Wireless Pers Commun (2008) 46:397–411
DOI 10.1007/s11277-007-9442-5

Queueing Delay Analysis for the Joint Scheduling Exploiting Multiuser Diversity over a Fading Channel

Fumio Ishizaki · Gang Uk Hwang

Published online: 11 January 2008

Abstract In this paper, we consider a joint packet scheduling algorithm for wireless networks and investigate its characteristics. The joint scheduling algorithm is a combination of the Knopp and Humblet (KH) scheduling, which fully exploits multiuser diversity, and the probabilistic weighted round-robin (WRR) scheduling, which does not use multiuser diversity at all. Under the assumption that the wireless channel process for each user is described by the Nakagami-*m* model, we develop a formula to estimate the tail distribution of the packet delay for an arbitrary user under the joint scheduling. Numerical results exhibit that under the joint scheduling, the ratio of the number of slots assigned for the WRR scheduling to that for the KH scheduling dominates the characteristics of the delay performance.

Keywords Multiuser diversity · Packet scheduling · QoS (Quality-of-Service) · Delay analysis

1 Introduction

In multiservice wireless networks, traffic is a mix of real-time multimedia traffic such as multimedia conferencing and non real-time data traffic such as file transfers, and the provision of quality-of-service (QoS) guarantees is strongly required for real-time traffic. This requirement, however, imposes a challenging issue on 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. Hence, scheduling algorithms taking those characteristics of wireless

F. Ishizaki (✉)
Department of Information and Telecommunication Engineering, Nanzan University, Seto, Japan
e-mail: fumio@ieee.org

G. U. Hwang
Department of Mathematical Sciences and Telecommunication Engineering Program,
Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
e-mail: guhwang@kaist.edu

channels into account are key components to the success of QoS guarantees in wireless networks.

Since the utilization of multiuser diversity [10] can increase the information theoretic capacity of the *overall* system, much attention has been paid to scheduling algorithms exploiting multiuser diversity [4,13,15]. Multiuser diversity is a diversity existing between the wireless channel states of different users, and this diversity comes from the fact that the wireless channel state processes of different users are usually independent for the same shared medium. Several scheduling algorithms exploiting multiuser diversity have been proposed. Among them, the Knopp and Humblet (KH) scheduling maximizes the throughput of the overall system by fully exploiting multiuser diversity, but it can cause serious unfairness between users [19]. At the expense of the throughput performance, the proportional fair scheduling provides a strict fairness between users by considering the normalized signal-to-noise ratio (SNR) of the users [19]. The joint scheduling studied in [18] is a simple combination of the KH scheduling and the round-robin (RR) scheduling. By combining the KH scheduling with the RR scheduling, the joint scheduling tries to achieve a desirable balance between the fairness and the throughput performances.

In this paper, we consider a joint scheduling which is a combination of the KH scheduling and the probabilistic weighted round-robin (WRR) scheduling, and we develop a formula to estimate the tail distribution of packet delay for an individual user under the joint scheduling. The joint scheduling considered in this paper may be viewed as an extension of the joint scheduling studied in Wu and Negi [18]. The aim of this paper is to examine the characteristics of the joint scheduling algorithm by analytical approach. Compared to the analysis of the information theoretic capacity of the overall system, the analysis of the tail distribution of packet delay for an individual user under a scheduling algorithm exploiting multiuser diversity has not been sufficiently provided, although the tail distribution of delay is one of the most important performance metrics in provisioning QoS for real-time traffic in multiservice wireless networks.

Kim and Kang [9] consider MBCS (Multi-users Best Channel Scheduling) which takes advantage of multiuser diversity and derive the analytical result of mean delay in the case of two users. Numerical results show that the mean delay performance of the MBCS is better than that of the SBCS (Single-user Best Channel Scheduling). Wu and Negi [18] consider the joint scheduling which is a simple combination of the KH scheduling and the RR scheduling. By using a technique developed in Wu and Negi [17], the scheduling algorithm estimates the tail distribution of delay for users (not for an individual user) in a *fluid* queueing model, and it 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. Simulation results show that their approach can substantially increase the delay-constrained capacity of a fading channel, compared to the pure RR scheduling, when delay constraints are not very tight. Ishizaki and Hwang [6] develop a formula to estimate the tail distribution of packet delay for an individual user under the most coarse version of the KH (CKH) scheduling and that under the RR scheduling. Numerical results exhibit that for the delay performance, 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 and/or the wireless channel condition is not good.

In this paper, we consider a discrete-time queueing model with packet-by-packet scheduling in order to examine the characteristics of the joint scheduling algorithm by analytical approach. Fluid queueing models have been considered in most of the previous analytical studies (e.g., [5,11,18,20]) on the performance evaluation of delay in wireless networks, due to their analytical tractability. However, queueing models with packet-by-packet scheduling are more suitable for the performance evaluation of scheduling algorithms than fluid queueing models, because multiple users are simultaneously served in fluid queueing models while users are served under packet-by-packet scheduling in real networks. The analysis provided in this paper is based on the theory of of effective bandwidth (EB) and an extension of the analysis presented in Ishizaki and Hwang [6]. Contrary to the study in Wu and Negi [17], our formula relies purely on the analytical results and it does not need the observation of the (actual or simulated) queueing dynamics to estimate the delay performance. Our formula is thus suitable to examine the characteristics of the joint scheduling algorithm in the problem of QoS provisioning for real-time traffic.

