Performance Analysis of a Service-dependant Handoff Scheme in Voice/Data Integrated Cellular Mobile Systems

Bo LI[1], Roberto Battiti[2], *Member, IEEE*, and Akira FUKUDA[3], *Member, IEEE*

[1]ISN National Key Lab, Xidian University, Xi'an, China, 710071
Email: luse1998@yahoo.com

[2]Department of Computer Science and Telecommunications
University of Trento, 38050 POVO, Trento, Italy
Email: battiti@dit.unitn.it

[3]Department of Electrical Engineering, Shizuoka University
Johoku 3-5-1, Hamamatsu, 432 Japan
Email: teafuku@ipc.shizuoka.ac.jp

Correspondence Author and Address:
Dr. LI Bo
P.O.Box 102, Xidian University, Xi'an, China, 710071
Email: luse1998@yahoo.com

Abstract

In this paper, we propose and analyze a service-dependent handoff and channel allocation scheme in voice and data integrated cellular mobile systems, which combines the ideas of "Variable Bandwidth" and "Preemptive Priority" together. In the scheme, voice and data traffic are considered. According to the variations of the offered traffic intensity at each cell, both a voice and a data call in service can occupy a full-rate channel or a half-rate channel. In order to guarantee the Quality of Service (QoS) for both voice and data traffic, channel resources are fairly shared between voice and data calls according to an optimal channel allocation scheme, which minimizes the difference between the average bandwidth of a voice call in service and that of a data call in service. To minimize the forced termination of a voice call, a voice call can preempt a data call in service if all the calls in the channel pool of the current cell are already assigned with half-rate service. The interrupted data call returns back to the queue specially prepared for data traffic. By analysis, we obtain the most important system performance measures. Comparisons with the scheme, which only supports "Preemptive Priority" without "Variable Bandwidth" supporting, shows that if the total arrival rate for originating calls is not very heavy, the new scheme can provide lower blocking probability and forced termination probability for both voice and data traffic, and shorter average total transmission time for a successfully completed data call.

Index Terms： Handoff scheme, variable bandwidth, preemptive priority, cellular mobile systems

I. INTRODUCTION

One of the central issues in performance characterization of cellular mobile and personal communication systems (PCS) is the problem of handoff [1]. With the penetration of PCS, microcell and hybrid cell (macro-, micro-, pico-) structures are exploited to support the drastically increased demand [2]. The smaller cell size and the variable propagation conditions in microcells cause much more frequent handoffs [3]. A poorly designed handoff strategy will generate very heavy signaling traffic and worse QoS. Handoff process usually starts from the handoff initiation phase [4]-[6]. The focus of this paper is on the handoff execution phase, and it is assumed that the handoff request detection and initiation procedures are perfect (i.e. all valid requests are detected and no invalid requests activate the handoff procedure).

Forced termination of ongoing voice calls is more undesirable than blocking of originating calls from the user's viewpoint. Therefore, several schemes giving priority to handoff requests have been proposed to reduce the forced termination probability. Traffic models and performance measures of systems adopting some priority schemes are discussed in [7]-[17]. The most popular way of giving priority to handoff requests is to reserve a fixed number of channels for them. That is, when the number of free channels in a cell is less than or equal to a predefined value, originating calls are blocked. The handoff calls can still gain access to the system until there are no available channels. This scheme is called "priority reservation scheme" or "cut-off priority scheme" [7]-[15]. Moreover, in order to ensure the handoff performance in microcell, the sub-rating scheme is presented in [16][17]: when a handoff call enters a busy cell (no channel available), an occupied full-rate channel can be divided into two half-rate channels to accommodate the arriving handoff call. However, in all of the above studies, only voice traffic is considered.

Future cellular mobile systems are required to accommodate multiple types of services, such as voice, data, and video. In order to meet the future demands, the handoff strategy needs to take different features of these services into account, i.e., the ideal handoff process is service-dependent. Recently, a lot of channel allocation and handoff schemes have been proposed based on the idea of "Variable Bandwidth", such as those proposed in [18]-[23]. The idea is that when a real-time user arrives at or handoff to a congested cell, the channels allocated to the real-time users already in the cell may need to be reduced to satisfy the minimum channel requirement of the arriving user. Moreover, in [24], authors propose a "Preemptive Priority" handoff scheme for integrated voice/data cellular mobile systems. In the scheme, the right to preempt the service of a data call in service is given to a handoff voice call if on arrival it finds no idle channels. Ideas of "Preemptive Priority" and "Priority Reservation" are further combined in [25]-[26], where some number of channels is exclusively reserved for handoff calls. However, in the scheme proposed in [24]-[26], "Variable Bandwidth" supporting is not considered. Moreover, fairness on channel allocation between voice and data traffic is also not considered. In this paper, based on our previous work proposed in [24], we propose a new service-dependent handoff scheme in voice and data integrated cellular systems, which combines the idea of "Variable Bandwidth" and "Preemptive Priority" together. Contributions

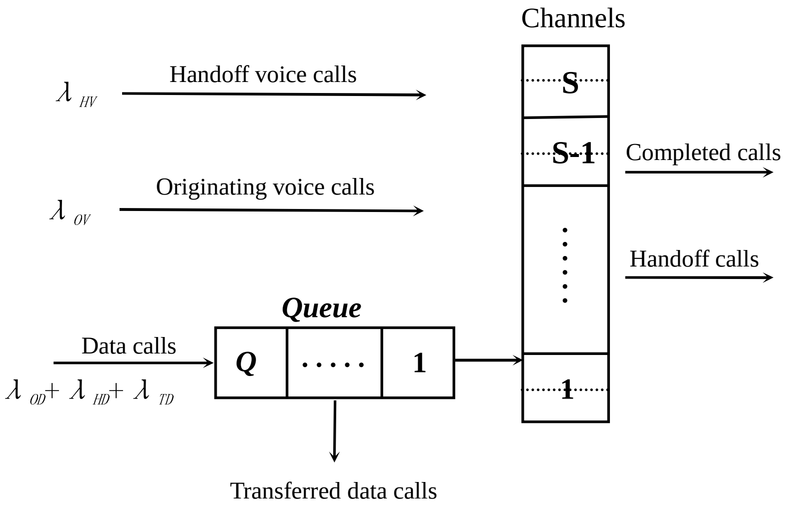


Fig. 1 System model

of the paper can be summarized as:

1. The proposed scheme supports not only the “Variable Bandwidth” for mobile users but also the “Preemptive Priority” for voice calls, which further decrease the blocking probability for voice calls.
2. In the scheme, in order to guarantee the QoS for both voice and data traffic, channel resources are fairly shared between voice and data calls based on an optimal channel allocation scheme, which guarantees that service for data traffic will not be starved by voice traffic.
3. Thorough performance analysis on the proposed scheme are given, which helps one to obtain deeper insight into the behavior of the whole systems.

The remainder of this paper is organized as follows: Section II introduces the proposed handoff scheme. In Section III, traffic model is described. Optimal channel allocation scheme is proposed in Section IV. System performance is analyzed in Section V. In Section VI, numerical results are presented and discussed.

II. DISCRIPTION OF THE HANDOFF SCHEME

We consider a system with many homogeneous cells each having S full-rate channels (see Fig. 1). Each full-rate channel can be further divided into two independent half-rate channels.

Moreover, in the base station of each cell, there is a queue with buffer size Q for data calls. In our analysis, we focus our attention on a single cell, which is called the marked cell. Moreover, we define N_V and N_D as the current number of voice and data calls in service at the marked cell, respectively. And N_Q is defined as the current number of data calls waiting in the queue. Note that in the paper we are interested in studying the performance of the combination of ideas of “Variable Bandwidth” and “Preemptive Priority”. Therefore, for simplicity, originating calls and handoff calls are treated in exactly the same way, which helps one to obtain deeper insight into the proposed basic idea. Moreover, the basic idea here can be easily extended to be adopted in cases where schemes, such as “Priority Reservation Scheme”, assigning priority to handoff calls over originating calls are required.

An arriving voice call (an originating voice call or a handoff voice call) gets a full-rate channel if $N_V + N_D < S$. Under the condition of $S \le N_V + N_D < 2S$, it gets a half-rate or a full-rate channel according to the channel allocation scheme proposed in section IV. Furthermore, if $N_V + N_D = 2S$ and $N_D > 0$ and $N_Q < Q$, it gets a half-rate channel by preempting a data call in service. The preempted data call returns back to the queue, and waits for a channel to be available. Otherwise, the arriving voice call is blocked by the system.

A data call waiting in the queue is transferred to the target cell when the mobile user moves out of the current cell before it gets a channel (see Fig. 1). An arriving data call (an originating data call or a handoff data call or a transferred data call) gets a full-rate channel if $N_V + N_D < S$. If $S \le N_V + N_D < 2S$, it gets a half-rate channel or a full-rate channel based on the channel allocation scheme explained in section IV. Furthermore, if $N_V + N_D = 2S$ and $N_Q < Q$, it waits in the queue. Otherwise, the arriving data call is blocked by the system.

III. Traffic Model

Let random variable T_{CV} be the call holding time of a voice call, i.e., the time a voice call lasts if it is not forced into termination. The call holding time T_{CV} of a voice call is assumed to have an exponential distribution with mean $E[T_{CV}](=1/\mu_{CV})$.

Let random variable L_D denote data length of a data call, which is measured in bits. It is assumed that L_D has an exponential distribution (Note that although L_D is a discrete random variable, for simplicity, its distribution is approximated as a continuous one). Let $R_{D,Full}$ and $R_{D,Half}$ (measured in bits per second) denote the data transmission rate with a full-rate and a half-rate channel, respectively. Moreover, it is assumed that $R_{D,Full} = 2R_{D,Half}$. Random variable $T_{CD,Full} = L_D \big/ R_{D,Full}$ and $T_{CD,Half} = L_D / R_{D,Half}$ denote the transmission time needed for a data call served by a full-rate channel and a half-rate channel, respectively. Obviously, both $T_{CD,Full}$ and $T_{CD,Half}$ have exponential distributions with means $E[T_{CD,Full}](=1/\mu_{CD,Full})$ and $E[T_{CD,Half}](=1/\mu_{CD,Half})$, respectively. And, we have

$$E[T_{CD,Half}] = 2E[T_{CD,Full}] \tag{1}$$

Let the random variable T_{dwell} be the dwell time of a mobile user in a cell. Again we assume that it has an exponential distribution with mean $E[T_{dwell}](=1/\mu_{dwell})$.

The arrival processes of originating voice calls and originating data calls in a cell are assumed to be Poisson, with rates λ_{OV} and λ_{OD}, respectively.

Because we assume an equilibrium homogeneous mobility pattern, the arrival rate of handoff requests at the marked cell is equal to the departure rate of handoff calls from the cell. Let λ_{HV} denote the arrival rate of handoff voice requests, and λ_{HD} denote the arrival rate of handoff data requests, both processes being Poisson. It is apparent from the above discussions that

$$\lambda_{HV} = E_V \cdot \mu_{dwell} \tag{2}$$

where E_V is the average number of voice calls holding channels at the marked cell, and

$$\lambda_{HD} = E_D \cdot \mu_{dwell} \tag{3}$$

where E_D is the average number of data calls holding channels at the marked cell.

A data channel request in the queue of a cell is transferred to the target cell when the user moves out of the cell before its getting a channel. The arrival rate λ_{TD} of the Poisson process regulating the transferred data requests at the marked cell is given by

$$\lambda_{TD} = E_Q \cdot \mu_{dwell} \tag{4}$$

where E_Q is the average number of data calls waiting in the queue at the marked cell.

