Uplink Connection Scheduling for Flat-Rate Data Services in the UMTS Network

Chung-Yung Chia and Ming-Feng Chang

***Abstract*—To attract more users to mobile packet services, universal mobile telecommunication system (UMTS) operators have been prompting flat-rate packet services. Since usage does not incur cost, flat-rate users tend to stay online longer and occupy most of the radio channel resources. We consider a UMTS network serving two types of user connections: normal user connections (NUCs) and flat-rate user connections (FRUCs). Our goal is to maximize the revenue of the operator by giving a priority to NUCs over FRUCs without discontenting the flat-rate users, in order not to lose the flat-rate users to other operators. Uplink FRUCs may be asked to subrate or suspend transmission when the radio network is fully utilized. Four combinations of scheduling techniques, including queueing, guard channels (GCs), preemption, and rate adaptation, have been studied, and analytic models using Markov processes were used to evaluate their performances. We proposed a cost function representing the revenue loss due to both blocked NUCs and lost flat-rate users. The system parameters used in our analysis are based on realistic operation data. Our analytic results indicate that the revenue loss can be minimized by using waiting queues (WQs) and preemption. Rate adaptation is ineffective in minimizing the revenue loss because subrated connections are less efficient in using the radio spectrum. GCs for NUCs are unnecessary when a WQ or preemption is used. This paper may be valuable for UMTS operators in serving flat-rate users.**

***Index Terms*—Connection scheduling, flat-rate service, universal mobile telecommunication system (UMTS).**

Nomenclature

Systems Notations

B	Size of the NUC waiting queue.
C	Cost function.
C_f	Monthly revenue loss due to lost flat-rate users.
$C_{\min}$	Minimum value of the cost function.
C_n	Monthly revenue loss due to blocked NUCs.
G	Number of guard channels.
G_{opt}	Optimum number of guard channels.
L_{QN}	Average NUC queue lengths.
L_{QF}	Average FRUC queue length.
N_F	Number of full-rate connections.
$N_F^*(y)$	Maximum number of full-rate connections when there are y half-rate connections.
$N_{\text{FF}}(i,j,k)$	Number of full-rate FRUCs in state (i,j,k).
N_H	Number of half-rate connections.
$N_H^*(z)$	Maximum number of half-rate connections when there are z full-rate connections.
$N_{\text{HF}}(i,j,k)$	Number of half-rate FRUCs in state (i,j,k).
$N_{\text{HF+FF}}(i,j,k)$	Total number of full-rate and half-rate FRUCs in state (i,j,k).
P_{BF}	Blocking probability of FRUCs.
P_{BN}	Blocking probabilities of NUCs.
P_F	Probability that the first event that occurs to a serving half-rate FRUC is being full rated.
P_{FC}	Probability that the first event that occurs to a serving full-rate FRUC is completed.
P_{FP}	Probability that the first event that occurs to a full-rate serving FRUC is being preempted.
P_{FPrm}	Probability that a serving full-rate FRUC is preempted before its completion.
P_{FS}	Probability that a serving full-rate FRUC is subrated before its completion or preemption.
P_{FST}	$= 1 - P_{\text{FC}} - P_S - P_{\text{FP}}$.
$P_{i,j,k}$	Stationary state probability of the network in state (i,j,k).
P_S	Probability that the first event that occurs to a serving full-rate FRUC is being subrated.
P_{SC}	Probability that the first event that occurs to a serving half-rate FRUC is completed.
P_{SP}	Probability that the first event that occurs to a serving half-rate FRUC is being preempted.
P_{SPrm}	Probability that a serving half-rate FRUC is preempted before its completion.
P_{SST}	$= 1 - P_{\text{SC}} - P_F - P_{\text{SP}}$.
P_{QF}	Queuing probability of FRUCs.
P_{QN}	Queuing probability of NUCs.
S_{All}	Scheduler with guard channels, waiting queues, rate adaptation, and a preemption scheduler.
S_G	Set of all existing transition states of S_{All}.
S_{NPrm}	Scheduler without preemption.
S_{NRA}	Scheduler without rate adaptation.
S_{NWQ}	Scheduler without the NUC waiting queues.
T_X	Average transmission rate of serving FRUCs.
Q	Size of the FRUC waiting queue.

Manuscript received March 4, 2007; revised May 31, 2008 and September 3, 2008. First published November 11, 2008; current version published May 11, 2009. This work was supported in part by the Nation Science Council under Contracts 95-2221-E-009-029 and 96-2752-E-009-005-PAE. The review of this paper was coordinated by Prof. J. Li.

C.-Y. Chia is with the National Chiao Tung University, Hsinchu 300, Taiwan, and also with the Telecommunication Laboratory, Chunghwa Telecom, Taoyuan 326, Taiwan (e-mail: yung88@cht.com.tw).

M.-F. Chang is with the Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan (e-mail: mfchang@csie.nctu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2008.2009114

W_{TF}	Waiting time of FRUCs.
W_{TN}	Waiting time of NUCs.
α	Cost weighting factor of flat-rate users.
α_F	Activity factor of full-rate connections.
α_H	Activity factor of half-rate connections.
β	Departure threshold of the FRUC blocking probability.
δ_F	Nominal capacity of a full-rate connection.
δ_H	Nominal capacity of a half-rate connection.
λf	Arrival rate of FRUCs.
λn	Arrival rate of NUCs.
$1/\mu\mathrm{f}$	Average service time of full-rate FRUCs.
$1/\mu\mathrm{n}$	Average service time of NUCs.
$\theta i, j, k$	Indicator to indicate whether state (i, j, k) exists or not.
ρn	Traffic load of NUCs.
ζ	Intercell interference factor for a cell.
$\Omega(\mathrm{NF, NH})$	Total transmission power that is received by the RNC in a cell.

I. Introduction

The UNIVERSAL mobile telecommunication system (UMTS) using wideband code division multiple access (WCDMA) radio technology represents an evolution in terms of capacity, data rates, and service capabilities from the global system for mobile communications/general packet radio service (GSM/GPRS) network [1]. It is an integrated solution for mobile voice and data with wide area coverage and high data rates. The UMTS network can provide packet data rates up to 384 kb/s in high-mobility situations and as high as 2 Mb/s for stationary users. The packet data usage of current UMTS users is not popular because of the lack of popular mobile data applications and the high cost of data transmission. To attract more packet data users, UMTS operators have begun to provide flat-rate packet services. Flat-rate users pay a fixed monthly charge for unlimited data packet transmission. Since usage incurs no extra charge, flat-rate users tend to keep data connections alive longer and occupy most of the network resources. Without special treatments for different classes of user connections, normal users who are charged by usage may be blocked from accessing the UMTS network.

Since blocked normal user connections (NUCs) result in revenue loss of the UMTS operator, to increase the revenue, normal users should be given priority on transmission. On the other hand, if flat-rate users experience blocked connections frequently, the discontent flat-rate users may switch to other service providers. The loss of flat-rate users also leads to revenue loss. Therefore, a balance needs to be found in allocating radio resources to normal and flat-rate users. In this paper, we propose a cost function representing the revenue loss due to both blocked normal users and lost flat-rate users. Since the revenue loss on both situations depends on the blocking probabilities, we investigate scheduling techniques, including queueing, guard channels (GCs), preemption, and rate adaptation, to keep the blocking probabilities of normal and flat-rate users at different levels and to minimize the cost function, i.e., the revenue loss. We consider the aforementioned four scheduling techniques, because they have been repeatedly used in giving transmission priorities in mobile networks. However, no one has investigated the effectiveness of these four scheduling techniques in maximizing the revenue of UMTS operators serving flat-rate and normal users.

Much research has been done on the mobile network in giving a transmission priority to a certain type of service. In mobile networks, terminating a handoff call is considered a higher cost than blocking a new call. When a handoff call arrives but there is no free channel in the cell, the handoff call can be placed in a queue, and handoff is delayed until free channels become available [2]. To further give a priority to handoff calls, a small number of free channels called GCs can be reserved for handoff calls. GCs significantly reduce the forced termination probability of handoff calls at the cost of blocking more new calls and reducing the system throughput [3]. To increase the total carried traffic and improve the perceived service quality, Guerin put originating calls in a queue when the network has very few free resources [4]. Zeng *et al.* also proposed that both the new and handoff calls can be queued and showed that the forced termination probability of handoff calls decreased drastically with only a small increase in the blocking probability of new calls [5]. For integrated voice and data communications, Zeng *et al.* presented a system with two queues for handoff calls, one for voice and the other for data. Their results showed that the forced termination probability of voice handoff calls and the average transmission delay of data connections decreased by increasing the size of handoff queues [6]. Leong *et al.* presented a system with two buffers for data calls, one for a new data call and the other for handoff. Their results indicated that the quality of service (QoS) can be guaranteed for both voice and data services in a multicell environment [7].

Preempting a low-priority call to free radio resources for high-priority calls is another effective way to ensure transmission priority. However, this approach usually preempts data calls only, because cutting off voice communications can be very annoying to the users. The high priority of real-time traffic, such as voice and video, can preempt non-real-time traffic (data). Several researchers have shown that the preemption of non-real-time data can guarantee QoS for real-time classes and achieve high channel utilization [8], [9]. Kim *et al.* proposed that high-priority voice calls can preempt low-priority voice calls. Voice calls that have low signal–interference ratios (SIRs) and long durations are considered low-priority calls, which can be preempted to improve the entire network performance [10].

