Queueing Delay Analysis for Packet Schedulers with/without Multiuser Diversity over a Fading Channel

Fumio Ishizaki, *Member, IEEE*, and Gang Uk Hwang, *Member, IEEE*

***Abstract*— In this paper, we focus on a packet scheduling algorithm exploiting multiuser diversity in wireless networks. We compare the delay performance of individual user under the scheduling algorithm exploiting multiuser diversity with that under the round-robin scheduling algorithm in order to reveal the characteristics of the scheduling algorithm exploiting multiuser diversity. For this purpose, we develop an approximate formula to estimate the tail distribution of packet delay for an arbitrary user under the scheduling algorithm exploiting multiuser diversity and that under the round-robin scheduling algorithm. Numerical results exhibit that in contrast to the throughput performance of the overall system, the scheduling algorithm exploiting multiuser diversity is not necessarily superior to the round-robin scheduling algorithm for the delay performance of individual user. More specifically, it is shown that the former is superior to the latter only when the system lies in a severe environment, e.g., when the arrival rate is large, the burstiness of the arrival process is strong or the average signal-to-noise ratio is low.**

***Index Terms*— Multiuser diversity, Queueing delay analysis, Packet scheduling, Quality of service (QoS)**

I. Introduction

With rapid adoption of wireless technology combined with the explosive growth of the Internet, it is promised that demand for wireless data services is continuously increasing. Traffic for wireless data services is expected to be a mix of real-time multimedia traffic such as multimedia conferencing and non real-time data traffic such as file transfers. In such a multiservice wireless environment, providing quality-of-service (QoS) such as delay and packet loss rate is critical for real-time traffic. This requirement, however, imposes a challenging issue in the design of wireless networks, because wireless channels have low reliability, and time varying signal attenuation (fading), which may cause severe QoS violations. In addition, the available bandwidth of wireless channel is severely limited. Therefore, scheduling or control for efficient bandwidth utilization is a key component to the success of QoS guarantees in wireless networks.

One way to achieve efficient bandwidth utilization of time-varying wireless channel is to exploit diversity. By using multiple independent signal paths yielded by diversity, higher channel capacity between the transmitter and the receiver can be achieved. Knopp and Humblet [1] introduced multiuser diversity, which is a diversity existing between the channel states of different users. This diversity comes from the fact that the wireless channel state processes of different users are usually independent for the same shared medium.

Recently several researchers have studied scheduling algorithms or control exploiting multiuser diversity (see, e.g., [2]–[4]), and they have reported that the utilization of multiuser diversity can substantially increase the information theoretic capacity or maximum throughput of the *overall* system. Contrary to these studies, Wu and Negi [5] focus on the delay performance of users under scheduling algorithms exploiting multiuser diversity as the problem of QoS provisioning in wireless networks. They consider the problem of QoS provisioning for K users over time-slotted Rayleigh fading down-link channel. They then develop an efficient scheduling algorithm which is a simple combination of the Knopp and Humblet (KH) scheduling, which exploits multiuser diversity, and the round-robin (RR) scheduling, which does not use multiuser diversity at all. Note here that contrary to the information theoretic capacity or maximum throughput of the overall system, the delay performance of individual user under the KH scheduling is not necessarily superior to that under the RR scheduling, because the RR scheduling is able to guarantee that a user can be served at every K slots where K denotes the number of users while the KH scheduling is not. By estimating the tail distribution of delay in a *fluid* queueing model by the technique developed in [6], their scheduling determines the optimal combination of the KH scheduling and the RR scheduling in advance. Although their technique is applicable to general physical layer channel models, the observation of the (actual or simulated) queueing dynamics at link layer is needed to predict the asymptotic constant and the asymptotic decay rate for the tail distribution of delay. Simulation results show that their approach can substantially increase the delay-constrained capacity of a fading channel, compared to the RR scheduling, when delay constraints are not very tight.

In this paper, we focus on a packet scheduling algorithm exploiting multiuser diversity in wireless networks. In particular, we consider the CKH scheduling algorithm, which most coarsely utilizes multiuser diversity and is considered as the most coarse version of KH scheduling, and we compare the packet delay performance of individual user under the CKH scheduling with that under the RR scheduling. It is expected

This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment). F. Ishizaki is with Department of Information and Telecommunication Engineering, Nanzan University, Seto, Japan (Email: fumio@ieee.org). G.U. Hwang is with Department of Mathematical Sciences and Telecommunication Engineering Program, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea (Email: guhwang@amath.kaist.ac.kr).

that queueing models with *packet-by-packet* scheduling are more suitable for the performance evaluation of scheduling algorithms than fluid queueing models, although fluid queueing models are tractable. The reason is that multiple users are simultaneously served in fluid queueing models while users are served under packet-by-packet scheduling in real networks. Contrary to the previous studies [5], [6], we therefore consider a discrete-time queueing model (with packet-by-packet scheduling) in this paper. We assume that the wireless channel process for each user is described by the Nakagami-m channel model [7] and we determine the effective bandwidth function of the service process under the CKH scheduling and that under the RR scheduling. We then analyze a discrete-time queueing model based on the theory of effective bandwidth. Based on the analytical results, we develop an approximate formula to estimate the tail distribution of packet delay for an arbitrary user under the CKH scheduling and that under the RR scheduling. Finally we provide numerical results to compare the delay performance under the CKH scheduling with that under the RR scheduling.

The main contribution of this paper is to reveal the characteristics of the scheduling algorithm utilizing multiuser diversity in the delay performance of individual user, through the numerical comparison between the delay performance under the CKH scheduling and that under the RR scheduling. Although an extensive numerical study to understand the characteristics of scheduling algorithm exploiting multiuser diversity in the delay performance is strongly needed, it has not been made due to the lack of analytical results. Our formulas developed in this paper are suitable for the numerical study, because contrary to the previous study [5], our formulas do not need the observation of the (actual or simulated) queueing dynamics to estimate the delay performance.

The remainder of this paper is organized as follows. Section II describes models studied in this paper. In Section III, we introduce the notion of effective bandwidth and present the analysis of our models based on the theory of effective bandwidth. Based on the analytical results, we develop a formula to estimate the tail distribution of packet delay for an arbitrary user under the CKH scheduling and that under the RR scheduling. Section IV provides numerical results to compare the delay performance under the CKH scheduling with that under the RR scheduling. Conclusion is drawn in Section V.

