Performance analysis of a channel allocation scheme for multi-service mobile cellular networks

Shensheng Tang* † and Wei Li‡

Department of Electrical Engineering and Computer Science, University of Toledo, Toledo, OH 43606, U.S.A.

SUMMARY

This paper presents a new channel allocation scheme, namely the dynamic partition with pre-emptive priority (DPPP) scheme, for multi-service mobile cellular networks. The system is modelled by a two-dimensional Markov process and analysed by the matrix-analytic method. A pre-emptive priority (PP) mechanism is employed to guarantee the quality of service (QoS) requirement of the real-time (RT) traffic at the expense of some degradation of non-real-time (NRT) traffic, while the victim buffer compensates the degradation and has no negative impact on the RT traffic. The complete service differentiation between new calls and handoff calls from different traffic classes is achieved by using the dynamic partition (DP) concept with the help of related design parameters. The performance analysis and numerical results show that the DPPP scheme, compared with the existing schemes, is effective and practical in multi-service environments. Copyright © 2006 John Wiley & Sons, Ltd.

Received 21 May 2005; Revised 11 March 2006; Accepted 15 March 2006

KEY WORDS: channel allocation; pre-emptive priority (PP); victim buffer; quality of service (QoS); handoff; RT traffic; NRT traffic; mobile cellular networks

1. INTRODUCTION

In the current wireless and mobile industry, a new trend is the shift from the primarily voice-centric service to data and multimedia applications. Different classes of traffic have different quality of service (QoS) requirements. For example, there are four traffic classes defined by 3GPP [1]: conversational class, streaming class, interactive class, and background class. The main distinguishing factor between these traffic classes is how delay sensitive the

*Correspondence to: Shensheng Tang, Department of Electrical Engineering and Computer Science, University of Toledo, Toledo, OH 43606, U.S.A.
†E-mail: stang2@utnet.utoledo.edu
‡E-mail: wli@eecs.utoledo.edu

Contract grant sponsor: National Science Foundation; contract/grant number: CCF-0515263

Copyright © 2006 John Wiley & Sons, Ltd.
traffic is. The conversational class is very delay-sensitive, while the background class is the most delay-insensitive class. The first two classes are mainly used to carry real-time (RT) traffic flows, such as voice and streaming video. The other two classes are mainly used to carry non-real-time (NRT) traffic, such as the traditional Internet applications like WWW, Email and FTP.

One of the results of data and multimedia applications is more bandwidth requirements per user, which limits the capacity of the system; while the bandwidth resource is always scarce and precious in wireless and mobile communications [2]. Hence, the problem formulation is to maximize the system capacity, and at the same time, to guarantee different QoS requirements for different service classes. To solve this problem, an effective and efficient channel allocation (or bandwidth allocation) scheme is extremely necessary.

Channel allocation strategies have been studied extensively in the literature. Chen and Fang [3] proposed a dynamic multiple-threshold bandwidth reservation scheme to support two QoS criteria, i.e. the system keeps the handoff dropping probability always less than a predefined QoS bound, while maintaining the relative priorities of different traffic classes in terms of blocking probability by dynamically adjusting bandwidth reservation thresholds. Huang et al. [4] proposed a system with adaptive resource allocation for multimedia QoS support under the constraint of limited and varying bandwidth resources. A service model consisting of three service classes with different handoff dropping requirements was presented. Appropriate call-admission control and dynamic resource allocation schemes were developed to allocate resources adaptively to the RT service classes, and the rate-adaptive feature of multimedia applications were used to further improve the efficiency of resource utilization. However, the performance analysis only focuses on the RT service class. The NRT applications, serviced by the best-effort model, are just involved in the simulation part.

Huang et al. [5] proposed an analytical model based on the movable-boundary scheme that dynamically (not adaptively) adjusts the number of channels for voice and data traffic to study the performance of an integrated voice/data network with a finite data buffer, which leads the bandwidth to be utilized efficiently while satisfying the QoS requirements for voice and data traffic. The limitation is that there is no priority for handoff calls, so no service performance can be distinguished between new calls and handoff calls. Li et al. [6] studied and compared the performance of two bandwidth allocation schemes for an integrated voice/data cellular system, namely dynamic partition (DP) and dual-threshold bandwidth reservation (DTBR) schemes. They demonstrated that both schemes can achieve comparable performance by proper manipulation of control parameters. The trade-off is that the DP scheme can more easily achieve the target QoS requirement at the expense of some over-provisioning, and thus potentially leads to lower channel efficiency when compared with the DTBR scheme. Compared with Reference [5], the handoff voice calls are distinguished from the new voice calls in Reference [6], but there is still no service differentiation for data traffic. In mobile cellular networks, since the system does not have sufficient resources to provide good service to all users, it should perform service differentiation which aims at providing better service to higher-priority users, such as handoff users. Thus, complete service differentiation is very important in designing a channel allocation scheme.

In this paper, we propose a new channel allocation scheme, called the dynamic partition with pre-emptive priority (DPPP) scheme, for a multi-service mobile cellular network. Compared with the adaptive bandwidth allocation schemes in References [3, 4] (which need tremendous
computational complexity and processing delay, and thus have an extremely high CPU processing cost in the base station), the DPPP scheme is more similar to the scheme in Reference [5] and the DP scheme in Reference [6], where the bandwidth allocation is not adaptive, i.e. it does not adapt instantaneously to the variable traffic load, and thus saves much computational cost. Furthermore, it can be adjusted manually to adapt to the variable load. The main features of the DPPP scheme are highlighted as follows:

(i) A pre-emptive priority (PP) mechanism is employed. The NRT traffic can use all the channels in the sharing region, but it is subject to the pre-emption of the RT traffic. In such a way, the QoS requirement of the RT traffic can be guaranteed in the high intensity of RT traffic; and the high channel utilization can be maintained in the low intensity of RT traffic. The analysis and comparison between the two schemes of References [5, 6] and the DPPP scheme will be elaborated upon later.

(ii) A dynamic partition (DP) concept [6] (which is extended from the movable-boundary scheme [5] to differentiate handoff voice calls) is extended to distinguish the performance between new RT calls and handoff RT calls, as well as new NRT calls and handoff NRT calls. Thus, the complete service differentiation can be achieved between new calls and handoff calls from different traffic classes. The service-differentiation capability is an important feature for future wireless mobile systems.

(iii) The performance analysis focuses on the RT and NRT traffic classes (rather than only RT traffic in Reference [3]). The NRT traffic has a fixed bandwidth (e.g. one channel), unlike the best-effort type; and it can be controlled through adjusting the size of the NRT-only-region.

The remainder of the paper is organized as follows: Section 2 proposes the model description and the relationship between the DPPP scheme and the existing schemes; Section 3 presents the detailed performance analysis of the DPPP scheme; Section 4 presents the numerical results of the DPPP scheme and the performance comparison between it and the existing schemes; and Section 5 concludes the paper.

2. MODEL DESCRIPTION

Consider a mobile network with a certain number of cells, in which a mobile sets up a call connection through the base station. Each base station keeps associated traffic classes, QoS profiles, and channel information, etc.

Suppose each cell has \( M \) channels that serve four types of traffic: new RT, handoff RT, new NRT and handoff NRT traffic. The \( M \) channels are dynamically assigned to three regions: the RT-only-region, the NRT-only-region and the channel sharing region, as shown in Figure 1. The RT-only-region has \( M_1 \) channels for the RT traffic exclusively, the NRT-only-region has \( M_2 \) channels for the NRT traffic exclusively, and the sharing region has \( M - M_1 - M_2 \) channels for both the RT and NRT traffic. For simplicity, we further assume that either a RT call or an NRT call occupies one channel. Readers are referred to a separate work [7] for the variable bandwidth case.