Subrating current calls to free radio resources for new or handoff calls is another way to reduce blocking probabilities. A serving full-rate channel can temporarily be divided into two half-rate channels when the network is fully utilized, one to serve the existing call and the other to serve the handoff call [11]. Chen *et al.* study GPRS networks where a data session can occupy more than one GPRS data channel. When there are no free channels upon the arrival of a voice call, one slot of an existing multislot GPRS data session is deallocated for the new voice arrival [12]. Their results show that the voice blocking probability can greatly be reduced, particularly at high GPRS traffic load.

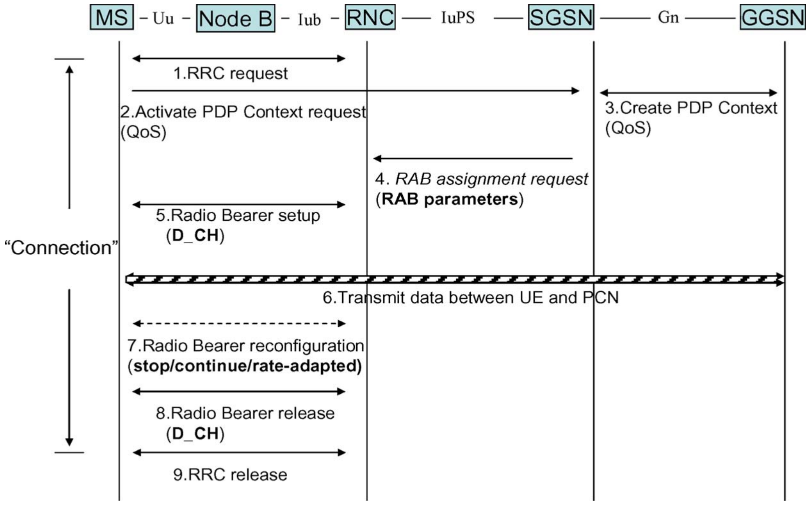

MS
Uu
Node B
Iub
RNC
IuPS
SGSN
Gn
GGSN
1.RRC request
2.Activate PDP Context request (QoS)
3.Create PDP Context (QoS)
4. RAB assignment request (RAB parameters)
5.Radio Bearer setup (D_CH)
"Connection"
6.Transmit data between UE and PCN
7.Radio Bearer reconfiguration (stop/continue/rate-adapted)
8.Radio Bearer release (D_CH)
9.RRC release

Fig. 1. RAB assignment procedure.

Most of the researches focus on reducing the blocking and forced termination probabilities of high-priority connections. However, very few studies have been done on maximizing the operator revenue for mobile networks serving flat-rate users as well as normal users. In this paper, we investigate combinations of the scheduling techniques aforementioned to maximize the operator revenue. We propose a cost function that represents the revenue loss of service providers providing both flat-rate and per-packet charging services. An iterative algorithm has been developed to determine the optimal number of GCs and the best combination of scheduling techniques in minimizing the revenue loss. This paper may be valuable for UMTS operators in serving flat-rate users.

II. System Models and Assumptions

A UMTS network consists of three interacting domains: core network (CN), UMTS terrestrial radio access network (UTRAN), and mobile stations (MSs). The UTRAN provides the air interface access method for MSs [13]. A base station is referred to as Node B; the control node for a group of Node Bs is called a radio network controller (RNC). Wideband CDMA technology was selected to be the air interface of the UTRAN. To be specific, we study the frequency division duplex WCDMA operation in this paper.

An RNC can allocate a physical dedicated radio channel (D_CH) to an MS by a radio access bearer (RAB) assignment procedure [13]–[15]. Fig. 1 shows the message flow of a D_CH assignment procedure. In Step 1, an MS establishes a radio resource control (RRC) connection with the RNC before creating a packet data protocol (PDP) context between the MS and gateway GPRS support node (GGSN). In Step 2, the MS sends an "Activate PDP Context Request" message to the serving GPRS support node (SGSN) with a QoS element indicating service class (conversional, streaming, interactive, or background data). In Step 4, the SGSN sends a "RAB assignment request" message with RAB parameters, which will be described in more details later, to the RNC to establish a RAB connection between the MS and SGSN. After the D_CH is established in Step 5, the MS can start to transmit/receive packets to/from the CN in Step 6. When necessary, the RNC can instruct the MS that the packet transmission of the connection should be stopped or continued or that the transmission rate on its assigned D_CH should be changed by a radio bearer (RB) reconfiguration procedure as indicated in Step 7. The MS should comply with the instructions. After the MS completes transmission, the RB and RRC of the MS can be released in Steps 8 and 9.

The RAB parameters sent from the SGSN to the RNC in Step 4 can be used to instruct the RNC regarding the scheduling policy of the data connection. The parameters include a priority level element, a preemption capability element indicating the capability to preempt lower priority RABs, a preemption vulnerability element indicating whether the D_CH is vulnerable to be preempted or not, and a queueing allowed element indicating whether the RAB request can be queued. In addition, maximum and guaranteed bit rate elements indicate the transmission rate of MSs. These RAB parameters can be used to instruct the RNC how to schedule the packet transmission.

A user connection starts at the establishment of an RRC between the MS and the RNC, as shown in Step 1 of Fig. 2, and ends at the disconnection of the RNC. We assume that there are two types of user connections in the UMTS network: NUCs, which are assigned a higher priority in transmission, and flat-rate user connections (FRUCs), which may be subrated or suspended when the network traffic load is high. When a connection is subrated, its transmission rate and power can be reduced, and thus, transmission power allowance is released for other connections. Since NUCs are charged by the volume of packet transmission, a NUC tends to be shorter, such as sending an e-mail or uploading short files. On the other hand, since FRUCs pay a fixed monthly fee no matter how many packets they transmit, a FRUC is generally longer, such as playing online games and using peer-to-peer applications.

In the UMTS R99 network, the uplink data transmission can only be scheduled on connection level, but not on packet level. This is because, after D_CHs are allocated to MSs, the MSs can start or pause data transmission anytime without notifying the RNC. However, the RNC can suspend or subrate the uplink connection as we have described. On the other hand, downlink

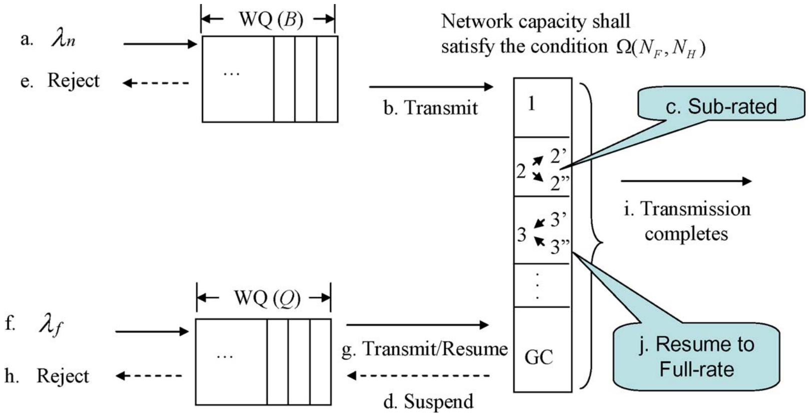


Fig. 2. System queueing model for a reference cell.

data transmission can be scheduled on packet level, because all downlink packets are stored and forwarded by the RNC. The RNC can determine priorities in forwarding different classes of packets. As a result, uplink and downlink transmissions may require different scheduling techniques. In this paper, we consider the scheduling for uplink data connections only. We use the CDMA uplink soft-capacity model to estimate the uplink total bandwidth of a cell a in system.

A. CDMA Uplink Capacity Model

The capacity of a CDMA network is not fixed; it has a so-called "soft capacity." Since a FRUC can be subrated, we consider two transmission rates of data services from MSs, full- and half-rate data connections. We can obtain the limit on the total transmission power received by the Node B in a cell from (1) [16]. α_F and α_H denote the activity factor of full- and half-rate data connections in a cell, respectively. N_F and N_H denote the numbers of MSs using full- and half-rate data connections, respectively. δ_F denotes the *nominal capacity* of a full-rate data connection, i.e., the portion of total transmission power received by the Node B in a cell; δ_H denotes that of a half-rate data connection [17]. ζ is the intercell interference factor for a cell, which can be obtained from measurements [18]

$$\Omega(N_F, N_H) = \alpha_F \times N_F \times \delta_F + \alpha_H \times N_H \times \delta_H < \frac{1}{(1+\zeta)}. \tag{1}$$

From (1), we can obtain the *Pole capacity* of MSs using full- and half-rate data connections in a cell as in (2). $N_F^*(y)$ denotes the maximum number of full-rate serving MSs in a cell when there are y half-rate serving MSs, and $N_H^*(z)$ denotes the maximum number of half-rate serving MSs in a cell when there are z full-rate serving MSs. In particular, $N_F^*(0)$ denotes the maximum number of full-rate serving MSs in a cell and $N_H^*(0)$ denotes the half-rate serving MSs

$$N_F^*(y) = \max\left\{N_F \left| \frac{1}{(1+\zeta)} \geq (\alpha_F \times N_F \times \delta_F + \alpha_H \times y \times \delta_H)\right.\right\}$$
$$N_H^*(z) = \max\left\{N_H \left| \frac{1}{(1+\zeta)} \geq (\alpha_F \times z \times \delta_F + \alpha_H \times N_H \times \delta_H)\right.\right\}. \tag{2}$$

In the following analysis, we assume that the spread spectrum bandwidth (W) of the WCDMA network is 5 MHz, the full-rate data transmission is 128 kb/s, and the half rate is 64 kb/s. The two data rates are the default uplink data rates provided by the Chunghwa Telecom (CHT) UMTS R99 network. According to the Third Generation Partnership Project (3GPP) specification [19], full- and half-rate data transmissions have different SIR requirements to achieve block error rate $< 10^{-2}$ in multipath fading conditions; for full rate, it is 8.4 dB, and for half-rate, 9.2 dB. From the desired SIR, we can obtain the nominal capacity $\delta_F = 0.177$ and $\delta_H = 0.106$. The activity factor for data services (α_F and α_H) is assumed to be 0.5 in busy hours, and the intercell interference factor (ζ) is assumed to be 0.1. These assumptions follow those in [18].