IV. Channel Allocation Scheme

First, we define the state s of the marked cell as a couple of non-negative integers $s:(N_V, N_{TD})$, where N_V denotes that the number of voice calls in service, N_{TD} is the sum of the number of data calls N_D in service and the number of data calls N_Q waiting in the queue. Since N_V, N_{TD}, N_D and N_Q are determined by state s, for clarity, in the rest part of the paper these parameters are denoted as $N_V(s)$, $N_{TD}(s)$, $N_D(s)$ and $N_Q(s)$, respectively. The states form a two-dimensional sample space Ω_s

$$\begin{aligned}\Omega_s \equiv \{s \mid 0 \le N_V(s) \le N_V(s) + N_{TD}(s) \le 2S \cup \\ 0 \le N_V(s) \le 2S < N_V(s) + N_{TD}(s) \le 2S + Q\}\end{aligned} \tag{5}$$

we have

$$N_D(s) = \begin{cases} N_{TD}(\mathrm{s}) & \text{if } N_V(s) + N_{TD}(s) \le 2S \\ 2S - N_V(\mathrm{s}) & \text{if } N_V(s) + N_{TD}(s) > 2S \end{cases} \tag{6}$$

Moreover, $C_V(s)$ and $C_D(s)$ are defined as the total number of equivalent half-rate channels occupied by voice calls and data calls at state s, respectively.

As we know, both voice and data calls in service can be allocated with full-rate or half-rate channels. Given current system state $s:(N_V(s), N_{TD}(s))$, channel allocation pattern (that is, $C_V(s)$ and $C_D(s)$) should be determined. For a given system state, different channel allocation schemes result in different allocation patterns $C_V(s)$ and $C_D(s)$. In order to find out what kind of channel allocation scheme should be used, channel allocation criterion should be determined first. There are two extreme criteria, which should be definitely avoided. One extreme is that channel resources are allocated to voice traffic as much as possible (try to guarantee that voice calls are allocated with full-rate channels as much as possible), which has the risk of starving data traffic. For the other extreme, channel resources are allocated to data traffic as much as possible (try to guarantee that data calls are allocated with full-rate channels as much as possible), which degrades the service quality for voice traffic. Therefore, fairness in channel allocation between voice and data traffic should be considered. We adopt fairness as our design criterion. In this paper, an optimized channel allocation scheme is proposed, which achieves fair bandwidth allocation between voice and data traffic.

When $N_V(s) + N_{TD}(s) \le S$, all the calls in service occupy full-rate channels. When $N_V(s) + N_{TD}(s) \ge 2S$, all the calls in service occupy half-rate channels. When $S < N_V(s) + N_{TD}(s) < 2S$, all the channel resources are occupied. That is, we have

$$C_V(s) + C_D(s) = 2S \qquad (7)$$

In this case, if $N_V(s) > 0$ and $N_D(s) > 0$, the goal of our channel allocation scheme is to minimize the value $\left|\frac{C_V(s)}{N_V(s)} - \frac{C_D(s)}{N_D(s)}\right|$. That is, channel resources are fairly shared between voice and data calls to minimize the difference between the average bandwidth of a voice call in service and that of a data call in service. To minimize $\left|\frac{C_V(s)}{N_V(s)} - \frac{C_D(s)}{N_D(s)}\right|$, it can be easily found that we have to minimize $\left|C_V(s) - \frac{2S \cdot N_V(s)}{[N_D(s) + N_V(s)]}\right|$. Therefore, the optimum solution can be given as

$$\begin{cases} C_V^{*}(s) = \left\lfloor \dfrac{2S \cdot N_V(s)}{N_D(s) + N_V(s)} + \dfrac{1}{2} \right\rfloor \\ C_D^{*}(s) = 2S - C_V^{*}(s) \end{cases} \qquad (8)$$

It should be pointed out that although equation (8) is derived under the condition $S < N_V(s) + N_{TD}(s) < 2S$, $N_V(s) > 0$ and $N_D(s) > 0$, it holds for all the cases under the condition $S \le N_V(s) + N_{TD}(s) \le 2S$.

In the following, we find out how to allocate channels when the system state transit from one to another. Note that in our discussions, we assume that both s and s' are permitted states.

Since the discussions on the arrival or the departure of data calls are almost the same as those of voice calls, we only present the cases for voice calls. The discussions are divided into three different cases:

Case 1: $N_V(s)+N_{TD}(s)<S$.

Case 1.1: Arrival of a voice call.

Assuming that system state transits from $s:(N_V(s),N_{TD}(s))$ to $s':(N_V(s)+1,N_{TD}(s))$. The accepted voice call gets a full-rate channel.

Case 1.2: Departure of a voice call.

Assuming that system state transits from $s':(N_V(s)+1,N_{TD}(s))$ to $s:(N_V(s),N_{TD}(s))$. A full-rate channel occupied by the voice call is released.

Case 2. $N_V(s)+N_{TD}(s)\geq 2S$.

Case 2.1: Arrival of a voice call.

Assuming that system state transits from $s:(N_V(s),N_{TD}(s))$ to $s':(N_V(s)+1,N_{TD}(s))$. The accepted voice call gets a half-rate channel by preempting a data call in service.

Case 2.2: Departure of a voice call.

Assuming that the system state transits from $s':(N_V(s)+1,N_{TD}(s))$ to $s:(N_V(s),N_{TD}(s))$. A half-rate channel occupied by the voice call is released and allocated to the data call waiting in the head of the queue.

Case 3. $S\leq N_V(s)+N_{TD}(s)<2S$.

Case 3.1: Arrival of a voice call.

Assuming that system state transits from $s:(N_V(s),N_{TD}(s))$ to $s':(N_V(s)+1,N_{TD}(s))$. From equation (8), we have

$$C_V{}^*(s')=\left\lfloor \frac{2S\cdot[N_V(s)+1]}{N_V(s)+1+N_D(s)}+\frac{1}{2}\right\rfloor \tag{9}$$

It can be easily proved that in this case $0\leq C_V{}^*(s')-C_V{}^*(s)\leq 2$. Therefore, $C_V{}^*(s')-C_V{}^*(s)$ can only be equal to 0, 1, or 2. Furthermore, we have the following three different cases:

Case 3.1.1: $S \le N_V(s) + N_{TD}(s) < 2S$ *and* $C_V^*(s') - C_V^*(s) = 0$.

In this case, it can be easily proved that $C_V^*(s) - N_V(s) \ge 1$. Therefore, there is *at least* one voice call occupying a full-rate channel at state s. A full-rate channel occupied by some voice call can be divided into two half-rate channels, and the newly arrived voice call is accommodated by one of these half-rate channels.

Case 3.1.2: $S \le N_V(s) + N_{TD}(s) < 2S$ *and* $C_V^*(s') - C_V^*(s) = 1$.

In this case, it can be easily proved that $C_D^*(s) - N_D(s) \ge 1$. Therefore, there is *at least* one data call occupying a full-rate channel at state s. A full-rate channel occupied by some data call can be divided into two half-rate channels, and the newly arrived voice call is accommodated by one of these half-rate channels.

Case 3.1.3: $S \le N_V(s) + N_{TD}(s) < 2S$ *and* $C_V^*(s') - C_V^*(s) = 2$.

In this case, it can be easily proved that $C_D^*(s) - N_D(s) \ge 2$. Therefore, there are *at least* two data calls occupying full-rate channels at state s. Two corresponding full-rate channels occupied by some two data calls can be divided into four half-rate channels, and two of these half-rate channels combine together to form a full-rate channel to accommodate the newly arrived voice call.

Case 3.2: Departure of a voice call.

We assume that the system state transits from $s':(N_V(s)+1, N_{TD}(s))$ to $s:(N_V(s), N_{TD}(s))$.

Case 3.2.1: When $C_V^*(s') - C_V^*(s) = 0$.

From the discussion presented in Case 3.1.1, it can be seen that there are at least two voice calls occupying half-rate channels at state s'. Therefore, if a voice call occupying a full-rate channel departs from the marked cell, the system divides the released full-rate channel into two half-rate channels and merges these two channels into some two voice calls occupying half-rate channels, respectively. Then these two voice call will be provided with full-rate channels. On the other hand, if a voice call occupying a half-rate channel departs, the system merges the released half-rate channel into some voice call occupying a half-rate channel. Then the voice call is provided with full-rate service.

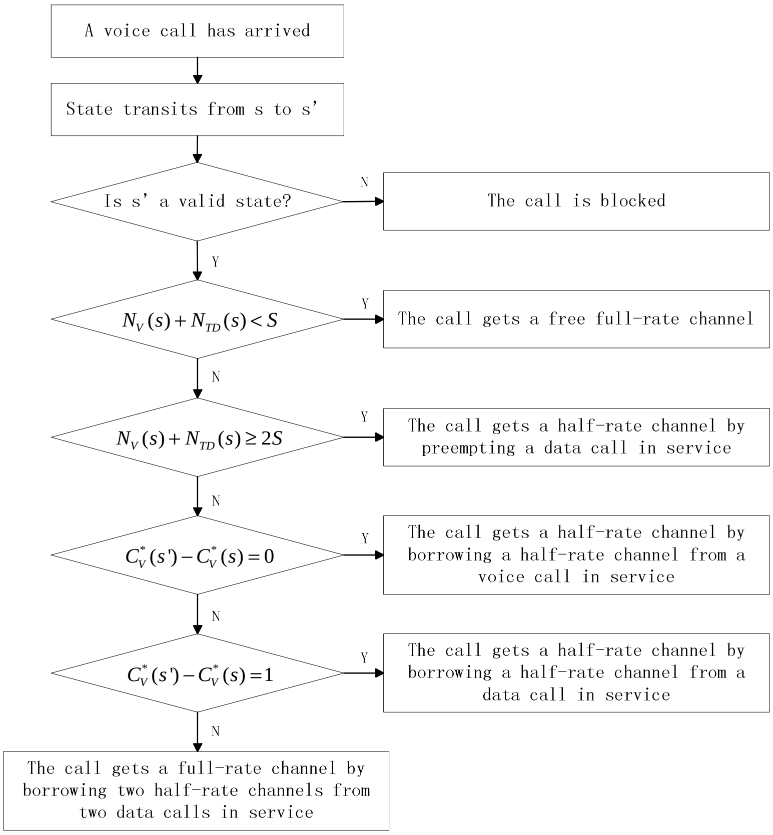

A voice call has arrived
State transits from s to s'
Is s' a valid state?
N
The call is blocked
Y
$N_V(s)+N_{TD}(s)<S$
Y
The call gets a free full-rate channel
N
$N_V(s)+N_{TD}(s)\ge 2S$
Y
The call gets a half-rate channel by preempting a data call in service
N
$C_V^*(s')-C_V^*(s)=0$
Y
The call gets a half-rate channel by borrowing a half-rate channel from a voice call in service
N
$C_V^*(s')-C_V^*(s)=1$
Y
The call gets a half-rate channel by borrowing a half-rate channel from a data call in service
N
The call gets a full-rate channel by borrowing two half-rate channels from two data calls in service

Fig. 2 Channel allocation strategy for an arriving voice call

Case 3.2.2: When $C_V{}^*(s')-C_V{}^*(s)=1$.

Based on the discussion in Case 3.1.2, it can be seen that there are at least one voice call and one data call occupying half-rate channels at state s'. Therefore, if a voice call occupying a full-rate channel departs from the cell, the system divides the released full-rate channel into two half-rate channels and merges these two channels into some voice call and some data call occupying half-rate channels, respectively. Then both the voice call and the data call are provided with full-rate channels. On the other hand, if a voice call occupying a half-rate channel departs, the system merges the released half-rate channel into some data call occupying a half-rate channel. Then the data call is provided with a full-rate channel.

Case 3.2.3: When $C_V{}^*(s')-C_V{}^*(s)=2$.

