
Stanford University
Stanford, CA 94305-9515
Email: {ekbal,cioffi}@stanford.edu

Abstract—Time variation of wireless channels can have a serious effect on the quality of service (QoS) experienced by the data streams. However, these channel processes cannot be easily characterized within a queueing theory framework. In this work, the burstiness properties of the wireless channel process are captured using its peakedness function. The channel process is then replaced by a renewal process with the same peakedness function to enable analysis based on queueing theory results. Simulations show that the approximate queue formed using peakedness matching is a very good predictor of the QoS behavior of the original queue.

I. INTRODUCTION

Unlike wire-line channels, wireless channels are very unreliable. The link quality of a wireless channel can vary quickly, for example, within a time duration of tens or hundreds of milliseconds. In many cases, especially in high speed links, the time scale of channel variation is comparable to the time scale of QoS parameters like average delay. The QoS parameters associated with the higher layer queues will then be closely linked to the nature of the physical layer output process. Simple channel models can then lead to erroneous conclusions. To accurately predict the performance of higher layer queues, the channel random process should be modeled appropriately to fit the queueing theoretic description of higher layer queues. However, there is significant difficulty in finding a good description.

Most of queueing theory deals with input inter-arrival and output service times that are independent renewal processes. In many cases, one or both of these processes are assumed to be i.i.d memoryless. Such assumptions enable simple techniques to compute the QoS parameters associated with the queue. For example, it is very easy to analyze the M/M/1 queue since it assumes that the inter-arrival and service times are i.i.d exponential (memoryless) processes. In reality, the processes may be neither renewal nor memoryless. Input traffic processes in high performance networks have significant inherent correlation structure, which leads to burstiness in the arrival process. The structure of these traffic processes does enable approximation by Markov-modulated Poisson processes (MMPP) and its variations. Several exact results are available using MMPP models for traffic arrival.

These MMPP based results still assume that the output process is a memoryless renewal process, or at best a Markov renewal process. Even though there is a lot of work on modeling of channel processes through first-order Markov chains [1], [2], later results in [3] show that such methods can be very erroneous in several cases of interest. More advanced hidden Markov models (HMM) are needed to accurately model the physical layer channel variation [4]. With these advanced models, performance evaluation can be complex.

However, a very precise model of the random process may not be necessary to predict the QoS parameters with reasonable accuracy. In this paper, the peakedness function of the channel random process is used to predict the QoS performance with good accuracy. This function is used in teletraffic analysis to characterize the burstiness of input traffic processes. But, the description works well even for wireless channel processes. Note that if the HMM model of the channel is available, certain queueing theoretic results can be applied for more exact analysis for some specific cases [5], [6]. However, the analysis methods can be quite complex for non-trivial input processes.

This rest of this paper is organized as follows. Section II describes the queueing model used in this work. Section III describes the parameter (peakedness function) used to characterize burstiness of the channel process. Section IV describes a method to approximate the complex wireless channel process by a renewal process convenient for queueing analysis. Section V gives simulation results confirming the validity of these approximations. Section VI concludes this paper.

II. SYSTEM MODEL

Consider a discrete-time FIFO queueing system $Q$ given by

$$x[n+1] = \{x[n] + a[n] - d[n]\}^+$$

(1)

where $x[n]$ is the number of packet units waiting in the queue at time $n$, $a[n]$ is the number of packet arrivals during time slot $n$, $d[n]$ is the number of packet departures during time slot $n$ and $\{x\}^+ = \max(x, 0)$. Packets arriving during a given slot is not allowed to depart in the same slot. Note that this assumption is just for consistency and is not critical in any way. Different elements of this queueing system are described in the following subsections.