From (2), we can obtain $N_F^*(0) = 10$ and $N_H^*(0) = 17$. Note that $N_H^*(0)$ is less than twice of $N_F^*(0)$ because more number of MSs transmitting leads to more signal interference. In other words, half-rate transmission is less efficient in using radio bandwidth.

B. Scheduler With All Four Features

The queueing model of the connection scheduler that implements waiting queues (WQ), GCs, preemption, and rate adaptation on the RNC is shown in Fig. 2. There are two WQs, one for new NUCs and the other for new and preempted FRUCs. When an ongoing uplink connection is put in a WQ, the Node B instructs the MS to stop packet transmission. Since there is no packet transmission, no storage space on Node B is needed for the uplink packets of a queued connection. GCs of dynamic sizes are reserved for NUCs. The number of GCs will be determined by an iterative algorithm described later to maximize the revenue. The connection scheduler works as follows. When a new NUC (line a) request arrives, it can immediately be served if the network is not fully utilized (line b). Otherwise, the RNC first tries to subrate serving FRUCs (line c) to accommodate the new NUC. If this is not possible, i.e., all serving FRUCs are subrated, the RNC preempts FRUCs into the WQ (dotted line d). If there is no serving FRUC, the new NUC is put into the NUC WQ; if the queue is full, it is rejected [see the dotted line in Fig. 2(e)].

When a new FRUC (line f) request arrives, it can immediately be served if there are free channels other than the reserved GCs (line g). Otherwise, serving full-rate FRUCs can be subrated (line c) to accommodate the new FRUC, if doing so satisfies the total power limit in (2). Otherwise, the new FRUC request can be put into the FRUC WQ; if the queue is full, it is rejected (dotted line h).

When a serving connection finishes, it releases radio channels (line i). The free channel will serve a waiting NUC first. If there is no waiting NUC, waiting FRUCs will be served (line g). If there is no waiting FRUC, serving half-rate FRUCs can resume full-rate transmission (line j).

The scheduler implementing all four scheduling techniques described previously will be referred to as $S_{\rm All}$. To evaluate the effectiveness of WQs, rate adaptation, and preemption, we also analyzed three additional schedulers, each of which omits one scheduling technique. Let $S_{\rm NWQ}$ denote the one without a NUC WQ, $S_{\rm NRA}$ without rate adaptation, and $S_{\rm NPrm}$ without preemption. All the schedulers implement GCs. Due to space limitation, the analytic models and the performance measure equations of $S_{\rm NRA}$, $S_{\rm NWQ}$, and $S_{\rm NPrm}$ are not presented in this paper.

III. Analytic Models

We can use the M/M/C/B Markov process to analyze the connection schedulers. Let B denote the size of the NUC WQ, G the number of GCs, and Q the size of the FRUC WQ. The new arrivals of NUCs and FRUCs were assumed to form Poisson processes with rates λ_n and λ_f, respectively. The service times of NUCs and full-rate FRUCs were assumed to be exponentially distributed with means $1/\mu_n$ and $1/\mu_f$, respectively. The assumption of Poisson arrivals can provide a good approximation when the user population is large; the assumption of exponential service time facilitates the analysis.

A. Analytic Model of $S_{\rm All}$

The analytic model for the scheduler with all features $(S_{\rm All})$ is described as follows. Let state (i,j,k) denote that there are i transmitting NUCs, j waiting NUCs, and k transmitting or waiting FRUCs. The exact numbers of full- and half-rate transmitting FRUCs can be determined by an algorithm shown in Fig. 3. Let $N_{\rm FF}(i,j,k)$ denote the number of full-rate FRUCs in state (i,j,k), $N_{\rm HF}(i,j,k)$ that of half rate, and $N_{\rm FF+HF}(i,j,k)$ the total number of full- and half-rate FRUCs in state (i,j,k).

Let S_G be the set of all existing transition states of $S_{\rm All}$. For each existing state (i,j,k), the number of serving NUCs cannot be more than $N_F^*(0)$, the number of queued NUCs cannot be more than B, and the number of FRUCs cannot be more than $N_{\rm FF+HF}(i,j,k)+Q$. S_G can be expressed as in (3). The maximum size of S_G should be limited to $[N_F^*(0)+1]*[B+1]*[N_H^*(0)+Q+1]$

$$S_G=\{(i,j,k)|\,[0\le i\le N_F^*(0), j=0, 0\le k\le N_H^*(i)+Q] \text{ or } [i=N_F^*(0), 0<j\le B, 0\le k\le Q]\}. \quad (3)$$

```
if [$i$ >=( $N_F^*(0)$-$G$)],
  all k FRUCs are queued, $N_{FF}(i,j,k)$=0 and $N_{HF}(i,j,k)$=0
else if [($i$+$k$) < ( $N_F^*(0)$ -$G$)], all FRUCs are served in full rate,
  $N_{FF}(i,j,k)$=$k$ and $N_{HF}(i,j,k)$=0
else if there exists a minimal $h$, such that [($i$+$k$) < ( $N_F^*(h)$ -$G$+$h$)],
  $N_{FF}(i,j,k)$=$k$-$h$ and $N_{HF}(i,j,k)$=$h$
else if $N_{FF}(i,j,k)$=0, $N_{HF}(i,j,k)$=$N_H^*(i+G)$, and ($k$-
  $N_H^*(i+G)$) FRUCs are queued
```

Fig. 3. Algorithm determines the numbers of full- and half-rate FRUCs in state (i,j,k).

Part of the state transition diagram is shown in Fig. 4. To handle the nonexisting states, an indicator $\theta_{i,j,k}$ is used to indicate whether state (i,j,k) exists or not. $\theta_{i,j,k}=1$ if state (i,j,k) belongs to S_G; otherwise, $\theta_{i,j,k}=0$. Let $P_{i,j,k}$ denote the steady-state probability of the network in state (i,j,k). For existing state (i,j,k), the output flows (lines 1–4), its input flows from other states (dotted lines 5–8), and the transition rate of each line is shown in Fig. 4.

The rate of the input flows of state (i,j,k) can be expressed in (4) and that of output flows in (5). When the system is in equilibrium, the rates are equal; the system equilibrium equation of state (i,j,k) can be expressed in (6)

$$Inflow(i,j,k)=\begin{cases}\lambda_n\cdot P_{i-1,j,k}\theta_{i-1,j,k}+\lambda_n\\ \cdot P_{i,j-1,k}\theta_{i,j-1,k}+\lambda_f\\ \cdot P_{i,j,k-1}\theta_{i,j,k-1}+(i+1)\mu_n\\ \cdot P_{i+1,j,k}\theta_{i+1,j,k}+i\mu_n\\ \cdot P_{i,j+1,k}\theta_{i,j+1,k}\\ +\left(N_{\rm FF}(i,j,k+1)\cdot\mu_n\right.\\ \qquad \left.+\left(N_{\rm HF}(i,j,k+1)\cdot\mu_n/2\right)\right)\\ \cdot P_{i,j,k+1}\theta_{i,j,k+1}\end{cases} \quad (4)$$

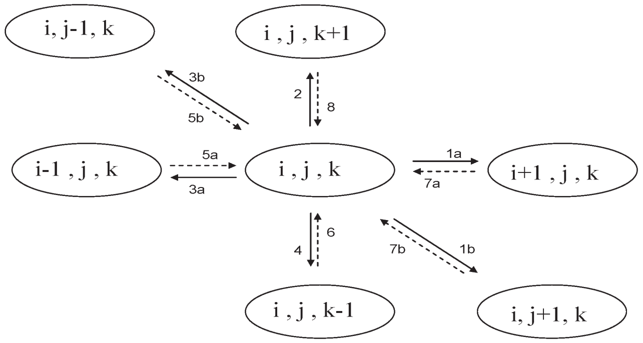
$$Outflow(i,j,k)=\begin{cases}\lambda_n\theta_{i+1,j,k}+\lambda_n\cdot\theta_{i,j+1,k}+\lambda_f\\ \cdot\theta_{i,j,k+1}+i\mu_n\cdot\theta_{i-1,j,k}\\ +i\mu_n\cdot\theta_{i,j-1,k}\\ +\left(N_{\rm FF}(i,j,k)\cdot\mu_n\right.\\ \qquad \left.+\left(N_{\rm HF}(i,j,k)\cdot\mu_n/2\right)\right)\\ \cdot\theta_{i,j,k-1}\end{cases} \quad (5)$$

$$P_{i,j,k}=Inflow(i,j,k)/Outflow(i,j,k). \quad (6)$$

From (3)–(6) and the constraint $\sum_{(i,j,k)\in S_G}P_{i,j,k}=1$, we can use the iterative algorithm proposed in [21] to obtain the stationary state probabilities $P_{i,j,k}$. The iterative algorithm of our system will be described in more details later. It is possible to extend the system model for an arbitrary number of rates, if the scheduling scheme can determine the numbers of connections served at each rate, given the numbers of NUCs and FRUCs.