Since the NRT traffic can usually tolerate some degree of service degradation while the RT traffic is more delay-sensitive, we introduce a PP mechanism for the model, i.e. a RT call can
randomly pre-empt an NRT call in the sharing region under certain conditions (see the detailed channel allocation strategy below). To compensate the NRT traffic, we also introduce a victim buffer\(^3\) to temporarily store the pre-empted NRT call and ensure that it goes back again if it can get an idle channel within its queuing time. Because the number of pre-empted calls is at most \(M - M_1 - M_2\), we set the victim buffer size as \(M - M_1 - M_2\).

In order to give some priority to handoff RT calls, we introduce two thresholds \(m_{\text{RT}}\) and \(m_{\text{NRT}}\) in the sharing region for the RT calls and NRT calls, respectively. When the RT-only-region is full and a new RT call arrives, if the number of ongoing RT calls in the sharing region is less than \(m_{\text{RT}}\), the new RT call is accepted; otherwise, it is blocked. The handoff RT call is always accepted unless all the channels in the sharing region are occupied by ongoing RT calls upon its arrival. Similarly, when the NRT-only-region is full and a new NRT call arrives, if the number of ongoing NRT calls in the sharing region is less than \(m_{\text{NRT}}\) and there is an idle channel available, the new NRT call is accepted; otherwise, it is blocked. When a handoff NRT call arrives, if only there is an idle channel in the sharing region, the handoff NRT call is accepted; otherwise, it is dropped. It is noteworthy that, in the above call-admission control process, \(m_{\text{RT}}\) is set from the point of view of the RT traffic and \(m_{\text{NRT}}\) is set from the point of view of the NRT traffic; both of them are set within the same channel pool (the sharing region). Also, \(m_{\text{RT}}\) and \(m_{\text{NRT}}\) are not specific channels, just the number. As a result, the RT traffic strictly complies with the above call-admission rules by ignoring the existence of the NRT traffic (due to the PP mechanism). However, the NRT traffic has to consider the existence of the RT traffic and may sometimes not be able to comply with the above rules. In other words, the NRT traffic can only comply with the rules in the range of those channels not occupied by the RT traffic. This unique feature can bring the following advantage that the existing schemes cannot provide: when there is less RT traffic or no RT traffic in the sharing region, the NRT traffic can use the idle channels as many as possible to improve the utilization of the system; after that, when more RT traffic arrives in the sharing region, the system does not need to worry that the RT traffic has no idle channels to use. Hence,

---

\(^3\)Here the ‘victim buffer/cache’ concept is borrowed from computer architecture, which is used to hold blocks that are thrown out of the main cache, the evicted entries have high likelihood to be accessed again in the near future.
the detailed call-admission control and channel allocation strategy for the DPPP scheme is as follows.

- **When a new RT call arrives,** it will first enter the RT-only-region if the RT-only-region is not full. If the RT-only-region is full, the new RT call will enter the sharing region by getting an idle channel or randomly pre-empting an ongoing NRT call as long as the number of channels occupied by ongoing RT calls in the system is less than \( M_1 + m_{RT} \). Otherwise, the new RT call will be blocked.

- **When a handoff RT call arrives,** it will first enter the RT-only-region if the RT-only-region is not full. If the RT-only-region is full, the handoff RT call will enter the sharing region by getting an idle channel or randomly pre-empting an ongoing NRT call as long as the ongoing RT calls have not yet occupied all the channels of the sharing region. Otherwise, it will be dropped.

- **When a new NRT call arrives,** it will first enter the NRT-only-region if the NRT-only-region is not full. If the NRT-only-region is full, the new NRT call will enter the sharing region as long as there is an idle channel there and the number of channels occupied by ongoing NRT calls in the system is less than \( M_2 + m_{NRT} \). Otherwise, it will be blocked.

- **When a handoff NRT call arrives,** it will first enter the NRT-only-region if the NRT-only-region is not full. If the NRT-only-region is full, the handoff NRT call will enter the sharing region as long as there is an idle channel in the sharing region. Otherwise, it will be dropped.

- **When a RT call is completed,** the system will check how many channels are used by RT calls at that instant. If the number is greater than or equal to \( M_1 \), the system will direct the released channel to the sharing region; otherwise, it will direct it to the RT-only-region. Similarly, when an NRT call is completed, if the number is greater than or equal to \( M_2 \), the system will direct the released channel to the sharing region; otherwise, it will direct it to the NRT-only-region. If a channel is released in the system and there are some pre-empted NRT calls in the victim buffer, then the head call of the queue in the buffer will connect back to the system according to FCFS (first come first served) discipline\(^*\). If the victim buffer is empty, the released channel will be made available for future requests.

**Remark 1**

It is noteworthy that the basic model can be extended to a number of variations such as using a buffer to further store the associated RT or NRT request when the requested channel is not available, which will further reduce the corresponding blocking or dropping probability [5, 8]. It is also worthy to note that the DPPP scheme is applied to the network layer and above and the ‘channel’ concept here refers to a generic logical channel, which could be a physical channel or a time slot.

\(^*\)Note that if there are some pre-empted NRT calls in the victim buffer, a RT call can only be released to the sharing region due to the channel-release mechanism; an NRT call may be released to the sharing region or NRT-only-region. Thus, the head call of the queue can connect back if only there is a released channel in the system.
Remark 2
If we remove the PP mechanism and the victim buffer, and let $m_{RT} = m_{NRT} = M - M_1 - M_2$, then the DPPP scheme (with a buffer to queue the NRT traffic) will become the same as the scheme in Reference [5]. The thresholds $m_{RT}$ and $m_{NRT}$ are used to give higher priority to handoff traffic. The PP mechanism is used to guarantee the QoS of the RT traffic, especially at the case of high NRT traffic intensity. The victim buffer is used to compensate the pre-empted NRT calls through temporarily holding them for possible re-connection to the system in the near future. As a result, compared with the scheme in Reference [5], the DPPP scheme is more flexible to variable traffic environments and has more advantages, such as the QoS-guarantee and complete service-differentiation capabilities.

Remark 3
If we remove the PP mechanism and the victim buffer, and let $m_{NRT} = M - M_1 - M_2$ (i.e. the service differentiation for the NRT traffic disappears), $M = C, M_2 = K_2$ and $M_1 = K_1$ (where $C$, $K_1$ and $K_2$ are the corresponding parameters in Reference [6]), then the DPPP scheme will become similar to the DP scheme in Reference [6] with the exception of the domain of the guard channels. In Reference [6], the guard channels are reserved in the RT-only-region, while in our case they are reserved in the sharing region. When there are no handoff RT requests, the guard channels in Reference [6] will be wasted, while in our case they may be exploited by the NRT requests. Thus, the guard channels in the RT-only-region would have lower channel utilization than that in the sharing region. Note that setting the guard channels in the sharing region may risk losing some QoS of handoff RT calls; however, this risk is removed in our case due to the PP mechanism. If we further let $K_3 = K_1$ (in DP scheme) and $m_{RT} = M - M_1 - M_2$ (in DPPP scheme), then the two schemes will become exactly the same.

3. PERFORMANCE ANALYSIS

In this section, we analyse the system performance by assuming that all cells are statistically identical, so we can focus our analysis on one cell.