From the discussion presented in Case 3.1.3, it can be seen that there are at least one voice call occupying a full-rate channel and two data calls occupying half-rate channels at state s'. Therefore, if a voice call occupying a full-rate channel departs from the cell, the system divides the released full-rate channel into two half-rate channels and merges these two channels into some two data calls occupying half-rate channels, respectively. Therefore, these two data calls are provided with full-rate service. On the other hand, if a voice call occupying a half-rate channel departs, the system merges the released half-rate channel into some data call occupying a half-rate channel. Furthermore, the system makes some voice call occupying a full-rate channel to be served by a half-rate channel, and the released half-rate channel is merged into another data call occupying a half-rate channel. Therefore, these two data calls are provided with full-rate service.

For clarity, the flowchart of channel allocation scheme is shown in Fig. 2. Because of the space limitation, only the case that a voice call has arrived is shown. Other cases can be easily derived based on the descriptions given in Section II and this section.

V.PERFORMANCE ANALYSIS

A. The System-State Probabilities

We can obtain the limiting probability of state s, denoted by $p(s)$, by solving the following stationary state-transition equations:

$$\begin{cases} \sum_{s'\in\Omega_s} p(s')\cdot q(s',s) = 0_s \quad (s\in\Omega_s) \\ \sum_{s\in\Omega_s} p(s) = 1 \end{cases} \tag{10}$$

where $q(s',s), s'\neq s$ denotes the transition rate from state s' to s, and $q(s,s)$ is the transition rate out of state s. $q(s,s)$ can be obtained by

$$q(s,s) = -\sum_{s'\in\Omega_s, s'\neq s} q(s,s') \tag{11}$$

Next, we determine the transition rate $q(s',s)$ from state $s':(N_V(s'), N_{TD}(s'))$ to state s.

Case 1) Arrival of A Data Call at the Marked Cell: Let $q_1(s',s)$ be the transition rate from state s' to state $s:(N_V(s'), N_{TD}(s')+1)$. If $0 \le N_V(s') + N_{TD}(s') < 2S$, the call is served immediately and $q_1(s',s) = \lambda_{HD} + \lambda_{TD} + \lambda_{OD}$. If $2S \le N_V(s') + N_{TD}(s') < 2S + Q$, the call is buffered in the queue and $q_1(s',s) = \lambda_{HD} + \lambda_{TD} + \lambda_{OD}$. Otherwise, the call is blocked and $q_1(s',s) = 0$.

Case 2) Arrival of A Voice Call at the Marked Cell: Let $q_2(s',s)$ be the transition rate from state

s' to state $s:(N_V(s')+1, N_{TD}(s'))$. If $0 \le N_V(s') + N_{TD}(s') < 2S$, the call is served immediately and $q_2(s',s) = \lambda_{HV} + \lambda_{OV}$. If $2S \le N_V(s') + N_{TD}(s') < 2S + Q$ and $N_D(s') > 0$, the arriving voice call preempts a data call and $q_2(s',s) = \lambda_{HV} + \lambda_{OV}$. Otherwise, the call is blocked and $q_2(s',s) = 0$.

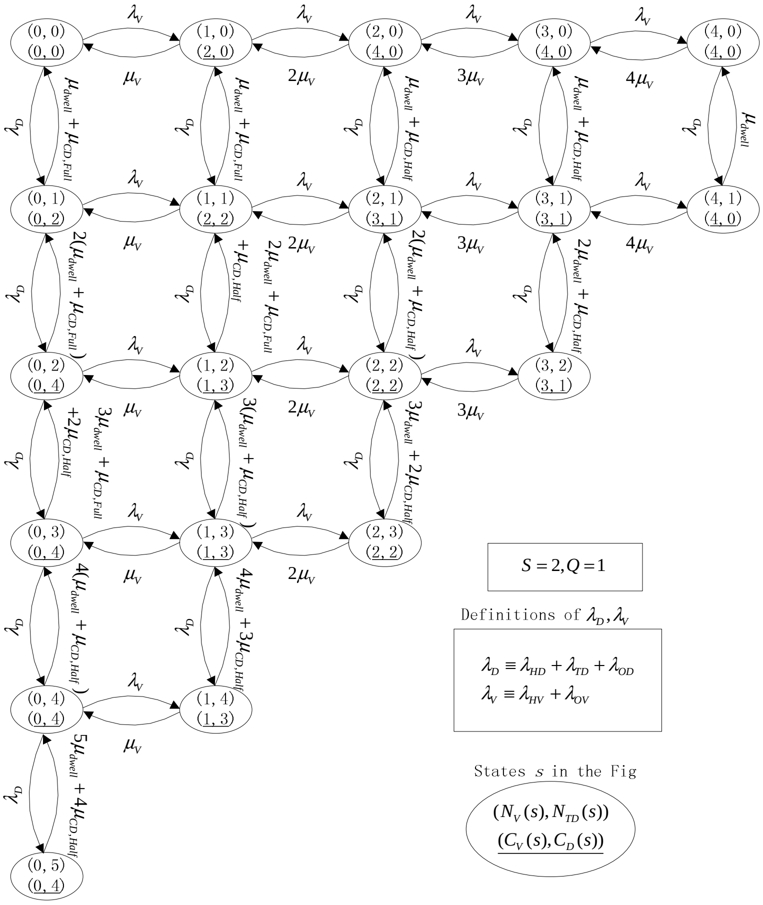
Case 3) Departure of A Voice Call from the Channel Pool: Let $q_3(s',s)$ be the transition rate from state s' to state $s:(N_V(s')-1, N_{TD}(s'))$. If $N_V(s') > 0$, the transition rate is $q_3(s',s) = N_V(s') \cdot (\mu_{dwell} + \mu_{CV})$. Otherwise, $q_3(s',s) = 0$.

Case 4) Departure of A Data Call from the Channel Pool: Let $q_4(s',s)$ be the transition rate from state s' to state $s:(N_V(s'), N_{TD}(s')-1)$. If $N_D(s') > 0$, the transition rate is $q_4(s',s) = [C_D(s') - N_D(s')] \cdot (\mu_{dwell} + \mu_{CD,Full}) + [2N_D(s') - C_D(s')](\mu_{dwell} + \mu_{CD,Half})$. Otherwise, $q_4(s',s) = 0$.

Case 5) Departure of A Waiting Data Call from the Data Queue: Let $q_5(s',s)$ be the transition rate from state s' to state $s:(N_V(s'), N_{TD}(s')-1)$. If $N_Q(s') > 0$, the transition rate is $q_5(s',s) = N_Q(s')\mu_{dwell}$. Otherwise, $q_5(s',s) = 0$.

An simple example Markov state transition diagram for the marked cell is shown in Fig. 3, where $S = 2, Q = 1$. After determining the state transition rates, we can solve the stationary state-transition equations in (10) to obtain all the state probabilities by using numerical method (refer to Appendix A).

B. Performance Measures for Voice calls



(0, 0) (0, 0)
λ_V
(1, 0) (2, 0)
λ_V
(2, 0) (4, 0)
λ_V
(3, 0) (4, 0)
λ_V
(4, 0) (4, 0)
μ_V
$2\mu_V$
$3\mu_V$
$4\mu_V$
λ_D
$\mu_{dwell}+\mu_{CD,Full}$
λ_D
$\mu_{dwell}+\mu_{CD,Full}$
λ_D
$\mu_{dwell}+\mu_{CD,Half}$
λ_D
$\mu_{dwell}+\mu_{CD,Half}$
λ_D
μ_{dwell}
(0, 1) (0, 2)
(1, 1) (2, 2)
(2, 1) (3, 1)
(3, 1) (3, 1)
(4, 1) (4, 0)
$2(\mu_{dwell}+\mu_{CD,Full})$
$2\mu_{dwell}+\mu_{CD,Full}$ $+\mu_{CD,Half}$
$2(\mu_{dwell}+\mu_{CD,Half})$
$2\mu_{dwell}+\mu_{CD,Half}$
(0, 2) (0, 4)
(1, 2) (1, 3)
(2, 2) (2, 2)
(3, 2) (3, 1)
$3\mu_{dwell}+\mu_{CD,Full}$ $+2\mu_{CD,Half}$
$3(\mu_{dwell}+\mu_{CD,Half})$
$3\mu_{dwell}+2\mu_{CD,Half}$
(0, 3) (0, 4)
(1, 3) (1, 3)
(2, 3) (2, 2)
$4(\mu_{dwell}+\mu_{CD,Half})$
$4\mu_{dwell}+3\mu_{CD,Half}$
(0, 4) (0, 4)
(1, 4) (1, 3)
$5\mu_{dwell}+4\mu_{CD,Half}$
(0, 5) (0, 4)
$S=2, Q=1$
Definitions of λ_D, λ_V
$\lambda_D \equiv \lambda_{HD}+\lambda_{TD}+\lambda_{OD}$
$\lambda_V \equiv \lambda_{HV}+\lambda_{OV}$
States s in the Fig
$(N_V(s), N_{TD}(s))$
$(C_V(s), C_D(s))$

Fig. 3 An example Markov state transition diagram for the case of $S=2, Q=1$

Based on the above discussions, blocking probability of a voice call can be given as

$$\begin{cases} P_{B,V} = \sum\limits_{s\in\Omega_{B,V}} p(s) \\ \Omega_{B,V} \equiv \{s \mid s\in\Omega_s, N_V(s)=2S\} \bigcup \{s \mid s\in\Omega_s, N_V(s)+N_{TD}(s)=2S+Q\} \end{cases} \tag{12}$$

Once a voice call is served, the voice call handoffs only to neighboring cells with probability $P_{H,V}$, which is given by

$$P_{H,V} \equiv \Pr ob\{T_{dwell} < T_{CV}\} = \mu_{dwell} / (\mu_{dwell}+\mu_{CV}) \tag{13}$$

The forced termination probability of a voice calls, denoted by $P_{FT,V}$, can be obtained by

$$P_{FT,V} = P_{H,V}\cdot[P_{B,V}+(1-P_{B,V})\cdot P_{FT,V}] = (P_{H,V}\cdot P_{B,V})/[1-P_{H,V}\cdot(1-P_{B,V})] \tag{14}$$

Moreover, let c_V denote the average bandwidth per voice call in service. We have

$$c_V \equiv \left(\sum_{s\in\{s \mid N_V(s)>0\}} \frac{C_V(s)}{N_V(s)} \cdot p(s) \right) \Bigg/ \left(\sum_{s\in\{s \mid N_V(s)>0\}} p(s) \right) \tag{15}$$