A. Traffic Process

The arrival traffic process $a[n]$ is typically described by an MMPP [7], [8]. In an MMPP model, packet arrival is a Poisson process whose rate is a deterministic function of the state of a discrete-time Markov chain (DTMC). Such a doubly stochastic process can adequately model the correlation structure of typical network traffic.
B. Channel Process

In general, the wireless channel in a high speed wireless network can be described by a multi-path Rayleigh fading model [9], [10]. The multiple paths are formed by reflections from different objects in the channel environment. For simplicity, consider a flat fading Rayleigh channel where the channel delay spread (maximum difference between the times of arrival of the paths) is negligible. In this case, the baseband-equivalent channel time response process is given by

$$s(t) = \alpha(t) \exp\{j\phi(t)\}\delta(t)$$  \hspace{1cm} (2)

where $\alpha(t)$ is a Rayleigh fading process and $\phi(t)$ is uniformly distributed over $[0, 2\pi]$. Note that a Rayleigh fading process is generated by the envelope $\sqrt{x(t)^2 + y(t)^2}$ where $x(t), y(t)$ are independent equal variance zero mean Gaussian random processes.

The time variation of the fading process is modeled in physical layer literature by the Doppler power spectrum (Fourier transform of the autocorrelation function) of the Gaussian processes that generate the Rayleigh envelope $\alpha(t)$. A popular model is the Jake’s spectrum [9], which is given by

$$S_x(f) = S_y(f) = \begin{cases} \frac{1}{1-(f/f_D)^2} & \text{if } |f| \leq f_D; \\ 0 & \text{otherwise.} \end{cases}$$  \hspace{1cm} (3)

where $f_D$ is called the Doppler frequency. Note that a stationary Gaussian process can be completely defined by its first and second order statistics. Experiments have revealed that fading spectrum (3) is a reasonable assumption in many high speed wireless environments [10]. There are also experimental results available for the exact nature of the power spectrum in specific environments.

The normalized doppler $f_D T_s$, where $T_s$ is the slot size of the queue in (1), indicates how fast a given channel is varying in time with respect to the queue evolution time scale. If $f_D T_s$ is too small, the channel is highly correlated in time resulting in a slow fading scenario. Similarly, a large normalized doppler indicates a fast fading scenario.

C. The Departure Process

The departure process $d[n]$ depends on the instantaneous quality of the channel process described in the previous subsection. Consequently, the nature of this process is quite different from the Markovian dynamics exhibited by processes considered within the framework of queuing theory. The exact mapping between the channel quality and $d[n]$ depends on the algorithm implemented in the physical layer.

For example, say, the physical layer employs adaptive transmission with a given packet error rate target $PER_t$ and $N$ symbols per slot. Then, the process $d[n]$ in bits/slot is given by

$$d[n] = \begin{cases} N \log_2 \left( 1 + \frac{[\alpha(n)]^2 SNR_0}{\Gamma} \right) & \text{with prob. } 1 - PER_t; \\ 0 & \text{with prob. } PER_t. \end{cases}$$  \hspace{1cm} (4)

where $\Gamma$ is a function of $PER_t$ computed as in [11]. Clearly, the QoS parameters are directly linked to the channel process and the physical layer algorithm.

D. Queueing Model and Bounded Random Walk

The queueing model given in (1) is in the form of a bounded random walk process. That leads directly to the following lemma.  

**Lemma 1:** Let $A = \{a(n)\}_{n=0}^\infty$ and $D = \{d(n)\}_{n=0}^\infty$ be the discrete-time random processes associated with $a(n)$ and $d(n)$ respectively in queue $Q$ described by (1). Consider a continuous time single server queue $Q'$ with inter-arrival time process $D$ and service time process $A$. Then, the queue length process of $Q'$ is equivalent, with probability 1, to the waiting time process of $Q'$.

The proof is straightforward, as the waiting time sequence of a continuous time single server queue has the same form [12] as (1) under the conditions of Lemma 1. This lemma essentially transforms the queue length evolution in (1) to the waiting time evolution process of a $G/G/1$ queue where inter-arrival and service time sequences are independent of each other, but, potentially, correlated in time. Not many results are available in queuing theory for solving such queues because of the very general nature of this problem. The next few sections show methods to tackle this problem for the kind of processes encountered in wireless channels.