B. Performance Measures

The performance measures we considered are described as follows. A NUC or FRUC is blocked when the WQ is full.



line 1a = $\lambda_n, if\left(\theta_{i+1,j,k} == 1\right)$; line 1b = $\lambda_n, if\left(\theta_{i,j+1,k} == 1\right)$ // lines 1.a and 1.b cannot co-exist

line 2 = $\lambda_f, if\left(\theta_{i,j,k+1} == 1\right)$

line 3a = $i \cdot \mu_n, if\left(\theta_{i-1,j,k} == 1\right)$; line 3b = $i \cdot \mu_n, if\left(\theta_{i,j-1,k} == 1\right)$ // lines 3.a and 3.b cannot co-exist

line 4 = $N_{FF}(i,j,k) \cdot \mu_f + N_{HF}(i,j,k) \cdot \mu_f / 2, if\left(\theta_{i,j,k-1} == 1\right)$

line 5a = $\lambda_n, if\left(\theta_{i-1,j,k} == 1\right)$; line 5b = $\lambda_n, if\left(\theta_{i,j-1,k} == 1\right)$ // lines 5.a and 5.b cannot co-exist

line 6 = $\lambda_f, if\left(\theta_{i,j,k-1} == 1\right)$

line 7a = $i \cdot \mu_n, if\left(\theta_{i+1,j,k} == 1\right)$; line 7b = $i \cdot \mu_n, if\left(\theta_{i,j+1,k} == 1\right)$ // lines 7.a and 7.b cannot co-exist

line 8 = $N_{FF}(i,j,k+1) \cdot \mu_f + N_{HF}(i,j,k+1) \cdot \mu_f / 2, if\left(\theta_{i,j,k+1} == 1\right)$

Fig. 4. State transition diagram of S_{All} and the rates of input/output flow.

Based on the Poisson arrivals see time averages property [20], the blocking probabilities of NUCs (P_{BN}) or FRUCs (P_{BF}) can be expressed as

$$P_{\text{BN}} = \sum_{(i,j,k)\in S_G,(j==B)} P_{i,j,k} \tag{7}$$

$$P_{\text{BF}} = \sum_{(i,j,k)\in S_G,(k-N_{\text{HF+FF}}(i,j,k)==Q)} P_{i,j,k}. \tag{8}$$

From the stationary state probabilities, we can obtain the average queue lengths of the NUC WQ (L_{QN}) and FRUC WQ (L_{QF}); they can be expressed as follows, respectively:

$$L_{\text{QN}} = \sum_{(i,j,k)\in S_G} j \cdot P_{i,j,k} \tag{9}$$

$$L_{\text{QF}} = \sum_{(i,j,k)\in S_G} [k - N_{\text{HF+FF}}(i,j,k)] \cdot P_{i,j,k}. \tag{10}$$

The average waiting times of NUCs (W_{TN}) and FRUCs (W_{TF}) in the WQ can be obtained using *Little's formula*; they can be expressed as follows, respectively:

$$W_{\text{TN}} = \frac{L_{\text{QN}}}{\lambda_n \cdot (1 - P_{\text{BN}})} \tag{11}$$

$$W_{\text{TF}} = \frac{L_{\text{QF}}}{\lambda_f \cdot (1 - P_{\text{BF}})}. \tag{12}$$

The queueing probability of a connection is defined as the probability that a connection cannot immediately be served upon its arrival. A new NUC is put into the WQ when the network is fully utilized by NUCs. The queueing probability of NUCs (P_{QN}) can be expressed in (13). A new FRUC is put into the WQ when all the serving FRUCs are subrated or subrating full-rate serving FRUCs cannot release enough bandwidth for the new FRUC to transmit in half rate. The queueing probability of FRUCs (P_{QF}) can be expressed as

$$P_{\text{QN}} = \sum_{(i,j,k)\in S_G, i==N_F^*(0),(j<B)} P_{i,j,k} \tag{13}$$

$$P_{\text{QF}} = \sum_{\substack{(i,j,k)\in S_G, 0<(k-N_{\text{HF+FF}}(i,j,k))<Q \\ N_{\text{HF+FF}}(i,j,k)==N_{\text{HF+FF}}(i,j,k+1)}} P_{i,j,k}. \tag{14}$$

To obtain the probability that a full-rate serving FRUC is subrated, consider the first event that occurs to a full-rate serving FRUC. The FRUC may be complete (with probability P_{FC}), subrated (with probability P_S), or preempted (with probability P_{FP}), or none of the aforementioned events occurs but a state transition occurs (with probability P_{FST}). The probabilities can be expressed as (15)–(18), shown at the bottom of the next page.

Let P_{FS} denote the probability that a full-rate serving FRUC is subrated before its completion or preemption. From the memoryless property of the Markov process, P_{FS} can be expressed as

$$P_{\text{FS}} = P_S + P_{\text{FST}} P_{\text{FS}} = P_S / (1 - P_{\text{FST}}). \tag{19}$$

In the same way, we consider the first event that occurs to a serving subrate FRUC. The FRUC may be complete (with probability P_{SC}), full-rated (with probability P_F), or preempted (with probability P_{SP}), or none of the aforementioned events occurs but a state transition occurs (with probability P_{SST}). The probabilities can be expressed as (20)–(23), shown at the bottom of the next page.

Let P_{FPrm} denote the probability that a full-rate serving FRUC is preempted before its completion and P_{SPrm} denote that of a subrated serving FRUC. From the memoryless property of the Markov process, they can be expressed as follows, respectively:

$$\begin{aligned} P_{\mathrm{FPrm}} &= P_{\mathrm{FP}} + P_S P_{\mathrm{SPrm}} + P_{\mathrm{FST}} P_{\mathrm{FPrm}} \\ &= (P_{\mathrm{FP}} + P_S P_{\mathrm{SPrm}})/(1 - P_{\mathrm{FST}}) \qquad (24) \\ P_{\mathrm{SPrm}} &= P_{\mathrm{SP}} + P_F P_{\mathrm{FPrm}} + P_{\mathrm{SST}} P_{\mathrm{SPrm}} \\ &= (P_{\mathrm{SP}} + P_F P_{\mathrm{FPrm}})/(1 - P_{\mathrm{SST}}). \qquad (25) \end{aligned}$$

From (22)–(25), we can obtain P_{FPrm} and P_{SPrm}; they can be expressed as follows, respectively:

$$\begin{aligned} P_{\mathrm{FPrm}} = {} & [(1 - P_{\mathrm{SST}})P_{\mathrm{FP}} + P_S P_{\mathrm{SP}}] \\ & / [(1 - P_{\mathrm{FST}})(1 - P_{\mathrm{SST}}) - P_S P_F] \qquad (26) \\ P_{\mathrm{SPrm}} = {} & [(1 - P_{\mathrm{FST}})P_{\mathrm{SP}} + P_F P_{\mathrm{FP}}] \\ & / [(1 - P_{\mathrm{FST}})(1 - P_{\mathrm{SST}}) - P_S P_F]. \qquad (27) \end{aligned}$$

From the stationary state probabilities, we can obtain the average transmission rate of serving FRUCs by

$$T_F = \sum_{(i,j,k)\in S_G, k>0} P_{i,j,k} \cdot \frac{N_{\mathrm{FF}}(i,j,k) \cdot 128k + N_{\mathrm{HF}}(i,j,k) * 64k}{N_{\mathrm{FF}}(i,j,k) + N_{\mathrm{HF}}(i,j,k)}. \qquad (28)$$