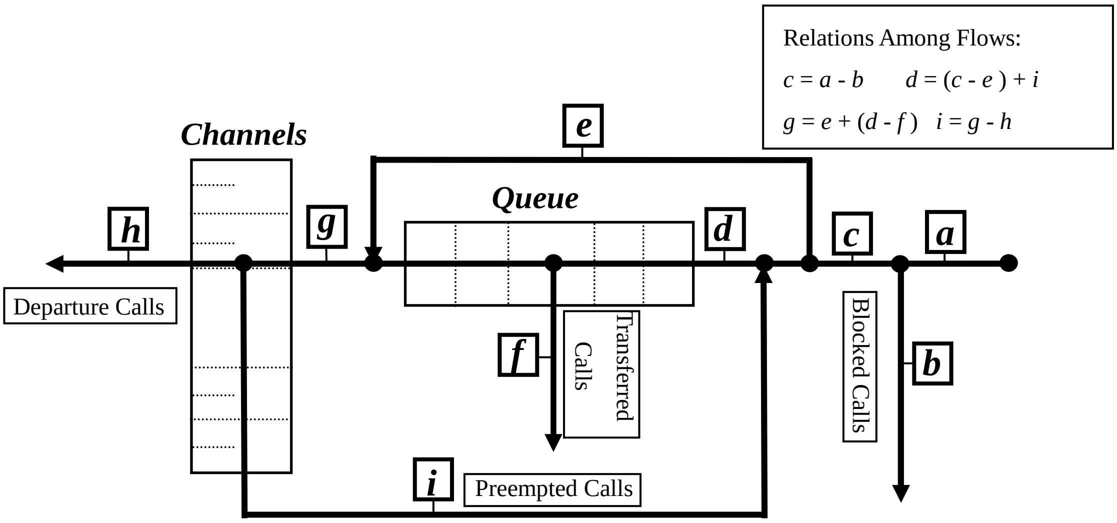


Fig. 4. The signal flowchart for data calls

C. Performance Measures for Data calls

Before analysis, it is helpful to describe the traffic flows for data calls at the marked cell, which is shown in Fig. 4. At point ‘*a*’ in Fig. 4, data calls arrive at the marked cell, and the traffic intensity at point ‘*a*’ is

$$\lambda_{D,a} = \lambda_{OD} + \lambda_{HD} + \lambda_{TD} \tag{16}$$

Next, some arriving data calls are blocked. Traffic flow for the blocked data calls is shown at point ‘*b*’. For a data call, which is not blocked at the marked cell , it gets a channel directly with probability P_C or waits in the queue with probability $1-P_C$. P_C can be given as

$$P_{C,D} = \left(\sum_{s\in\{s|s\in\Omega_s, N_V(s)+N_{TD}(s)<2S\}} p(s) \right) \Big/ \left(1-P_{B,D}\right) \tag{17}$$

where $P_{B,D}$ is the blocking probability for a data call, We have

$$P_{B,D} = \sum_{s\in\{s|s\in\Omega_s, N_V(s)+N_{TD}(s)=2S+Q\}} p(s) \tag{18}$$

The traffic intensity at point ‘*c*’ can be given as

$$\lambda_{D,c} = (\lambda_{HD} + \lambda_{TD} + \lambda_{OD}) \cdot (1-P_{B,D}) \tag{19}$$

For data calls waiting in the queue, some of them are transferred to target cells before they get channels at the marked cell. The transferred data traffic intensity is $\lambda_{D,f} = \lambda_{TD}$ (see point ‘*f*’ in Fig. 4). From point '*h*' and '*i*' in Fig. 4, it can be seen that part of data calls in service are preempted by voice calls. The traffic intensity for data calls leaving the marked cell because of the completion of their calls or moving out of the marked cell without the completion of their calls can be expressed as

$$\lambda_{D,h} = E_D \cdot \mu_{dwell} + E_{D,Full} \cdot \mu_{CD,Full} + E_{D,Half} \cdot \mu_{CD,Half} \tag{20}$$

where $E_{D,Full}$ and $E_{D,Half}$ denote the average number of data calls occupying full-rate and half-rate channels, respectively. The traffic intensity $\lambda_{D,i}$ at point '*i*' in Fig. 4 can be given as

$$\lambda_{D,i} = \lambda_{D,g} \cdot P_{PD} \tag{21}$$

where $\lambda_{D,g}$ is the traffic intensity at point '*g*' in Fig. 4, and P_{PD} is the preemption probability for a data call in service. Therefore, traffic intensity $\lambda_{D,d}$ at point '*d* ' is

$$\lambda_{D,d} = \lambda_{D,c} \cdot (1 - P_C) + \lambda_{D,i} \tag{22}$$

Based on the above descriptions on traffic flows for data calls, we can analyze some of the most important performance measures. In the following, it is assumed that $\lambda_{OD} > 0$.

C.1) Preemption Probability for A Data Call in Service: From Fig. 4, we have

$$\lambda_{D,g} = \lambda_{D,h} / (1 - P_{PD}) = (E_D \cdot \mu_{dwell} + E_{D,Full} \cdot \mu_{CD,Full} + E_{D,Half} \cdot \mu_{CD,Half}) / (1 - P_{PD}) \tag{23}$$

On the other hand, we have

$$\begin{aligned} \lambda_{TD} &= \lambda_{D,i} \cdot P_{Tr,PD} + \lambda_{D,c} \cdot (1 - P_C) \cdot P_{Tr,AD} \\ &= \lambda_{D,g} \cdot P_{PD} \cdot P_{Tr,PD} + (\lambda_{HD} + \lambda_{TD} + \lambda_{OD}) \cdot (1 - P_{B,D}) \cdot (1 - P_C) \cdot P_{Tr,AD} \end{aligned} \tag{24}$$

where $P_{Tr,PD}$ is the probability that a preempted data call, waiting in the queue, is transferred to some target cell before getting its channel again at the marked cell. $P_{Tr,AD}$ is the probability for an arriving data call, which is not blocked by the marked cell and waits in the queue, being transferred to some target cell before getting its channel. The derivations of these two probabilities are given in Appendix B. From equation (24), we have

$$\lambda_{D,g} = \frac{\lambda_{TD} - (\lambda_{HD} + \lambda_{TD} + \lambda_{OD}) \cdot (1 - P_{B,D}) \cdot (1 - P_C) \cdot P_{Tr,AD}}{P_{PD} \cdot P_{Tr,PD}} \tag{25}$$

By combining equation (23) and (25), P_{PD} can be obtained as

$$\begin{cases} P_{PD} = \dfrac{\Theta_1}{\Theta_1 + \Theta_2} \\ \Theta_1 = \dfrac{1}{E_D \cdot \mu_{dwell} + E_{D,Full} \cdot \mu_{CD,Full} + E_{D,Half} \cdot \mu_{CD,Half}} \\ \Theta_2 = \dfrac{P_{Tr,PD}}{\lambda_{TD} - (\lambda_{HD} + \lambda_{TD} + \lambda_{OD}) \cdot (1 - P_{B,D}) \cdot (1 - P_C) \cdot P_{Tr,AD}} \end{cases} \tag{26}$$

In the above derivations about P_{PD}, it is assumed that $Q > 0$. Moreover, when $Q = 0$, we define $P_{PD} \equiv 0$.

C.2) Forced Termination Probability of An Accepted Data Call: Once a data call is served in some cell, the call handoffs to neighboring cells with probability $P_{H,D}$, which can be given as

$$P_{H,D} = \frac{E_D \cdot \mu_{dwell}}{E_D \cdot \mu_{dwell} + E_{D,Full} \cdot \mu_{CD,Full} + E_{D,Half} \cdot \mu_{CD,Half}} \tag{27}$$

Next we define $P_{HT,DS}$ as the probability for a data call in service leaving the marked cell before the completion of its conversation. In this case, the data call leaves the marked cell from the channel pool and initiates a handoff at the target cell, or it departs from the data queue and is transferred to the target cell. Again, with reference to Fig. 4, we have

$$P_{HT,DS} = (1 - P_{PD}) \cdot P_{H,D} + P_{PD} \cdot P_{Tr,PD} + P_{PD} \cdot (1 - P_{Tr,PD}) \cdot P_{HT,DS} = \frac{(1 - P_{PD}) \cdot P_{H,D} + P_{PD} \cdot P_{Tr,PD}}{1 \text{-} P_{PD} \cdot (1 - P_{Tr,PD})} \tag{28}$$

Moreover, $P^*_{FT,D}$ is defined as the forced termination probability of an ongoing data call after it arrives at the target cell. Finally, we define $P_{FT,D}$ as the forced termination probability of an accepted data call (once a data call obtain a channel service in some cell, we say that the call is accepted by the system). We have the following relations

$$\begin{cases} P_{FT,D} = P_{HT,DS} \cdot P^*_{FT,D} \\ P^*_{FT,D} = P_{B,D} \ + (1 - P_{B,D}) \cdot P_C \cdot P_{HT,DS} \cdot P^*_{FT,D} + (1 - P_{B,D}) \cdot (1 - P_C) \cdot P_{Tr,AD} \cdot P^*_{FT,D} \\ \qquad\qquad + \ (1 - P_{B,D}) \cdot (1 - P_C) \cdot (1 - P_{Tr,AD}) \cdot P_{HT,DS} \cdot P^*_{FT,D} \end{cases} \tag{29}$$

Then $P_{FT,D}$ can be expressed as

$$P_{FT,D} = \frac{P_{HT,DS} \cdot P_{B,D}}{1 - (1 - P_{B,D}) \cdot [P_C \cdot P_{HT,DS} + (1 - P_C) \cdot (1 - P_{Tr,AD}) \cdot P_{HT,DS} + (1 - P_C) \cdot P_{Tr,AD}]} \tag{30}$$

C.3) Average Total Waiting Time in Queues for Successfully Completed Data Calls: From Fig. 4, we can see that during the whole course of the communication of a successfully completed data call (including its waiting time in queues before its managing to get a channel and start its communication), it may spend some time waiting in queues at different cells. T_{TW} is defined as the average total waiting time in queues for a *successfully completed* data call.

As for a preempted data call, it enters the queue and waits for a channel to be available. Let $T_{W,PD}$ denote the average total time of a preempted data call waiting in the queue at the marked cell from the moment that it enters the queue. We have

$$\begin{aligned} T_{W,PD} &= P_{Tr,PD} \cdot T_{WTr,PD} + (1-P_{Tr,PD}) \cdot (1-P_{PD}) \cdot T_{WC,PD} + (1-P_{Tr,PD}) \cdot P_{PD} \cdot (T_{WC,PD} + T_{W,PD}) \\ &= [P_{Tr,PD} \cdot T_{WTr,PD} + (1-P_{Tr,PD}) \cdot T_{WC,PD}] \big/ [1-P_{PD} \cdot (1-P_{Tr,PD})] \end{aligned} \tag{31}$$

where $T_{WC,PD}$ denotes the average waiting time of a preempted data call in the queue at the marked cell from the moment that it enters the queue to the moment it gets a channel again, under the condition that it gets a channel successfully. $T_{WTr,PD}$ denotes the average waiting time of a preempted data call in the queue at the marked cell from the moment it enters the queue to the moment it is transferred to another cell, under the condition that it is transferred to another cell before its getting a channel at the marked cell. The derivation of $T_{WC,PD}$ is in Appendix C. The derivation of $T_{WTr,PD}$ can be found in Appendix D.

As for an unblocked arriving data call that waits in the queue at the marked cell, we define $T_{W,AD}$ as the average total time for the call waiting in the queue at the marked cell from the moment that it enters the queue. We have

$$\begin{aligned} T_{W,AD} &= P_{Tr,AD} \cdot T_{WTr,AD} + (1-P_{Tr,AD}) \cdot (1-P_{PD}) \cdot T_{WC,AD} + (1-P_{Tr,AD}) \cdot P_{PD} \cdot (T_{WC,AD} + T_{W,PD}) \\ &= P_{Tr,AD} \cdot T_{WTr,AD} + (1-P_{Tr,AD}) \cdot (T_{WC,AD} + P_{PD} \cdot T_{W,PD}) \end{aligned} \tag{32}$$

where $T_{WC,AD}$ denotes the average waiting time of a queued arriving data call in the queue at the marked cell from the moment that it enters the queue to the moment that it gets its channel under the condition that it gets a channel. $T_{WTr,AD}$ denotes the average waiting time of a queued arriving data call in the queue at the marked cell from the moment that it enters the queue to the moment that it is transferred to another cell under the condition that it is transferred to another cell before its getting a channel at the marked cell. The derivation of $T_{WC,AD}$ is in Appendix C.

The derivation of $T_{WTr,AD}$ can be found In Appendix D.