III. BURSTINESS PARAMETERS

**Burstiness** is a term often used in traffic analysis to represent time correlation in arrival statistics. If the arrival process is highly positively correlated in time, packet arrivals will be bunched together, i.e. in other words a bursty arrival process. The effect of a fading process on the output of the queue can also be described in a similar fashion. For a highly positively correlated fading process like a Rayleigh process with low normalized Doppler, the departure process will be bursty. The theme of this paper is to extract appropriate parameters from the fading process so that it can be represented in a queuing theoretic framework by an approximate renewal process with similar burstiness. Traffic analysis studies use a number of different parameters to describe the burstiness of a process [13], [14], [15], [16]. Of those, **peakedness function** was found to be very important in capturing the effects of the fading process.

A. Peakedness Function

Consider a traffic stream with inter-arrival times given by the discrete-time random process $X = \{x(n)\}_{n=0}^\infty$ with mean $m_X$ and variance $\sigma_X^2$. Let this stream be the input to an infinite-server group as in Fig. 1 with i.i.d exponential service times and service rate per server $\mu_S$, i.e. mean service time $\mu_S^{-1}$. Let $S(t)$ be the number of servers occupied at any instant $t$. Then, the peakedness function [17] of the traffic stream is defined as

$$z_X(\mu_S) = \lim_{t \to \infty} \frac{\text{var}(S(t))}{E(S(t))}$$  \hspace{1cm} (5)
If the HMM model of the channel is available, this function can be computed analytically [17]. Otherwise, simulations can be used to evaluate the function. Note that simulations cannot be used to compute \( z_X(\mu_S) \) for a set of service rates close to zero.

One of the following methods have to be used:

- Use Newton expansion method as in [16] after obtaining \( z_X(\mu_S) \) for a set of service rates close to zero.
- If the fading process’ correlation coefficient function \( \rho_X[n] \) is known or easy to find, use the following relation from [16]

\[
z_X(0) = 1 + \left( \frac{\sigma_X}{m_X} \right)^2 \sum_{n=-\infty}^{\infty} \rho_X[n] \]

This function is used in traffic analysis to characterize a traffic arrival process with inter-arrival time process \( X \). Hence, typical use of this function is in contexts where \( X \) is generated from stationary renewal or Markov renewal point processes.

In situations where direct analysis of such queues are too complex because of a complicated arrival traffic distribution, peakedness function is used to give good approximate results for average delay, delay distribution and blocking probabilities with or without finite buffers [18].

In this paper, Lemma 1 is used to convert the queue system \( Q \) in (1) to generate a G/G/1 queue \( Q' \). In \( Q' \), the inter-arrival time process \( D \) is generated from an ergodic stationary fading process. The mathematical basis of peakedness analysis can run into problems in this case. For example, the nature of the adaptive physical layer algorithm may result in situations where \( d[n] = 0 \) leading to batch arrivals in \( Q' \). This can lead to technical difficulties when coupled with the non-renewal nature of the fading process. These difficulties, fortunately, were tackled in a few purely theoretical notes made in the unpublished memo by A. Eckberg (1976) [17] which shows that the function is still meaningful and all convergence properties necessary for computing (5) are still intact. The detailed implications of those results will not be discussed in this work. Suffice it to say that without those results, the rest of the paper would be on shaky mathematical grounds.

IV. EQUIVALENT RENEWAL QUEUE

After the peakedness function is evaluated, the next step is to approximate the departure process in \( Q \) using an appropriate renewal process that captures the information contained in the peakedness function. The peakedness matching technique from traffic analysis literature [17], [18] is used for this purpose.

A. Matching Peakedness Function

Consider the G/G/1 queue \( Q' \) created from \( Q \) according to the Lemma 1. Note that the departure fading process \( D \) in \( Q \) is the inter-arrival time process in \( Q' \). Consider a renewal process \( D_R \) with the same peakedness function as \( D \), i.e.

\[
z_{D_R}(\mu_S) = z_D(\mu_S), \quad \mu_S \geq 0 \tag{7}
\]

Consider a new queue with the same service time distribution as in \( Q' \), but with the inter-arrival time process \( D_R \). Denote this queue by \( Q_R \). Unlike \( Q' \) which has a non-renewal fading process as its inter-arrival process, this new queue has a renewal arrival process. From renewal theory [17], it can be shown that the Laplace transform of the inter-arrival time distribution of \( Q_R \) is given by

\[
A_R(\mu_S) = 1 - \frac{1}{z_D(\mu_S) + \frac{1}{m_D \mu_S}}, \quad \mu_S \geq 0 \tag{8}
\]

where \( m_D \) is the mean of the process \( D \).