II. System model

We begin with the description of the system model. Fig. 1 shows the system model for multiuser traffic over wireless channel. We assume that in the model, time is divided into equal intervals T_f referred to as slots and the service time of a packet is equal to one slot. The model can be considered as a down link in a cellular wireless network where a base station transmits data to K $(K \geq 1)$ mobile user terminals.

A. Channel Model

We assume that the wireless channel process for each user is described by the general Nakagami-m model [7]. The

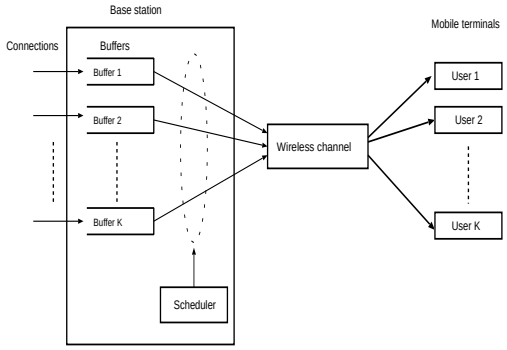

Fig. 1. System model for multiuser traffic

Nakagami-m model is applicable to a broad class of fading channels. It includes the Rayleigh channel as a special case when the Nakagami fading parameter $m = 1$. Also it well approximates the Ricean channels by one-to-one mapping between the Ricean factor K and the Nakagami fading parameter m [7].

Let $\{Z_n^{(k)}\}_{n=0}^{\infty}$ denote the wireless channel process for the kth $(k = 1, \ldots, K)$ user where $Z_n^{(k)}$ denotes the received signal-to-noise ratio (SNR) for the kth user at the beginning of the nth slot. We assume that all the wireless channel processes are independent with each other and they are stationary. We also assume that the wireless channel processes are homogeneous in their parameters.

Suppose that for the utilization of multiuser diversity, the scheduler partitions the entire SNR range into L grades with boundary points denoted by $\{\gamma_l\}_{l=0}^{L}$ with $\gamma_0 = 0$, $\gamma_l < \gamma_{l+1}$ $(l = 0, \ldots, L-1)$ and $\gamma_L = \infty$. For $k = 1, \ldots, K$ and $n = 0, 1, \ldots$, we define a random variable $L_n^{(k)}$ on $\mathcal{L} = \{0, \ldots, L-1\}$ by $L_n^{(k)} = l$ if $\gamma_l \leq Z_n^{(k)} < \gamma_{l+1}$. $L_n^{(k)}$ is considered as the channel grade of the kth user in the nth slot when the number of grades is L, and L may be considered as the granularity of the measured SNR or the granularity of the utilization of multiuser diversity.

As in [8], we assume that the channel grade process $\{L_n^{(k)}\}_{n=0}^{\infty}$ $(k = 1, \ldots, K)$ is well described by a finite-state Markov chain (FSMC). The state transitions of the FSMC happen only between adjacent states. Under slow fading conditions and a small value of T_f, this assumption is natural. Let $\boldsymbol{P} = (p_{i,j})$ $(i, j \in \mathcal{L})$ denote the transition matrix of the FSMC. The transition probabilities are determined as follows (for the detailed derivation of the transition probabilities, see [8]). From the assumption made in this subsection, for $i, j \in \mathcal{L}$, we have

$$p_{i,j} = 0, \qquad |i - j| \geq 2. \tag{1}$$

The adjacent-state transition probabilities are determined by [9]

$$p_{i,i+1} = \frac{\chi(\gamma_{i+1})T_f}{\pi_i}, \qquad i = 0, \ldots, L-2, \tag{2}$$

$$p_{i,i-1} = \frac{\chi(\gamma_i)T_f}{\pi_i}, \qquad i = 1, \ldots, L-1, \tag{3}$$

where $\chi(\gamma)$ denotes the level cross-rate (LCR) at an instanta-

neous SNR γ in the Nakagami-m model and it is given by

$$\chi(\gamma) = \frac{\sqrt{2\pi} f_d}{\Gamma(m)} \left(\frac{m\gamma}{\overline{\gamma}}\right)^{m-\frac{1}{2}} \exp\left(-\frac{m\gamma}{\overline{\gamma}}\right).$$

Here, f_d denotes the mobility-induced Doppler spread, $\overline{\gamma} = \mathrm{E}[\gamma]$ is the average received SNR, $\Gamma(m) = \int_0^\infty t^{m-1}\exp(-t)dt$ is the Gamma function, and π_i ($i \in \mathcal{L}$) denotes the stationary probability that the FSMC is in state i and it is given by

$$\pi_i = \frac{\Gamma(m, m\gamma_i/\overline{\gamma}) - \Gamma(m, m\gamma_{i+1}/\overline{\gamma})}{\Gamma(m)}, \tag{4}$$

where $\Gamma(m,x) = \int_x^\infty t^{m-1}\exp(-t)dt$ is the complementary incomplete Gamma function. With the normalizing condition $\sum_{j=0}^{L-1} p_{i,j} = 1$ for all i, (1), (2) and (3) yield

$$p_{i,i} = \begin{cases} 1 - p_{i,i+1} - p_{i,i-1} & (i = 1, \ldots, L-2), \\ 1 - p_{i,i+1} & (i = 0), \\ 1 - p_{i,i-1} & (i = L-1). \end{cases} \tag{5}$$

(1), (2), (3) and (5) determine the transition matrix $\boldsymbol{P}$ of the FSMC, whose stationary distribution is given by (4).

B. Utilization of Multiuser Diversity

In this subsection, we describe the scheduler employing the Knopp and Humblet (KH) scheduling. Since the channel processes of the users are assumed to be independent with each other, we can potentially utilize multiuser diversity. Under the KH scheduling, the base station is assumed to know the current value of the channel grade $L_n^{(k)}$ for all k. In order to increase the capacity of the overall system with multiuser diversity, among all the users, the scheduler first selects users whose channel grades are the highest, i.e., users whose channel grades are equal to τ_n^* where τ_n^* ($n = 0, 1, \ldots$) is defined by

$$\tau_n^* = \max_{k \in \{1,\ldots,K\}} L_n^{(k)}.$$

Among the selected users, the scheduler randomly selects one user and assigns the slot to transmit the packet of the selected user. However, to avoid deep channel fades, packet will not be transmitted if $L_n^{(k)} = 0$ for all k. We assume that if $L_n^{(k)} > 0$, packets are always successfully transmitted over the wireless channel and correctly received at the user.