Then based on equations (31) and (32), we can calculate the average total waiting time $T_{TW,1}$ of an unblocked arriving data call in the queue at the marked cell as follows

$$T_{TW,1} = (1 - P_C) \cdot T_{W,AD} + P_C \cdot P_{PD} \cdot T_{W,PD} \tag{33}$$

Moreover, $T_{TW,1}$ can also be expressed as

$$T_{TW,1} = P_{O,AD} \cdot T_{OW,1} + (1 - P_{O,AD}) \cdot T_{OW,1}' \tag{34}$$

where $P_{O,AD}$ is the probability that an arriving data call, which is not blocked by the marked cell, exits the marked cell before the completion of its call. It can be derived as follows (see Fig. 4)

$$P_{O,AD} = \left(\lambda_{D,h} \cdot P_{H,D} + \lambda_{TD}\right) / \lambda_{D,c} \tag{35}$$

$T_{OW,1}$ denotes the average total waiting time of an unblocked arriving data call, which leaves the marked cell before the completion of its call, in the queue at the marked cell. And $T_{OW,1}'$ denotes the average total waiting time of an unblocked arriving data call, which completes its call at the marked cell, in the queue at the marked cell. Finally, T_{TW} is obtained as

$$\begin{aligned} T_{TW} &= \sum_{\mathrm{i}=1}^{\infty} \left(P_{O,AD}\right)^{i-1} \cdot (1 - P_{O,AD}) \cdot [(i-1) \cdot T_{OW,1} + T_{OW,1}'] = \frac{P_{O,AD} \cdot T_{OW,1} + (1 - P_{O,AD}) \cdot T_{OW,1}'}{1 - P_{O,AD}} \\ &= \frac{T_{TW,1}}{1 - P_{O,AD}} = \frac{(1 - P_C) \cdot T_{W,AD} + P_C \cdot P_{PD} \cdot T_{W,PD}}{1 - P_{O,AD}} \end{aligned} \tag{36}$$

C.4) Average Total Transmission Time of Successfully Completed Data Calls: Let T_{Trans} denote the average total transmission time of a *successfully completed* data call, which includes the waiting time of the call in queues. Using Little's formula, we can get the average transmission time $T_{Trans,1}$ of a successfully completed data call in *one* cell as follows:

$$T_{Trans,1} = \frac{E_D + E_Q}{(\lambda_{HD} + \lambda_{TD} + \lambda_{OD}) \cdot (1 - P_{B,D})} \tag{37}$$

Then, we have

$$T_{Trans} = \frac{T_{Trans,1}}{1 - P_{O,AD}} \tag{38}$$

D. Other Performance Measures of the System

Based on the calculation of the state probabilities, we can easily obtained the following performance measures:

$$E_V = \sum_{s\in\Omega_s} N_V(s)p(s) \tag{39}$$

$$E_D = \sum_{s\in\Omega_s} N_D(s)p(s) \tag{40}$$

$$E_{D,Full} = \sum_{s\in\Omega_s} [C_D(s) - N_D(s)]\cdot p(s) \tag{41}$$

$$E_{D,Half} = \sum_{s\in\Omega_s} [2N_D(s) - C_D(s)]\cdot p(s) \tag{42}$$

$$E_Q = \sum_{s\in\Omega_s} N_Q(s)p(s) \tag{43}$$

VI. Numerical Results and Discussions

Because the new scheme is a preemptive priority handoff scheme with full-rate and half-rate service support, we abbreviate it as FHPS (Full-rate and Half-rate Preemptive Scheme). In this section, FHPS is compared with two extreme schemes. In one scheme, which we call FPS (Full-rate only Preemptive Scheme) for short, except for the fact that only full-rate channel is supported for both voice and data traffic by the system, the other part of the scheme are exactly the same as those of FHPS. FPS has been proposed and studied in [24]. In another extreme scheme, which we call HPS (Half-rate only Preemptive Scheme) for short, only half-rate channel is supported for both voice and data traffic, all the other aspects of the scheme are the same as those of FHPS. We define $\lambda_O \equiv \lambda_{OD} + \lambda_{OV}$ (calls/second) as the total arrival rate of originating calls at the marked cell. Moreover, we define $\gamma = \lambda_{OD} / \lambda_O$ as the relative rate of originating data calls. In numerical examples, parameters are set as follows: $S = 10$, $Q = 10$, $\gamma = 0.5$, $E[T_{CV}] = 100.0s$, $E[T_{CD,Full}] = 40.0s$, $E[T_{CD,Half}] = 80.0s$, $E[T_{dwell}] = 60.0s$.

Fig. 5 shows the forced termination probabilities $P_{FT,V}$ for voice traffic in these three schemes versus λ_O. Fig. 6 shows the forced termination probabilities $P_{FT,D}$ for data traffic in these three schemes versus λ_O. Fig. 7 shows the preemption probabilities P_{PD} of a data call in service in these three schemes versus λ_O. Fig. 8 and Fig. 9 show the total average transmission time T_{Trans} and the total average waiting time T_{TW} of successfully completed data calls of these three schemes versus λ_O, respectively. The average bandwidths c_V for a voice call in service of these three schemes are shown in Fig. 10.

In FPS, only full-rate channel is supported, both voice and data calls have much smaller chances to get access to channel resources, which brings about higher forced termination probabilities for both voice and data calls relative to the other two schemes (refer to Figs. 5 and 6). From Figs. 5 and 6, it can be also seen that $P_{FT,V}$ (forced termination probability of voice calls), and $P_{FT,D}$ (forced termination probability of data calls) in FHPS are even smaller than those in HPS. This is mainly because that in FHPS, both full and half-rate channels are supported, and the system tries to allocate full-rate channels to voice and data calls as much as possible if there are enough channel resources, which implies a more efficient utilization of channel resources in comparison with HPS. The above reason can also be used to explain why P_{PD} (the preemption probability of data calls in service), T_{Trans} (the total average transmission time of successfully completed data calls), and T_{TW} (the total average waiting time of successfully completed data calls) in FHPS are smaller than those in HPS are (see Figs. 7, 8, and 9).

The average total transmission time T_{Trans} for a data call consists of two parts: one is the time for a data call spent while a channel is allocated, and the other one is the time for a data call waiting in queues. From Fig. 9, we can see that T_{TW} in FPS are significantly larger than those in FHPS and HPS. Therefore, in FHPS and HPS, a data call spends less waiting time in queues. Since only half-rate channel is provided in HPS, T_{Trans} in HPS are larger than those in FHPS. Because of the introduction of the efficient channel allocation scheme, in most cases, shortest T_{Trans} can be achieved by using FHPS (see Fig. 8).

From Figs. 5 to 9, we can see that by using FHPS, smaller forced termination probabilities for both voice and data calls and shorter average total transmission time for data calls can be obtained compared with the other two candidate schemes. However, there are no free meals and the communication quality for voice calls is degraded by adopting sub-rating service, see Fig. 10. Extensive numerical experiments show that if the total arrival rate λ_O for originating calls to the system is not heavy (say for the case that λ_O is smaller then 0.15 calls/second), the degradation of communication quality for voice calls is small so that its influence on system performance can be omitted.

VII. CONCLUSIONS

We propose and analyze a service-dependent handoff scheme in voice and data integrated cellular mobile systems, which combines the ideas of “Variable Bandwidth” and “Preemptive Priority” together. Comparisons with the scheme, which only supports “Preemptive Priority” without “Variable Bandwidth” supporting, shows that if total arrival rate for originating calls is not very heavy, the new scheme can provide lower blocking probability and forced termination probability for both voice and data traffic, and shorter average total transmission time for a successfully completed data call. The improved performance mainly relies on two factors: one is that apart from the idea of “Preemptive Priority”, “Variable Bandwidth” is introduced, which further decreases the blocking probabilities for both voice and data traffic. Moreover, in the proposed scheme, “Variable Bandwidth” functions before “Preemptive Priority”, which decreases the probability for data calls in service being preempted by voice calls. The second factor lies in the fact that fairer channel allocation scheme is considered, which further avoids the starvation of the service for data traffic by voice calls. However, there are no free meals and the communication quality for voice calls is degraded by adopting sub-rating service. Extensive numerical experiments show that if the total arrival rate for originating calls to the system is not heavy, the degradation of communication quality for voice calls is small so that its influence on system performance can be omitted.

In the future work, we are interested in considering some extensions for the proposed scheme. In this paper, for simplicity, priority of handoff calls over originating calls is not considered. Therefore, a very natural extension is to consider schemes which provide some priority for handoff calls. For example, the idea of “Priority Reservation” can be combined with the proposed scheme. In this paper, only voice and data calls are considered. Extending the proposed handoff and channel allocation scheme into more general cases, where arbitrary types of traffic with more flexible bandwidth requirements are considered, is a very interesting work. Finally, real implementation issues of the proposed scheme should also be considered. In this paper, attentions are focus on the performance analysis of the proposed handoff and channel allocation scheme. However, the implementation of an efficient signaling scheme, which is of importance for the real implementation, is not discussed. As we all know, complicated signaling introduces more processing overhead for the system. Therefore, in our future work, we will find ways to design an efficient signaling scheme which brings about moderate processing overhead for the system. To minimize the introduced processing overhead, one possible way is to limit the number of calls (say only handoff calls) having the rights to enjoy the "Variable Bandwidth" and "Preemptive Priority".

ACKNOWLEDGMENT

We thank the anonymous reviewers for their comments to improve the quality of this paper.

APPENDIX A. CALCULATION OF STATE PROBABILITIES

If we assume that λ_{HV}, λ_{HD} and λ_{TD} are constant, the flow balance equations in (10) can be solved as a set of linear simultaneous equations. However, actually λ_{HV}, λ_{HD} and λ_{TD} are not constant, but dependent on the state probabilities $p(s)$'s through equations (2)-(4). Thus we use the following iteration procedure to obtain all the state probabilities:

Step 1. Select arbitrary initial (positive) values for λ_{HV}, λ_{HD} and λ_{TD}.

Step 2. Compute all the state probabilities $p(s)$'s in equation (10) by using SOR method [27].

Step 3. Compute the average number of voice calls holding channels, the average number of data calls holding channels and the average number of data calls waiting in the queue by using equations (39), (40) and (43).

Step 4. Compute new λ_{HV}, λ_{HD} and λ_{TD} using equations (2)-(4). If $\left|\frac{new\,\lambda_{HV} - old\,\lambda_{HV}}{old\,\lambda_{HV}}\right| < \varepsilon$, $\left|\frac{new\,\lambda_{HD} - old\,\lambda_{HD}}{old\,\lambda_{HD}}\right| < \varepsilon$ and $\left|\frac{new\,\lambda_{TD} - old\,\lambda_{TD}}{old\,\lambda_{TD}}\right| < \varepsilon$, then stop the iteration (ε is a very small positive number to check the convergence). Otherwise, set $\lambda_{HV} \Leftarrow old\lambda_{HV} + \omega \cdot (new\lambda_{HV} - old\lambda_{HV})$, $\lambda_{HD} \Leftarrow old\lambda_{HD} + \omega \cdot (new\lambda_{HD} - old\lambda_{HD})$, $\lambda_{TD} \Leftarrow old\lambda_{TD} + \omega \cdot (new\lambda_{TD} - old\lambda_{TD})$ (in our calculations, ω is set to be 0.5), and go back to **step 2** again.

APPENDIX B. CALCULATION OF $P_{Tr,AD}$ AND $P_{Tr,PD}$

In the following part of the paper, we focus our attention on the data call waiting at the k th position in the queue at the marked cell, and call it as ‘the marked data call’. Given that the system is at state $s_q \in \Omega_{s_q}$,

$$\Omega_{s_q} \equiv \{s \mid s \in \Omega_s, N_Q(s) > 0\} \tag{B.1}$$

and the marked data call is queued at the k th position ($0 < k \le N_Q(s_q)$) in the queue, then we define $p_{tr,D}(s_q, k)$ as the probability of the marked data call being transferred to the target cell before its leaving the queue and getting a channel at the marked cell. In the following, we use quasi-system states, which are determined by the current system state and the position of the marked data call in the queue, to describe state transitions related to the changing of the position of the marked data call. Therefore, the quasi-system states related to the marked data call can be expressed as (s_q, k), and $s_q \in \Omega_{s_q}, 0 < k \le N_Q(s_q)$. Assuming that the current state is (s_q, k), then the transitions of quasi-system states are driven by the occurrence of the following events:

1) *Departure of the Marked Data Call from the Queue:* Denote $r_1(s_q, k)$ to be the transition rate from system state $s_q : (N_V(s_q), N_{TD}(s_q))$ to $s_{q,1} : (N_V(s_q), N_{TD}(s_q) - 1)$, and $r_1(s_q, k) = \mu_{dwell}$.