C. Cost Function Scheme

In this paper, we consider a mobile data operator's revenue that consists of the transmission fee of normal users and the monthly fee of flat-rate users. Instead of calculating the total revenue, we propose a cost function representing the revenue loss due to blocked NUCs and the loss of flat-rate users, since NUCs are charged by the volume of packets transmitted. In a fully utilized network, retransmitting blocked NUCs only leads

$$P_{\mathrm{FC}} = \sum_{(i,j,k)\in S_G} P_{i,j,k} \cdot \frac{\mu_n}{i\mu_n + N_{\mathrm{FF}}(i,j,k)\cdot\mu_f + N_{\mathrm{HF}}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \qquad (15)$$

$$\begin{aligned} P_S = {} & \sum_{\substack{(i,j,k)\in S_G, N_{\mathrm{FF}}(i,j,k)!=0, \\ N_{\mathrm{HF}}(i+1,j,k)>N_{\mathrm{HF}}(i,j,k)}} P_{i,j,k} \cdot \frac{\lambda_n \cdot [N_{\mathrm{HF}}(i+1,j,k) - N_{\mathrm{HF}}(i,j,k)]/N_{\mathrm{FF}}(i,j,k)}{i\mu_n + N_{\mathrm{FF}}(i,j,k)\cdot\mu_f + N_{\mathrm{HF}}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \\ & + \sum_{\substack{(i,j,k)\in S_G, N_{\mathrm{FF}}(i,j,k)!=0, \\ N_{\mathrm{HF}}(i,j,k+1)>N_{\mathrm{HF}}(i,j,k)}} P_{i,j,k} \cdot \frac{\lambda_f \cdot [N_{\mathrm{HF}}(i,j,k+1) - N_{\mathrm{HF}}(i,j,k) - 1]/N_{\mathrm{FF}}(i,j,k)}{i\mu_n + N_{\mathrm{FF}}(i,j,k)\cdot\mu_f + N_{\mathrm{HF}}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \qquad (16) \end{aligned}$$

$$P_{\mathrm{FP}} = \sum_{\substack{(i,j,k)\in S_G, N_{\mathrm{FF}}(i,j,k)!=0 \\ N_{\mathrm{HF}}(i,j,k)==N_{\mathrm{HF}}(i+1,j,k) \\ N_{\mathrm{FF}}(i,j,k)>N_{\mathrm{FF}}(i+1,j,k)}} P_{i,j,k} \cdot \frac{\lambda_n \cdot [N_{\mathrm{FF}}(i,j,k) - N_{\mathrm{FF}}(i+1,j,k)]/N_{\mathrm{FF}}(i,j,k)}{i\mu_n + N_{\mathrm{FF}}(i,j,k)\cdot\mu_f + N_{\mathrm{HF}}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \qquad (17)$$

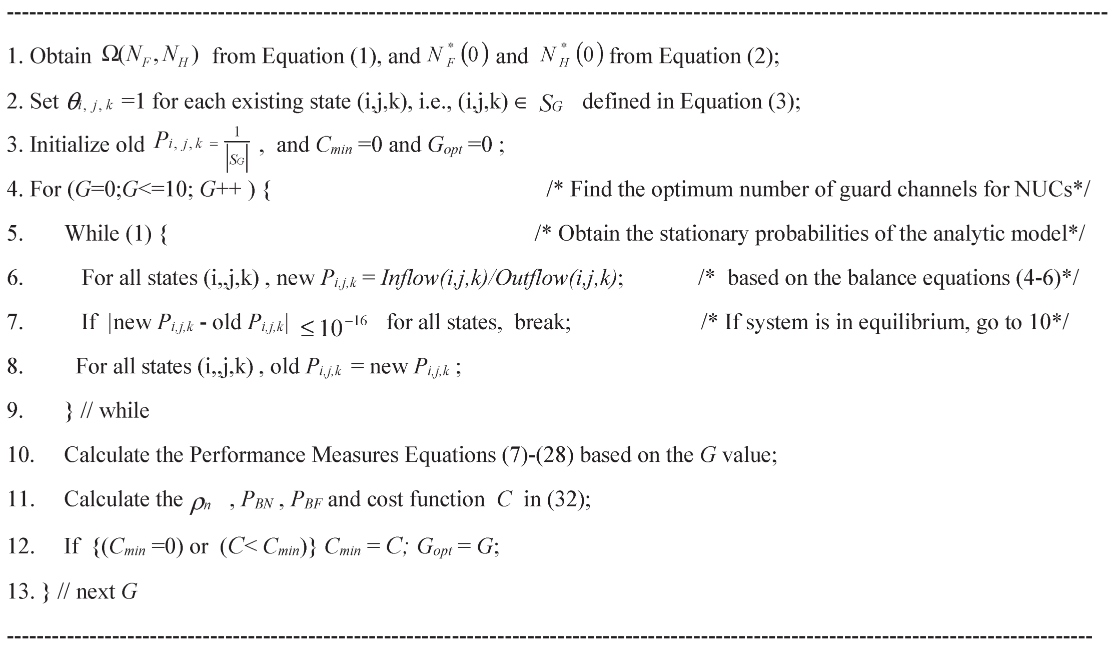
$$P_{\mathrm{FST}} = 1 - P_{\mathrm{FC}} - P_S - P_{\mathrm{FP}} \qquad (18)$$

$$P_{\mathrm{SC}} = \sum_{(i,j,k)\in S_G} P_{i,j,k} \cdot \frac{\frac{\mu_f}{2}}{i\mu_n + N_{\mathrm{FF}}(i,j,k)\cdot\mu_f + N_{\mathrm{HF}}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \qquad (20)$$

$$\begin{aligned} P_F = {} & \sum_{\substack{(i,j,k)\in S_G, N_{\mathrm{HF}}(i,j,k)!=0, \\ N_{\mathrm{FF}}(i-1,j,k)>N_{\mathrm{FF}}(i,j,k)}} P_{i,j,k} \cdot \frac{i\mu_n \cdot [N_{\mathrm{FF}}(i-1,j,k) - N_{\mathrm{FF}}(i,j,k)]/N_{\mathrm{HF}}(i,j,k)}{i\mu_n + N_{\mathrm{FF}}(i,j,k)\cdot\mu_f + N_{\mathrm{HF}}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \\ & + \sum_{\substack{(i,j,k)\in S_G, N_{\mathrm{HF}}(i,j,k)!=0 \\ N_{\mathrm{FF}}(i,j,k-1)\geq N_{\mathrm{FF}}(i,j,k)}} P_{i,j,k} \cdot \frac{N_{\mathrm{FF}}(i,j,k)\cdot\mu_f \cdot [N_{\mathrm{FF}}(i,j,k-1) - N_{\mathrm{FF}}(i,j,k) + 1]/N_{\mathrm{HF}}(i,j,k)}{i\mu_n + N_{\mathrm{FF}}(i,j,k)\cdot\mu_f + N_{\mathrm{HF}}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \\ & + \sum_{\substack{(i,j,k)\in S_G, N_{\mathrm{HF}}(i,j,k)!=0 \\ N_{\mathrm{FF}}(i,j,k-1)>N_{\mathrm{FF}}(i,j,k)}} P_{i,j,k} \cdot \frac{N_{\mathrm{HF}}(i,j,k)\cdot\frac{\mu_f}{2} \cdot [N_{\mathrm{FF}}(i,j,k-1) - N_{\mathrm{FF}}(i,j,k)]/N_{\mathrm{HF}}(i,j,k)}{i\mu_n + N_{\mathrm{FF}}(i,j,k)\cdot\mu_f + N_{\mathrm{HF}}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \qquad (21) \end{aligned}$$

$$P_{\mathrm{SP}} = \sum_{\substack{(i,j,k)\in S_G, N_{\mathrm{HF}}(i,j,k)!=0 \\ N_{\mathrm{FF}}(i,j,k)==N_{\mathrm{FF}}(i+1,j,k) \\ N_{\mathrm{HF}}(i,j,k)>N_{\mathrm{HF}}(i+1,j,k)}} P_{i,j,k} \cdot \frac{\lambda_n \cdot [N_{\mathrm{HF}}(i,j,k) - N_{\mathrm{HF}}(i+1,j,k)]/N_{\mathrm{HF}}(i,j,k)}{i\mu_n + N_{\mathrm{FF}}(i,j,k)\cdot\mu_f + N_{\mathrm{HF}}(i,j,k)\cdot\frac{\mu_f}{2} + \lambda_n + \lambda_f} \qquad (22)$$

$$P_{\mathrm{SST}} = 1 - P_{\mathrm{SC}} - P_F - P_{\mathrm{SP}} \qquad (23)$$



```
1. Obtain $\Omega(N_F, N_H)$ from Equation (1), and $N_F^*(0)$ and $N_H^*(0)$ from Equation (2);
2. Set $\theta_{i,j,k}$ =1 for each existing state (i,j,k), i.e., (i,j,k) ∈ $S_G$ defined in Equation (3);
3. Initialize old $P_{i,j,k} = \frac{1}{|S_G|}$ , and $C_{min}$ =0 and $G_{opt}$ =0 ;
4. For (G=0;G<=10; G++ ) {                                   /* Find the optimum number of guard channels for NUCs*/
5.     While (1) {                                           /* Obtain the stationary probabilities of the analytic model*/
6.       For all states (i,,j,k) , new $P_{i,j,k}$ = Inflow(i,j,k)/Outflow(i,j,k);     /* based on the balance equations (4-6)*/
7.       If |new $P_{i,j,k}$ - old $P_{i,j,k}$| $\leq 10^{-16}$ for all states, break;   /* If system is in equilibrium, go to 10*/
8.      For all states (i,,j,k) , old $P_{i,j,k}$ = new $P_{i,j,k}$ ;
9.     } // while
10.    Calculate the Performance Measures Equations (7)-(28) based on the G value;
11.    Calculate the $\rho_n$ , $P_{BN}$ , $P_{BF}$ and cost function C in (32);
12.    If {($C_{min}$ =0) or (C< $C_{min}$)} $C_{min}$ = C; $G_{opt}$ = G;
13. } // next G
```

Fig. 5. Iterative algorithm minimizing the cost function.

