An Adaptive Bandwidth Allocation Scheme with Preemptive Priority for Integrated Voice/Data Mobile Networks

Shensheng Tang, *Student Member, IEEE* and Wei Li, *Member, IEEE*

Abstract— **This paper presents an adaptive bandwidth allocation scheme, called Complete Sharing with Preemptive Priority (CSPP) scheme, for integrated voice/data mobile networks. It uses the complete sharing approach to maximize the bandwidth utilization efficiency by allowing data traffic to share all the bandwidth resource of a cell and introduces a preemptive priority mechanism to maintain the high performance of voice traffic while the injury of data traffic is compensated through a victim buffer. A degradation/compensation mechanism is applied to the data traffic which has elastic characteristics. The model of the CSPP scheme is analyzed by a two dimensional Markov process. The steady state probability vector is obtained recursively by the matrix-analytic method and many important performance measures are then determined such as different types of probabilities, total channel utilization, transmission delay, busy period time and blocking period time. From the performance analysis and comparison as well as the numerical results, we can conclude that the CSPP scheme can achieve complete service differentiation for different traffic classes and very high and stable channel utilization under various traffic loads. By adjusting the defined constraint factors, we can not only maintain higher priority for handoff voice over new voice calls and voice calls over data calls but also provide necessary protection for data calls in hot-spot data traffic situations.**

Index Terms— **Adaptive bandwidth allocation, Complete sharing (CS), Preemptive priority (PP), Handoff, Integrated voice/data mobile networks.**

I. Introduction

CURRENTLY, wireless and mobile networks are fast evolving from the primary voice-centric communications to many applications of data and multimedia. 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 traffic is. Conversational class (eg. voice) and streaming class (eg. streaming video) are mainly used to carry real-time traffic flows. Interactive and background classes are mainly used by traditional Internet applications like WWW, Email and FTP, and they may have elastic traffic characteristics, i.e., they can take bandwidth from a varying range of values.

One of the results of these applications is leading to more bandwidth requirement per user, which limits the capacity of the system. The trend in cellular networks is to shrink cell size to improve the system capacity through effective frequency reuse. This potentially results in more frequent handoffs and makes connection-level QoS more difficult to achieve. Handoff calls have an important effect on system performance and are usually given higher priority to maintain lower dropping probability. This, however, usually comes at the expense of high call blocking probability and potentially poor channel utilization [2]. Moreover, the scarce bandwidth resource has always been the bottleneck of wireless and mobile communications. Thus, the problem formulation is to maximize the resource utilization and at the same time to guarantee different QoS requirements for different service classes. To solve this problem, an effective and efficient bandwidth allocation strategy is necessary.

Bandwidth allocation strategies have been studied extensively in literature for single service. The authors in [3] provide a comprehensive survey of a large number of published papers in this area. One of the challenges in moving to a multi-service system is that the limited bandwidth has to be shared among multiple traffic classes. The 2.5G and 3G cellular systems employ radio resource management (RRM) strategies that allow them not only to assign variable bandwidth to incoming voice or data calls but also to dynamically vary the bandwidth of ongoing calls. These RRM strategies are known as flexible (adaptive) resource allocation (FRA) strategies and their ultimate end is to find a trade-off between the existing capacity and the QoS. They may or may not use degradation/compensation of calls in progress. Degradation involves the gradual reduction of the bandwidth allocated to an ongoing call, while compensation is the reverse process [18].

In [4], Huang et al. propose a scheme based on the movable-boundary that dynamically adjusts the number of channels for voice and data traffic, which allows the bandwidth to be utilized efficiently and satisfies the QoS requirements for voice and data traffic. The limitation is that there is no priority for handoff (both voice and data) calls, so no service differentiation can be distinguished between new calls and handoff calls. In [5], the capacity maximization in multi-service mobile networks by making use of the multiple fractional channel reservation (MFCR) strategy is investigated. An algorithm to determine the maximum cell capacity and the optimum amounts of reserved resources is addressed. A guide on the

Manuscript received September 12, 2004; revised October 11, 2005 and February 10, 2006; accepted March 6, 2006. The associate editor coordinating the review of this paper and approving it for publication was H. Yanikomeroglu.

The authors are with the Department of Electrical Engineering and Computer Science, The University of Toledo, Toledo, OH 43606 USA (e-mail: stang@eng.utoledo.edu, wli@eecs.utoledo.edu).