For comparison, we also consider a scheduler employing the round-robin (RR) scheduling, which does not utilize multiuser diversity at all. It assigns a slot for the users in turn, irrespective of the wireless channel processes.

C. Queueing Models

In this subsection, we consider the queueing dynamics at the buffer of an arbitrary user under the KH scheduling and those under the RR scheduling. Without loss of generality, we assume that the first user is an arbitrary user, and we hereafter call the arbitrary user the tagged user.

First we describe the queueing behavior at the buffer of the tagged user under the KH scheduling. Let X_n ($n = 0, 1, \ldots$) denote a random variable representing the queue length (i.e., the number of packets in the buffer of the tagged user) at the beginning of the nth slot. Let A_n ($n = 0, 1, \ldots$) denote a random variable representing the number of packets arriving at the buffer of the tagged user in the nth slot. Let C_n ($n = 0, 1, \ldots$) denote a random variable representing the number of packets which can be served at the buffer of the tagged user in the nth slot under the KH scheduling. We here define an auxiliary i.i.d. (independent and identically distributed) stochastic sequence $\{V_n\}_{n=0}^\infty$ according to the uniform distribution on $[0, 1]$. We also define a random variable ν_n^* ($n = 0, 1, \ldots$) by $\nu_n^* = \sum_{k=1}^K I(L_n^{(k)} = \tau_n^*)$ where $I(\cdot)$ denotes the indicator function. Note that ν_n^* denotes the number of users (including the tagged user) being in the highest grade in the nth slot. C_n ($n = 0, 1, \ldots$) is then given by

$$C_n = \begin{cases} 1 & (L_n^{(1)} = \tau_n^* > 0 \text{ and } V_n \le 1/\nu_n^*), \\ 0 & (\text{otherwise}), \end{cases} \tag{6}$$

We are now ready to describe the queueing dynamics at the buffer of the tagged user under the KH scheduling. The queueing process $\{X_n\}_{n=0}^\infty$ at the buffer of the tagged user evolves according to the following recursion:

$$X_{n+1} = (X_n - C_n)^+ + A_n, \tag{7}$$

where $(x)^+$ denotes $\max(0, x)$. Note that $\{C_n\}$ becomes a stationary Markov modulated process, because $\{L_n^{(k)}\}$ ($k = 1, \ldots, K$) are stationary Markov chains and $\{V_n\}$ is an i.i.d. sequence.

Next we describe the queueing behavior at the buffer of the tagged user under the RR scheduling. The queueing process at the buffer of the tagged user under the RR scheduling also follows the same recursion (7), but C_n under the RR scheduling is given by

$$C_n = \begin{cases} 1 & (L_n^{(1)} > 0 \text{ and } n \bmod K = 0), \\ 0 & (\text{otherwise}). \end{cases} \tag{8}$$

Note that $\{C_n\}$ under the RR scheduling also becomes a stationary Markov modulated process, because $\{L_n^{(1)}\}$ is a stationary Markov chain.

At closing this section, we consider the maximum throughput of the tagged user under the saturation condition that there always exists a packet at the buffer of the tagged user. From (6), the maximum throughput s_{KH} of the tagged user under the KH scheduling is given by

$$\begin{aligned} s_{\mathrm{KH}} &= \mathrm{E}[C_n] \\ &= \sum_{l=1}^{L-1}\sum_{m=1}^{K} \mathrm{P}(C_n = 1, \tau_n^* = L_n^{(1)} = l, \nu_n^* = m) \\ &= \sum_{l=1}^{L-1}\sum_{m=1}^{K} \mathrm{P}(C_n = 1 | \tau_n^* = L_n^{(1)} = l, \nu_n^* = m) \\ &\quad \cdot \mathrm{P}(L_n^{(1)} = l, L_n^{(2)} \le l, \cdots, L_n^{(K)} \le l, \nu_n^* = m) \\ &= \sum_{l=1}^{L-1}\sum_{m=1}^{K} \frac{1}{m}\binom{K-1}{m-1} \pi_l^m \Big(\sum_{j=0}^{l-1}\pi_j\Big)^{K-m} \\ &= \frac{1}{K}\sum_{l=1}^{L-1}\sum_{m=1}^{K} \binom{K}{m} \pi_l^m \Big(\sum_{j=0}^{l-1}\pi_j\Big)^{K-m} \end{aligned}$$

$$= \frac{1}{K}\sum_{l=1}^{L-1}\left[(\sum_{j=0}^{l}\pi_j)^K - (\sum_{j=0}^{l-1}\pi_j)^K\right]$$
$$= \frac{1}{K}(1-\pi_0^K). \quad (9)$$

On the other hand, from (8), the maximum throughput $s_{\rm RR}$ of the tagged user under the RR scheduling is given by

$$s_{\rm RR} = \mathrm{E}[C_n] = \frac{1}{K}(1-\pi_0). \quad (10)$$

Comparing (9) and (10), we see that $s_{\rm KH}$ is always greater than $s_{\rm RR}$ and the multiuser diversity gain $s_{\rm KH}/s_{\rm RR}$ for the maximum throughput is given by

$$\frac{s_{\rm KH}}{s_{\rm RR}} = \frac{1-\pi_0^K}{1-\pi_0}. \quad (11)$$

From (11), we see that when the condition of the wireless channel is not good, the multiuser diversity gain for the maximum throughput is greater. A similar discussion has been made in [5].

III. Analysis based on effective bandwidth

In this section, we present the analysis based on the theory of effective bandwidth. The theory of effective bandwidth has been extensively studied for wireline packet networks and has been widely accepted as a basis of connection admission control (CAC) and resource allocation (for detailed and theoretical descriptions of the effective bandwidth approach, see, e.g., [10], [11] and references therein). Recently the theory of effective bandwidth has been studied for wireless packet networks, too (see, e.g., [5], [6], [12]–[15]).