Note that the departure process in \( Q_R \) is defined by the arrival process in the discrete time queue \( Q \). This process can also be approximated similarly by the peakedness function. Since the arrival process \( A \) in \( Q \) has Markovian dynamics (like an MMPP), it can be dealt with within the constructs of queueing theory. Hence, issues involved in modeling \( A \) is ignored in this paper. In fact, the methods used for modeling the fading process are equally valid for \( A \). But, the structure of process \( A \) can be exploited to develop better approximations.

For the rest of this paper, the departure process in \( Q_R \) is assumed to be adequately modeled by a renewal process. The Laplace transform of the corresponding service time distribution in \( Q_R \) will be indicated by \( D_R(s) \).

B. Approximating Peakedness Function

In cases where the peakedness function can be analytically computed from the fading process model (like in the case of an HMM channel), finding \( z_D(\mu_S) \) for all \( \mu_S \geq 0 \) is easy. If peakedness function has to be estimated based on simulations, finding the full description may be time consuming. Hence, it is necessary to find methods to approximate the peakedness function with reasonable accuracy. One approximation that worked very well was based on results in [16]. In that recent contribution, the authors approximated the peakedness function using two parameters \( \alpha \) and \( \beta \) as

\[
z(\mu_S) \approx 1 + \frac{\alpha \beta}{\mu_S + \alpha} \tag{9}
\]

This approximation was shown in [16] to work well for a class of traffic arrival models. Surprisingly, even though the fading processes have more long term correlation structures,
our results indicate that the approximation works very well for fading processes also. However, another set of approximations in [16] based on phase-type processes did not work all that well without significant complexity increase.

The two parameters in the approximation (9) is computed as follows. The parameter $\beta$ satisfies the relation $\beta = z(0) - 1$. Hence, computing $z(0)$ as described in Section III-A is enough to get $\beta$. To compute $\alpha$, $z(\mu_s)$ is computed for a set of moderately small service rates $\mu_s = \{\mu_s, i\}_{i=1}^{n}$ and a scaled average is computed as follows

$$\alpha = \frac{1}{(1 - \rho^2)n} \sum_{i=1}^{n} \mu_s, i \left(1 - \frac{z(\mu_s, i)}{z(\mu_s, i) - z(0)}\right)$$

(10)

where $\rho$ is lag-1 correlation coefficient of the process. The additional bias factor $(1 - \rho^2)^{-1}$ was found to be necessary in [16].

C. Approximating Average Delay

Average delay of the queue $Q$ can be found from the average delay of $Q'$ and an application of Little’s law. The renewal GI/G/1 queue $Q_R$ is used to approximate the average delay of $Q'$. From queueing theory [12], [16] the delay process in $Q_R$ is asymptotically given by

$$\text{Prob(waiting time} \geq x) \approx \gamma \exp(-\theta x)$$  

(11)

resulting in an approximate average delay of $\bar{\gamma}$.

The parameter $\theta$ in (11) is the smallest positive root of the equation

$$A_R(\theta)D_R(-\theta) = 1$$  

(12)

The parameter $\gamma$ in (11) is computed as follows

- Define $K(s) = A_R(-s)D_R(s)$. Let $k(x)$ be the inverse Laplace transform of $K(s)$.
- Split $K(s) = K_-(s) + K_+(s)$ where

$$K_-(s) = \int_{-\infty}^{0} k(x) \exp\{-sx\} dx$$  

(13)

Note that this can be done using partial fraction expansion for rational $K(s)$.