Digital Object Identifier 10.1109/TWC.2006.04632.

selection of the optimal prioritization order was given due to the huge computational complexity involved in the selection. In [6], the authors propose a call admission control (CAC) policy using the limited fractional guard channel mechanism for a cellular system supporting voice and data services. In [7], a resource management strategy comprising both CAC management and bandwidth management is proposed for heterogeneous adaptive-rate traffic. The CAC management strategy extends the guard channel strategy to obtain priority-based CAC management for adaptive-rate sources; the bandwidth management strategy allows channels to be assigned proportionally to the throughput window declared by users. All the CAC strategies used in [5]–[7] are based on guard channel strategy or its variants. It is well known that, in guard channel strategy, the reservation of "guard channels" directly leads to poor channel utilization. Thus, other CAC strategies that can both satisfy the QoS requirements of higher priority traffic classes and maintain good channel utilization are expected to be developed. In [8], a novel mathematical approach is proposed to evaluate the performance of FRA strategy with uniform QoS provisioning in multi-service cellular networks in terms of the call blocking probabilities and transmission delay. The FRA strategy improves channel utilization by dynamically adjusting user transmission rates. The bandwidth offered to users can be adjusted in accordance with the elasticities of their service types. However, in the cellular systems of [8] no handoff prioritization strategy is used in the traffic analysis due to complexity. In [9] and [10], Li et al. proposed a dual trunk bandwidth reservation (DTBR) scheme, which can provide the necessary service guarantee and service differentiation for voice and data traffic and effectively utilize bandwidth by using a complete sharing approach. It also assumes the elastic characteristics of data traffic. But it still does not give service differentiation between handoff data and new data calls, and the total channel utilization can still be further improved (which will be shown later in this paper).

In this paper, we propose a novel adaptive bandwidth allocation scheme, called complete sharing with preemptive priority (CSPP) scheme, for integrated voice/data mobile networks. It takes into consideration the service differentiation between new calls and handoff calls (both voice and data traffic), and uses the complete sharing approach to further improve the bandwidth utilization by allowing data calls to share all the channels of a cell (note in [9] and [10] data calls can only share some of the channels of a cell due to the existence of contention between data and voice traffic). A degradation/compensation mechanism in [18] is utilized in our scheme to capture elastic data traffic. In order to protect the delay-sensitive voice traffic, we introduce a preemptive priority (PP) mechanism[1] for voice calls (which removes the contention and makes it feasible for data traffic to share all the bandwidth resource). The injury of data traffic by the PP mechanism is compensated through a victim buffer[2]. By adjusting the defined constraint factors, we can not only maintain higher priority for handoff voice over new voice calls and voice calls over data calls but also provide the necessary protection for data calls if required.

The rest of the paper is organized as follows. Section II proposes the novel model and the call admission and bandwidth allocation strategy. Section III presents the detailed performance analysis including the development of a large number of important performance measures. In Section IV we present the numerical results of the CSPP scheme and the performance comparison between CSPP and DTBR schemes as well as the optimal design issues. Finally we conclude the paper in Section V.

II. Model Description and Adaptive Bandwidth Allocation

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 class, QoS profiles, and channel information, etc.

Suppose each cell has N bandwidth units that serve four types of traffic: new voice, handoff voice, new data and handoff data. Assume a voice call takes only one bandwidth unit (e.g., a channel), and a data call takes bandwidth from a range of bandwidth units between b_m and b_M. For simplicity and without loss of generality, we further assume $b_m = 1$.

Since data traffic can usually tolerate some degree of service degradation while voice is more delay-sensitive [9], [10], we introduce a PP mechanism for voice traffic, i.e., a voice call can randomly preempt a data call (new data or handoff data) under certain conditions (see details in the call admission and bandwidth allocation strategy). To compensate data traffic, we introduce a victim buffer to temporarily store the preempted data calls and ensure that they will go back again if there is available bandwidth in the system before its maximum queuing time runs out. More precisely, two constraint factors σ_1 and σ_2 are set for voice calls ($0 \leq \sigma_1 \leq \sigma_2$) so that new voice calls can at most use $\lfloor \sigma_1 N \rfloor$ and handoff voice calls can at most use $\lfloor \sigma_2 N \rfloor$ out of N bandwidth units (where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x). It is worthy to note that a voice call can either obtain a free channel or obtain an occupied channel by a data call through preemption. Sometimes in the hot-spot data traffic area, we need to reserve some channels exclusively for data traffic, thus we use σ_1 and σ_2 to constrain both new and handoff voice calls). We assign $\sigma_1 < \sigma_2$ to maintain some advantage for handoff voice calls. For simple notation, we set $N_1 = \lfloor \sigma_1 N \rfloor$ and $N_2 = \lfloor \sigma_2 N \rfloor$. Because data calls are preempted by at most N_2 voice calls, we set the victim buffer size as N_2.

To improve the channel utilization of the system, we allow data calls to use all the N channels of a cell. This is feasible in the proposed scheme because of the introduction of the PP mechanism (however, it is impossible in the other schemes mentioned previously that do not use the PP mechanism due to

[1] The PP mechanism has been used widely in queuing theory and its applications in cellular systems for example [11], [12] and [13]. In [11], a PP policy is studied in hierarchical cellular network by simulation, but no theoretical analysis is addressed. A PP mechanism provided for only handoff voice calls in [12] and for both new voice and handoff voice calls in [13] is proposed for integrated voice/data cellular systems. However, both the traffic models do not consider any elastic data traffic (i.e., in these models, each data call occupies one channel).

[2] 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.