The remainder of this paper is organized as follows. Section 2 describes a model studied in this paper. In Sect. 3, we introduce the notion of effective bandwidth function (EBF) and analyze the queueing model based on the theory of EB. Based on the analytical results, we develop a formula to estimate the tail distribution of packet delay for an arbitrary user under the joint scheduling. Section 4 provides numerical results to examine the characteristics of the joint scheduling algorithm. Conclusion is drawn in Sect. 5.

2 System Model

We begin with the description of the system model. Figure 1 shows the system model for multiuser traffic over wireless channel. We assume that in the model, time is divided into equal intervals of unit time 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 downlink in a cellular wireless network where a base station (BS) transmits data to K $(K \geq 1)$ mobile user terminals.

2.1 Channel Model

We assume that the wireless channel process for each user is described by the general Nakagami-m model [16]. The 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 Ricean channels by one-to-one mapping between

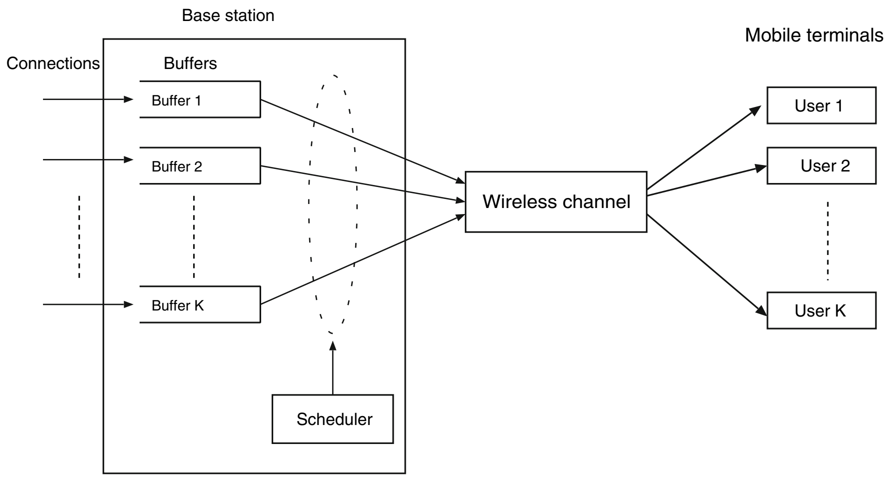


Fig. 1 System model for multiuser traffic

the Ricean factor K and the Nakagami fading parameter m [16]. Let $\left\{Z_n^{(k)}\right\}_{n=0}^{\infty}$ denote the wireless channel process for the kth ($k = 1, \ldots, K$) user where $Z_n^{(k)}$ denotes the received 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 L is the granularity of the measured SNR or the granularity of the utilization of multiuser diversity. That is, 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 $\gamma_l \le Z_n^{(k)} < \gamma_{l+1}$. $L_n^{(k)}$ is considered as the channel grade of the kth user in the nth slot.

As in Liu et al. [12], we assume that the channel grade process $\left\{L_n^{(k)}\right\}_{n=0}^{\infty}$ of the kth user ($k = 1, \ldots, K$) is well described by a finite-state Markov chain (FSMC). The state transitions of the FSMCs 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 FSMCs. The transition probabilities are determined as follows (for the detailed derivation of the transition probabilities, see Liu et al. [12]). From the assumption made in this subsection, for $i, j \in \mathcal{L}$, we have

$$p_{i,j} = 0 \quad (|i-j| \ge 2). \tag{1}$$

The adjacent-state transition probabilities are determined by Razavilar et al. [14]

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

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

where $\chi(\gamma)$ denotes the level cross-rate at an instantaneous 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 $\pi_i\ (i \in \mathcal{L})$ 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 FSMCs, whose stationary probabilities are given by (4).

2.2 Joint Scheduling Utilizing Multiuser Diversity

In this subsection, we describe a scheduler employing the joint scheduling which is a combination of the Knopp and Humblet (KH) scheduling and the probabilistic weighted round-robin (WRR) scheduling. The joint scheduling considered in this paper may be viewed as an extension of the joint scheduling proposed by Wu and Negi [18].

First, we describe the KH scheduling. Under the KH scheduling, the BS 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 current slot to transmit the packet of the user. However, to avoid deep channel fades, packet will not be transmitted if $L_n^{(k)} = 0$ for all k.

Next, we describe the RR scheduling, which does not utilize multiuser diversity at all. It assigns slots for users in turn, irrespective of the wireless channel processes. However, to avoid deep channel fades, packet will not be transmitted if the scheduler selects the kth user for service in the nth slot and $L_n^{(k)} = 0$.