2) *Departure of a Data Call Waiting behind the Marked Data Call from the Queue:* Let $r_2(s_q, k)$ be the transition rate from system state $s_q : (N_V(s_q), N_{TD}(s_q))$ to $s_{q,2} : (N_V(s_q), N_{TD}(s_q) - 1)$. After the occurrence of the event, the position of the marked data call in the queue remains to be k, and $r_2(s_q, k) = [N_Q(s_q) - k] \cdot \mu_{dwell}$.

3) *Departure of a Data Call Waiting before the Marked Data Call from the Queue or Departure of a Data Call in Service from the Channel Pool:* Let $r_3(s_q, k)$ be the transition rate from system state $s_q : (N_V(s_q), N_{TD}(s_q))$ to $s_{q,3} : (N_V(s_q), N_{TD}(s_q) - 1)$. After the occurrence of the event, the position of the marked data call in the queue is $k - 1$ (specifically, if $k - 1 = 0$, it indicates that the marked data call gets a channel), and $r_3(s_q, k) = (k-1) \cdot \mu_{dwell} + N_D(s_q) \cdot (\mu_{dwell} + \mu_{CD,Half})$.

4) *Departure of a Voice Call in Service from the Channel Pool:* Let $r_4(s_q, k)$ be the transition rate from system state $s_q : (N_V(s_q), N_{TD}(s_q))$ to $s_{q,4} : (N_V(s_q) - 1, N_{TD}(s_q))$. After the occurrence of the event, the position of the marked data call in the queue is $k - 1$. If $s_{q,4} \in \Omega_s$, $r_4(s_q, k) = N_V(s_q) \cdot (\mu_{dwell} + \mu_{CV})$. Otherwise, $r_4(s_q, k) = 0$.

5) *Arrival of a Data Call:* Let $r_5(s_q, k)$ be the transition rate from system state $s_q : (N_V(s_q), N_{TD}(s_q))$ to $s_{q,5} : (N_V(s_q), N_{TD}(s_q) + 1)$. After the occurrence of the event, the position of the marked data call in the queue is remained to be k. If $s_{q,5} \in \Omega_s$, $r_5(s_q, k) = \lambda_{HD} + \lambda_{TD} + \lambda_{OD}$. Otherwise, $r_5(s_q, k) = 0$.

6) Arrival of a Voice Call: Let $r_6(s_q,k)$ be the transition rate from system state $s_q:(N_V(s_q),N_{TD}(s_q))$ to $s_{q,6}:(N_V(s_q)+1,N_{TD}(s_q))$. In this case, one of the data calls in service is preempted by the arriving voice call. After the occurrence of the event, the position of the marked data call in the queue is remained to be k. If $s_{q,6}\in\Omega_s$, $r_6(s_q,k)=\lambda_{HV}+\lambda_{OV}$. Otherwise, $r_6(s_q,k)=0$.

Therefore, the total transition rate out of the state (s_q,k) can be expressed as

$$r(s_q,k)=\sum_{i=1}^{6}r_i(s_q,k) \tag{B.2}$$

Moreover, we define

$$p_{tr,D}(s,k)\equiv 0 \quad \text{if } s\notin\Omega_{s_q} \text{ or } k\le 0 \text{ or } k>N_Q(s) \tag{B.3}$$

Then we have the following set of linear equations

$$p_{tr,D}(s_q,k)=\frac{r_1(s_q,k)}{r(s_q,k)}+\sum_{i=2,5,6}\frac{r_i(s_q,k)}{r(s_q,k)}\cdot p_{tr,D}(s_{q,i},k)+\sum_{i=3,4}\frac{r_i(s_q,k)}{r(s_q,k)}\cdot p_{tr,D}(s_{q,i},k-1) \tag{B.4}$$

where $s_q\in\Omega_{s_q},0<k\le N_Q(s_q)$. By solving the above linear equations, we can get all the values for $p_{tr,D}(s_q,k)$'s. And then, $P_{Tr,AD}$ can be given as follows

$$\begin{cases} P_{Tr,AD}=\left(\sum\limits_{s\in\Omega_{Tr,AD}}p(s)\cdot p_{tr,D}(s',N_Q(s'))\right)\Big/\left(\sum\limits_{s\in\Omega_{Tr,AD}}p(s)\right) \\ \Omega_{Tr,AD}\equiv\{s\,|\,s\in\Omega_s, N_V(s)+N_D(s)=2S, N_Q(s)<Q\} \end{cases} \tag{B.5}$$

where $N_V(s')=N_V(s),N_{TD}(s')=N_{TD}(s)+1$. Moreover, $P_{Tr,PD}$ can be expressed as

$$\begin{cases} P_{Tr,PD}=\left(\sum\limits_{s\in\Omega_{Tr,PD}}p(s)\cdot p_{tr,D}(s',N_Q(s'))\right)\Big/\left(\sum\limits_{s\in\Omega_{Tr,PD}}p(s)\right) \\ \Omega_{Tr,PD}\equiv\{s\,|\,s\in\Omega_s, N_V(s)+N_D(s)=2S, N_D(s)>0, N_Q(s)<Q\} \end{cases} \tag{B.6}$$

where $N_V(s')=N_V(s)+1,N_{TD}(s')=N_{TD}(s)$.

APPENDIX C. CALCULATION OF $T_{WC,AD}$ AND $T_{WC,PD}$

Given that the system is at state $s_q\in\Omega_{s_q}$, and we know that the marked data call waiting at the k th position ($0<k\le N_Q(s_q)$) in the queue *can* get a channel at the marked cell, then we

define $t_{WC}(s_q, k)$ as the average waiting time for the marked data call before it gets a channel. We use quasi-system states for the marked cell to describe state transitions related to the marked data call. The state transitions of quasi-system states have been described in Appendix B.

First, let us start from Fig. C.1. In the figure, a general tree structure of the quasi-state transition is shown. We assume that when the system is at state s_0 (see the root node in the tree structure) the marked data call is waiting in the queue. T_0 is the average holding time of the system at state s_0. Because of the occurrence of some events, the system state changes. If the next state is s_i ($1 \le i \le n-1$, $n > 1$), then the marked data call gets a channel. Let $q_{0,i}$ ($1 \le i \le n-1$, $n > 1$) be the transition probability from state s_0 to s_i ($1 \le i \le n-1$, $n > 1$). Specifically, for the case of $n = 1$, we say that it is impossible for the marked data call to get a channel with only *one step* state transition. If the next state of s_0 is s_i $(n \le i \le m)$, although the marked data call still waits in the queue, can it finally get its channel if the system state transits from s_i to $s_{i,l}$ $(n \le i \le m, 1 \le l \le k_i)$ without its being transferred to other cells before getting a channel. It should be noted that although from state s_i $(n \le i \le m)$ there are total of k_i different

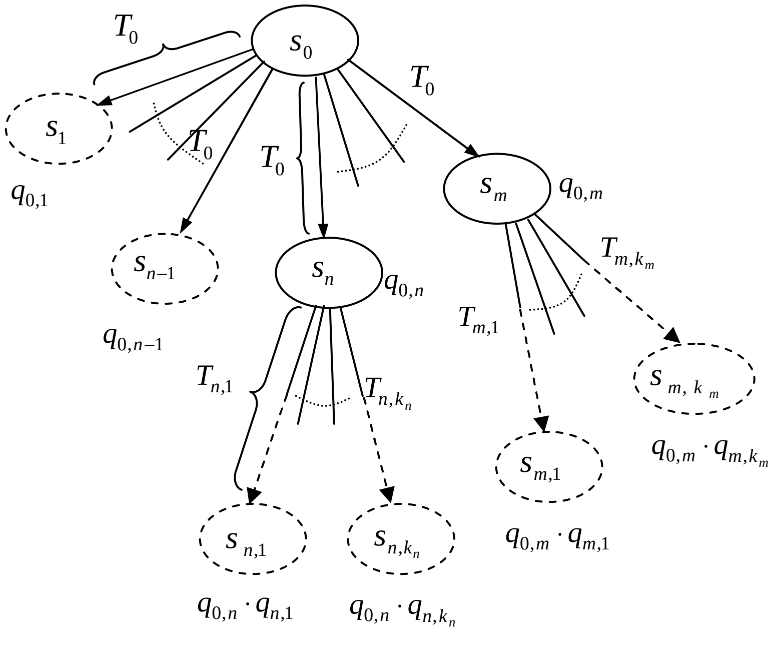


Fig. C.1 Tree structure of the quasi-state transition

state transition routes, which makes it possible for the marked data call to get a channel, we can not draw a conclusion that if $i1 \neq i2$ or $l1 \neq l2$ then $s_{i1,l1} \neq s_{i2,l2}$ $(n \leq i1, i2 \leq m; 1 \leq l1 \leq k_{i1}; 1 \leq l2 \leq k_{i2})$ always holds. That is, subscriptions in the symbol $s_{i,l}$ are not used to distinguish different states but to distinguish different state transition routes, which makes it possible for the marked data call to get a channel. Furthermore, It is assumed that the probabilities $q_{0,i}$ $(n \leq i \leq m)$ and $q_{i,l}$ $(n \leq i \leq m, 1 \leq l \leq k_i)$ denote the transition probabilities from state s_0 to s_i $(n \leq i \leq m)$ and those from state s_i to $s_{i,l}$, respectively. And let $T_{i,l}$ $(n \leq i \leq m, 1 \leq l \leq k_i)$ denote the average waiting time for the marked data call in the queue under the condition that the system state transits from state s_i to $s_{i,l}$. Therefore, assuming that the original system state is s_i $(n \leq i \leq m)$, the average waiting time $T(s_i)$ $(n \leq i \leq m)$ for the marked data call in the queue before its getting a channel can be expressed as

$$T(s_i) = \left(\sum_{l=1}^{k_i} q_{i,l} \cdot T_{i,l} \right) \Bigg/ \left(\sum_{l=1}^{k_i} q_{i,l} \right) \tag{C.1}$$

As for state s_0, the average waiting time $T(s_0)$ for the marked data call in the queue before its getting a channel can be expressed as

$$\begin{aligned} T(s_0) &= \frac{\sum_{i=1}^{n-1} q_{0,i} T_0 + \sum_{i=n}^{m} \sum_{l=1}^{k_i} q_{0,i} q_{i,l} (T_0 + T_{i,l})}{\sum_{i=1}^{n-1} q_{0,i} + \sum_{i=n}^{m} \sum_{l=1}^{k_i} q_{0,i} q_{i,l}} = \frac{\sum_{i=1}^{n-1} q_{0,i} T_0 + \sum_{i=n}^{m} \sum_{l=1}^{k_i} q_{0,i} q_{i,l} T_0 + \sum_{i=n}^{m} q_{0,i} \cdot \left[T(s_i) \cdot \sum_{l=1}^{k_i} q_{i,l} \right]}{\sum_{i=1}^{n-1} q_{0,i} + \sum_{i=n}^{m} \sum_{l=1}^{k_i} q_{0,i} q_{i,l}} \\ &= \frac{\sum_{i=1}^{n-1} q_{0,i} T_0 + \sum_{i=n}^{m} q_{0,i} \cdot \left(\sum_{l=1}^{k_i} q_{i,l} \right) \cdot [T(s_i) + T_0]}{\sum_{i=1}^{n-1} q_{0,i} + \sum_{i=n}^{m} \sum_{l=1}^{k_i} q_{0,i} q_{i,l}} = \frac{\sum_{i=1}^{n-1} q_{0,i} T_0 + \sum_{i=n}^{m} q_{0,i} \cdot q_c(s_i) \cdot [T(s_i) + T_0]}{q_c(s_0)} \end{aligned} \tag{C.2}$$

where $q_c(s)$ is the probability of the marked data call getting its channel successfully given that the original system state is s. From the above equation, we can see that $T(s_0)$ can be expressed as the linear combinations of T_0 and $T(s_i)$ $(n \leq i \leq m)$.