To keep the presentation of the analysis compact, we assume in the analysis that $L = 2$, i.e., the scheduler most coarsely utilizes multiuser diversity. Assuming that the PER (packet error rate) is determined by encoding scheme and received SNR as shown in [16], we select the boundary γ_1 such that the PER is negligible when the received SNR is greater than γ_1. Then, we may consider that the packet transmission is always successful when the received SNR is greater than γ_1. Accordingly, the choice of $L = 2$ in the analysis is natural in practice, and it is at least acceptable for the purpose to understand the characteristics of the scheduling algorithm exploiting multiuser diversity. Hereafter we call the KH scheduling with $L = 2$ the CKH scheduling (most Coarse version of the KH scheduling).

A. EBF of Arrival Process

We begin with the notion of the Gärtner-Ellis (GE) limit (or the asymptotic decay rate function). Let $\Lambda_A(\theta)$ denote the GE limit for cumulative arrival process $\tilde{A}_n$ of general arrival process, where $\tilde{A}_n$ is the input of work from the source during the time interval $[0, n)$. $\Lambda_A(\theta)$ is defined by $\Lambda_A(\theta) = \lim_{n\to\infty} n^{-1}\log \mathrm{E}\exp(\theta\tilde{A}_n)$, provided that the limit exists. We then define the function $\xi_A(\theta)$ of θ by $\xi_A(\theta) = \Lambda_A(\theta)/\theta$, which is called the *effective bandwidth function* (EBF) of the arrival process. It is known (see. e.g., [10]) that the EBF $\xi_A(\theta)$ is increasing in θ, and it converges to the average rate of the arrival process as $\theta \downarrow 0$ and to the peak rate of the arrival process as $\theta \uparrow \infty$.

We now return to our model. In this paper, we assume that the arrival process $\{A_n\}_{n=0}^{\infty}$ is generated by an on-off source, which can incorporate the bursty behavior of the arrival process. In any slot, the on-off source is in one of the two different states: on-state and off-state. In off-state, it does not generate a packet, and in on-state, it generates one packet with probability λ. The transition probability from on-state (resp. off-state) to off-state (resp. on-state) is denoted by $1-\alpha$ (resp. $1-\beta$), where $0 \le \alpha, \beta \le 1$. The following parameters are used to characterize the on-off source: the mean on-period B_{on}, the mean off-period B_{off} and the average rate ρ. These parameters are expressed in terms of α, β and λ as follows:

$$B_{on} = \frac{1}{1-\alpha}, \quad B_{off} = \frac{1}{1-\beta}, \quad \rho = \frac{\lambda(1-\beta)}{2-\alpha-\beta}.$$

It is known that the GE limit $\Lambda_A(\theta)$ of the arrival process in our model and its EBF $\xi_A(\theta)$ are given by (see, e.g., [10])

$$\Lambda_A(\theta) = \log\delta_A(\theta), \qquad \xi_A(\theta) = \frac{\log\delta_A(\theta)}{\theta}, \quad (12)$$

where $\delta_A(\theta)$ is given by $\delta_A(\theta) = \zeta(\theta) + \sqrt{\zeta(\theta)^2 - b\phi(\theta)}$, $\phi(\theta) = 1-\lambda+\lambda e^{\theta}$, $\zeta(\theta) = (\alpha\phi(\theta)+\beta)/2$, and $b = \alpha+\beta-1$. Although we assume that $\{A_n\}_{n=0}^{\infty}$ is generated by the on-off source, the analysis presented in this section is applicable to any arrival processes whose GE limits exist.

B. EBF of Service Process under CKH Scheduling

In this subsection, we first define the notion of the EBF for general service processes. We then provide a useful expression for the GE limit for the service process under the CKH scheduling and that for its EBF.

We start with the GE limit for general service process. Let $\tilde{C}_n$ $(n = 0, 1, \ldots)$ denote a random variable representing the cumulative service process during the time interval $[0, t)$. Let $\Lambda_C(\theta)$ denote the GE limit of the cumulative service process $\tilde{C}_n$. Similar to the GE limit for the arrival process, $\Lambda_C(\theta)$ is defined by $\Lambda_C(\theta) = \lim_{n\to\infty} n^{-1}\log \mathrm{E}\exp(\theta\tilde{C}_n)$, provided that the limit exists. We now define the function $\xi_C(\theta)$ of θ by

$$\xi_C(\theta) = -\frac{\Lambda_C(-\theta)}{\theta}, \quad (13)$$

which is called the EBF of the service process. Note here that from the definition (13), we have $\Lambda_{-C}(\theta) = \Lambda_C(-\theta)$, where $\Lambda_{-C}(\theta)$ denote the GE limit of $-C(t)$. Thus, the EBF of the service process can be also written as $\xi_C(\theta) = -\Lambda_{-C}(\theta)/\theta$ [10]. It is known (see. e.g., [15]) that the EBF $\xi_C(\theta)$ is decreasing in θ, and it converges to the average service rate as $\theta \downarrow 0$ and to the minimum service rate as $\theta \uparrow \infty$.

We now return to our model. Note that under the CKH scheduling and the homogeneous wireless channel setting, the service process $\{C_n\}$ is a Markov modulated process whose underlying Markov chain is $\{(L_n^{(1)}, S_n)\}$ where S_n is defined by $S_n = \sum_{k=2}^{K} L_n^{(k)}$ for all n. To show a useful expression for the GE limit and EBF of the service process under the CKH

scheduling in our model, we need to define some matrices. We first define a $K \times K$ matrix $\boldsymbol{R}$ by

$$[\boldsymbol{R}]_{i,j} = \sum_{k=\max(0,i+j-K+1)}^{\min(i,j)} \binom{i}{k} p_{1,1}^{k} p_{1,0}^{i-k} \cdot \binom{K-1-i}{j-k} p_{0,1}^{j-k} p_{0,0}^{K-1-i-j+k},$$

where $[\boldsymbol{R}]_{i,j}$ $(i,j = 0,\ldots,K-1)$ denotes the (i,j)th element of $\boldsymbol{R}$. Note that $[\boldsymbol{R}]_{i,j}$ denotes the conditional probability that j channel grade processes among the $(K-1)$ channel grade processes are in state 1 in the current slot given that i channel grade processes among the $(K-1)$ channel grade processes were in state 1 in the previous slot. We then define a $2K \times 2K$ matrix $\boldsymbol{Q}_{\rm KH}$ by $\boldsymbol{Q}_{\rm KH} = \boldsymbol{P} \otimes \boldsymbol{R}$, where $\otimes$ denotes the Kronecker product. We next define a $2K \times 2K$ diagonal matrix $\boldsymbol{D}_{\rm KH}(\theta)$ by

$$\boldsymbol{D}_{\rm KH}(\theta) = \mathrm{diag}(\overbrace{1,\cdots,1}^{K}, e^{\theta}, \frac{1+e^{\theta}}{2}, \cdots, \frac{K-1+e^{\theta}}{K}).$$

Finally we define a $2K \times 2K$ matrix $\boldsymbol{C}_{\rm KH}(\theta)$ by $\boldsymbol{C}_{\rm KH}(\theta) = \boldsymbol{Q}_{\rm KH}\boldsymbol{D}_{\rm KH}(\theta)$.