to more NUCs blocked. Therefore, we assume that blocked NUCs in a fully utilized network will not be retransmitted and thus represent revenue loss. The revenue loss of blocked NUCs is proportional to the blocking probability (P_{BN}) and the traffic load of NUCs $(\rho_n = \lambda_n/\mu_n)$. The monthly revenue loss due to blocked NUCs can be expressed in (29), where D denotes the transmission charge of a NUC per busy hour, E the number of busy hours per month, and F the number of cells

$$C_n = D \cdot E \cdot F \cdot \rho_n \cdot P_{\mathrm{BN}}. \tag{29}$$

The revenue loss due to the loss of flat-rate users also depends on the blocking probability. Since flat-rate users are not charged by the volume of packet transmission, blocked FRUCs do not result in direct revenue loss. However, when the blocking probability is above a departure threshold β, flat-rate users may become discontent and start to switch to other operators. We assume that the number of flat-rate users lost per month is proportional to the discrepancy of the blocking probability (P_{BF}) above β. The monthly revenue loss due to lost flat-rate users can be expressed in (30), where X denotes the total number of flat-rate users, Y the percentage of flat-rate users lost due to each percentage increase of blocking probability above β, and Z the monthly charge of a flat-rate user

$$C_f = \begin{cases} X \cdot Y \cdot Z \cdot (P_{\mathrm{BF}} - \beta) \cdot 100, & \text{if } (P_{\mathrm{BF}} > \beta) \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

The total monthly loss is C_n plus C_f. Dividing the monthly revenue loss by D, E, and F, we obtain the cost function, as shown in (31)–(33), where α represents the cost weighting factor of flat-rate connections

$$\begin{aligned} C &= C_n + C_f \\ &= \begin{cases} D \cdot E \cdot F \cdot \rho_n \cdot P_{\mathrm{BN}} \\ +X \cdot Y \cdot Z \cdot (P_{\mathrm{BF}} - \beta) \cdot 100, & \text{if } (P_{\mathrm{BF}} > \beta) \\ D \cdot E \cdot F \cdot \rho_n \cdot P_{\mathrm{BN}}, & \text{otherwise} \end{cases} \end{aligned} \tag{31}$$

$$C = \begin{cases} \rho_n \cdot P_{\mathrm{BN}} + \alpha \cdot (P_{\mathrm{BF}} - \beta), & \text{if } (P_{\mathrm{BF}} > \beta) \\ \rho_n \cdot P_{\mathrm{BN}}, & \text{otherwise} \end{cases} \tag{32}$$

$$\alpha = \frac{X \cdot Y \cdot Z}{D \cdot E \cdot F} \cdot 100. \tag{33}$$

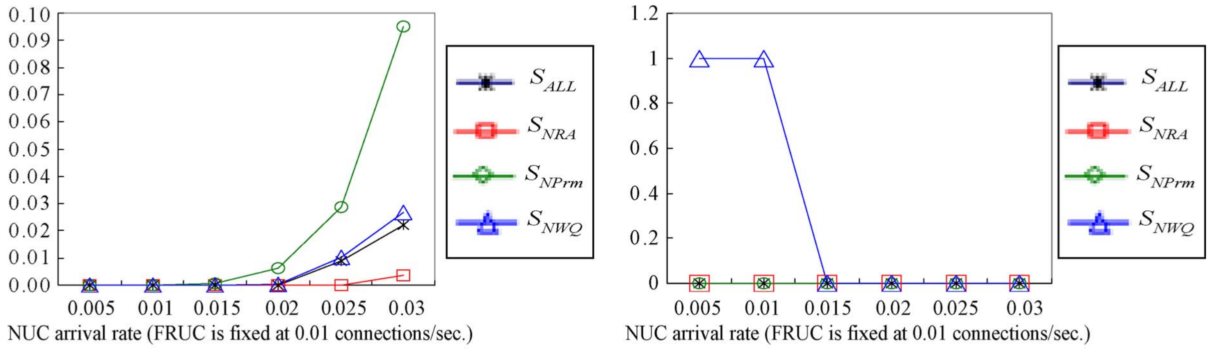
When the cost weighting factor of FRUCs is less than that of NUCs $(\alpha < \rho_n)$, the scheduler should give priority to NUCs without considering the FRUC blocking probability. On the other hand, when α is larger than ρ_n, the scheduler should give priority to NUCs when the FRUC blocking probability is below the departure threshold (β), but it should give priority to FRUCs when the FRUC blocking probability is above β. Note that β should be chosen to reflect the beginning of user dissatisfaction as the blocking probability increases; its proper value may be obtained from the past operation data.

D. Iterative Algorithm

To minimize the cost function, an iterative algorithm, as shown in Fig. 5, was developed to obtain the stationary state probabilities, optimum number of GCs, and performance measures. The iterative algorithm first initializes system input parameters, such as the power limit in a cell, the maximum numbers of serving NUCs and FRUCs in a cell, etc., in Steps 1–3. The for loop in Step 4 determines the optimum number of GCs. The while loop in Step 5 determines the iterations that obtain the stationary probabilities of existing states. In Steps 10 and 11, based on the stationary state probabilities, we can obtain the performance measures and the cost function. In Step 12, we obtain the minimum value of the cost function.

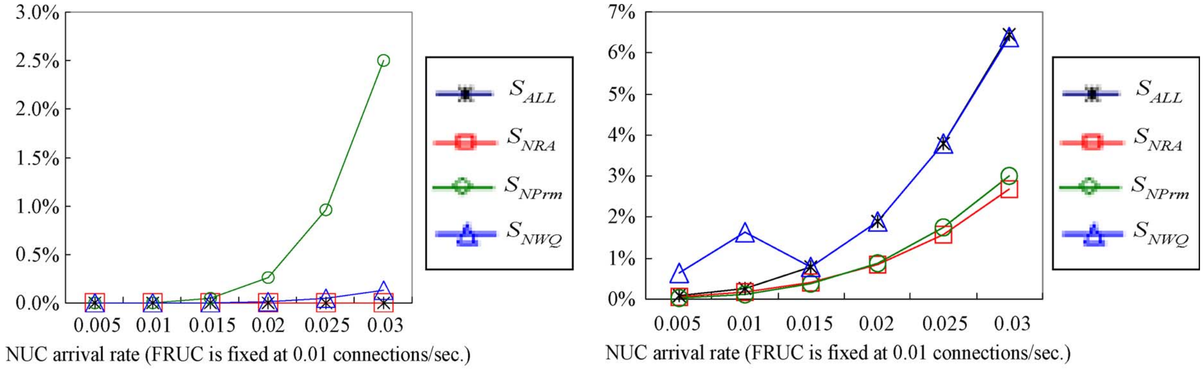
IV. Numerical Results and Discussions

In the following analysis, we assume that the spread spectrum bandwidth (W) of the WCDMA network is 5 MHz, the uplink full-rate transmission is 128 kb/s, and the half rate is



(a) (b)

Fig. 6. Cost function and the numbers of GCs with $\alpha = 0.504$ and $\beta = 0.02$ $(B = 4, Q = 10)$. (a) Cost function values. (b) Optimal numbers of GCs.



(a) (b)

Fig. 7. Average blocking probabilities $(B = 4, Q = 10)$. (a) NUC blocking probabilities (P_{BN}). (b) FRUC blocking probabilities (P_{BF}).

64 kb/s. The size of the NUC WQ (B) is 4, and the size of FRUC WQ (Q) is 10. The mean service time of NUCs $(1/\mu_n)$ is assumed to be 2 min, and the mean service time of FRUCs $(1/\mu_f)$ is assumed to be 10 min. The arrival rate of FRUCs (λ_f) is fixed at 0.01 connections/s, and that of NUC (λ_n) varies in the range of 0.005–0.03 connections/s, i.e., ρ_n varies in the range of 0.6–3.6 connections. We compare four connection schedulers: S_{All}, S_{NRA}, S_{NPrm}, and S_{NWQ}. The iterative algorithm for each scheduler has been developed in C language. The program was run on a laptop PC with a 1.6-GHz Pentium CPU and 512-MB RAM. For each traffic load, the stationary state probabilities can converge in less than 1 min.

To choose a suitable α value for the cost function in (31), we use the operation data from CHT in Taiwan and make assumptions if operation information is unavailable. In (29), D (the 128-kb/s transmission charge per hour) is NT\$562.5, $E = 60$ (i.e., the number of busy hours per day equals to two), and F (the number of cells) is 1000. The number of flat-rate users X is 200 000, Y is assumed to be 0.001 (i.e., 1 out of 1000 users would quit per month due to a percentage increase of blocking probability above β), and Z (the monthly fee of a flat-rate user) is NT\$850. β should be chosen to reflect the level of user dissatisfaction; it was chosen to be 0.02, which is the target blocking probability for flat-rate subscribers of CHT. Given that, we can obtain the factor $\alpha = 0.504$. Note that α is less than ρ_n (0.6–3.6) in our experiments, i.e., the cost weighting factor of FRUCs is less than that of NUCs.

Fig. 6(a) shows the cost function as NUC traffic increases. The FRUC traffic is fixed at 0.01 connections/s. The cost function represents the revenue loss of the operator; the lesser, the better. The results indicate that S_{NRA} has the least amount of revenue loss among all schedulers. When the NUC traffic is less than 0.015 connections/s, the revenue losses of S_{All}, S_{NWQ}, and S_{NPrm} are as small as that of S_{NRA}, but the losses rise rapidly as the NUC traffic increases above 0.02 connections/s, particularly for S_{NPrm}. This indicates that when the system traffic load is high, WQs and preemption are necessary, but rate adaptation is not. This is because subrated connections are less "bandwidth efficient" and results in system throughput reduction and revenue loss. S_{NPrm} suffers the biggest revenue loss when the NUC traffic load is high. This indicates that preemption is essential in reducing the revenue loss.