In each cell, the arrivals of new RT, handoff RT, new NRT and handoff NRT traffic are assumed to be Poisson distribution with rate $\lambda^{h}_{RT}$, $\lambda^{h}_{RT}$, $\lambda^{n}_{NRT}$ and $\lambda^{h}_{NRT}$, respectively. Thus, the total arrival rate for the RT traffic is $\lambda_{RT} = \lambda^{h}_{RT} + \lambda^{n}_{RT}$, and for the NRT traffic is $\lambda_{NRT} = \lambda^{n}_{NRT} + \lambda^{h}_{NRT}$. The requested call holding time and cell residence time are assumed to be exponentially distributed with mean $1/\mu_{RT}$ and $1/\nu_{RT}$ for the RT traffic, and $1/\mu_{NRT}$ and $1/\nu_{NRT}$ for the NRT traffic, respectively. Thus, it is easy to show that the channel occupancy time is exponentially distributed with mean $1/\mu_{RT} = 1/(h_{RT} + r_{RT})$ for the RT traffic and $1/\mu_{NRT} = 1/(h_{NRT} + r_{NRT})$ for the NRT traffic, respectively. The above assumptions have been found to be valid for a wide range of systems under certain conditions [9], and have been widely used in the literature [3–6, 8, 10–13]. Note that $\lambda^{n}_{RT}$ and $\lambda^{h}_{NRT}$ are determined by the new arrival rates and related probabilities and will be elaborated upon later.

The scheme can be modelled by a Markov process. Let $X(t)$ be the number of RT calls being served and $Y(t)$ be the number of NRT calls in the system (including the NRT calls being served and those in the buffer) at time $t$. We know that $(X(t), Y(t))$ is a two-dimensional Markov process with state-space $\{(n_1, n_2) | 0 \leq n_1 \leq M - M_2, 0 \leq n_2 \leq M - M_1\}$. Let $TR(n_1, n_2; n_1', n_2')$
denote the probability transition rate from state \((n_1, n_2)\) to \((n'_1, n'_2)\), then we have
\[
\begin{align*}
\text{TR}(n_1, n_2; n_1 + 1, n_2) &= \begin{cases}
\lambda_{RT}, & 0 \leq n_1 < M_1 + m_{RT}, \quad 0 \leq n_2 \leq M - M_1 \\
\lambda^b_{RT}, & M_1 + m_{RT} \leq n_1 < M - M_2, \quad 0 \leq n_2 \leq M - M_1 \\
0, & \text{otherwise}
\end{cases} \\
\text{TR}(n_1, n_2; n_1 - 1, n_2) &= n_1 \mu_{RT}, \quad 1 \leq n_1 \leq M - M_2, \quad 0 \leq n_2 \leq M - M_1
\end{align*}
\]

\[
\begin{align*}
\text{TR}(n_1, n_2; n_1, n_2 + 1) &= \begin{cases}
\lambda_{NRT}, & 0 \leq n_1 \leq M - M_2, \quad 0 \leq n_2 < \min(M - n_1, M_2 + m_{NRT}) \\
\lambda^b_{NRT}, & 0 \leq n_1 < M - M_2 - m_{NRT} \\
M_2 + m_{NRT} \leq n_2 < \min(M - n_1, M - M_1) \\
0, & \text{otherwise}
\end{cases} \\
\text{TR}(n_1, n_2; n_1, n_2 - 1) &= \begin{cases}
n_2 \mu_{NRT}, & 0 \leq n_1 \leq M - M_2 \\
(M - n_1) \mu_{NRT} + (n_2 - M + n_1) \mu_{NRT}, & M_1 < n_1 \leq M - M_2 \\
M - n_1 < n_2 \leq M - M_1 & \text{otherwise}
\end{cases}
\end{align*}
\]

Let \(\pi(n_1, n_2)\) denote the steady state probability with state \((n_1, n_2)\). The steady state probability vector (ordered lexicographically) is then partitioned as \(\pi = (\pi_0, \pi_1, \ldots, \pi_{M-M_2})\), where \(\pi_n = (\pi(n,0), \pi(n,1), \ldots, \pi(n, M - M_1))\), \(0 \leq n \leq M - M_2\). The vector \(\pi\) is the solution of equations \(\pi Q = 0\) and \(\pi e = 1\), where \(e\) and \(0\) are vectors of all ones and zeros, respectively, and \(Q\) is the infinitesimal generator of the two-dimensional Markov process given by

\[
Q = \begin{bmatrix}
E_0 & B_0 & 0 & \cdots & 0 & 0 & 0 \\
D_1 & E_1 & B_1 & \cdots & 0 & 0 & 0 \\
& \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & D_{M-M_2-1} & E_{M-M_2-1} & B_{M-M_2-1} \\
0 & 0 & 0 & \cdots & 0 & D_{M-M_2} & E_{M-M_2}
\end{bmatrix}
\]

where each submatrix has dimension \((M - M_1 + 1) \times (M - M_1 + 1)\), and is expressed as

\[
B_i = \begin{cases}
\lambda_{RT} I_{M-M_1+1}, & 0 \leq i < M_1 + m_{RT} \\
\lambda^b_{RT} I_{M-M_1+1}, & M_1 + m_{RT} \leq i < M - M_2 \\
\end{cases}
\]

\[
D_i = i \mu_{RT} I_{M-M_1+1}, \quad 1 \leq i \leq M - M_2
\]
where $I_n$ is an $n \times n$ identity matrix. If we denote by $A_{i}(j,k)$ the $j$th row and $k$th column element of matrix $A_i$, $0 \leq j, k \leq M - M_1$, and define $A_{i}(j,k) \equiv 0$ for $j,k < 0$ or $j,k > M - M_1$, then we have

$$A_{i}(j,k) = \begin{cases} 
\hat{\lambda}_{\text{NRT}}, & 0 \leq i \leq M - M_2, \ 0 \leq j < \min(M - i, M_2 + m_{\text{NRT}}), \ k = j + 1 \\
\hat{\lambda}_{\text{NRT}}^h, & 0 \leq i < M - M_2 - m_{\text{NRT}}, \ M_2 + m_{\text{NRT}} \leq j < \min(M - i, M - M_1), \ k = j + 1 \\
(\mu_{\text{NRT}} - (j - M + i)\hat{\lambda}_{\text{NRT}}), & 0 \leq i \leq M - M_2, \ 1 \leq j \leq \min(M - i, M - M_1), \ k = j - 1 \\
-[A_{i}(j - 1) + A_{i}(j + 1)], & 0 \leq i \leq M - M_2, \ 0 \leq j \leq M - M_1, \ k = j \\
0, & \text{otherwise}
\end{cases}$$

By the result of References [12, 14], we obtain the steady state probability of a cell

$$\pi_n = \pi_{n-1} B_{n-1} (-C_n)^{-1} = \pi_0 \prod_{i=1}^{n} [B_{i-1} (-C_i)^{-1}], \ 1 \leq n \leq M - M_2$$

(2)

where $\pi_0$ satisfies $\pi_0 C_0 = 0$ and

$$\pi_0 \left[ I + \sum_{n=1}^{M-M_2} \prod_{i=1}^{n} [B_{i-1} (-C_i)^{-1}] \right] e = 1$$

(3)

and $C_i (0 \leq i \leq M - M_2)$ can be recursively determined by $C_{M-M_2} = E_{M-M_2}$ and

$$C_i = E_i + B_i (-C_{i+1})^{-1} D_{i+1}, \ 0 \leq i \leq M - M_2 - 1$$

(4)

**Remark 4**

As a special case, we remove the sharing region and its corresponding channels, i.e. the handoff RT and new RT traffic have the same priority level, so do the handoff NRT and new NRT traffic. In other words, both new RT calls and handoff RT calls can only use the $M_1$ channels in the RT-only-region, both new NRT calls and handoff NRT calls can only use the $M_2$ channels in the NRT-only-region. Hence, submatrix $B_i$, $D_i$ and $E_i$ become simpler.