Finally, we describe the joint scheduling. Without loss of generality, we hereafter assume that the first user is an arbitrary user, and we call the arbitrary user the tagged user. Under the joint scheduling, successive T slots compose a frame. To describe a service pattern in a frame, we introduce a bivariate deterministic sequences $\{(m_n, h_n)\}_{n=0}^{T-1}$ where for $n = 0, \ldots, T-1$, $m_n \in \{0, 1\}$ and $h_n \in [0, 1]$, $M \overset{\triangle}{=} \sum_{n=0}^{T-1} m_n$, $H \overset{\triangle}{=} \sum_{n=0}^{T-1} h_n I(m_n = 0)$ and $I(\cdot)$ denotes the indicator function. The deterministic sequence $\{(m_n, h_n)\}_{n=0}^{T-1}$ determines a scheduling algorithm employed in the nth slot of a frame as follows: If $m_n = 1$, the KH scheduling is employed in the nth slot of a frame. If $m_n = 0$, the tagged user is selected for service in the nth slot of a frame with probability h_n. Thus, M and H denote the expected number of slots for the KH scheduling in a frame and that for the probabilistic WRR scheduling in a frame, respectively. Note that when $T = K$, $m_n = 0$ for all n and $h_0 = 1, h_1 = \cdots = h_{T-1} = 0$, the joint scheduling reduces to the RR scheduling. Also note that it reduces to the KH scheduling when $m_n = 1$ for all n. We hereafter call the sequence $\{(m_n, h_n)\}_{n=0}^{T-1}$ the service sequence.

We assume that for the three schedulers mentioned above, if the kth user is selected for service in the nth slot and $L_n^{(k)} > 0$, the packet of the kth user (if any) is always successfully transmitted over the wireless channel and correctly received at the user.

2.3 Queueing Model

In this subsection, we consider the queueing dynamics at the buffer of the tagged user under the joint 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 under the joint scheduling. 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 joint 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

$\nu_n^* = \sum_{k=1}^{K} I(L_n^{(k)} = \tau_n^*)$. 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

$$C_n = \begin{cases} 1 & (m_{n \bmod T} = 1, L_n^{(1)} = \tau_n^* > 0, V_n \le 1/\nu_n^*) \\ & \text{ or } (m_{n \bmod T} = 0, L_n^{(1)} > 0, V_n \le h_{n \bmod T}), \\ 0 & (\text{otherwise}). \end{cases}$$

The queueing process $\{X_n\}_{n=0}^{\infty}$ at the buffer of the tagged user then evolves according to the following recursion:

$$X_{n+1} = \max(0, X_n - C_n) + A_n. \tag{6}$$

At closing this section, we consider the maximum throughput of the tagged user under the saturation condition that there always exist packets at the buffer of the tagged user. The maximum throughput s_{J} of the tagged user under the joint scheduling is given by

$$s_{\mathrm{J}} = \frac{M}{T} s_{\mathrm{KH}} + \frac{KH}{T} s_{\mathrm{RR}},$$

where s_{KH} and s_{RR} denote the maximum throughput of the tagged user under the KH scheduling and that under the RR scheduling, respectively, and they are given by Ishizaki and Hwang [6]

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

3 Analysis Based on EB

The theory of EB has been extensively studied for wireline packet networks and has been widely accepted as a basis of connection admission control (CAC) and resource allocation. Recently the theory of EB has been studied for wireless packet networks, too (see, e.g., [5,11,17,20]). For detailed and theoretical descriptions on the theory of EB, see [1–3,7,8] and references therein.

In this section, we present the analysis based on the theory of EB. We first introduce the notion of the EBF for general arrival and service processes. We then provide useful expressions for the EBFs in our model.

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 Liu et al. [12], 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 joint scheduling algorithm. Hereafter we call the KH scheduling with $L = 2$ the most coarse version of the KH (CKH) scheduling. Similarly, the resulting joint scheduling with the CKH scheduling is called the most coarse version of the joint (CJ) scheduling. Accordingly, the channel grade process of each user is described by a 2-state Markov chain with transition probability matrix $\boldsymbol{P} = (p_{i,j})_{0 \le i,j \le 1}$ for the CJ scheduling.

3.1 EBF of Arrival Process

In this subsection, we first define the notion of the EBF for general arrival processes. 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 the cumulative arrival process $\tilde{A}_n$ of a general arrival process, i.e., $\tilde{A}_n$ is the number of packets arriving from the source during the time interval $[0, n)$ and it is given by $\tilde{A}_n = \sum_{t=0}^{n-1} A_t$. $\Lambda_A(\theta)$ is defined by

$$\Lambda_A(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathrm{E} \exp(\theta \tilde{A}_n),$$

provided that the limit exists. We then define the function $\xi_A(\theta)$ of θ by

$$\xi_A(\theta) = \frac{\Lambda_A(\theta)}{\theta},$$

which is called the EBF of the arrival process. It is known (see. e.g., [1]) 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 $\{A_n\}_{n=0}^{\infty}$ of the tagged user is generated by an on–off source, which can incorporate the burst 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 may be 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: $B_{on} = (1-\alpha)^{-1}$, $B_{off} = (1-\beta)^{-1}$, and $\rho = \lambda(1-\beta)/(2-\alpha-\beta)$. It is known that the GE limit $\Lambda_A(\theta)$ of the on-off source in our model and its EBF $\xi_A(\theta)$ are given by (see, e.g., [1])

$$\Lambda_A(\theta) = \log \delta_A(\theta),$$
$$\xi_A(\theta) = \frac{1}{\theta} \log \delta_A(\theta),$$

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.