We are now ready to provide a useful expression for the GE limit for the service process under the CKH scheduling and that for its EBF (For the proof of Proposition 1, e.g., see [10], [17]).

Proposition 1. *The GE limit $\Lambda_C(\theta)$ for the service process under the CKH scheduling is given by*

$$\Lambda_C(\theta) = \log \delta_C(\theta), \tag{14}$$

where $\delta_C(\theta)$ is the Perron-Frobenius (PF) eigenvalue of $\boldsymbol{C}_{\rm KH}(\theta)$. Thus, the EBF $\xi_C(\theta)$ of the service process under the CKH scheduling is given by

$$\xi_C(\theta) = -\frac{\log \delta_C(-\theta)}{\theta}. \tag{15}$$

C. EBF of Service Process under RR Scheduling

In this subsection, we provide an expression for the GE limit for the service process under the RR scheduling and that for its EBF. Note that under the RR scheduling, the service process $\{C_n\}$ is a Markov modulated process whose underlying Markov chain is $\{(S_n, L_n^{(1)})\}$ where S_n is defined by $S_n = n \bmod K$ for all n.

First we define a $K \times K$ matrix $\boldsymbol{U}$ by

$$\boldsymbol{U} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ 1 & 0 & \cdots & \cdots & 0 \end{pmatrix}.$$

We then define a $2K \times 2K$ matrix $\boldsymbol{Q}_{\rm RR}$ by $\boldsymbol{Q}_{\rm RR} = \boldsymbol{U} \otimes \boldsymbol{P}$. Let $\boldsymbol{D}_{\rm RR}(\theta)$ denote a $2K \times 2K$ diagonal matrix given by

$$\boldsymbol{D}_{\rm RR}(\theta) = \mathrm{diag}(\overbrace{1,\cdots,1}^{2K-1}, e^{\theta}).$$

We finally define a $2K \times 2K$ matrix $\boldsymbol{C}_{\rm RR}(\theta)$ by $\boldsymbol{C}_{\rm RR}(\theta) = \boldsymbol{Q}_{\rm RR}\boldsymbol{D}_{\rm RR}(\theta)$. We then have the following proposition, which gives an explicit expression for the PF eigenvalue of the matrix $\boldsymbol{C}_{\rm RR}(\theta)$. We provide the proof of Proposition 2 in Appendix.

Proposition 2. *Let $\tilde{\boldsymbol{C}}_{\rm RR}(\theta)$ denote a 2×2 matrix defined by $\tilde{\boldsymbol{C}}_{\rm RR}(\theta) = \boldsymbol{P}^K \mathrm{diag}(1, e^{\theta})$. The PF eigenvalue $\delta_C(\theta)$ of the matrix $\boldsymbol{C}_{\rm RR}(\theta)$ is then given by*

$$\delta_C(\theta) = \left[\eta(\theta) + \sqrt{\eta(\theta)^2 + \kappa(\theta)}\right]^{1/K}, \tag{16}$$

where

$$\eta(\theta) = \frac{\tilde{c}_{00}(\theta) + \tilde{c}_{11}(\theta)}{2},$$
$$\kappa(\theta) = -\tilde{c}_{00}(\theta)\tilde{c}_{11}(\theta) + \tilde{c}_{01}(\theta)\tilde{c}_{10}(\theta),$$

and $\tilde{c}_{ij}(\theta)$ $(i,j = 0,1)$ denotes the (i,j)th element of the matrix $\tilde{\boldsymbol{C}}_{\rm RR}(\theta)$.

Then, the GE limit for the service process under the RR scheduling and its EBF are also expressed as (14) and (15), respectively, but $\delta_C(\theta)$ is the PF eigenvalue of $\boldsymbol{C}_{\rm RR}(\theta)$ and given by (16).

D. Approximations Based on the Theory of EB

The theory of EB can be used to obtain approximation formulas for the tail distribution of the queue length in steady state and that of the queueing delay. In this subsection, we provide such approximation formulas.

Let X_∞ denote a random variable representing the queue length evolved by (7) in steady state. It is known that under some conditions, the tail distribution $\mathrm{P}(X_\infty > x)$ of the queue length in steady state is approximately given by [10]

$$\mathrm{P}(X_\infty > x) \approx \exp(-\theta^* x),$$

where θ^* is the unique real solution of the equation

$$\Lambda_A(\theta) + \Lambda_C(-\theta) = 0. \tag{17}$$

Similarly, let D denote a random variable representing the delay of a randomly chosen packet from the tagged user. It is known that under some conditions, the tail distribution $\mathrm{P}(D > t)$ of the delay of a randomly chosen packet is approximately expressed as [14]

$$\mathrm{P}(D > t) \approx \exp(\Lambda_C(-\theta^*)t), \tag{18}$$

where θ^* is the unique real solution of the equation (17).

IV. Numerical results

In this section, we provide numerical results to compare the delay performance under the CKH scheduling with that under the RR scheduling. Throughout this section, we assume that the (maximum) service rate of wireless channel and the packet size are 2Mbps and 250bytes, respectively, and the number of users K is equal to 10. Under this setting, the length of one slot is equal to 1msec. We also assume the Nakagami fading parameter $m = 1$ (i.e., the Rayleigh fading channel) and the Doppler frequency $f_d = 10$Hz.