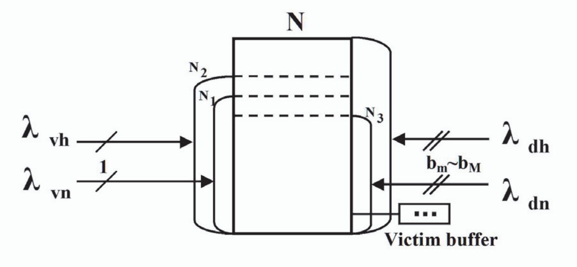

Fig. 1. Block diagram of the basic CSPP scheme.

the contention between different types of traffic). Similarly, to give some higher priority for handoff data calls, we introduce another constraint factor σ_3 ($0 \leq \sigma_3 \leq 1$) so that new data calls can at most use $\lfloor \sigma_3 N \rfloor$ out of N channels while handoff data calls can use all the N channels in the system. We set $N_3 = \lfloor \sigma_3 N \rfloor$. Of course both handoff data and new data calls are subjected to voice calls' preemption under some conditions (see the detailed call admission and bandwidth allocation strategy). Therefore, the prioritization order (from high to low) of different traffic types can be roughly determined as handoff voice calls - new voice calls - handoff data calls - new data calls. In addition, the CSPP scheme can be flexibly applied to various traffic environments by adjusting the three constraint factors. For example, in the hot-spot data traffic situations, the CSPP scheme can properly decrease σ_1 and/or σ_2 to provide necessary protection for data traffic. A simple block diagram for the CSPP scheme is given in Figure 1.

Ideally, it is assumed that each data connection always attempts to get the maximum bandwidth b_M if the available bandwidth can accommodate it; otherwise, the data calls will equally share the bandwidth left over from voice calls through the use of resource reallocations (triggered by call arrivals or departures), but the allowed bandwidth should be at least b_m. This assumption is reasonable and has been used widely in literature [9], [15], [17]–[19]. Thus, the actual allowed bandwidth b for a data call can vary between b_m and b_M. By adjusting the value of b_m and b_M, this elastic data traffic model can be applied in a wide range of applications defined in [1].

We also assume that the rate control of the elastic data traffic in progress is ideal. That is, the data calls in progress can readjust their current bandwidth after an infinitesimal amount of time when any system state changes (call arrivals or departures). This kind of assumption has also been used by other research for analyzing similar problems [9], [15], [19].

Therefore, the detailed call admission and bandwidth allocation strategy for CSPP scheme is highlighted as follows:

- When a new data call arrives, it will enter the system with maximum bandwidth b_M if there is enough bandwidth available and the total occupied bandwidth in the system is less than N_3. If not, it will "squeeze" into the system to get degradation service with actual bandwidth b by elastic characteristic as long as $b \geq b_m$ and the total occupied bandwidth in the system is less than N_3. Otherwise, it will be blocked.
- When a handoff data call arrives, it will enter the system with maximum bandwidth b_M if there is enough bandwidth available. If not, it will "squeeze" into the system to get degradation service with actual bandwidth b by elastic characteristic as long as $b \geq b_m$. Otherwise, it will be dropped.
- When a new (respectively handoff) voice call arrives, it will enter the system if there is at least one channel available and the number of channels occupied by voice calls in progress is less than N_1 (respectively N_2). If not, it will "squeeze" into the system by pushing data calls into degradation service as long as the data calls in progress can get at least the minimum bandwidth b_m. If it cannot "squeeze" into the system, it will enter the system by randomly preempting a data call in progress if the number of channels occupied by voice calls in progress is less than N_1 (respectively N_2). Otherwise, it will be blocked (respectively dropped).
- When a voice call or data call is completed, the system will release all the bandwidth it used. If there are some preempted data calls queuing in the buffer at that instant, then the preempted data calls will connect back to the system according to FCFS (first come first served) discipline. Otherwise, all other data calls still in progress will get compensation service by sharing the released bandwidth equally. If all the data calls in progress have the maximum bandwidth b_M, the remaining bandwidth will be made available for future requests.

Remark 1: It is noteworthy that, i) the basic model can be extended from the one unit bandwidth voice traffic to multi-unit bandwidth streaming class traffic [1] such as streaming video, which would cause more complicated mathematical analysis. ii) the basic model can be extended to a number of variations such as using a buffer to further store the associated voice or data request when the requested bandwidth is not available, which will further reduce the corresponding blocking or dropping probability [4], [14], [16].