$$B_i = \hat{\lambda}_{\text{RT}} I_{M_1} = B, \ 0 \leq i < M_1$$

$$D_i = i \mu_{\text{RT}} I_{M_1} = i D, \ 1 \leq i \leq M_1$$

$$E_i = \begin{cases} 
A - i D - B, & 0 \leq i < M_1 \\
A - i D, & i = M_1
\end{cases}$$
If we denote by \( A_{i}(j,k) \) the \( j \)th row and \( k \)th column element of matrix \( A \), then

\[
A(j,k) =
\begin{cases}
\hat{\lambda}_{\text{NRT}}, & 0 \leq j < M_2, \ k = j + 1 \\
j\mu_{\text{NRT}}, & 1 \leq j \leq M_2, \ k = j - 1 \\
\hat{\lambda}_{\text{NRT}}\delta(j - M_2) - j\mu_{\text{NRT}}, & 0 \leq j \leq M_2, \ k = j \\
0, & \text{otherwise}
\end{cases}
\]

where \( \delta(i) = \begin{cases} 1, & i \neq 0 \\ 0, & i = 0 \end{cases} \)

From Equations (2) and (4), we have

\[
\pi_n = \pi_{n-1} \left[ \frac{1}{n} BD^{-1} \right] \quad \text{and} \quad \pi_n A = 0, \quad (1 \leq i \leq M_1)
\]

which are determined by the fact that \( \pi_0 C_0 = 0 \) and \( \pi_n [C_n + nD] = 0 \). From the above two equations, we have

\[
\pi_n = \frac{1}{n} \left( \frac{\hat{\lambda}_{\text{RT}}}{\mu_{\text{RT}}} \right) \pi_{n-1} = \frac{1}{n!} \left( \frac{\hat{\lambda}_{\text{RT}}}{\mu_{\text{RT}}} \right)^n \pi_0
\]

and

\[
\pi(n,j) = \frac{1}{j} \left( \frac{\hat{\lambda}_{\text{NRT}}}{\mu_{\text{NRT}}} \right) \pi(n,j-1) = \frac{1}{j!} \left( \frac{\hat{\lambda}_{\text{NRT}}}{\mu_{\text{NRT}}} \right)^{j-1} \pi(n,0) = \frac{1}{j!n!} \left( \frac{\hat{\lambda}_{\text{RT}}}{\mu_{\text{RT}}} \right)^n \left( \frac{\hat{\lambda}_{\text{NRT}}}{\mu_{\text{NRT}}} \right)^{j-1} \pi(0,0)
\]

\[
(0 \leq n \leq M_1, 0 \leq j \leq M_2)
\]

The steady state probability of the special case obtained by Equation (6) is consistent with the result derived by other methods given in References [15, 16].

After obtaining the steady state probabilities, it is easy to determine various performance measures, such as the blocking probabilities of the new RT and NRT traffic, the dropping probability of the handoff RT and NRT traffic, the loss probability of pre-empted NRT traffic, the total channel utilization, the mean waiting time in the buffer, and the busy period time.

### 3.1. Some basic performance measures

- **The blocking probability of the new RT traffic.** It is the probability that, upon arrival of a new RT call, the RT-only-region is full and \( m_{\text{RT}} \) channels in the sharing region are occupied by ongoing RT calls, and given by

\[
P_{\text{RT}}^n = \sum_{n_1=M_1+m_{\text{RT}}}^{M-M_2} \sum_{n_2=0}^{M-M_1} \pi(n_1,n_2) = \sum_{n_1=M_1+m_{\text{RT}}}^{M-M_2} \pi_{n_1} e = \pi_0 \sum_{n_1=M_1+m_{\text{RT}}}^{M-M_2} \prod_{i=1}^{n_1} [B_{i-1}(C_i)^{-1}] \cdot e
\]

- **The blocking probability of the new NRT traffic.** It is the probability that, upon arrival of a new NRT call, the NRT-only-region is full, and at the same time, \( m_{\text{NRT}} \) channels in the sharing region are occupied by ongoing NRT calls or there is no idle channel at all in the sharing region, and given by

\[
P_{\text{NRT}}^n = \sum_{n_1=0}^{M-M_2} \sum_{n_2=\min(M-M_1,M_2+m_{\text{NRT}})}^{M-M_1} \pi(n_1,n_2)
\]
The dropping probability of the handoff RT traffic. It is the probability that, upon arrival of the handoff RT call, the RT-only-region is full and all the channels in the sharing region are occupied by ongoing RT calls, and given by

$$P_{RT}^b = \sum_{n_2=0}^{M-M_2} \pi(M-M_2, n_2) = \pi_{M-M_2} \cdot e = \pi_0 \prod_{i=1}^{M-M_2} [B_{i-1}(-C_i)^{-1}] \cdot e$$

The dropping probability of the handoff NRT traffic. It is the probability that, upon arrival of a handoff NRT call, the NRT-only-region is full and all the channels in the sharing region are occupied by ongoing RT and/or NRT calls, and given by

$$P_{NRT}^b = \sum_{n_1=0}^{M-M_1} \sum_{n_2=\min(M-n_1, M-M_1)}^{M-M_2} \pi(n_1, n_2)$$

The total channel utilization. The total channel utilization is defined as the ratio between the mean number of channels serving RT and NRT calls and the total number of channels in the system.

$$\eta = \frac{1}{M} \left\{ \sum_{n_1=0}^{M-M_1} \sum_{n_2=0}^{M-M_2} (n_1 + n_2)\pi(n_1, n_2) + \sum_{n_1=0}^{M-M_1} \sum_{n_2=M_2+1}^{M-M_1} (n_1 + n_2)\pi(n_1, n_2) \\
+ \sum_{n_1=M_1+1}^{M-M_2} \sum_{n_2=M_2+1}^{M-M_2} (n_1 + n_2)\pi(n_1, n_2) + \sum_{n_1=M_1+1}^{M-M_2} \sum_{n_2=M-n_1+1}^{M-M_2} M\pi(n_1, n_2) \right\}$$

3.2. The loss probability of pre-empted data traffic

Sometimes we are interested in the probability that the pre-empted NRT calls in the victim buffer are eventually lost before they connect back to the system because of the mobility, which is called the loss probability of pre-empted NRT traffic. We know that if a RT call arrives to find that there are \( M + j (0 \leq j \leq M - M_1 - M_2 - 1) \) calls in the system (all \( M \) channels are being used by RT/NRT calls and \( j \) pre-empted NRT calls in the buffer) and it satisfies the pre-emption condition (i.e. for a new RT call, the number of channels occupied by RT traffic is less than \( m_{RT} \) and there is at least one NRT call in the sharing region; for a handoff RT call, there is at least one NRT call in the sharing region), then it randomly pre-empts an ongoing NRT call to the buffer, which leads to another state that there are all \( M \) channels being used by RT/NRT calls and \( j + 1 \) pre-empted calls in the buffer. Furthermore, if a pre-empted NRT call arrives to the buffer with all \( M \) RT/NRT calls in the system and \( j \) pre-empted calls in the buffer, then it could connect back to the cell according to FCFS discipline only if \( j + 1 \) calls leave the cell within its queuing time, i.e. its cell residence time (since the only reason for a pre-empted call to be lost is that the mobile moves out of the cell). Therefore, if we denote by \( \tau \) the queuing time of the data call in the buffer, then \( \tau \) is of exponential distribution with rate \( \varphi_{NRT} \). If we denote by \( \varphi_j \) \((0 \leq j \leq M - M_1 - M_2 - 1)\) the interval from the epoch that a pre-empted call arrives to find that \( j \) pre-empted calls already waiting in the buffer, \( n_1 \) RT calls and \( M - n_1 \) NRT calls being served in the system \((M_1 + 1 \leq n_1 \leq M - M_2, 0 \leq j \leq M - M_1 - M_2 - 1)\), to the epoch that one of the
calls leaves the cell (either leaves a channel or leaves the buffer), then either an ongoing RT call’s completion or an ongoing NRT call’s completion will lead the head call of the queue in the buffer to connect back to the system, and the departure of the head call will lead each of the remaining calls in the buffer to move forward one step. Hence, \( \varphi_j \) is of exponential distribution with rate \( n_1 \mu_{RT} + (M - n_1) \mu_{NRT} + j \rho_{NRT} \). Specifically, if the loss probability of pre-empted NRT traffic, \( P_{\text{pre}} \), is defined as the percentage of those pre-empted calls eventually lost to the total NRT calls in the system, then we have

\[
P_{\text{pre}} = \frac{\sum_{n_1=M_1+1}^{M-M_2} \sum_{j=0}^{n_1-M_1-1} (j+1) \pi(n_1, M - n_1 + j + 1) P\{ \tau < \varphi_0 + \varphi_1 + \cdots + \varphi_j \}}{\sum_{n_1=0}^{M-M_2} \sum_{n_2=1}^{M-M_1} n_2 \pi(n_1, n_2)}
\]