Before comparing the delay performance under the CKH scheduling with that under the RR scheduling, we first examine

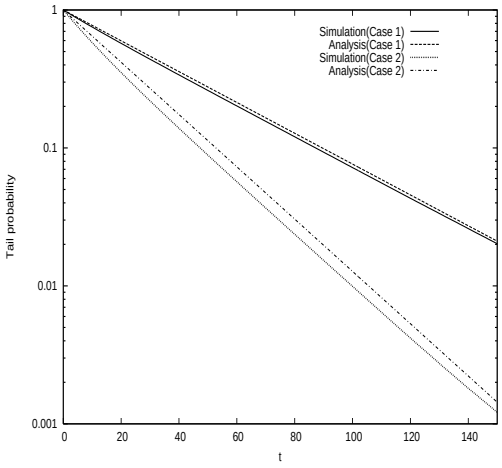

Fig. 2. Tail probabilities of delay under the CKH scheduling

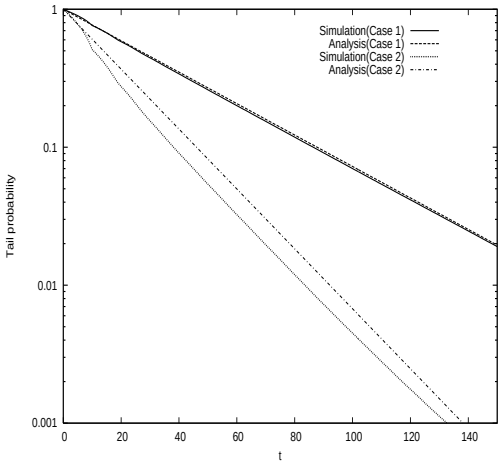

Fig. 3. Tail probabilities of delay under the RR scheduling

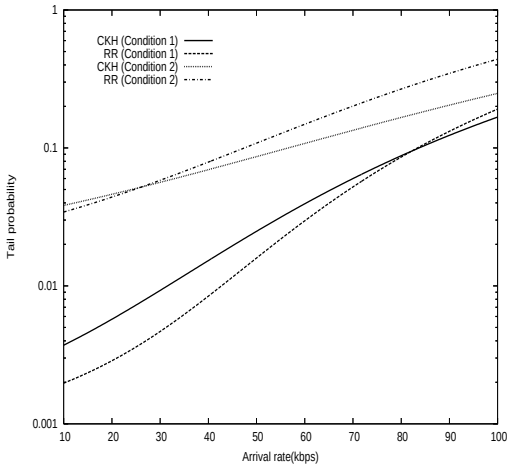

Fig. 4. Delay performance as a function of arrival rate

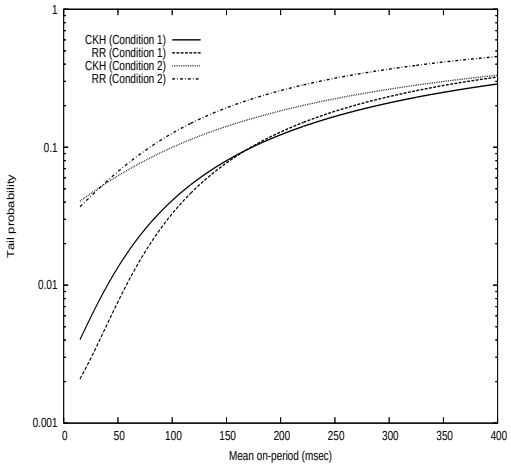

Fig. 5. Delay performance as a function of mean on-period length

the accuracy of our approximation formula (18) based on the analysis presented in Section III. Figs. 2 and 3 show the tail probabilities of delay of a randomly chosen packet from the tagged user under the CKH scheduling and those under the RR scheduling. The tail probabilities estimated by our approximation formulas are denoted by "Analysis" and those estimated by simulation are denoted by "Simulation" in the figures. In Case 1 of both figures, the parameters of the arrival process from the tagged user are set as $\alpha = 0.900$, $\beta = 0.992$, and $\lambda = 0.300$. Under this setting, the average rate, the mean on-period and the mean off-period are 42.0kbps, 10.0msec and 133msec, respectively. In Case 2 of both figures, the parameters of the arrival process from the tagged user are set as $\alpha = 0.800$, $\beta = 0.980$, and $\lambda = 0.200$. Under this setting, the average rate, the mean on-period and the mean off-period are 36.4kbps, 5.00msec and 50.0msec, respectively. In both figures, we set $\gamma_1 = 7$dB and $\overline{\gamma} = 16$dB. In Figs. 2 and 3, we observe that for both scheduling algorithms, the tail probabilities estimated by our approximation formulas are close to those estimated by simulation.

In what follows, we compare the delay performance under the CKH scheduling with that under the RR scheduling. First we compare the effect of the arrival rate from the tagged user on the delay performance under the CKH scheduling with that under the RR scheduling. For this purpose, we change the parameter λ of the on-off source while fixing the other parameters α and β. Fig. 4 shows the probability that the delay of a randomly chosen packet from the tagged user is greater than 100msec under the CKH scheduling and that under the RR scheduling as a function of the arrival rate. These probabilities are estimated by approximation formula (18). We set the parameters of the on-off source as $\alpha = 0.800$ and $\beta = 0.980$. Under this setting, mean on-period and mean off-period are 5msec and 50msec, respectively. For wireless channel, we consider the following two conditions. For Condition 1 (resp. Condition 2), the parameters for the channel are set as $\gamma_1 = 7$dB and $\overline{\gamma} = 16$dB (resp. $\gamma_1 = 7$dB and $\overline{\gamma} = 12$dB).