- Then, $\gamma$ is given by

$$\gamma = 2 \frac{K_+(0) - K_+(\theta)}{K_+(-\theta) - K_+(\theta)}$$  

(14)

The $\gamma$ computation is straightforward if $A_R(s)$ and $D_R(s)$ are rational. For example, if the arrival process in $Q$ is i.i.d exponential with service rate $\mu$ and $A_R(s)$ is rational, then

$$\gamma = \mu - \theta$$  

(15)

V. Simulations

This section considers three instances of the Rayleigh flat fading process described in Section II-B with normalized doppler $f_DT_s = .15, .1$ and .05. Since the normalization is with respect to the slot time, these are realistic values for a broadband wireless system. SNR based continuous rate adaptation with constant coding gap of $\Gamma = 5$dB is used by the physical layer so that the number of bits/slot is given by (4). The simulated process has average SNR of 15dB. A packet with uncoded 4-QAM (2N bits/slot) is considered a single packet unit for counting the queue size. These three cases result in peakness function $\gamma < 1$, $\approx 1$ and $\gamma > 1$ respectively. Note that the peakness function of the memoryless (i.i.d exponential) random process that arises everywhere in queuing theory is

$$z_M(\mu_s) = 1, \quad \mu_s \geq 0$$  

(16)

The traffic arrival process $A$ of queue $Q$ is assumed to be i.i.d exponential process with mean $m_A$ to isolate the effect of fading process modeling on the queue analysis.

In the first scenario ($f_DT_s = .15$), $z_D(\mu_s)$ did not exhibit much variation for moderately small values of $\mu_s$. So, a further assumption was made to neglect $\alpha$ in approximation (9), i.e. assume $\alpha \gg 1$. Following the method in Section IV-C, the average queue length of queue $Q$ is then given by

$$E_Q[X] \approx \frac{z_D(0)m_Dm_A}{m_D - m_A} - m_A$$  

(17)

Fig. 2 shows the average queue size in packet units for moderate system loads $\rho = \frac{m_A}{m_D}$. Note that the approximation in (17) is almost perfect. The figure also compares the results with an ‘M/M/1’ queue $Q_M$ that has the same evolution as in (1), but with i.i.d exponential $A$ and $D$. In this case,

$$E_{Q_M}[X] \approx \frac{\rho}{1 - \rho}$$  

(18)

Note again that the average queue size of the process in (1) is the average waiting time of the corresponding $Q_R$. The average delay of (1) can then be found using Little’s law.

In the second scenario ($f_DT_s = .1$), the simulated peakness function $z(\mu_s)$ was very close to 1 for moderately small $\mu_s$. Hence, the fading process $D$ was replaced by an i.i.d exponential process. As shown in Fig. 3, the approximation is very good. Even though the fading process has a complex description, peakness function shows that as far as queueing

Fig. 2. Average queue size comparison for $f_DT_s = .15$ : peakness < 1
analysis is concerned, it can be approximated by a simple process. Hence, peakedness analysis helps to extract the most important aspects of the fading process.

In the third scenario ($f_{D}T_s = 0.05$), the peakedness of this process was approximated using (9) with $\alpha = 0.075$ and $\beta = 0.84$. Fig. 4 gives the queue size comparison for this case. The effect of $\alpha$ is significant in this case and a $\beta$ based approximation as in Fig. 2 will not work.

Note that, in this case, the queue size is larger than the corresponding M/M/1 prediction. This is the case for processes with peakedness function $> 1$, indicating an increased burstiness. In the first example, the peakedness function was $< 1$, resulting in queue sizes that were lower than the M/M/1 prediction.

VI. CONCLUSION

Accurate modeling of channel behavior is very important in the analysis of queueing behavior in wireless channel. The main difficulty in modeling wireless channel behavior in a queueing theory framework is that the wireless channels does not typically follow Markov dynamics. In this work, the peakedness function of the channel process was used to approximate the effects of channel process. Results show that indeed the information contained in peakedness function is sufficient to give good average delay prediction for a wide variety of fading processes. In addition, other QoS parameters like the queue length distribution, delay distribution, packet drop rate etc. also can be predicted well enough, though those results are not included in the current contribution due to space limitations.

ACKNOWLEDGMENT

This material is based upon work supported under a Stanford Graduate Fellowship. The authors would also like to thank the anonymous reviewers for their helpful comments.

REFERENCES