3.2 EBF of Service Process Under CJ 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 CJ 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 $[0, n)$, i.e., $\tilde{C}_n$ is given by $\tilde{C}_n = \sum_{t=0}^{n-1} C_t$. Let $\Lambda_C(\theta)$ denote the GE limit of the cumulative service process $\tilde{C}_n$. Similar to the GE limit for arrival process, $\Lambda_C(\theta)$ is defined by

$$\Lambda_C(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathrm{E} \exp(\theta \tilde{C}_n),$$

provided that the limit exists. We define the function $\xi_C(\theta)$ of θ by

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

which is called the EBF of the service process. It is known (see. e.g., [20]) 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. To show a useful expression for the GE limit and the EBF of the service process under CJ scheduling in our model, we need to define some matrices. We first define a $K \times K$ matrix $\boldsymbol{R}$ by

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

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 excluding the one of the tagged user 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 $\boldsymbol{Q}_{\mathrm{KH}}$ by

$$\boldsymbol{Q}_{\mathrm{KH}} = \boldsymbol{P} \otimes \boldsymbol{R},$$

where $\otimes$ denotes the Kronecker product. We next define a $T \times T$ matrix $\boldsymbol{U}$ by

$$[\boldsymbol{U}]_{i,j} = \begin{cases} 1 & (j = (i+1) \bmod T), \\ 0 & (\text{otherwise}), \end{cases}$$

where $[\boldsymbol{U}]_{i,j}$ $(i, j = 0, \ldots, T-1)$ denotes the (i, j)th element of $\boldsymbol{U}$. We then define a $2KT \times 2KT$ matrix $\boldsymbol{Q}_{\mathrm{CJ}}$ by

$$\boldsymbol{Q}_{\mathrm{CJ}} = \boldsymbol{U} \otimes \boldsymbol{Q}_{\mathrm{KH}} = \boldsymbol{U} \otimes \boldsymbol{P} \otimes \boldsymbol{R}.$$

Next we define a $1 \times 2K$ vector $\boldsymbol{d}_n(\theta) (n = 0, 1, \ldots, T-1)$ by

$$\boldsymbol{d}_n(\theta) = \begin{cases} \left(\overbrace{1, \cdots, 1}^{K}, e^{\theta}, \frac{1+e^{\theta}}{2}, \cdots, \frac{K-1+e^{\theta}}{K} \right) & (m_n = 1), \\ \left(\overbrace{1, \cdots, 1}^{K}, \overbrace{1 - h_n + h_n e^{\theta}, \cdots, 1 - h_n + h_n e^{\theta}}^{K} \right) & (m_n = 0). \end{cases}$$

We then define a diagonal $2KT \times 2KT$ matrix $\boldsymbol{D}_{\mathrm{CJ}}(\theta)$ by

$$\boldsymbol{D}_{\mathrm{CJ}}(\theta) = \mathrm{diag}(\boldsymbol{d}_0(\theta), \cdots, \boldsymbol{d}_{T-1}(\theta)).$$

Finally we define a $2KT \times 2KT$ matrix $\boldsymbol{C}_{\mathrm{CJ}}(\theta)$ by

$$\begin{aligned}\boldsymbol{C}_{\mathrm{CJ}}(\theta) &= \boldsymbol{Q}_{\mathrm{CJ}} \boldsymbol{D}_{\mathrm{CJ}}(\theta) \\ &= (\boldsymbol{U} \otimes \boldsymbol{P} \otimes \boldsymbol{R})\ \mathrm{diag}(\boldsymbol{d}_0(\theta), \cdots, \boldsymbol{d}_{T-1}(\theta)).\end{aligned}$$

Since the matrix $\boldsymbol{C}_{\mathrm{CJ}}(\theta)$ is primitive, i.e., $[\boldsymbol{C}_{\mathrm{CJ}}(\theta)]^k > \boldsymbol{O}$ for some $k \geq 1$, we can obtain the EBF of the service process by computing the Perron-Frobenius (PF) eigenvalue $\delta_C(\theta)$

of the $2KT \times 2KT$ matrix $\boldsymbol{C}_{\mathrm{CJ}}(\theta)$ (see [1]). When the value of $2KT$ is large, it is not easy to compute the PF eigenvalue of $\boldsymbol{C}_{\mathrm{CJ}}(\theta)$. However, as shown below, we can obtain the PF eigenvalue of $\boldsymbol{C}_{\mathrm{CJ}}(\theta)$ from the PF eigenvalue of a $2K \times 2K$ matrix. To do so, we need to define a $2K \times 2K$ matrix $\tilde{\boldsymbol{C}}_{\mathrm{CJ}}(\theta)$ by

$$\tilde{\boldsymbol{C}}_{\mathrm{CJ}}(\theta) = \boldsymbol{Q}_{\mathrm{KH}}\ \mathrm{diag}(\boldsymbol{d}_0(\theta)) \cdots \boldsymbol{Q}_{\mathrm{KH}}\ \mathrm{diag}(\boldsymbol{d}_{T-1}(\theta)).$$

We then have the following proposition. The proof of the proposition is provided in Ishizaki and Hwang [6].