In Fig. 4, we observe the following. First, although the CKH scheduling is always superior to the RR scheduling for the maximum throughput, this is not the case for the delay performance. In fact, for Condition 1 (resp. Condition 2), the CKH scheduling is superior to the RR scheduling only when the arrival rate is greater than 82kbps (resp. 26kbps). Thus, the CKH scheduling is superior to the RR scheduling only

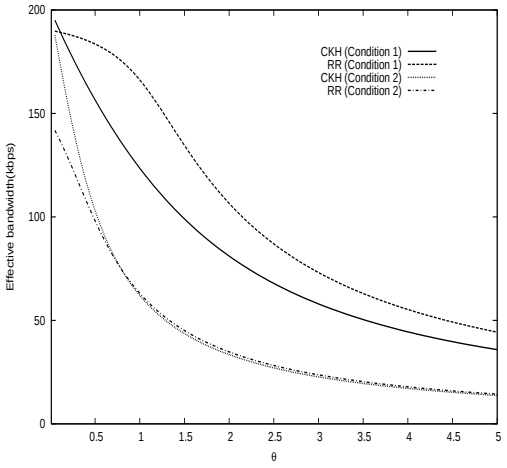

Fig. 6. EBF of service processes

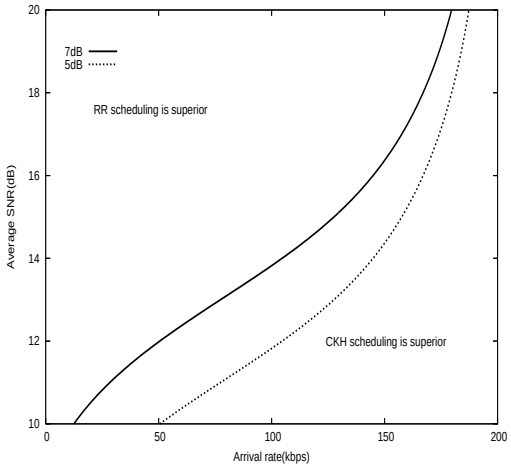

Fig. 7. Region where the CKH scheduling is superior

when the arrival rate is greater than a threshold. The threshold varies with the average SNR, and the region where the CKH scheduling is superior to the RR scheduling becomes narrow when the average SNR is high.

Next we examine how the burstiness of the arrival process affects the delay performance under the CKH scheduling and that under the RR scheduling. For this purpose, we change the mean on-period of the on-off source while fixing the average arrival rate (i.e., we change α and β while fixing ρ and λ). Then, a large value of the mean on-period means the strong burstiness of the arrival process. Fig. 5 shows the probability that the delay of a randomly chosen packet from the tagged user is greater than 100msec under the CKH scheduling and that under the RR scheduling as a function of the mean on-period. We set the parameter $\lambda = 0.250$ and fix the arrival rate to 20kbps. For wireless channel, we consider the same two conditions as in Fig. 4. In Fig. 5, we observe that for both scheduling algorithms, the burstiness of the arrival process has a strong impact on the delay performance. We also see that the CKH scheduling is superior to the RR scheduling only when the mean-on period is greater than a threshold. The threshold varies with the average SNR, and the region where the CKH scheduling is superior to the RR scheduling becomes narrow when the average SNR is high. In fact, for Condition 1 (resp. Condition 2), the CKH scheduling is superior to the RR scheduling when the mean on-period is greater than 168msec (resp. 36msec).

We investigate the observations in Figs. 4 and 5 in more detail. Fig. 6 shows the EBF $\xi_C(\theta)$ of the service process under the CKH scheduling and that under the RR scheduling as a function of θ for Conditions 1 and 2. In Fig. 6, we observe the following. When θ is small, the EBF under the CKH scheduling is greater than that under the RR scheduling for both channel conditions. However, with the increase in the value of θ, the EBF under the CKH scheduling more rapidly decrease than that under the RR scheduling. As a result, when θ is large, the EBF under the CKH scheduling is less than that under the RR scheduling. From the above observation, we see that while the CKH scheduling is superior to the RR scheduling for a small value of θ, the RR scheduling is superior to the CKH scheduling for a large value of θ. Note here that the EBF $\xi_C(\theta)$ at large value of θ denotes the service capacity as the QoS constraint is stringent, and vice versa [15]. Hence, only when the required QoS for delay is not stringent, the service capacity under the CKH scheduling is greater than that of the RR scheduling. In the simulation results in [5], a similar observation has been made. Another interpretation of Fig. 6 is that when the resulting delay performance is bad, the CKH scheduling is superior to the RR scheduling; otherwise the RR scheduling is superior to the CKH scheduling.

Finally we examine the effect of the boundary γ_1 on the delay performance under the CKH scheduling and that under the RR scheduling. For this purpose, for $\gamma_1 = 5, 7$dB, we show which scheduling is superior for various values of the average SNR $\overline{\gamma}$ and the arrival rate. A small γ_1 means that the system has a strong error tolerance against low average SNR. In Fig. 7, we consider the delay of a randomly chosen packet from the tagged user is greater than 100msec as the performance measure for comparison, and we change λ while fixing α and β to 1.000 and 0.000, respectively (i.e., the arrival process is a Bernoulli process). We see in Fig. 7 that if the system has a strong error tolerance, the region where the CKH scheduling is superior to the RR scheduling becomes narrow.

V. Conclusion

In this paper, we focus on the CKH scheduling algorithm, which most coarsely utilizes multiuser diversity. We then compare the packet delay performance of individual user under the CKH scheduling with that under the RR scheduling in order to reveal the characteristics of the scheduling algorithm exploiting multiuser diversity. For this purpose, we develop an approximate formula to estimate the tail distribution of packet delay for an arbitrary user under the CKH scheduling and that under the RR scheduling. Numerical results exhibit that in contrast to the throughput performance of the overall system, for the delay performance of individual user, the CKH scheduling is superior to the RR scheduling only when the system lies in a severe environment, e.g., when the arrival rate is large, the burstiness of the arrival process is strong or

the average SNR is low. We also see that if the system has a strong error tolerance against low average SNR, the region where the CKH scheduling is superior to the RR scheduling becomes narrow.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for valuable comments which significantly improved the quality of this paper.

APPENDIX

A. Proof of Proposition 2

For the proof of Proposition 2, we provide Theorem 1 below, which shows a more general result than Proposition 2. Then, Proposition 2 will immediately follow from Theorem 1.