Remark 2: Due to the introduction of the PP mechanism and victim buffer in the CSPP scheme, more capacity of different traffic types can be achieved as compared to the schemes without the PP mechanism, for example, the DTBR scheme. In the CSPP scheme, the new voice and handoff voice calls can use bandwidth resource from zero to N_1 and N_2 respectively, where $N_1 \leq N_2 \leq N$. The new data and handoff data calls can use bandwidth resource from zero to $min\{N_3, N - \mathit{the\ occupied\ bandwidth\ resource\ by\ voice\ calls}\}$ and $(N - \mathit{the\ occupied\ bandwidth\ resource\ by\ voice\ calls})$ respectively. That is, the maximum bandwidth resource is N_3 for the new data calls and N for handoff data calls. On the other hand, in the DTBR scheme, the new voice and handoff voice calls can use channels from zero to K_1 and N respectively (which is similar to the CSPP scheme). Nevertheless, the data calls can only share the K_2 channels with voice calls, where there must be the relationship $K_2 < K_1 < N$ if service differentiation among handoff voice, new voice and data calls is required. While in the CSPP scheme, the selection of N_3 is not affected by N_1 and N_2, i.e., N_3 can be smaller or larger than or equal to N_1 or N_2. Thus, data calls can access more bandwidth resource in the CSPP scheme, which means higher resource utilization can be achieved (this will also be shown in Section IV by numerical results).

Another advantage of the CSPP scheme is that high resource utilization is achieved by the complete sharing approach without affecting the performance of higher prioritized traffic types (e.g., voice calls). This is because that the voice traffic can ignore the existence of data traffic by the PP mechanism. On the contrary, in the DTBR scheme, the QoS of voice calls must be affected by data calls due to the existence of contention between them, which is an inevitable problem in the complete sharing approach, i.e., complete sharing can improve system utilization but cannot effectively control the QoS of the shared traffic types. Thus, if the QoS requirements (alternatively, the capacity parameters) of different traffic types are given, by using the CSPP scheme, we can first properly design the two constraint factors of voice traffic, σ_1 and σ_2, which lead to the specific capacity parameters for new and handoff voice traffic respectively. Next, we design the constraint factor σ_3 to schedule the priority relation between new data and handoff data traffic. Note that the influence of the PP mechanism on data traffic should be properly considered.

Remark 3: The CSPP scheme presented here is applicable to a circuit switched FDMA/TDMA based wireless systems, such as GSM [20], HSCSD (High Speed Circuit Switched Data) [21] and EDGE (Enhanced Data rate for GSM Evolution) [22]. From the application viewpoint, this scheme is feasible, since the PP mechanism is commonly used in operating systems, such as eCos, WinCE, VxWorks, QNX, uC/OS, etc., that run on wireless embedded devices such as handheld multimedia nodes [23]. It is noteworthy that the assumption of ideal rate control is used in the proposed scheme, i.e., the data calls in progress can readjust their current bandwidth very quickly when the system state changes due to call arrivals or departures. However, this assumption is used only for tractable analysis, and has also been used by other researchers [9], [15], [19]. In practice, this assumption should not severely affect the system performance except requiring a little more readjusting time than the ideal situation. In addition, the general concepts of the traffic analysis are not restricted to systems with FDMA and/or TDMA based schemes and can also manage circuit or virtual circuit switched services in CDMA networks [8]. The traffic analysis presented here applies to work being done in many areas such as MPLS, ATM, TCP/IP/RSVP, etc. The 3G systems are based on combined FDMA/wideband CDMA multiple access techniques [24]. Their core network is divided into circuit switched and packet switched domains due to reasons of migration and service continuity. The circuit switched core network enables the support of different transports (e.g. ATM or IP) in a bearer-independent fashion [25]. For example, if the ATM is used for 3G UMTS core transmission, the adaptation layer type 2 (AAL2) can handle circuit switched connection and AAL5 can be used for data delivery. Consequently, the proposed scheme could probably be applied to 2G, 2.5G and 3G (with circuit switched) systems. In other words, the scheme can be easily modified to fit different wireless systems (good scalability).

Remark 4: It is worthy to note that the CSPP scheme is applied to the network layer and above and the “channel" concept here refers to a generic logical channel, which could be one or several physical channels or time slots for communication in the up- and down-link directions with a generic multiple access method. In a FDMA based system, spectrum is divided up into frequencies (channels) and then assigned to users. One user at any given time is assigned to a channel. In the GSM system, each frequency is splitted into time slots. Each user is allowed to access the entire frequency channel at respective time slot. Different users share this same frequency channel at different time slots [20]. The HSCSD and EDGE solutions enable higher rates by simultaneously using multiple time slots [22]. The HSCSD can offer data rates up to 115 kbits/s. The EDGE can offer data rates up to 384 kbits/s due to using 8PSK modulation mode to replace normal GMSK mode. The 3G UMTS system radio interface has logical channels, which are mapped to transport channels, which are again mapped to physical channels. In the radio interface, different users can simultaneously transmit at different data rates and data rates can even vary in time [24]. 3GPP specifications enable the user of a single terminal to establish and maintain several connections simultaneously. A multimedia service (could be a circuit switched or packet switched multimedia call) may involve several connections and flexibility is required in order to add and delete resources [26]. Also note that the “channel" defined here refers to traffic channel, rather than control channel (which usually involves a lot of signaling overhead). The main burden caused by the CSPP scheme is the requirements of associated traffic class, QoS profiles, channel occupancy status and information of call arrivals and departures. Fortunately, these records have already kept in the base station. Thus, the overhead involved by the CSPP scheme is not very high.