Fig. 6(b) shows the optimum number of GCs for each scheduler under different traffic loads. S_{All}, S_{NRA}, and S_{NPrm} do not need any GCs. The results indicate that the NUC WQ plus either preemption or rate adaptation is effective in giving NUCs priority. GCs may reduce the system throughput and thus the revenue. In contrast, S_{NWQ} needs one GC when the NUC traffic load is low because it has no NUC WQ. However, when the traffic load is high, no GC is needed because of the same reason that GCs lead to system throughput reduction and revenue loss.

Fig. 7(a) shows the blocking probabilities of NUCs as the NUC arrival rate increases. All schedulers provide very low blocking probabilities for NUCs, except S_{NPrm}; the NUC

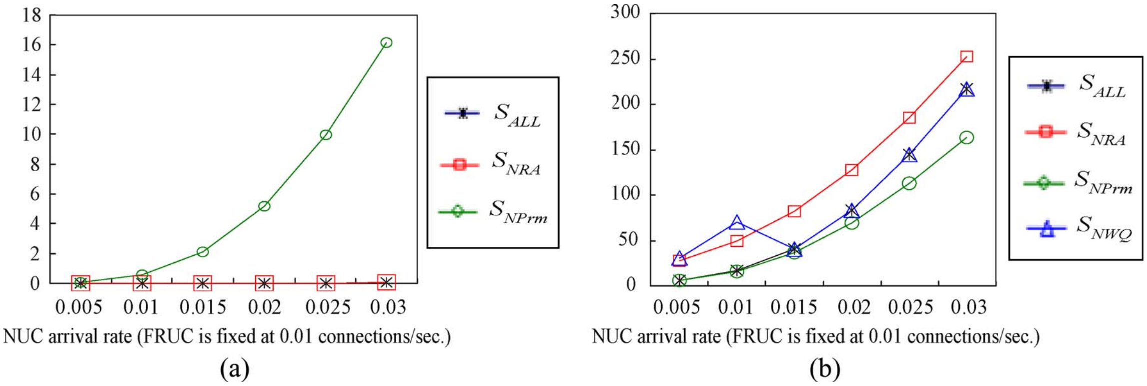


Fig. 8. Average waiting times $(B = 4, Q = 10)$. (a) NUC waiting times $(W_{\rm NT})$. (b) FRUC waiting times $(W_{\rm TF})$.

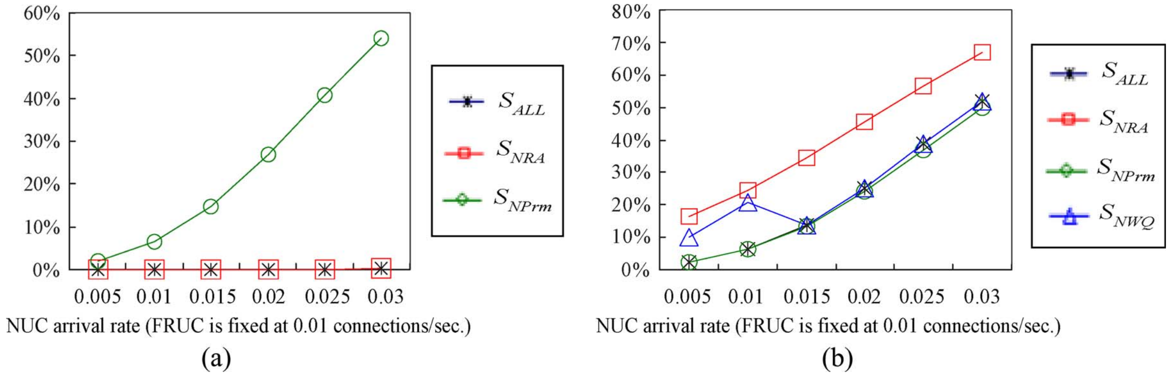


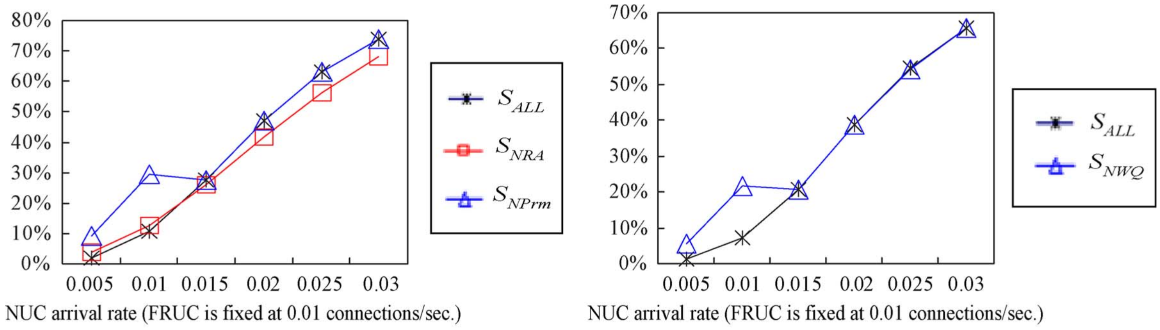
Fig. 9. Average queueing probabilities $(B = 4, Q = 10)$. (a) NUC queueing probabilities $(P_{\rm QN})$. (b) FRUC queueing probabilities $(P_{\rm QF})$.

blocking probability of $S_{\rm NPrm}$ increases more rapidly as NUC traffic increases. This is because subrated FRUCs are less efficient in using a spectrum. If FRUCs can only be subrated, but cannot be preempted, there would be more subrated FRUCs when the system traffic load is high. As a result, the overall system throughput decreases, and more NUCs are blocked. Therefore, preempting FRUCs is essential in reducing the blocking probability of NUCs. When the NUC traffic is high and no NUC WQ is used (as in $S_{\rm NWQ}$), the blocking probability slightly rises. This indicates that the NUC WQ is necessary when the system load is close to its capacity.

Fig. 7(b) shows the blocking probabilities of FRUCs as the NUC arrival rate increases. The results indicate that the FRUC blocking probabilities of $S_{\rm NRA}$ and $S_{\rm NPrm}$ are about the same; $S_{\rm NRA}$ outperforms $S_{\rm NPrm}$ by a small margin. Even though FRUCs cannot be preempted in $S_{\rm NPrm}$, the blocking probability of $S_{\rm NPrm}$ is still higher than that of $S_{\rm NRA}$. This is also because subrated FRUCs are less "bandwidth efficient." In addition, the FRUC blocking probabilities of $S_{\rm All}$ and $S_{\rm NWQ}$ are higher and rise more rapidly as NUC traffic increases, because FRUCs are impaired by both preemption and subrating. The fluctuations of FRUC blocking probabilities in $S_{\rm NWQ}$, when the NUC traffic increases from 0.01 to 0.015 connections/s, are caused by the change in the number of GCs.

Fig. 8(a) shows the average waiting times (i.e., queueing times) of NUCs as the NUC traffic increases. The waiting times of NUCs in schedulers $S_{\rm All}$ and $S_{\rm NRA}$ are very insignificant under all traffic loads, i.e., NUCs are rarely queued. This is because serving FRUCs can be preempted to free radio resources. If FRUCs cannot be preempted, such as in $S_{\rm NPrm}$, the average waiting time of NUCs increases steadily as the traffic of NUCs increases. Fig. 8(b) shows the average waiting times of FRUCs as the NUC traffic increases. The waiting times of all schedulers show the same trend of rising as the NUC traffic increases. Even when the system traffic is low, the average waiting time of FRUCs in $S_{\rm NRA}$ is as large as 60 s, which is unacceptable for real-time applications. $S_{\rm NPrm}$ provides the shortest waiting time, while $S_{\rm NRA}$ the longest. The difference can be as high as 100 s when the NUC traffic is 0.03 connections/s. Note that the fluctuations of FRUC waiting times in $S_{\rm NWQ}$, when the NUC traffic increases from 0.01 to 0.015 connections/s, are also caused by the change in the number of GCs. This change of GCs also results in fluctuations of $S_{\rm NWQ}$ results in later figures.

Fig. 9(a) shows the probability that a NUC is queued. The results indicate that NUCs in $S_{\rm All}$ and $S_{\rm NRA}$ are very rarely put into the WQ because FRUCs can be preempted to free radio resources. On the other hand, NUCs in $S_{\rm NPrm}$ are more likely to be queued. The probability that a new NUC is queued increases steadily and rapidly as the traffic load increases. The NUC queueing probability in $S_{\rm NPrm}$ can be as high as 50%. This indicates that preempting FRUCs is critical in reducing the queueing probability of NUCs. Fig. 9(b) shows the queueing probabilities of FRUCs as the NUC traffic increases. In general, the probability that a FRUC is queued increases as the traffic load increases. $S_{\rm NRA}$ has the largest FRUC queueing probability, because FRUCs cannot be subrated. Other schedulers provide about the same queueing probabilities under all traffic loads.



(a) (b)

Fig. 10. Average preempted probabilities of serving full- and half-rate FRUCs $(B = 4, Q = 10)$. (a) Full rate (P_{FPrm}). (b) Half rate (P_{SPrm}).