Moreover, we define

$$t_{WC}(s,k) \equiv 0 \quad \text{if } s \notin \Omega_{s_q} \text{ or } k \leq 0 \text{ or } k > N_Q(s) \tag{C.3}$$

Based on the state transitions of quasi-system states described in Appendix B and the result shown in equation (C.2), we can get the following linear equations

$$t_{WC}(s_q,k)=\sum_{i=3,4}\frac{\frac{r_i(s_q,k)}{r(s_q,k)}\cdot[1-p_{tr,D}(s_{q,i},k-1)]\cdot[t_w(s_q,k)+t_{WC}(s_{q,i},k-1)]}{1-p_{tr,D}(s_q,k)} + \sum_{i=2,5,6}\frac{\frac{r_i(s_q,k)}{r(s_q,k)}\cdot[1-p_{tr,D}(s_{q,i},k)]\cdot[t_w(s_q,k)+t_{WC}(s_{q,i},k)]}{1-p_{tr,D}(s_q,k)} \tag{C.4}$$

where $s_q\in\Omega_{s_q},0<k\le N_Q(s_q)$, and $t_w(s_q,k)$ is the average holding time at state (s_q,k)

$$t_w(s_q,k)=\frac{1}{r(s_q,k)} \tag{C.5}$$

In equation (C.4), the term $\frac{r_i(s_q,k)}{r(s_q,k)}$ corresponds to the term $q_{0,i}$ in equation (C.2). The term $[1-p_{tr,D}(s_{q,i},k)]$ and $[1-p_{tr,D}(s_{q,i},k-1)]$ in equation (C.4) correspond to the term $q_c(s_i)$ in equation (C.2), and $1-p_{tr,D}(s_q,k)$ corresponds to $q_c(s_0)$ in equation (C.2). Moreover, $[t_w(s_q,k)+t_{WC}(s_{q,i},k)]$ and $[t_w(s_q,k)+t_{WC}(s_{q,i},k-1)]$ correspond to the term $[T_0+T(s_i)]$ in equation (C.2). By solving the above linear equations, we can get all the values of $t_{WC}(s_q,k)$'s. Then $T_{WC,AD}$ can be given as follows

$$T_{WC,AD}=\left(\sum_{s\in\Omega_{Tr,AD}}p(s)\cdot[1-p_{tr,D}(s',N_Q(s'))]\cdot t_{WC}(s',N_Q(s'))\right)\Bigg/\left(\sum_{s\in\Omega_{Tr,AD}}p(s)\cdot[1-p_{tr,D}(s',N_Q(s'))]\right) \tag{C.6}$$

where $N_V(s')=N_V(s),N_{TD}(s')=N_{TD}(s)+1$. Moreover, $T_{WC,PD}$ can be expressed as

$$T_{WC,PD}=\left(\sum_{s\in\Omega_{Tr,PD}}p(s)\cdot[1-p_{tr,D}(s',N_Q(s'))]\cdot t_{WC}(s',N_Q(s'))\right)\Bigg/\left(\sum_{s\in\Omega_{Tr,PD}}p(s)\cdot[1-p_{tr,D}(s',N_Q(s'))]\right) \tag{C.7}$$

where $N_V(s')=N_V(s)+1,N_{TD}(s')=N_{TD}(s)$.

APPENDIX D. CALCULATION OF $T_{WTr,AD}$ AND $T_{WTr,PD}$

Given that the system is at state $s_q\in\Omega_{s_q}$, and we know that the marked data call waiting at the k th position ($0 < k \le N_Q(s_q)$) in the queue *can not* get a channel at the marked cell, i.e., the data call is transferred to the target cell before getting a channel at the marked cell, then we define $t_{WTr}(s_q,k)$ as the average waiting time in the queue for the data call before its being transferred to the target cell. The derivation of $t_{WTr}(s_q,k)$ is just the same as that of $t_{WC}(s_q,k)$ in Appendix C. Therefore, by referring the results in Appendix C, we have the following linear equations directly

$$\begin{aligned} t_{WTr}(s_q,k) = {} & \frac{\frac{r_1(s_q,k)}{r(s_q,k)} \cdot t_w(s_q,k)}{p_{tr,D}(s_q,k)} + \sum_{i=2,5,6} \frac{\frac{r_i(s_q,k)}{r(s_q,k)} \cdot p_{tr,D}(s_{q,i},k) \cdot [t_w(s_q,k) + t_{WTr}(s_{q,i},k)]}{p_{tr,D}(s_q,k)} \\ & + \sum_{i=3,4} \frac{\frac{r_i(s_q,k)}{r(s_q,k)} \cdot p_{tr,D}(s_{q,i},k-1) \cdot [t_w(s_q,k) + t_{WTr}(s_{q,i},k-1)]}{p_{tr,D}(s_q,k)} \end{aligned} \tag{D.1}$$

where $s_q \in \Omega_{s_q}, 0 < k \le N_Q(s_q)$. Moreover, we define

$$t_{WTr}(s,k) \equiv 0 \quad \text{if } s \notin \Omega_{s_q} \text{ or } k \le 0 \text{ or } k > N_Q(s) \tag{D.2}$$

By solving the above linear equations, we can get all the values of $t_{WTr}(s_q,k)$'s. Then $T_{WTr,AD}$ can be given as follows

$$T_{WTr,AD} = \left(\sum_{s \in \Omega_{Tr,AD}} p(s) \cdot p_{tr,D}(s', N_Q(s')) \cdot t_{WTr}(s', N_Q(s')) \right) \Bigg/ \left(\sum_{s \in \Omega_{Tr,AD}} p(s) \cdot p_{tr,D}(s', N_Q(s')) \right) \tag{D.3}$$

where $N_V(s') = N_V(s), N_{TD}(s') = N_{TD}(s) + 1$. Moreover, $T_{WTr,PD}$ can be expressed as

$$T_{WTr,PD} = \left(\sum_{s \in \Omega_{Tr,PD}} p(s) \cdot p_{tr,D}(s', N_Q(s')) \cdot t_{WTr}(s', N_Q(s')) \right) \Bigg/ \left(\sum_{s \in \Omega_{Tr,PD}} p(s) \cdot p_{tr,D}(s', N_Q(s')) \right) \tag{D.4}$$

where $N_V(s') = N_V(s) + 1, N_{TD}(s') = N_{TD}(s)$.

REFERENCES

[1] S.Tekinay and B.Jabbbari, “Handover and channel assignment in mobile cellular networks,” IEEE Commun. Magazine, Nov. 1991, pp.42-46.

[2] Victor O.K.LI and Xiao Xin Qiu, “Personal communication systems (PCS),” Proceedings of the IEEE, vol. 83, No. 9, Sept, 1995, pp. 1208-1243.

[3] Q.A.Zeng, K.Mukumoto, and A.Fukuda, “Influence of cell radius, moving speed, and duration of calls on

handoff rate in cellular mobile radio systems," Proc. Wireless'95, J-2, June 1995, pp.511-520.

[4] A.Murase, I.C.Symington, and E.Green, "Handover criterion for macro and microcellular systems," IEEE Vehicular Technology Conference'1991, pp.524-530.

[5] M.D.Austin and G.L.Stuber, "Direction biased handoff algorithms for urban microcells," IEEE Vehicular Technology Conference' 1994, pp.101-105.

[6] K.G.Cornett, "Bit error rate estimation techniques for digital land mobile radios," IEEE Vehicular Technology Conference'1991, pp.543-548.

[7] HONG,D., and S.S.Rappaport "Traffic model and performance analysis for cellular mobile radiotelephone systems with prioritized and non-prioritized hand-off procedures," IEEE Trans. On Vehicular Technology, 1986, vol. 35, pp.77-92.

[8] R.Guerin "Queueing-Blocking Systems with two arrival Stream and Guard Channels," IEEE Trans. On Commun., vol 36, No. 2, Feb. 1988, pp.153-163.

[9] Q.A.Zeng, K.Mukumoto, and A.Fukuda, "Performance analysis of mobile cellular radio system with priority reservation handoff procedures," Proc. IEEE Vehicular Technology Conference'1994, vol.3, pp.1829-1833.

[10] Q.A.Zeng, K. Mukumoto, and A.Fukuda, "Performance analysis of mobile cellular radio systems with two-level priority reservation handoff procedure," IEICE Trans on commun, vol. E80-B, No.4, April 1997, pp.598-607.

[11] Hong, D., and S.S.Rappaport "Priority oriented channel access for cellular systems serving vehicular and portable radio telephones," IEE Proc. I, 1989, CSV-136, No. 5, pp.339-346.

[12] S.S.Rappaport "The multiple-call hand-off problem in high-capacity communications systems," IEEE Trans. On Vehicular Technology, 1991, vol. 40, No. 3, pp.546-557.

[13] S.S.Rappaport "Models for call hand-off schemes in cellular communication networks," in Nanda, S., and Goodman, D. J. (Eds.):'Third generation wireless information networks'(Kluwer Academic Publishers, Boston, 1992), pp.163-185.

[14] S.S.Rappaport "Blocking, hand-off and traffic performance for cellular communication systems with mixed platforms," IEE Proc. I, 1993, CSV-140, No. 5, pp.389-401.

[15] C.Purzynski and S.S.Rappaport "Multiple call hand-off problem with queued hand-offs and mixed platform types," IEE Proc., I, 1995, CSV-142, No. 1, pp.31-39.

[16] Y. -B. Lin, S. Mohan, and A. Noerpel, "Analyzing the trade off between implementation costs and performance: PCS channel assignment strategies for handoff and initial access," IEEE Personal

Communications Magazine, vol. 1, No. 3, pp. 47-56, 1994.

[17] Y. -B. Lin, A. Noerpel, and D. A. Harasty, “The sub-rating channel assignment strategy for PCS hand-offs,” IEEE Trans. On Vehicular Technology, vol. 45, No. 1, Feb 1996, pp. 122-130.

[18] M. Naghshineh and M. Willebeek-LeMair, “End-to-end QoS provisioning multimedia wireless/mobile networks using an adaptive framework,” IEEE Communications Magazine, vol. 35, No. 11, 1997, pp. 72-81.

[19] V. Bharghavan, K. Lee, S. Lu, S. Ha, J. Li, and D. Dwyer, “The TIMELY adaptive resource management architecture,” IEEE Personal Communications Magazine, vol. 5, No. 4, 1998, pp. 20-31.

[20] B. Noble, “System support for mobile, adaptive applications,” IEEE Personal Communications Magazine, vol. 7, No. 1, 2000, pp. 44-49.

[21] M. Mirhakkak, N. Schult and D. Thomson, “Dynamic bandwidth management and adaptive applications for a variable bandwidth wireless environment,” IEEE Journal on Selected Areas in Communications, vol. 19, No. 10, 2001, pp. 1984-1997.

[22] M. Ei-Kadi, S. Olariu and H. Abdel-Wahab, “A rate-based borrowing scheme for QoS provisioning in multimedia wireless networks ,” IEEE Trans. Parallel and Distributed Systems, vol. 13, No. 2, 2002, pp. 156-166.

[23] Y. Xiao, C.L.P. Chen, and B. Wang, “Bandwidth degradation QoS provisioning for adaptive multimedia in wireless/mobile networks,” Computer Communications, vol. 25, No. 13, 2002, pp. 1153-1161.

[24] Bo LI, Qing An-ZENG, Kaiji MUKUMOTO and Akira FUKUDA, “A Preemptive Priority Handoff Scheme in Integrated Voice and Data Cellular Mobile Systems,” IEICE Trans. On Communications, vol. E82-B, No. 10, Oct 1999, pp. 1633-1642.