III. Performance Analysis

In this section, we analyze the system performance by considering a homogeneous mobile cellular network, where all cells are assumed statistically identical. In each cell, the arrivals of new voice, handoff voice, new data and handoff data traffic are assumed to be Poisson processes with rate λ_{vn}, λ_{vh}, λ_{dn} and λ_{dh}respectively. Thus, the total arrival rate is for voice traffic $\lambda_v = \lambda_{vn} + \lambda_{vh}$, and for data traffic $\lambda_d = \lambda_{dn} + \lambda_{dh}$. Note that λ_{vh} and λ_{dh} are determined by the new arrival rates and related probabilities and will be elaborated later.

Assume the call holding time for voice and data traffic are exponentially distributed with mean $1/h_v$ and $1/h_d$, and the cell residence time for voice and data traffic are also exponentially distributed with mean $1/r_v$ and $1/r_d$, respectively. These assumptions have been widely used in literature [4]–[10], [12]–[14], [16]–[18]. Thus, the channel occupancy time is exponentially distributed with mean $1/\mu_v = 1/(h_v+r_v)$ for voice traffic, but the mean for data traffic is state-dependent because of its elastic characteristic (to be analyzed below).

Let $X(t)$ and $Y(t)$ be the number of voice calls and data calls (including the data calls being served and those in the buffer) in the system at time t, respectively. We know that $(X(t), Y(t))$ is a two dimensional Markov process with the state space $S = \{(n_v, n_d)|0 \le n_v \le N_2, 0 \le n_d \le N\}$. It is noteworthy that if there is no preempted data call in the buffer, each data call in progress will occupy state-dependent bandwidth (please see equation (1) below). If, however, there are some preempted data calls in the buffer, each data call in progress will occupy minimum bandwidth b_m (since the proposed CAC strategy performs the “squeeze" operation before "preempt" operation), and the preempted data calls in the buffer will not occupy any bandwidth.

From previous analysis, we know that elastic data calls can use the maximum bandwidth when there is enough bandwidth available and use the actual allowed bandwidth when they cannot get the maximum bandwidth. When a data call has the maximum bandwidth b_M, its mean service rate is the maximum service rate $\mu_{d,max} = h_d + r_d$ [9], [15]. When a data call is served with actual allowed bandwidth b, its service rate is proportional to the actual bandwidth of each call [9], [15], [18], [19] (that is, the more the actual bandwidth, the less the service time. For instance, the file-downloading would take half the time with double granted bandwidth). In total, when the system is in state (n_v, n_d), each data call will occupy the bandwidth

$$b(n_v, n_d) = \begin{cases} \min\{b_M, \max\{1, \frac{N-n_v}{n_d}\}\}, \\ \ 0 \le n_v \le N_2, 1 \le n_d \le N, (n_v \ne N) \\ 0, \quad otherwise. \end{cases} \tag{1}$$

Note that above equation does have a condition: $n_v \ne N$. When $n_v = N$ (which only happens at the case $N_2 = N$), all N channels are occupied by voice calls only, there is either no data call at all or a few data calls in the buffer. In either case, the bandwidth occupied by data traffic $b(n_v, n_d)$ is zero. Assume that the mean cell residence time of data calls $1/r_d$ does not vary with the change of serving bandwidth [9], then the state-dependent data call service rate $\mu^d_{n_v,n_d}$ is:

$$\mu^d_{n_v,n_d} = \frac{b(n_v, n_d)}{b_M} h_d + r_d. \tag{2}$$

In equilibrium the total transition rates of data calls from state (n_v, n_d) to state $(n_v, n_d - 1)$, and to state $(n_v, n_d + 1)$ are respectively calculated as

$$TR^d_{(n_v,n_d)\to(n_v,n_d-1)} = \begin{cases} n_d \mu^d_{n_v,n_d}, \quad 0 \le n_v \le N_2, 1 \le n_d \le N - n_v, \\ (N - n_v)\mu^d_{n_v,N-n_v} + (n_d - N + n_v) r_d, \\ \qquad 1 \le n_v \le N_2, N - n_v + 1 \le n_d \le N. \end{cases} \tag{3}$$

$$TR^d_{(n_v,n_d)\to(n_v,n_d+1)} = \begin{cases} \lambda_d, \quad 0 \le n_v \le N_2, 0 \le n_d < \min\{N - n_v, N_3\}, \\ \lambda_{dh}, \quad 0 \le n_v \le N - N_3 - 1, N_3 \le n_d < N - n_v, \\ 0, \quad otherwise, \end{cases} \tag{4}$$

where $\max\{N - N_2 - 1, 0\} \le N_3 \le N$ to ensure the correct range of n_v and N_3.