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

$$\Lambda_C(\theta) = \frac{1}{T} \log \tilde{\delta}_C(\theta),$$

where $\tilde{\delta}_C(\theta)$ is the PF eigenvalue of $\tilde{\boldsymbol{C}}_{\mathrm{CJ}}(\theta)$. Thus, the EBF $\xi_C(\theta)$ of the service process under the CJ scheduling is given by

$$\xi_C(\theta) = -\frac{1}{\theta T} \log \tilde{\delta}_C(-\theta).$$

3.3 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 (6) in steady state. It is known (see, e.g., [1]) that under some condition, the tail distribution $\mathrm{P}(X_\infty > x)$ of the queue length in steady state is approximately given by $\mathrm{P}(X_\infty > x) \approx \exp(-\theta^* x)$, where θ^* is the unique real solution of the equation

$$\Lambda_A(\theta) + \Lambda_C(-\theta) = 0. \tag{7}$$

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 condition, the tail distribution $\mathrm{P}(D > t)$ of the delay of a randomly chosen packet is approximately expressed as [5]

$$\mathrm{P}(D > t) \approx \exp(\Lambda_C(-\theta^*)t), \tag{8}$$

where θ^* is the unique real solution of the Eq.7.

4 Numerical Results

In this section, we provide numerical results to investigate the characteristics of the CJ scheduling and its effect on the delay performance. In the numerical results, for simplicity, we consider only a class of the CJ scheduling algorithms where $h_n \in \{0, 1\}$ for all n and the RR scheduling is employed when $m_n = 0$. Throughout this section, we assume that the (maximum) service rate of wireless channel and the packet size are 2 Mbps and 250 bytes, respectively, and the number of users K is equal to 10. Under this setting, the length of one slot is equal to 1 ms. We also assume the Nakagami fading parameter $m = 1$ (i.e., the Rayleigh fading channel) and the Doppler frequency $f_d = 10$ Hz.

Before investigating the characteristics of the CJ scheduling and its effect on the delay performance, we first check the accuracy of our approximation formula (8) based on the analysis presented in Sect. 3. Figures 2 and 3 show the tail probabilities of delay of a randomly

Fig. 2 Tail probabilities of packet delay under the CJ scheduling

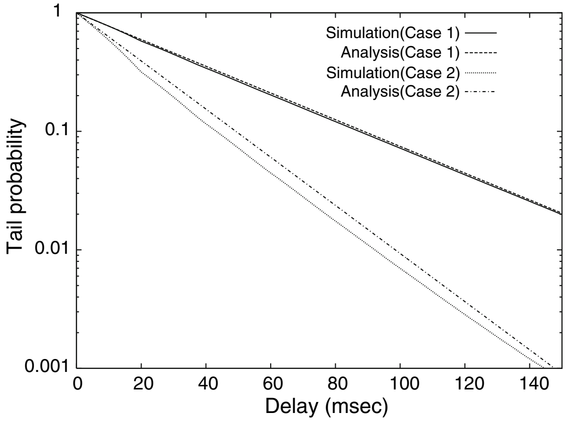
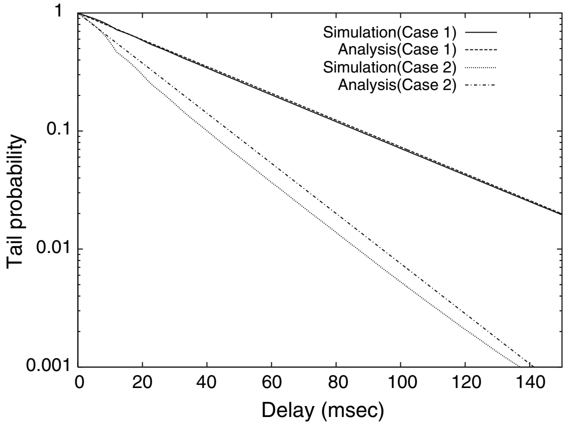


Fig. 3 Tail probabilities of packet delay under the CJ scheduling



chosen packet from the tagged user under the CJ scheduling. The tail probabilities estimated by our approximation formula 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, the mean off-period are 42.0 kbps, 10.0 ms and 133 ms, 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, the mean off-period are 36.4 kbps, 5.00 ms and 50.0 ms, respectively. In both figures, we set $\gamma_1 = 7$ dB and $\overline{\gamma} = 16$ dB. For Cases 1 and 2 in Fig. 2, we set $M = 10$, $T = 20$ and the service sequence $\{(m_n, h_n)\}_{n=0}^{T-1}$ as

$$m_n = \begin{cases} 0 & (n \bmod 2 = 0), \\ 1 & (\text{otherwise}), \end{cases} \tag{9}$$

$$h_n = \begin{cases} 1 & (n = 0), \\ 0 & (\text{otherwise}). \end{cases} \tag{10}$$

For Cases 1 and 2 in Fig. 3, we set $M = 2,\ T = 12$ and the service sequence $\{(m_n, h_n)\}_{n=0}^{T-1}$ as (10) and

$$m_n = \begin{cases} 1 & (n = 10, 11), \\ 0 & (\text{otherwise}). \end{cases}$$

In Figs. 2 and 3, we observe that the tail probabilities estimated by our approximation formula are very close to those estimated by simulation. In particular, for Case 1 in both figures, the tail probabilities estimated by our approximation formula are almost identical to those estimated by simulation. We thus confirm that the accuracy of our approximation formula is excellent.

We now examine the effect of service sequence on the delay performance. In the numerical results, we focus on the value of $M/(KH)$. We hereafter call it the KH/RR ratio, because it denotes the ratio of the average number of slots where the tagged user is served by the CKH scheduling to that by the RR scheduling.