If we denote by \( f_j(\cdot) \) the density function of \( \varphi_j \) and by \( f^*(s) \) the Laplace transform of the function \( f(\cdot) \), by the independent assumption of the random variables, we have

\[
P\{ \tau < \varphi_0 + \varphi_1 + \cdots + \varphi_j \} = 1 - \prod_{i=0}^{j} f_i^*(r_{NRT}) = \frac{(j+1) \rho_{NRT}}{n_1 \mu_{RT} + (M - n_1) \mu_{NRT} + (j+1) \rho_{NRT}}
\]

where

\[
f_i^*(r_{NRT}) = \frac{n_1 \mu_{RT} + (M - n_1) \mu_{NRT} + i \rho_{NRT}}{n_1 \mu_{RT} + (M - n_1) \mu_{NRT} + (j+1) \rho_{NRT}}, \quad 0 \leq i \leq j
\]

From the above results, we obtain

\[
P_{\text{pre}} = \frac{\sum_{n_1=M_1+1}^{M-M_2} \sum_{j=0}^{n_1-M_1-1} (j+1)^3 \rho_{NRT}}{\sum_{n_1=0}^{M-M_2} \sum_{n_2=1}^{M-M_1} n_2 \pi(n_1, n_2)} \pi(n_1, M - n_1 + j + 1)
\]

(12)

3.3. The mean waiting time of the pre-empted calls in the victim buffer

We have known that if a RT call arrives to find that there are \( M + j \) (\( 0 \leq j \leq M - M_1 - M_2 - 1 \)) calls in the system (all M channels are being used by RT/NRT calls and \( j \) pre-empted calls in the buffer) and it satisfies the pre-emption condition, then it can randomly pre-empt an ongoing NRT call to the buffer, the pre-empted call could connect back with one released channel according to FCFS discipline only if \( j + 1 \) calls leave the cell (either leaves a channel or leaves the buffer) within its queuing time. In steady state, the mean number of pre-empted data calls in the victim buffer and the mean arrival rate to the buffer can be easily calculated as

\[
EL_b = \sum_{n_1=M_1+1}^{M-M_2} \sum_{n_2=M-n_1+1}^{M-M_1} (n_2 - M + n_1) \pi(n_1, n_2)
\]

\[
E\lambda_b = \lambda_{RT} \sum_{n_1=M_1}^{M-M_1+m_{RT}-1} \sum_{n_2=M-n_1}^{M-M_1} \pi(n_1, n_2) + \lambda_{RT}^h \sum_{n_1=M_1+m_{RT}}^{M-M_1} \sum_{n_2=M-n_1}^{M-M_1} \pi(n_1, n_2)
\]

By Little’s law, we determine the mean waiting time of the pre-empted data calls in the buffer as

\[
EW_b = \frac{EL_b}{E\lambda_b}
\]

(13)
3.4. The handoff rates

From the previous analysis, we have obtained the new NRT blocking probability $P_{\text{NRT}}^n$, handoff NRT dropping probability $P_{\text{NRT}}^h$, and the loss probability of pre-empted NRT traffic $P_{\text{pre}}$. Since the NRT traffic is pre-empted randomly by the RT traffic, it is obvious that the pre-empted NRT calls include new NRT and handoff NRT calls. From the statistical viewpoint, the total new call blocking probability and handoff dropping probability for the NRT traffic, denoted by $P_{\text{NRT}}^N$ and $P_{\text{NRT}}^H$, should be, respectively,

$$P_{\text{NRT}}^N = P_{\text{NRT}}^n + (1 - P_{\text{NRT}}^n) \frac{\lambda_{\text{NRT}}^n}{\lambda_{\text{NRT}}^n + \lambda_{\text{NRT}}^h} P_{\text{pre}}$$ (14)

$$P_{\text{NRT}}^H = P_{\text{NRT}}^h + (1 - P_{\text{NRT}}^h) \frac{\lambda_{\text{NRT}}^h}{\lambda_{\text{NRT}}^n + \lambda_{\text{NRT}}^h} P_{\text{pre}}$$ (15)

Since the probabilities that an accepted RT and NRT call will attempt handoff operations are $a_{\text{RT}} = r_{\text{RT}}/(h_{\text{RT}} + r_{\text{RT}})$ and $a_{\text{NRT}} = r_{\text{NRT}}/(h_{\text{NRT}} + r_{\text{NRT}})$, respectively, the rates of the handoff RT and NRT calls departing from a cell, denoted by $\lambda_{\text{RT}}^h$ and $\lambda_{\text{NRT}}^h$, are, respectively,

$$\lambda_{\text{RT}}^h = a_{\text{RT}}[(1 - P_{\text{RT}}^n)\lambda_{\text{RT}}^n + (1 - P_{\text{RT}}^h)\lambda_{\text{RT}}^h]$$

and

$$\lambda_{\text{NRT}}^h = a_{\text{NRT}}[(1 - P_{\text{NRT}}^n)\lambda_{\text{NRT}}^n + (1 - P_{\text{NRT}}^h)\lambda_{\text{NRT}}^h]$$

where $\lambda_{\text{RT}}^h$ and $\lambda_{\text{NRT}}^h$ are the handoff arrival rates. We know that the handoff arrival rate to a cell consists of the sum of all the handoff departure rates from its neighbouring cells. For a homogeneous cellular network, the handoff departure rate from each cell is supposed to be statistically identical, and users leave a cell to each of its neighbouring cells with equal probability. Hence, the handoff arrival rate into a cell is equal to handoff departure rate out of the cell in steady state, i.e. $\lambda_{\text{RT}}^h = \lambda_{\text{RTC}}$, and $\lambda_{\text{NRT}}^h = \lambda_{\text{NRTC}}$.