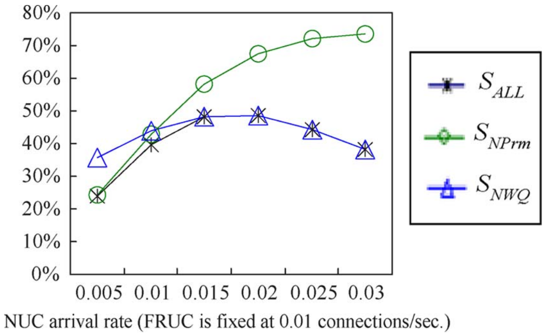


Fig. 11. Average subrated probabilities of serving full-rate FRUCs (P_{FS}) $(B = 4, Q = 10)$.

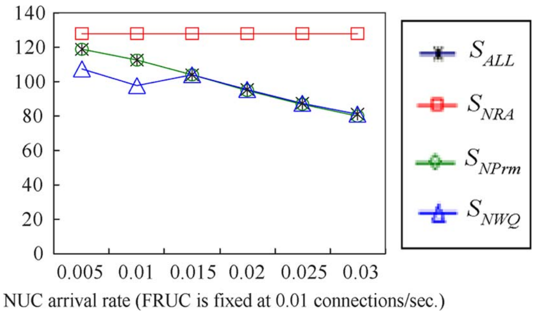


Fig. 12. Average transmission rate of serving FRUCs (T_F) $(B=4, Q=10)$.

Fig. 10(a) and (b) shows the probabilities that a serving full- and half-rate FRUC would be preempted before completion. All schedulers display the same trend of rising preemption probabilities as the traffic load increases. The preemption probability of S_{NRA} is lower than other schemes by a small margin. This is because subrated FRUCs are less bandwidth efficient.

Fig. 11 shows the probabilities that a serving full-rate FRUC would be subrated. As the NUC traffic increases, the subrating probabilities of S_{All} and S_{NWQ} first rise and then decline. The decline is because, when the system traffic is high, FRUCs are more likely to be preempted. On the other hand, the subrating probability of S_{NPrm} increases more rapidly and saturates later as the traffic load increases. This is because as the NUC traffic increases, S_{NPrm} cannot preempt FRUCs; it can only subrate more FRUCs.

Fig. 12 shows the average transmission rate of FRUCs. Since FRUCs may be subrated and/or preempted, the average transmission rate of FRUCs is reduced. In S_{NRA}, no FRUCs are subrated. In S_{All} and S_{NWQ}, a FRUC can be subrated and preempted; the average transmission rate is reduced to as much as 70% of the full-rate transmission when the system traffic is heavy.

V. Conclusion

In this paper, we have investigated four combinations of scheduling techniques, i.e., queueing, GCs, preemption, and rate adaptation, on their effectiveness in scheduling UMTS R99 uplink connections to reduce the revenue loss of the operators serving both normal and flat-rate users. We proposed a cost function representing the revenue loss due to both blocked NUCs and lost flat-rate users. The optimum numbers of GCs was determined by an iterative algorithm. The analytic results indicate that, when α, which is the cost weighting factor of flat-rate users, is less than ρ_n, queueing and preemption are essential for connection scheduling to maximize the revenue. Rate adaptation is ineffective, because half-rate connections are less bandwidth efficient. Subrating FRUCs reduced the system throughput and the operator revenue. In addition, no GC is needed if queueing and preemption are used because GCs increase the blocking probability of FRUCs and reduces system throughput.

In this paper, we consider uplink connection scheduling only. We did not study downlink traffic scheduling, which can be done on the packet level. In this paper, the cost weighting factor of flat-rate users (α) is less than that of normal users (ρ_n). Further study is needed for UMTS networks with α's larger than ρ_n, which is possible when the number of flat-rate users increases or the normal user traffic decreases. In this situation, a more sophisticated scheduler is needed. The scheduler should give priority to NUCs when the FRUC blocking probability is below the departure threshold β. When FRUC blocking probability is above the threshold, FRUCs should have priority.

References

[1] 3GPP TS 23.002, *Network Architecture*. Ver. 3.6, Rel. 1999.

[2] Y.-B. Lin, S. Mohan, and A. Noerpel, "Queueing priority channel assignment strategies for PCS hand-off and initial access," *IEEE Trans. Veh. Technol.*, vol. 43, no. 3, pp. 704–712, Aug. 1994.

[3] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans. Veh. Technol.*, vol. VT-35, no. 3, pp. 77–92, Aug. 1986.

[4] R. Guerin, “Queuing-blocking system with two arrival streams and guard channels,” *IEEE Trans. Commun.*, vol. 36, no. 2, pp. 153–163, Feb. 1998.

[5] Q.-A. Zeng, K. Mukumoto, and A. Fukuda, “Performance analysis of mobile cellular radio system with priority reservation handoff procedures,” in *Proc. IEEE VTC*, Jun. 1994, vol. 3, pp. 1829–1833.

[6] Q.-A. Zeng and D. P. Agrawal, “Performance analysis of a handoff scheduler in integrated voice/data wireless networks,” in *Proc. IEEE VTC*, Sep. 2000, pp. 1986–1992.

[7] C. W. Leong, W. Zhuang, Y. Cheng, and L. Wang, “Call admission control for integrated on/off voice and best-effort data services in mobile cellular communications,” *IEEE Trans. Commun.*, vol. 52, no. 5, pp. 778–790, May 2004.

[8] J. Wang, Q.-A. Zeng, and D. P. Agrawal, “Performance analysis of a preemptive and priority reservation handoff scheme for integrated service-based wireless mobile networks,” *IEEE Trans. Mobile Comput.*, vol. 2, no. 1, pp. 65–75, Jan.–Mar. 2003.

[9] M. S. Do, Y. Park, and J. Y. Lee, “Channel assignment with QoS guarantees for a multiclass multicode CDMA system,” *IEEE Trans. Veh. Technol.*, vol. 51, no. 5, pp. 935–948, Sep. 2002.

[10] S. Kim and P. K. Varshney, “An integrated adaptive bandwidth-management framework for QoS-sensitive multimedia cellular networks,” *IEEE Trans. Veh. Technol.*, vol. 53, no. 3, pp. 847–864, May 2004.

[11] Y.-B. Lin, A. Noerpel, and D. Harasty, “The sub-rating channel assignment strategy for PCS hand-offs,” *IEEE Trans. Veh. Technol.*, vol. 45, no. 1, pp. 122–130, Feb. 1996.

[12] W.-Y. Chen, J.-L. C. Wu, and L. L. Lu, “Performance comparisons of dynamic resource allocation with/without channel de-allocation in GSM/GPRS networks,” *IEEE Commun. Lett.*, vol. 7, no. 1, pp. 10–12, Jan. 2003.

[13] 3GPP TS 25.413, *UTRAN Iu Interface RANAP Signaling*. Ver. 3.14, Rel. 1999.

[14] 3GPP TS 25.331, *Radio Resource Control (RRC) Protocol Specification*. Ver. 3.21, Rel. 1999.

[15] 3GPP TS 23.060, *General Packet Radio Service (GPRS) Service Description; Stage 2*. Ver. 3.a.0, Rel. 1999.

[16] D. Niyato and E. Hossain, “Call-level and packet-level quality of service and user utility in rate-adaptive cellular CDMA networks: A queuing analysis,” *IEEE Trans. Mobile Comput.*, vol. 5, no. 12, pp. 1749–1763, Dec. 2006.

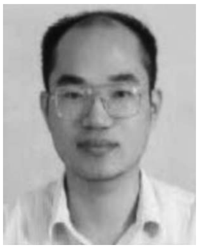
[17] L. Xu, X. Shen, and J. W. Mark, “Dynamic fair scheduling with QoS constraints in multimedia wideband CDMA cellular networks,” *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 60–73, Jan. 2004.

[18] J. Laiho and A. Wacker, “Radio network planning process and methods for WCDMA,” *Ann. Telecommun.*, vol. 56, no. 5/6, pp. 317–331, May/Jun. 2001.

[19] 3GPP TS 25.104, *BS Radio Transmission and Reception (FDD)*. Ver. 3.13, Rel. 1999.

[20] R. W. Wolff, “Poisson arrivals see time averages,” *Oper. Res.*, vol. 30, no. 2, pp. 223–231, Mar./Apr. 1982.

[21] Y.-B. Lin, “Performance modeling for mobile telephone networks,” *IEEE Netw.*, vol. 11, no. 6, pp. 63–68, Nov./Dec. 1997.



Chung-Yung Chia received the B.S. and M.S. degrees in computer engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1987 and 1989, respectively, where he is currently working toward the Ph.D. degree in computer science.

Since 1989, he has been with the Telecommunication Laboratory, Chunghwa Telecom, Taoyuan, Taiwan, where he is currently a Researcher and a Project Manager. His research interests include the design and analysis of wireless communications networks, the development of telecommunication scheduling and monitoring systems, and performance modeling.

Ming-Feng Chang received the B.S. and M.S. degrees in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1982 and 1984, respectively, and the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign in 1991.

From 2003 to 2005, he was the Chairman of the Department of Computer Science, National Chiao Tung University (NCTU), Hsinchu, Taiwan, where he is currently a Professor. His current research interests include the design and analysis of Internet communications, personal communications networks, mobile payment, and performance modeling.