Table 1 shows the probability that the delay of a randomly chosen packet from the tagged user is greater than 100 ms under the CJ scheduling for various service sequences. For Condition 1 (resp. Condition 2) in Table 1, the parameters for the wireless channel are set as $\gamma_1 = 7$ dB and $\overline{\gamma} = 16$ dB (resp. $\gamma_1 = 7$ dB and $\overline{\gamma} = 12$ dB). For service sequence 0, we set $M = 1,\ T = 1$ and $m_0 = 1$. In other words, the CJ scheduling with service sequence 0 is the CKH scheduling. Note that the KH/RR ratio is equal to infinity for the service sequence 0. For service sequence 1, we set $M = 10,\ T = 20$, (9) and (10). For service sequence 2, we set $M = 10$, $T = 20$, (10) and

$$m_n = \begin{cases} 0 & (n = 0, \ldots, 4 \text{ or } n = 10, \ldots, 14), \\ 1 & (\text{otherwise}). \end{cases}$$

For service sequence 3, we set $M = 10,\ T = 20$, (10) and

$$m_n = \begin{cases} 0 & (n = 0, \ldots, 9), \\ 1 & (\text{otherwise}). \end{cases} \tag{11}$$

Note that for the service sequences 1–3, the KH/RR ratios are equal to one. Among them, slots for the CKH scheduling are most evenly distributed in a frame for the service sequence 1 and least evenly distributed for the service sequence 3. For service sequence 4, we set $M = 2,\ T = 12$ and the service sequence $\{(m_n, h_n)\}$ is expressed as (10) and (11). For service sequence 5, we set $M = 2,\ T = 12$, (10) and

$$m_n = \begin{cases} 0 & (n = 0, \ldots, 4 \text{ or } n = 6, \ldots, 10), \\ 1 & (\text{otherwise}). \end{cases}$$

Table 1 Effect of service sequence on the tail probability of packet delay

Service seq.	Condition 1	Condition 2
0 (CKH)	1.997×10^{-2}	7.835×10^{-2}
1	1.565×10^{-2}	8.533×10^{-2}
2	1.564×10^{-2}	8.536×10^{-2}
3	1.571×10^{-2}	8.573×10^{-2}
4	1.312×10^{-2}	9.094×10^{-2}
5	1.303×10^{-2}	9.054×10^{-2}
6 (RR)	1.191×10^{-2}	9.381×10^{-2}

Note that for the service sequences 4 and 5, the KH/RR ratios are equal to 0.2. Between them, slots for the CKH scheduling are most evenly distributed in a frame for the service sequence 4. For service sequence 6, we set $M = 0,\ T = 10,\ m_n = 0$ for all n and (10). In other words, the CJ scheduling with service sequence 6 is the RR scheduling. Note that the KH/RR ratio for the service sequence 6 is equal to zero. For all the service sequences, the parameters of the arrival process from the tagged user are set as $\alpha = 0.800,\ \beta = 0.980$, and $\lambda = 0.250$. Under this setting, the average rate, the mean on-period, the mean off-period are 45.5 kbps, 5.00 ms and 50.0 ms, respectively.

We observe the following in Table 1. First, for the delay performance, the CKH scheduling is superior to the RR scheduling when the average SNR is low. In contrast, when the average SNR is high, the RR scheduling is superior to the CKH scheduling. The same observation has been made in Ishizaki and Hwang [6], too. Similarly, when the average SNR is low, the group of the CJ scheduling with the service sequences 1–3, where the KH/RR ratios are equal to one, is superior to the group of the CJ scheduling with the service sequences 4–5, where the KH/RR ratios are equal to 0.2. In contrast, when the average SNR is high, the latter group is superior to the former group. We thus see that the CJ scheduling with large KH/RR ratio is superior to that with small one when the average SNR is low, and vice versa. Second, different service sequences result in different delay performances. However, the KH/RR ratio has greater effect on the delay performance than how distribute slots for the CKH scheduling in a frame. The reason that the distribution of slots for the CKH scheduling does not have great impact on the delay performance is explained as follows: The probability that the delay is greater than a large value of threshold is dominated by the mean and higher moments of the service capacity over a long-term period. Indeed, how distribute slots for the CKH scheduling in a frame affects the mean and higher moments of the service capacity over a short-term period, but not much affect the ones over a long-term period.

We investigate the observations in Table 1 in more detail. Table 2 displays the EBFs $\xi_C(\theta)$ of the service processes under the service sequences 0–6, where the parameters for the wireless channel are set as $\gamma_1 = 7$ dB and $\overline{\gamma} = 16$ dB. We observe the following in Table 2. When θ is small, the EBFs under the CJ scheduling algorithms with large KH/RR ratio are greater than those with small one. However, with the increase in the value of θ, the EBFs under the CJ scheduling algorithms with large KH/RR ratio more rapidly decrease than those with small one. As a result, when θ is large, the EBFs under the CJ scheduling algorithms with large KH/RR ratio are less than those with small one. From the above observation, we see that while the CJ scheduling algorithms with large KH/RR ratio are superior to those with small one for a small value of θ, and vice versa for a large value of θ. Note that the

Table 2 EBF (kbps) of service process

Service seq.	θ					
	0.01	0.10	0.40	1.0	1.40	2.00
0 (CKH)	199	188	153	99.4	77.5	57.1
1	187	180	156	106	82.5	60.6
2	187	180	156	106	82.5	60.6
3	187	180	156	106	82.4	60.5
4	180	176	158	110	85.2	62.2
5	180	176	158	110	85.3	62.4
6 (RR)	176	173	159	112	86.5	63.1

EBF of service process at large value of θ denotes the service capacity as the QoS constraint is stringent [20]. Hence, only when the required QoS for delay is not stringent, the service capacities under the CJ scheduling algorithms with large KH/RR ratio are greater than those with small one.