From the above equations, we determine

$$\lambda_{\text{RT}}^h = \frac{r_{\text{RT}}(1 - P_{\text{RT}}^n)\lambda_{\text{RT}}^n}{h_{\text{RT}} + r_{\text{RT}}P_{\text{RT}}^h}$$ (16)

and

$$\lambda_{\text{NRT}}^h = \frac{r_{\text{NRT}}(1 - P_{\text{NRT}}^n)\lambda_{\text{NRT}}^n}{h_{\text{NRT}} + r_{\text{NRT}}P_{\text{NRT}}^h}$$ (17)

We observe that the handoff rates obtained by Equations (16) and (17) have the following relationship: on one hand, $\lambda_{\text{RT}}^h$ is a function of $P_{\text{RT}}^n$ and $P_{\text{RT}}^h$, $\lambda_{\text{NRT}}^h$ is a function of $P_{\text{NRT}}^n$ and $P_{\text{NRT}}^h$. On the other hand, $P_{\text{RT}}^n$, $P_{\text{RT}}^h$, $P_{\text{NRT}}^n$ and $P_{\text{NRT}}^h$ are functions of $\lambda_{\text{RT}}^h$ and $\lambda_{\text{NRT}}^h$, respectively. Therefore, it is not easy to calculate $\lambda_{\text{RT}}^h$ and $\lambda_{\text{NRT}}^h$ independently. We extend the iterative technique proposed in Reference [13] to calculate the handoff rate $\lambda_{\text{RT}}^h$ and $\lambda_{\text{NRT}}^h$ jointly, and then calculate different performance measures. Figure 2 shows an algorithm example for computing the handoff rates and some performance measures.
3.5. The busy period time in the cell

Here we introduce a new performance measure, i.e. the busy period time of each type of traffic. The busy period time of the RT traffic, denoted by $B_{v}$, is the time period for the RT traffic to stay in the system, i.e. the duration starting from the epoch that a cell without any RT traffic is connected with a RT call (either new call or handoff call) to the epoch that the cell is without any RT traffic again. Intuitively, the busy period time of each type of traffic is interesting and is a practical measure for the network providers. The network providers may require this information to evaluate the cost-effectiveness and design a system, which allows an appropriate number for each type of traffic in the system.

Let $N_{R}(t)$ be the number of RT calls being served and $J_{R}(t)$ be the number of NRT calls in the cell at time $t$. It is easy to prove that $(N_{R}(t), J_{R}(t))$ is a two-dimensional absorbing Markov process with absorbing state $\{(0, j) | j = 0, 1, \ldots, M - M_{1}\}$ and the infinitesimal generator matrix

\[
Q_{B} = \begin{bmatrix}
0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
D_{1} & E_{1} & B_{1} & \cdots & 0 & 0 & 0 \\
0 & 0 & 0 & \cdots & D_{M-M_{2}-1} & E_{M-M_{2}-1} & B_{M-M_{2}-1} \\
0 & 0 & 0 & \cdots & 0 & D_{M-M_{2}} & E_{M-M_{2}}
\end{bmatrix}
\]

(18)
where each submatrix is the same as that in Equation (1). The busy period time of the RT traffic in the cell is just the first absorbing time of the Markov process \((N_B(t), J_B(t))\) starting from the initial state \((0, \theta_1)\) to the absorbing state, where \(\theta_1\) is a row vector given by 
\[
\theta_1 = (\pi_1 / \pi(e, 0, \ldots, 0)).
\]
If we denote by 
\[
TB = \begin{bmatrix}
E_1 & B_1 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & D_{M-M_2-1} & E_{M-M_2-1} & B_{M-M_2-1} \\
0 & 0 & \cdots & 0 & D_{M-M_2} & E_{M-M_2}
\end{bmatrix}
\]
from References [17, Lemma 2.2.2] and [12], we can determine the distribution of \(B_x\)
\[
P(B_x \leq x) = 1 - \theta_1 \exp(T_B x) e \quad \text{for } x \geq 0
\]
and the non-central moments, \(EB^k_x\), is
\[
EB^k_x = (-1)^k k! (\theta_1 T_B^{-k} e) \quad \text{for } k \geq 0
\]
The busy period time of the NRT traffic can be found by using the same method.

4. NUMERICAL RESULTS AND DISCUSSIONS

In this section, we present some numerical results to study the performance of the DPPP scheme and to compare it with the existing schemes. The system configuration is set as follows. The total channel number in each cell \(M\) is 30, \(h_{RT} = 0.004\) s\(^{-1}\), \(r_{RT} = 0.001\) s\(^{-1}\), \(h_{NRT} = 0.0016\) s\(^{-1}\), \(r_{NRT} = 0.0004\) s\(^{-1}\). The handoff arrival rates of the RT and NRT traffic are calculated by the iterative algorithm given in Section 3.4. The intensity of the RT and NRT traffic can be varied between 4 and 12, i.e. \(\rho_{RT} = 4 - 12, \rho_{NRT} = 4 - 12\).

We know that in mobile multi-service applications, different traffic classes have different QoS requirements. One of the key QoS measures is the blocking or dropping probability. The blocking or dropping probability in our scheme can be adjusted through the different design parameters \(M_1, M_2, m_{RT}\) and \(m_{NRT}\). However, the QoSs are not usually satisfied at the condition of dynamic traffic load, the PP mechanism is a good selection to guarantee the QoS of some traffic class with higher priority\(^1\) (we will show this in Section 4.3). Specifically, we partition our study into three parts: the impact of \(M_1\) and \(M_2\) on the system performance, the impact of \(m_{RT}\) and \(m_{NRT}\) on the system performance, and the comparison of the DPPP with the existing schemes.

4.1. The impact of \(M_1\) and \(M_2\) on the system performance

To accurately target the impact of \(M_1\) and \(M_2\) on the system performance, we classify the traffic into the RT and NRT traffic classes without distinguishing the handoff case, i.e. \(m_{RT} = m_{NRT} = M - M_1 - M_2\). As an example, we set the target of \(P_{RT}\) (blocking probability

\(^1\)In general, RT traffic has higher priority over NRT traffic; handoff request has higher priority over new request.
of the RT traffic) to be on the order of $10^{-2}$ at the maximum traffic intensity (i.e. $\rho_{RT} = \rho_{NRT} = 12$), and study the system performance by the different probabilities.

Figures 3–5 show how the different probabilities depend on the change of $M_2$ in different RT traffic intensities when $M_1$ is fixed. Figure 3 presents that $P_{RT}$ increases with the increase of $M_2$; Figure 4 presents that $P_{NRT}$ decreases slightly with the increase of $M_2$. The reason is that when $M_2$ increases, the RT traffic will use fewer channels in the system, and the NRT traffic will use more channels in the NRT-only-region, but fewer channels in the sharing region. Since the NRT traffic in the NRT-only-region is not pre-empted by the RT traffic, the final result is that $P_{NRT}$ decreases a little. Figure 5 presents that $P_{pre}$ decreases with the increase of $M_2$. This can be explained as: when $M_2$ increases, the channels possibly used by the NRT traffic in the sharing region will be reduced, which leads to fewer chances for the NRT traffic to be pre-empted.

The above figures indicate that, in order to maintain the target of $P_{RT}$ at the maximum traffic intensity, a certain amount of channels ($M-M_2$) need to be designated appropriately for the RT traffic. It is noteworthy that, due to the introduction of PP mechanism, the channels in the sharing region can be shared by both RT and NRT traffic to improve the utilization, at the same time, the QoS of the RT traffic can be maintained. As far as the target of $P_{RT}$ is concerned, the appropriate value of $M_2$ can be 6, 8, 10, and 12. But the larger the value of $M_2$, the more benefits the NRT traffic will obtain. Hence, we choose $M_2 = 12$ to study the impact of $M_1$ on the system performance in Figures 6–8.

Figure 6 shows that $P_{RT}$ does not change with the change of $M_1$ when $M_2$ is fixed. This is because that the RT traffic can use every channel in the RT-only-region and the sharing region. Thus, the change of $M_1$ has no impact on $P_{RT}$. However, $M_1$ does have impact on $P_{NRT}$. We can observe from Figure 7 that $P_{NRT}$ decreases with the decrease of $M_1$. The reason is that when
Figure 4. Blocking probability of NRT traffic $P_{\text{NRT}}$ vs $M_2$.

Figure 5. Loss probability of pre-empted traffic $P_{\text{pre}}$ vs $M_2$. 
Figure 6. Blocking probability of RT traffic $P_{RT}$ vs $M_1$.