To state Theorem 1, we need the definitions of some matrices . We consider a $NJ \times NJ$ matrix $\boldsymbol{C}$ given by

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{L}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \boldsymbol{L}_{N-2} & \boldsymbol{O} \\ \boldsymbol{O} & \cdots & \cdots & \boldsymbol{O} & \boldsymbol{L}_{N-1} \\ \boldsymbol{L}_0 & \boldsymbol{O} & \cdots & \cdots & \boldsymbol{O} \end{pmatrix}, \tag{19}$$

where N and J are positive integers, $\boldsymbol{L}_n$ $(n = 0, 1, \cdots, N-1)$ denotes any $J \times J$ matrix and $\boldsymbol{O}$ denotes a $J \times J$ zero matrix. Let δ denote the PF eigenvalue of $\boldsymbol{C}$ in (19). We define a $J \times J$ matrix $\tilde{\boldsymbol{C}}$ as

$$\tilde{\boldsymbol{C}} = \boldsymbol{L}_0 \boldsymbol{L}_1 \cdots \boldsymbol{L}_{N-1}. \tag{20}$$

Theorem 1. *The PF eigenvalue δ of $\boldsymbol{C}$ in (19) is given by*

$$\delta = \tilde{\delta}^{1/N}, \tag{21}$$

where $\tilde{\delta}$ is the PF eigenvalue of $\tilde{\boldsymbol{C}}$.

PROOF: We denote the ith $(i = 0, \ldots, NJ-1)$ row vector of $\boldsymbol{C}$ by $\boldsymbol{c}_i$. We also denote an eigenvalue of $\boldsymbol{C}$ by σ. To obtain the determinant of $\boldsymbol{C} - \sigma\boldsymbol{I}$, we apply the following manipulations to $\boldsymbol{C} - \sigma\boldsymbol{I}$ from the 0th step to the $(N-2)$nd step. In the kth $(k = 0, \ldots, N-2)$ step,

1) define the $1 \times NJ$ vectors $\boldsymbol{a}_{kJ}, \ldots, \boldsymbol{a}_{(k+1)J-1}$ by

$$\begin{pmatrix} \boldsymbol{a}_{kJ} \\ \vdots \\ \boldsymbol{a}_{(k+1)J-1} \end{pmatrix} = \sigma^{-k-1} \boldsymbol{L}_0 \cdots \boldsymbol{L}_k \begin{pmatrix} \boldsymbol{c}_{kJ} \\ \vdots \\ \boldsymbol{c}_{(k+1)J-1} \end{pmatrix}.$$

2) for $j = 0, \ldots, J-1$, add $\boldsymbol{a}_{kJ+j}$ to the $[(N-1)J+j]$th row of $\boldsymbol{C}$.

Note that through the manipulations, the determinant is invariant. Thus, after the $(N-2)$nd step, we have

$$|\boldsymbol{C} - \sigma\boldsymbol{I}| = \begin{vmatrix} -\sigma\boldsymbol{I} & \boldsymbol{L}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \boldsymbol{L}_{N-2} & \boldsymbol{O} \\ \boldsymbol{O} & \cdots & \cdots & -\sigma\boldsymbol{I} & \boldsymbol{L}_{N-1} \\ \boldsymbol{O} & \cdots & \cdots & \boldsymbol{O} & \sigma^{-N+1}\tilde{\boldsymbol{C}} - \sigma\boldsymbol{I} \end{vmatrix}. \tag{22}$$

From (22), we obtain

$$|\boldsymbol{C} - \sigma\boldsymbol{I}| = (-\sigma)^{(N-1)J} |\sigma^{-N+1}\tilde{\boldsymbol{C}} - \sigma\boldsymbol{I}|. \tag{23}$$

In (23), we see that if and only if $|\tilde{\boldsymbol{C}} - \sigma^N \boldsymbol{I}| = 0$, we have $|\sigma^{-N+1}\tilde{\boldsymbol{C}} - \sigma\boldsymbol{I}| = 0$ and thus $|\boldsymbol{C} - \sigma\boldsymbol{I}| = 0$. Therefore, the eigenvalue σ of $\boldsymbol{C}$ is given by $\sigma = \tilde{\sigma}^{1/N}$, where $\tilde{\sigma}$ is an eigenvalue of $\tilde{\boldsymbol{C}}$. This completes the proof. ■

REFERENCES

[1] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," *Proc. of IEEE ICC '95*, pp.331–335, 1995.

[2] R. Ferrús, L. Alonso, A. Umbert, X. Revés, J. Pérez-Romero, and F. Casadevall, "Cross-layer scheduling strategy for UMTS downlink enhancement," *IEEE Radio Commun.*, vol.2, no.2, pp.24–28, 2005.

[3] X. Qin and R. Berry, "Exploiting multiuser diversity for medium access control in wireless networks," *Proc. of IEEE INFOCOM '03*, pp.1084–1094, 2003.

[4] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Commun. Mag.*, vol.41, no.10, pp.74–80, 2003.

[5] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE Trans. Veh. Technol.*, vol.54, no.3, pp.1198-1206, 2005.

[6] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol.2, pp.630–643, 2003.

[7] G. L. Stüber, *Principles of mobile communication*, 2nd ed., Kluwer, 2001.

[8] Q. Liu, S. Zhou and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol.4, no.3, pp.1142–1153, 2005.

[9] J. Razavilar, K. J. R. Liu, and S. I. Marcus, "Jointly optimized bit-rate/delay control policy for wireless packet networks with fading channels," *IEEE Trans. Commun.*, vol.50, no.3, pp.484–494, 2002.

[10] C.-S. Chang, *Performance guarantees in communication networks*, Springer-Verlag, 2000.

[11] A. I. Elwalid and D. Mitra, "Effective bandwidths of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Netw.*, vol.1, pp.329–343, 1993.

[12] J. G. Kim and M. M. Krunz, "Bandwidth allocation in wireless networks with guaranteed packet-loss performance," *IEEE/ACM Trans. Netw.*, vol.8, pp.337–349, 2000.

[13] M. M. Krunz and J. G. Kim, "Fluid analysis of delay and packet discard performance of QoS support in wireless networks," *IEEE J. Sel. Areas in Commun.*, vol.19, pp.384–395, 2001.

[14] M. Hassan, M. M. Krunz and I. Matta, "Markov-based channel characterization for tractable performance analysis in wireless packet networks," *IEEE Trans. Wireless Commun.*, vol.3, pp.821–831, 2004.

[15] X. Zang, J. Tang, H.-H. Chen, S. Ci and M. Guizani, "Cross-layer-based modeling for quality of service guarantees in mobile wireless networks," *IEEE Commun. Mag.*, vol.44, no.1, pp.100–106, 2006.

[16] Q. Liu, S. Zhou and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol.3, pp.1746–1755, 2004.

[17] J. A. Bucklew, *Large deviation techniques in decision, simulation, and estimation*, Wiley, 1990.