Of course the mean service rate of voice calls in any state is μ_v, and in equilibrium the total transition rates of voice calls from state (n_v, n_d) to state $(n_v - 1, n_d)$, and to state $(n_v + 1, n_d)$ are

$$TR^v_{(n_v,n_d)\to(n_v-1,n_d)} = n_v \mu_v, 1 \le n_v \le N_2, 0 \le n_d \le N. \tag{5}$$

$$TR^v_{(n_v,n_d)\to(n_v+1,n_d)} = \begin{cases} \lambda_v, \quad 0 \le n_v < N_1, 0 \le n_d \le N; \\ \lambda_{vh}, \quad N_1 \le n_v < N_2, 0 \le n_d \le N; \\ 0, \quad otherwise. \end{cases} \tag{6}$$

If we denote by $\pi(n_v, n_d)$ the steady state probability when the system is in state (n_v, n_d), the steady state probability vector (ordered lexicographically) is then partitioned as $\boldsymbol{\pi} = (\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \cdots, \boldsymbol{\pi_{N_2}})$, where the vector $\boldsymbol{\pi_n} = (\pi(n, 0), \pi(n, 1), \cdots, \pi(n, N))$ for $0 \le n \le N_2$. The vector $\boldsymbol{\pi}$ is the solution of equations $\boldsymbol{\pi} Q = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{e} = 1$, where $\mathbf{e}$ and $\mathbf{0}$ are vectors of all ones and zeros respectively, and the infinitesimal generator of the two dimensional Markov process Q is given by

$$Q = \begin{bmatrix} E_0 & B_0 & 0 & \cdots & 0 & 0 & 0 \\ D_1 & E_1 & B_1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & D_{N_2-1} & E_{N_2-1} & B_{N_2-1} \\ 0 & 0 & 0 & \cdots & 0 & D_{N_2} & E_{N_2} \end{bmatrix} \tag{7}$$

where each submatrix has size $(N+1)$ by $(N+1)$ and defined by

$$B_i(j, k) = \begin{cases} \lambda_v, \quad 0 \le i < N_1, 0 \le j \le N, k = j, \\ \lambda_{vh}, \quad N_1 \le i < N_2, 0 \le j \le N, k = j, \\ 0, \qquad otherwise, \end{cases}$$

$$D_i(j, k) = \begin{cases} i\mu_v, \quad 1 \le i \le N_2, 0 \le j \le N, k = j, \\ 0, \qquad otherwise, \end{cases}$$

where $B_i(j, k)$ and $D_i(j, k)$ are the j-th row and k-th column element of the matrix B_i and D_i, respectively. If we denote by $A_i(j, k)$ the j-th row and k-th column element of the matrix A_i, and define $A_i(j, k) = 0$ for $j, k < 0$ or $j, k > N$, then $A_i(j, k)$ can be expressed as

$$\begin{cases} \lambda_d, \quad 0 \le i \le N_2, 0 \le j < \min\{N - i, N_3\}, k = j + 1, \\ \lambda_{dh}, \quad 0 \le i \le N - N_3 - 1, N_3 \le j < N - i, k = j + 1, \\ j\mu^d_{i,j}, \quad 0 \le i \le N_2, 1 \le j \le N - i, k = j - 1, \\ (N - i)\mu^d_{i,N-i} + (j - N + i) r_d, \quad 1 \le i \le N_2, \\ \qquad N - i + 1 \le j \le N, k = j - 1, \\ -[A_i(j, j - 1) + A_i(j, j + 1)], \quad 0 \le i \le N_2, \\ \qquad 0 \le j \le N, k = j, \\ 0, \quad otherwise. \end{cases}$$

Then the matrix E_i can be calculated by

$$E_i = \begin{cases} A_i - B_i, & i = 0, \\ A_i - D_i - B_i, & 1 \le i < N, \\ A_i - D_i, & i = N. \end{cases}$$

By the results of [27] and [28], we determine the steady state probabilities as

$$\boldsymbol{\pi_n} = \boldsymbol{\pi_{n-1}} B_{n-1} \left(-C_n\right)^{-1} = \boldsymbol{\pi_0} \textstyle\prod_{i=0}^{n} [B_{i-1} \left(-C_i\right)^{-1}], \tag{8}$$

where $1 \le n \le N_2$ and $\boldsymbol{\pi_0}$ satisfies $\boldsymbol{\pi_0} C_0 = \mathbf{0}$ and

$$\boldsymbol{\pi_0} \Big[I + \textstyle\sum_{n=1}^{N_2} \prod_{i=1}^{n} [B_{i-1} \left(-C_i\right)^{-1}] \Big] \mathbf{e} = 1 \tag{9}$$

The C_i $(0 \le i \le N_2)$ can be recursively determined by $C_{N_2} = E_{N_2}$ and

$$C_i = E_i + B_i \left(-C_{i+1}\right)^{-1} D_{i+1}. \quad (0 \le i \le N_2 - 1) \tag{10}$$

After obtaining the steady state probabilities, it is easy to determine various performance measures, such as the new call blocking probabilities and handoff dropping probabilities for voice and data traffic, the loss probability of preempted data traffic, the total channel utilization, the transmission delay of data traffic, the busy period time and blocking period time in the cell.