Figure 7. Blocking probability of NRT traffic $P_{NRT}$ vs $M_1$. 
$M_1$ decreases, the NRT traffic can use more channels in the sharing region (though it is still subject to the RT traffic’s pre-emption), which leads to the decrease of $P_{NRT}$. Figure 8 shows that $P_{pre}$ increases with the decrease of $M_1$. This can be explained as: when $M_1$ decreases, the channels possibly used by the NRT traffic in the sharing region will be increased, which leads to more chances for the NRT traffic to be pre-empted.

From Figures 3–8, we observe that the probabilities increase with the increase of the RT traffic intensity. This is obvious, since when the RT traffic intensity increases, the channels in the RT-only-region and the sharing region are more quickly to be occupied, and thus the RT and NRT traffic are more quickly to be rejected. In addition, we also observe that $P_{pre}$ is negligible by comparing with $P_{NRT}$. This validates the effectiveness of using a victim buffer to compensate the degradation of the NRT traffic in our scheme. Note that the victim buffer is implemented by software and its size can be easily adjusted according to the change of $M_1$ and $M_2$. The main cost is some memory space that needs to be partitioned by the memory management program of the system. Moreover, the resource allocation entity managing the victim buffer and associated DPPP scheme is developed in the base station (rather than the terminal). This extra cost is trivial compared with the great benefit the victim buffer brings.

4.2. The impact of $m_{RT}$ and $m_{NRT}$ on the system performance

From the previous analysis, we know that the performance measures of the system are related to all the parameters $M_1$, $M_2$, $m_{RT}$ and $m_{NRT}$, where $m_{RT}$ and $m_{NRT}$ are used to provide higher priority to the handoff RT and handoff NRT traffic, respectively. In order to accurately target the impact of $m_{RT}$ and $m_{NRT}$ on the system performance and without loss of generality, we
choose a fixed set of $M_1$ and $M_2$, say $(M_1, M_2) = (8, 12)$, and change the values of $m_{RT}$ and $m_{NRT}$ to study the system performance. In addition, we also add the special case in Remark 4 as an intuitive comparison to validate the probability analysis in Section 3.

Figures 9–13 illustrate how the different probabilities depend on the change of $m_{RT}$ when $m_{NRT}$ is fixed. When $m_{RT}$ increases, the blocking probability of new RT traffic $P^b_{RT}$ decreases (shown in Figure 9); while the blocking probability of new NRT traffic $P^b_{NRT}$ increases (shown in Figure 10). The reason is that when $m_{RT}$ increases, the new RT traffic can use more channels in the sharing region, and the channels in the sharing region are more quickly to be used up, and thus the new NRT traffic is more quickly to be rejected. Figure 11 shows that the dropping probability of handoff RT traffic $P^h_{RT}$ increases with the increase of $m_{RT}$. This is because when $m_{RT}$ increases, more new RT traffic can enter the sharing region to contend with the handoff RT traffic. Figure 12 shows that the dropping probability of handoff NRT traffic $P^h_{NRT}$ increases with the increase of $m_{RT}$. The reason is the same as that in Figure 10.

From Figures 9–12, we also observe that the probabilities are far worse in the case of Remark 4 than in other cases. This is intuitive, since at Remark 4 case the new RT/NRT traffic and handoff RT/NRT traffic cannot use any channel in the sharing region.

Figure 13 shows that the loss probability of pre-empted NRT traffic $P_{pre}$ increases with the increase of $m_{RT}$. This is because when $m_{RT}$ increases, the channels in the sharing region are more quickly to be used up, and the NRT traffic is more quickly to be pre-empted, and thus the pre-empted calls have a larger probability to be lost. Moreover, when $m_{RT}$ increases, the pre-empted calls are more difficult to connect back to system because of the larger amount of RT calls in the sharing region. Anyway, by comparing with $P^b_{NRT}$ in Figure 10 and $P^h_{NRT}$ in Figure 12, we can find that $P_{pre}$ is negligible. For instance, in the case of $m_{RT} = 6$ and $m_{NRT} = 5$, we have

![Figure 9. Blocking probability of new RT traffic.](image-url)
Figure 10. Blocking probability of new NRT traffic.

Figure 11. Dropping probability of handoff RT traffic.
Figure 12. Dropping probability of handoff NRT traffic.

Figure 13. Loss probability of pre-empted NRT traffic.
\( P_{\text{NRT}} = 0.0485, \quad P_{\text{NRT}}^h = 0.0181, \quad \text{while} \quad P_{\text{pre}} = 2.0758 \times 10^{-6} \). This again validates the effectiveness of the victim buffer. In addition, to quantitatively study the compensation performance of the victim buffer, we calculate the mean waiting time of the pre-empted calls in the victim buffer under the specific system configuration. As shown in Figure 14, the mean waiting time is tolerable for the delay-insensitive NRT traffic.

Figure 15 shows that the total channel utilization increases with the increase of RT traffic intensity and \( m_{\text{RT}} \), respectively. It is obvious that when the RT traffic intensity increases, more channels in the system will be occupied per unit time. As \( m_{\text{RT}} \) increases, the channel utilization just slightly increases. The reason is that the channel utilization is only concerned about the mean number of channels occupied in the system, no matter which of the channels is occupied by RT/NRT traffic or new/handoff traffic. Hence, the impact of \( m_{\text{RT}} \) is not so significant on the channel utilization.

Figure 16 shows the impact of \( m_{\text{RT}} \) on the busy period time of the RT traffic at different \( \rho_{\text{RT}} \)'s. We use the mean of the busy period time as the measure and normalize it through dividing each value by the sum of all the values corresponding to different \( m_{\text{RT}} \)'s at every point of x-axis. It can be seen that, under the specific system configuration, the mean of \( B_v \) increases slightly with the increase of \( m_{\text{RT}} \) and does not change in low \( \rho_{\text{RT}} \)'s. This may be explained as: as \( m_{\text{RT}} \) increases, more new RT traffic can enter the sharing region; at the same time, less handoff RT traffic can enter the sharing region. As a result, there is only a slight increase for the duration of the RT traffic staying in the system. In general, the impact of \( m_{\text{RT}} \) on \( B_v \) is negligible, especially at the low RT traffic intensity.

Similar results can be obtained for the change of \( m_{\text{NRT}} \). Due to length limitation, we do not present details here.
Figure 15. The total channel utilization.

Figure 16. The busy period time of the RT traffic.
4.3. The comparison of the DPPP scheme with the existing schemes

In this subsection, we compare the performance of the DPPP scheme and the existing schemes, i.e. the scheme in Reference [5] and the DP scheme in Reference [6]. For the sake of fair comparison and simple analysis, we do not distinguish the handoff case. This will not affect the comparison since the handoff performance is obtained by simply reserving some guard channels in the DP and DPPP scheme and there is no consideration of handoff calls at all in Reference [5]. As a result, we have $K_3 = K_1$ in the DP scheme in Reference [6] (which leads to the same as that in Reference [5]), and $m_{RT} = m_{NRT} = M - M_1 - M_2$ in the DPPP scheme. For simplicity, we further consider the schemes without queuing, since the purpose of queuing is just to further reduce the associated blocking or dropping probability [5, 8]. In the following figures, ‘with PP’ corresponds to the DPPP scheme; ‘without PP’ corresponds to the DP scheme (or the scheme in Reference [5]).

Figure 17 shows how the blocking probability of the RT traffic $P_{RT}$ depends on the change of the NRT traffic intensity $\rho_{NRT}$ for the two schemes in different conditions. We can observe that, in all cases, as $\rho_{NRT}$ increases, the $P_{RT}$ will increase in the DP scheme and not change in the DPPP scheme. The reason is that when $\rho_{NRT}$ increases, more NRT (e.g. data) calls will contend with RT (e.g. voice) calls in the sharing region, which leads to the increase of $P_{RT}$ in the DP scheme. However, $P_{RT}$ in the DPPP scheme will not be affected because of the PP mechanism. In other words, the DPPP scheme can guarantee the QoS of RT traffic at the condition of dynamic load while the DP scheme cannot. As mentioned in Section 4.1, the extra cost for the DPPP scheme is trivial, since it only requires some extra memory space in the base station. Moreover, the DPPP scheme does not cause any extra computational complexity. It even has less complexity than those in References [5, 6], since it does not need to be adjusted frequently in the variable traffic load (please see the following Remark 5 for details).