We next confirm that the KH/RR ratio dominates the delay performance. Figure 4 shows the probability that the delay of a randomly chosen packet from the tagged user is greater than 100 ms as a function of H while fixing the KH/RR ratio $(= 1, 0.2)$. These probabilities are estimated by approximation formula (8). In Fig. 4, the parameters for the wireless channel are set as $\gamma_1 = 7$ dB and $\overline{\gamma} = 16$ dB and the parameters of the arrival process from the tagged user are set as $\alpha = 0.800$, $\beta = 0.980$, and $\lambda = 0.300$. For all the results shown in Fig. 4, the service sequences are set as (10) and (11). We observe in this figure that changing the value of H while fixing the KH/RR ratio does not much affect the delay performance. We thus confirm that the delay performance is dominated by the KH/RR ratio.

We finally observe the effect of the arrival rate on the delay performance for various service sequences. Figure 5 shows the probability that the delay of a randomly chosen packet from the tagged user is greater than 100 ms for various service sequences as a function of the arrival rate of the tagged user. These probabilities are estimated by approximation formula (8). In Fig. 5, the parameters for the wireless channel are set as $\gamma_1 = 7$ dB and $\overline{\gamma} = 16$ dB, and the parameters of the arrival process from the tagged user are set as $\alpha = 1.00$, $\beta = 0.00$

Fig. 4 Delay performance as a function of H

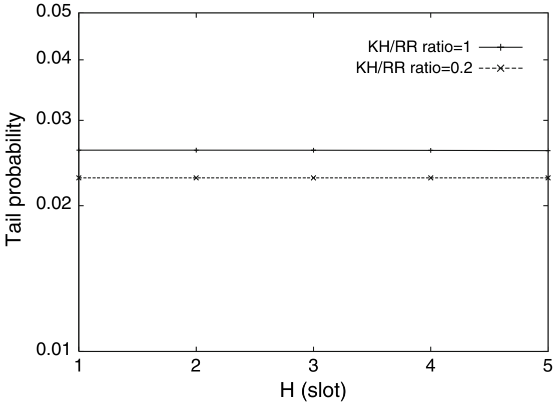
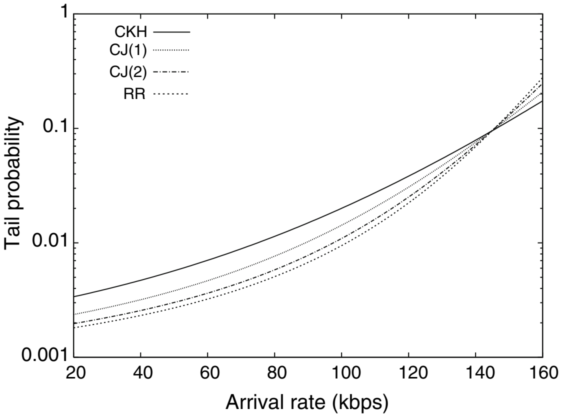


Fig. 5 Delay performance as a function of arrival rate



(i.e., the arrival process is a Bernoulli process). The service sequences of "CKH", "CJ(1)", "CJ(2)" and "RR" in Fig. 5 are the same ones as the service sequences 0, 3, 4 and 6 in Table 1, respectively. In Fig. 5, we observe that for all the service sequences, the tail probabilities of the packet delay increase with the increase of the arrival rate, and the increasing rate becomes greater when the KH/RR ratio is small. In other words, the CJ scheduling with larger KH/RR ratio is more tolerable against the increase of the arrival rate. We also see that when the arrival rate is small, the delay performance is improved with the decrease of the KH/RR ratio. In contrast, the delay performance is degraded with the decrease of the KH/RR ratio, when the arrival rate is large.

5 Conclusion

In this paper, we extend the analysis presented in Ishizaki and Hwang [6] and develop a formula to estimate the tail distribution of packet delay for an arbitrary user under the CJ scheduling algorithm. In the numerical results, we focus on a class of the CJ scheduling algorithms where $h_n \in \{0, 1\}$ for all n and the RR scheduling is employed when $m_n = 0$. We then observe that the characteristics of the CJ scheduling is mainly dominated by the KH/RR ratio. For instance, the service capacities under the CJ scheduling algorithms with large KH/RR ratio are greater than those with small one when the required QoS for delay is not stringent. We also see that the CJ scheduling algorithms with large KH/RR ratio are superior to those with small one when the arrival rate is high or the average SNR is low, and vice versa. In other words, the CJ scheduling algorithms with large KH/RR ratio are more tolerable when the system lies in sever environment.

Acknowledgements 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). Also this research was supported by Nanzan University Pache Research Subsidy I-A-1 for the 2007 academic year.