A. Basic performance measures

- **New voice blocking and handoff voice dropping probabilities**
  The new voice blocking (respectively handoff voice dropping) probability, denoted by P_{vn} (respectively P_{vh}), is defined as the probability that upon an arrival of a new voice (respectively handoff voice) call the occupied channels by voice calls are not less than N_1 (respectively N_2) and the arrival request has to be lost. Thus, we have

$$\begin{aligned} P_{vn} &= \textstyle\sum_{n_v=N_1}^{N_2} \sum_{n_d=0}^{N} \pi(n_v, n_d) \\ &= \textstyle\sum_{n_v=N_1}^{N_2} \boldsymbol{\pi_{n_v}} \mathbf{e} \\ &= \boldsymbol{\pi_0} \textstyle\sum_{n_v=N_1}^{N_2} \prod_{i=1}^{n_v} [B_{i-1} (-C_i)^{-1}] \mathbf{e}. \end{aligned} \tag{11}$$

$$\begin{aligned} P_{vh} &= \textstyle\sum_{n_d=0}^{N} \pi(N_2, n_d) \\ &= \boldsymbol{\pi_{N_2}} \cdot \mathbf{e} \\ &= \boldsymbol{\pi_0} \textstyle\prod_{i=1}^{N_2} [B_{i-1} (-C_i)^{-1}] \mathbf{e}. \end{aligned} \tag{12}$$

- **New data blocking and handoff data dropping probabilities**
  The new data blocking (respectively handoff data dropping) probability, denoted by P_{dn} (respectively P_{dh}), is defined as the probability that upon an external arrival of new data (respectively handoff data) call the occupied channels in the system are not less than N_3 (respectively N) and the arrival request has to be lost. We have

$$P_{dn} = \textstyle\sum_{n_v=0}^{N_2} \sum_{n_d=\min\{N-n_v, N_3\}}^{N} \pi(n_v, n_d) \tag{13}$$

$$P_{dh} = \textstyle\sum_{n_v=0}^{N_2} \sum_{n_d=N-n_v}^{N} \pi(n_v, n_d) \tag{14}$$

- **Total channel utilization η**
  The total channel utilization is defined as the ratio between the mean number of occupied channels and the total number of channels, i.e.,

$$\begin{aligned} \eta = \tfrac{1}{N} \textstyle\sum_{n_v=0}^{N_2} \sum_{n_d=0}^{N} \{ & n_v + \min\{n_d, N-n_v\} \\ & \cdot b(n_v, n_d)\} \pi(n_v, n_d). \end{aligned} \tag{15}$$

B. Loss probability of preempted data traffic

Sometimes we are interested in the probability that the preempted data calls in the victim buffer are lost before they connect back to the system because of the mobility, which is called the loss probability of preempted data traffic. We know that if a voice call arrives to find that there is no available bandwidth for it to enter or "squeeze" into the system and it satisfies the preemption condition, then it enters by randomly preempting a data call. Furthermore, if a preempted data call arrives to the buffer with all N voice/data calls being served in the system and j data calls in the buffer, then it could connect back to the cell according to FCFS discipline only if $j+1$ calls leave the cell before its maximum queuing time, i.e., its cell residence time, since the only reason for a preempted data call to be lost is that the mobile terminal moves out of the cell. Therefore, if we denote by φ_j $(0 \le j \le N_2 - 1)$ the interval from the epoch that a data call is preempted to the buffer with all N voice/data calls being served in the system and j data calls in the buffer to the epoch that one of the calls leaves the cell (either leaves a channel or leaves the buffer), and by τ the queuing time of the data call in the buffer, then τ is of exponential distribution with rate r_d. Now if a preempted data call arrives to find that j preempted data calls waiting in the buffer, n_v voice calls and $N-n_v$ data calls being served in the system $(1 \le n_v \le N_2, 0 \le j \le N_2 - 1)$, then either a voice call's completion or an ongoing data call's completion will lead the head call of the queue in the buffer to connect back to the system, and the leaving of the head call of the queue will lead each of the remaining data calls in the buffer to move forward one step. Thus φ_j is also of exponential distribution with rate $n_v\mu_v + (N-n_v)\mu^d_{n_v,N-n_v} + jr_d$. Specifically, if we define that the loss probability of a preempted data call is the percentage of those preempted data calls eventually lost to the total data calls in the system, denoted by P_{pre}, then we will have [13]:

$$P_{pre} = \frac{\sum_{n_v=1}^{N_2} \sum_{j=0}^{n_v-1} (j+1)\pi(n_v, N-n_v+j+1) P\{\tau < \varphi_0 + \cdots + \varphi_j\}}{\sum_{n_v=0}^{N_2} \sum_{n_d=1}^{N} n_d \pi(n_v, n_d)}.$$