![Figure 17. Comparison of blocking probability of RT traffic.](image-url)
The QoS-guarantee of the RT traffic in the DPPP scheme is achieved at the expense of some degradation of the NRT traffic, which can be seen in Figure 18. Fortunately, the NRT traffic is usually delay insensitive. Moreover, the loss probability of the pre-empted calls is negligible when compared with the blocking probability of NRT traffic, which means that most of the pre-empted calls could eventually go back to the system. This can be shown by comparing Figures 18 and 19. For instance, at $\rho_{\text{NRT}}=12$, $P_{\text{NRT}}$ is 0.0347, 0.0646 and 0.0561 for the three ‘with PP’ cases (through calculation), and $P_{\text{pre}}$ is $1.6058 \times 10^{-5}$, $3.8784 \times 10^{-5}$ and $2.6003 \times 10^{-5}$, respectively.

We compare the channel utilization in Figure 20 in the cases of $(M_1, M_2) = (10, 10)$ and $(8, 12)$. The channel utilization is the same for the two schemes. This is obvious, since the purpose of the PP mechanism is to make the RT call pre-empt the ongoing NRT call; as far as the channel utilization is concerned, it is not concerned that the channel is occupied by the RT call or NRT call. However, if the handoff traffic is differentiated, the DPPP scheme will have a little higher channel utilization than the DP scheme. This is determined by the fact that the guard channels reserved for handoff RT traffic in the DP scheme are from the RT-only-region; while in the DPPP scheme they are from the sharing region. The reason has been explained in Remark 3.

From the previous analysis and performance comparison, we can obtain a key feature, besides the complete service differentiation, of the DPPP scheme in the following.

**Remark 5**
At the condition of dynamic traffic load, the DPPP scheme can guarantee the QoS of the high-priority traffic while the DP scheme cannot. That is, to guarantee the QoS of RT traffic, the DP
Figure 19. Loss probability of pre-empted NRT traffic.

Figure 20. Comparison of the channel utilization.
scheme needs to be manually adjusted frequently to adapt to the variable load, while the DPPP scheme does not. In a real network, the traffic is possibly dynamic, but it is obviously impossible to manually adjust the parameters too frequently. On the other hand, although the adaptive bandwidth allocation schemes, such as References [3, 4], can adapt instantaneously to the variable traffic, the tremendous computational complexity and processing delay (from the instantaneous adaptation) limit their wide applications in mobile cellular networks. On the contrary, the DPPP scheme does not need to frequently adjust the parameters; it does only if the current configuration plus the PP mechanism cannot keep the QoS of the high-priority traffic, which seldom happens unless the high-priority traffic abruptly expands a lot in the system.

5. CONCLUSIONS

We have developed a new channel allocation scheme, namely the dynamic partition with pre-emptive priority (DPPP) scheme, for multi-service mobile cellular networks. The system performance focused on the RT and NRT traffic classes (including new calls and handoff calls) is analysed by the matrix-analytic method through a two-dimensional Markov process. A PP mechanism is employed to guarantee the QoS requirement of the RT traffic at the expense of degrading some NRT traffic, and the victim buffer compensates the degradation and has no negative impact on the RT traffic. A DP concept is used (with the help of two thresholds $m_{RT}$ and $m_{NRT}$) to achieve the complete service differentiation between new calls and handoff calls from different traffic classes.

From the performance analysis and numerical results, we can get the main conclusions: (a) the DPPP scheme can guarantee the QoS of the higher-priority traffic, while the existing ones cannot; (b) there is a complete service differentiation in the DPPP scheme, while there is not in the existing ones. Hence, the DPPP scheme, compared with the existing schemes, is an effective and practical scheme for multi-service mobile cellular networks.

There are a number of issues in future research: (a) consider variable bandwidth requirements for each type of call (i.e. RT and NRT traffic) in the DPPP scheme due to the trend of multimedia applications in mobile cellular networks; (b) Basically, the RT and NRT traffic in the DPPP scheme is analysed based on circuit-switched services. It is interesting to study the performance based on packet-switched services (e.g. wireless Internet traffic). The modelling and analysis methods of wireless Internet traffic might be referred to Reference [18], where the conceptual issues of wireless Internet traffic source, two traffic models (e.g. binomial or dynamic traffic models) associated with their assumptions and limitations, four channel allocation algorithms based on physical layer, and the throughput performance analysis under different traffic models are discussed.

NOMENCLATURE

- $EW_b$: the mean waiting time of the pre-empted data calls in the victim buffer
- $m_{RT}, m_{NRT}$: the threshold for RT and NRT traffic in the sharing region, respectively
- $M, M_1, M_2$: the number of channels in the system, RT-only-region and NRT-only-region, respectively
- $P_{pre}$: the loss probability of the pre-empted NRT traffic
- $p^R_{RT}, p^R_{NRT}$: the handoff dropping probability for RT and NRT traffic, respectively
The new call blocking probability for RT and NRT traffic, respectively
the mean of the requested call holding time for RT and NRT traffic, respectively
the mean of the cell residence time for RT and NRT traffic, respectively
the total channel utilization
the arrival rate for new and handoff NRT traffic, respectively
the arrival rate for new and handoff RT traffic, respectively
the mean of the channel occupancy time for RT and NRT traffic, respectively

ACKNOWLEDGEMENTS
The authors would like to thank the reviewers for their careful reading of the original manuscript. Their comments and suggestions have led to a much better improvement and presentation of the paper. The work of Wei Li was supported in part by the National Science Foundation under Grant CCF-0515263.

REFERENCES
AUTHORS’ BIOGRAPHIES

**Shensheng Tang** received his BS degree from the Tianjin University and MS degree from the China Academy of Telecommunications Technology (CATT), both in Electrical Engineering. He is presently pursuing his PhD degree in the department of Electrical Engineering and Computer Science at The University of Toledo, Ohio, U.S.A. He has over 8 years of industrial experience in electronics and telecommunications areas, which involves the development of electronic instruments and mobile communication equipments as well as the research related to 3G TD-SCDMA and WCDMA radio technologies. His current research interests focus on wireless communications and networking, including wireless LANs, mobile ad hoc networks, wireless sensor networks, 3G and future generations of wireless cellular networks.

**Wei Li** received his PhD degree from the Chinese Academy of Sciences in 1994. He is currently an Associate Professor in the Department of Electrical Engineering and Computer Science at the University of Toledo, U.S.A. He was once an Assistant Professor, Endowed Chair Professor in Bell South Mobility/BORSF in Telecommunications at the University of Louisiana at Lafayette, and also a recipient of US Air Force Summer Faculty Fellowship. Dr Li's research interests are in the wireless and mobile networks, mobile ad hoc networks, wireless sensor networks and queueing networks, etc. Dr Li has published over 60 peer-reviewed papers in professional journals, over 20 referred papers in the proceedings of professional conferences and 3 books. Dr Li is currently serving as an Editor for EURASIP Journal on Wireless Communications and Networking, for International Journal of Computer and Their Applications, for International Journal of High Performance Computing and Networking, and for International Journal of Sensor Networks. He has also served as a TPC member for some professional conferences such as IEEE ICC'05,04,02, IEEE GlobCom'05,03, IEEE WCNC'05,04,00, IEEE WirelessCom'05, Qshine'05,04, IEEE VTC'03, etc.