References

1. Chang, C.-S. (2000). *Performance guarantees in communication networks*. Springer-Verlag.
2. Chang, C. S., & Thomas, J. A. (1995). Effective bandwidths in high-speed digital networks. *IEEE Journal on Selected Areas in Communications, 3*, 1091–1100.
3. Elwalid, A. I., & Mitra, D. (1993). Effective bandwidths of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Transactions on Networks, 1*, 329–343.
4. Ferrús, R., Alonso, L., Umbert, A., Revés, X., Pérez-Romero, J., & Casadevall, F. (2005). Cross-layer scheduling strategy for UMTS downlink enhancement. *IEEE on Radio Communications, 2*(2), 24–28.
5. Hassan, M., Krunz, M. M., & Matta, I. (2004). Markov-based channel characterization for tractable performance analysis in wireless packet networks. *IEEE Transactions on Wireless Communications, 3*, 821–831.
6. Ishizaki, F., & Hwang, G. U. (2007). Queuing delay analysis for packet schedulers with/without multiuser diversity over a fading channel. *IEEE Transactions on Vehicular Technology, 56*, 3220–3227.
7. Kelly, F. P. (1991). Effective bandwidths at multi-class queues. *Queueing Systems, 9*, 5–16.
8. Kesidis, G., Walrand, J., & Chang, C.-S. (1993). Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networks, 1*, 424–428.
9. Kim, S. K., & Kang, C. G. (2005). Delay analysis of packet scheduling with multi-users diversity in wireless CDMA systems. *Wireless Networks, 11*, 235–241.
10. Knopp, R., & Humblet, P. A. (1995). Information capacity and power control in single-cell multiuser communications. Paper presented at ICC '95.

11. Krunz, M. M., & Kim, J. G. (2001). Fluid analysis of delay and packet discard performance of QoS support in wireless networks. *IEEE Journal on Selected Areas in Communications, 19*, 384–395.
12. Liu, Q., Zhou, S., & Giannakis, G. B. (2005). Queuing with adaptive modulation and coding over wireless links: Cross-layer analysis and design. *IEEE Transactions on Wireless Communications, 4*, 1142–1153.
13. Qin, X., & Berry, R. (2003). Exploiting multiuser diversity for medium access control in wireless networks. Paper presented at INFOCOM '03.
14. Razavilar, J., Liu, K. J. R., & Marcus, S. I. (2002). Jointly optimized bit-rate/delay control policy for wireless packet networks with fading channels. *IEEE Transactions on Communications, 50*, 484–494.
15. Shakkottai, S., Rappaport, T. S., & Karlsson, P. C. (2003). Cross-layer design for wireless networks. *IEEE Communications Magazine, 41*(10), 74–80.
16. Stüber, G. L. (2001). *Principles of mobile communication*, 2nd ed., Kluwer.
17. Wu, D., & Negi, R. (2003). Effective capacity: a wireless link model for support of quality of service. *IEEE Transactions on Wireless Communications, 2*, 630–643.
18. Wu, D., & Negi, R. (2005). Utilizing multiuser diversity for efficient support of quality of service over a fading channel. *IEEE Transactions on Vehicular Technology, 54*, 1198–1206.
19. Yang, L., Kang, M., & Alouini, M.-S. (2007). On the capacity-fairness tradeoff in multiuser diversity systems. *IEEE Transactions on Vehicular Technology, 56*, 1901–1907.
20. Zang, X., Tang, J., Chen, H.-H, Ci, S., & Guizani, M. (2006). Cross-layer-based modeling for quality of service guarantees in mobile wireless networks. *IEEE Communications Magazine, 44*(1), 100–106.

Author Biographies

Fumio Ishizaki received B.E., M.E., and Dr.Eng. from Kyoto University, Kyoto, Japan, in 1990, 1992 and 1996, respectively. From 1993 to 1995, he was an Assistant Professor in the Department of Electronic Engineering at Kinki University, Higashiosaka, Japan. From 1995 to 2000, he was an Assistant Professor in the Department of Information Science and Intelligent Systems at the University of Tokushima, Tokushima, Japan. From May 1998 to October 1998, he was a Visiting Researcher in the Department of Information and Computer Science, University of California, Irvine, CA, USA. From 2000 to 2005, he was an Associate Professor in the Department of Information and Telecommunication Engineering at Nanzan University, Seto, Japan. From August 2001 to July 2002, he was an Invited Member of Engineering Staff at Electronics and Telecommunication Research Institute (ETRI), Taejon, Korea. From 2005 to 2006, he was a Visiting Associate Professor in the Department of Mathematics and Telecommunication Mathematics Research Center at Korea University, Seoul, Korea. Since 2007, he has been an Associate Professor in the Department of Information and Telecommunication Engineering at Nanzan University, Seto, Japan. His research interests include queueing theory, control of communication networks and cross-layer design for wireless networks.

Gang Uk Hwang received his B.S., M.S., and Ph. D. degrees in Mathematics (Applied Probability) from KAIST, Daejeon, Republic of Korea, in 1991, 1993 and 1997, respectively. From February 1997 until March 2000, he was a senior member of research staff at Electronics and Telecommunications Research Institute (ETRI) where he was involved in developing ATM-based access node systems and MPLS systems. From March 2000 to February 2002, he was at the School of Interdisciplinary Computing and Engineering in University of Missouri—Kansas City as a visiting scholar. In March 2002, he joined the Department of Mathematical Sciences and Telecommunication Engineering Program at KAIST, where he is currently an Associate Professor. His research interests include teletraffic theory, QoS for next generation communication networks and the cross-layer design for wireless communication networks.