[25] Qing-An Zeng and D. P. Agrawal, “An analytical modeling of handoff for integrated voice/data wireless networks with priority reservation and preemptive priority procedures,” Proc. Workshop on Wireless Networks and Mobile Computing in conjunction with ICPP’2000, Aug 2000, pp. 523-530.

[26] Qing-An Zeng and D. P. Agrawal, “Modeling and efficient handling of handoffs in integrated wireless mobile networks,” IEEE Trans. On Vehicular Technology, vol. 51, No. 11, Nov. 2002, pp. 1469-1478.

[27] G. Bolch, S. Greiner, H. de Meer, and K.S. Trivedi, Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications, A Wiley-Interscience Publication, 1998, pp. 140-144.

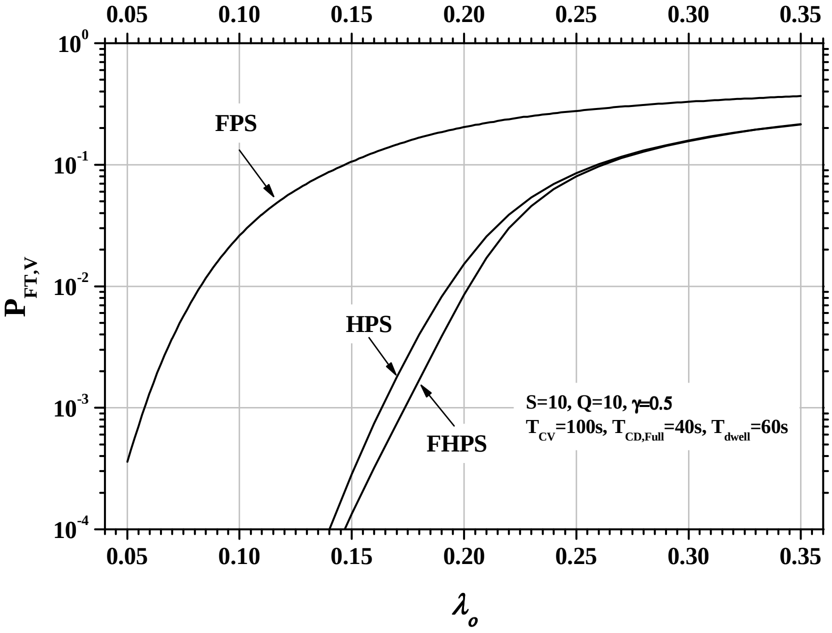


Fig. 5 $P_{FT,V}$ for voice traffic versus λ_o

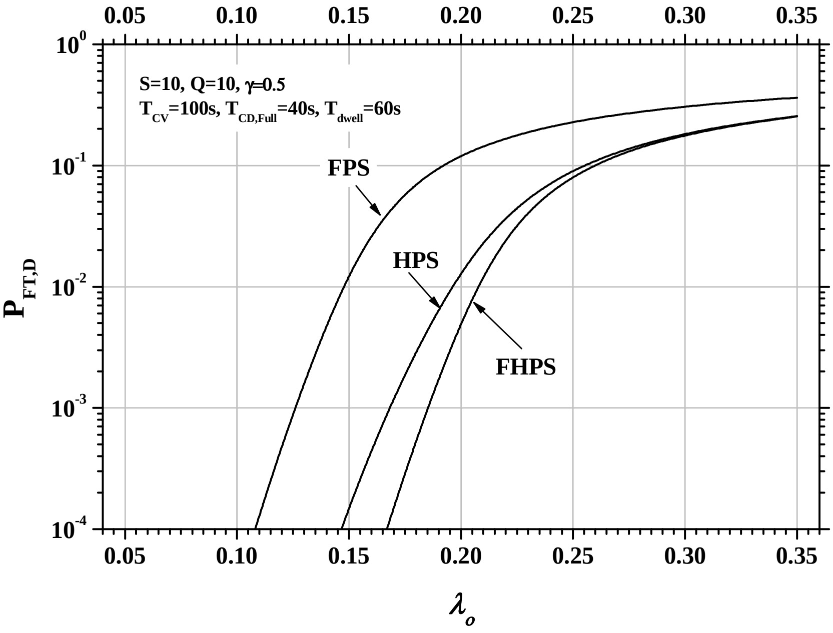


Fig. 6 $P_{FT,D}$ for data traffic versus λ_o

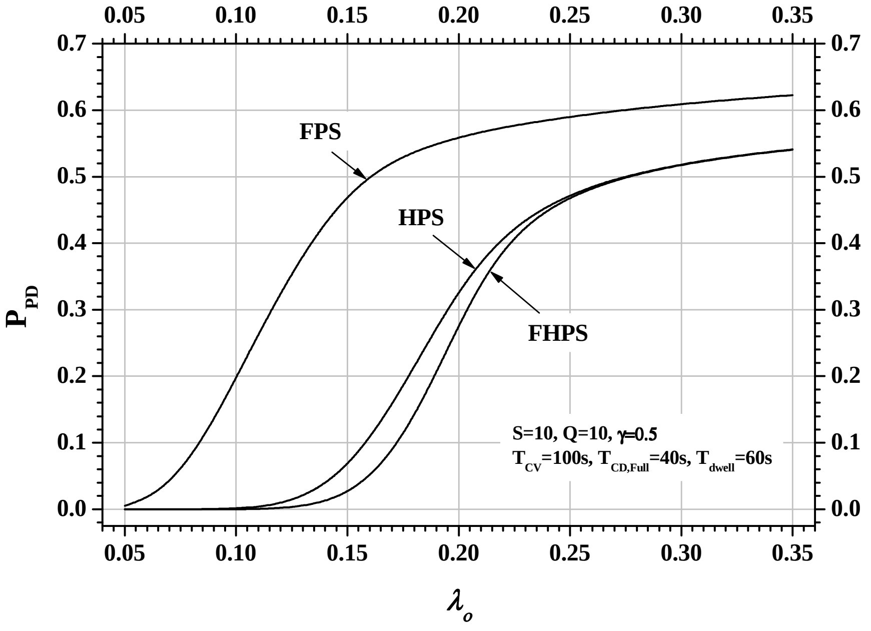


Fig. 7 Preemption probability P_{PD} versus λ_o

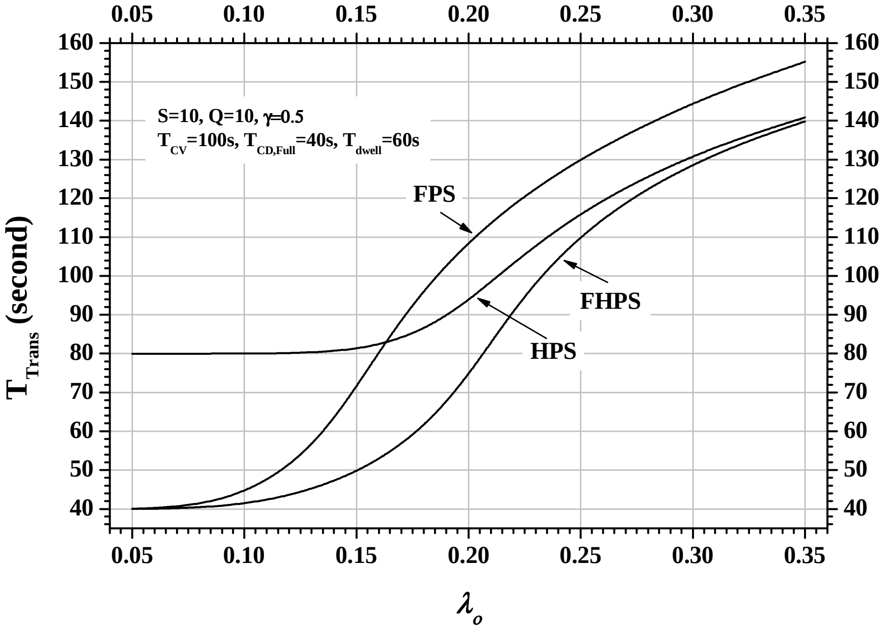


Fig. 8 Average total transmission time T_{Trans} of data calls versus λ_o

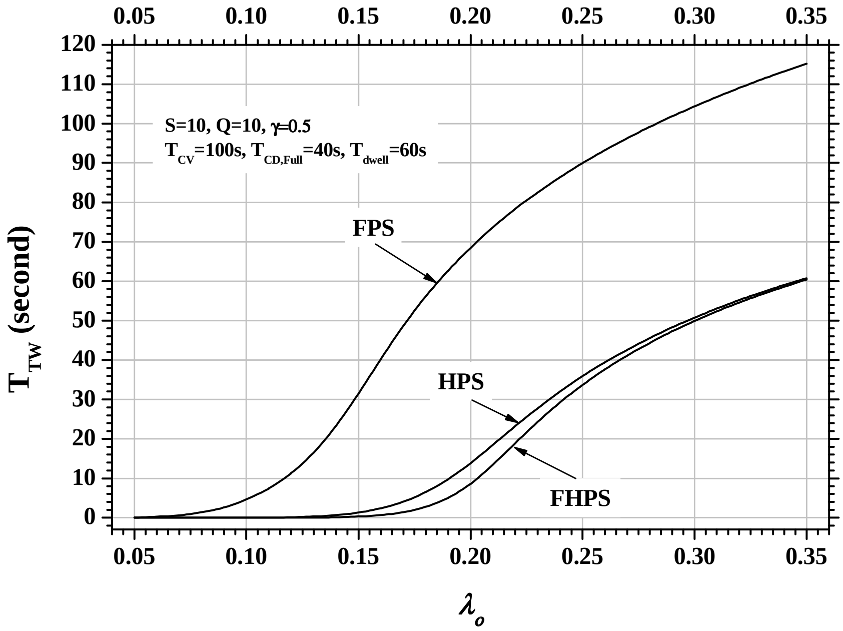


Fig. 9 Average total waiting time T_{TW} of data calls versus λ_o

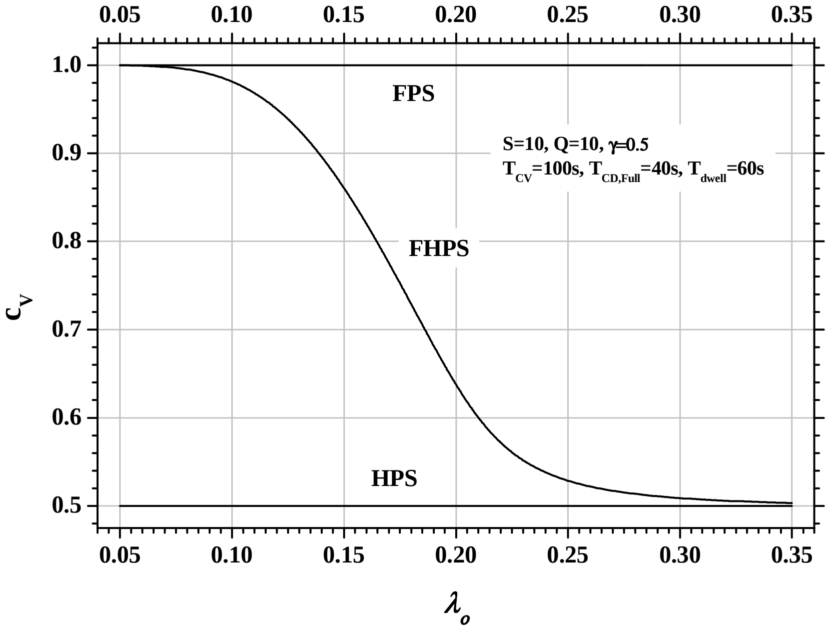


Fig. 10 Average bandwidth c_V per voice call in service versus λ_o