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

$$\begin{aligned} P\{\tau < \varphi_0 + \cdots + \varphi_j\} &= 1 - \textstyle\prod_{i=0}^{j} f_i^*(r_d) \\ &= \frac{(j+1)r_d}{n_v\mu_v + (N-n_v)\mu^d_{n_v,N-n_v} + (j+1)r_d}, \end{aligned} \tag{16}$$

where the last equation was obtained by the fact that

$$f_i^*(r_d) = \frac{n_v\mu_v + (N-n_v)\mu^d_{n_v,N-n_v} + ir_d}{n_v\mu_v + (N-n_v)\mu^d_{n_v,N-n_v} + (i+1)r_d},$$

where $0 \le i \le j$. After a few mathematical operations, we obtain

$$P_{pre} = \frac{\sum_{n_v=1}^{N_2} \sum_{j=0}^{n_v-1} \frac{(j+1)^2 r_d}{n_v\mu_v + (N-n_v)\mu^d_{n_v,N-n_v} + (j+1)r_d} \pi(n_v, N-n_v+j+1)}{\sum_{n_v=0}^{N_2} \sum_{n_d=1}^{N} n_d \pi(n_v, n_d)}. \tag{17}$$

In addition, we obtain the percentage of the data calls without suffering preemption to the total data calls in the system, q_{nonpre}, and the percentage of the preempted data calls which are eventually connected back to the system to the total data calls in the system, q_{back}, respectively, which will be used in the next subsection.

$$q_{nonpre} = \frac{\sum_{n_v=0}^{N_2} \sum_{n_d=1}^{N-n_v} n_d \pi(n_v, n_d)}{\sum_{n_v=0}^{N_2} \sum_{n_d=1}^{N} n_d \pi(n_v, n_d)}. \tag{18}$$

$$\begin{aligned} q_{back} &= \frac{\sum_{n_v=1}^{N_2} \sum_{j=0}^{n_v-1} (j+1)\pi(n_v, N-n_v+j+1) P\{\tau > \varphi_0 + \cdots + \varphi_j\}}{\sum_{n_v=0}^{N_2} \sum_{n_d=1}^{N} n_d \pi(n_v, n_d)} \\ &= \frac{\sum_{n_v=1}^{N_2} \sum_{j=0}^{n_v-1} \frac{(j+1)[n_v\mu_v + (N-n_v)\mu^d_{n_v,N-n_v}]}{n_v\mu_v + (N-n_v)\mu^d_{n_v,N-n_v} + (j+1)r_d} \pi(n_v, N-n_v+j+1)}{\sum_{n_v=0}^{N_2} \sum_{n_d=1}^{N} n_d \pi(n_v, n_d)}. \end{aligned} \tag{19}$$

C. Transmission delay of data traffic

Transmission delay is one of the most important performance measurements of interactive and background service classes [8]. These service classes have elastic traffic characteristics. They can equally share the bandwidth left over for them through the use of resource reallocations (triggered by call arrivals or departures). The elastic traffic characteristics have also been addressed in [8] by the FRA strategies with uniform QoS provisioning, from which the transmission delay for the data traffic, D, can be approximated by subtracting the optimal or minimum service time X from the actual service time Y , i.e.,

$$D = Y - X \cong X\left(\frac{b_M}{B_{avg}} - 1\right),$$

where random variable X represents the minimum call holding time of the data traffic (i.e., operating at the maximum bandwidth), which has been assumed exponentially distributed with mean $1/h_d$ (from the first term of equation (2)). Random variable Y represents the actual service time with almost exponential distribution with mean $b_M/(B_{avg}h_d)$ [8]. B_{avg} represents the average bandwidth of data calls,$b_m \leq B_{avg} \leq b_M$. Here B_{avg} can be easily calculated by

$$B_{avg} = \sum_{n_v=1}^{N_2}\sum_{n_d=0}^{N} b(n_v, n_d)\pi(n_v, n_d). \tag{20}$$

Thus, the mean transmission delay $E[D]$ is approximated by

$$E[D] \cong \frac{1}{h_d}\left(\frac{b_M}{B_{avg}} - 1\right). \tag{21}$$

However, in the proposed CSPP scheme, the transmission delay of data traffic involves more complicated analysis due to the introduction of the PP mechanism and victim buffer. The successfully completed data calls include two parts: i) the unencumbered data calls without suffering preemption, which has the same transmission delay as that in [8], i.e., $E[D]$; and ii) the preempted data calls that eventually connect back to the system due to the released bandwidth available, which have an additional transmission delay due to the waiting time in the victim buffer, $E[W_b]$, besides the transmission delay $E[D]$.

We know that if a voice call arrives to find that there is no available bandwidth for it to enter or "squeeze" into the system and it satisfies the preemption condition, then it enters by randomly preempting a data call. The preempted call could connect back according to FCFS discipline if it can get available bandwidth before its maximum queuing time in the buffer. In steady state, the mean number of preempted data calls in the victim buffer and the mean arrival rate to the buffer can be calculated as

$$E[L_b] = \sum_{n_v=1}^{N_2}\sum_{j=0}^{n_v-1}(j+1)\pi(n_v, N - n_v + j + 1),$$

$$E[\lambda_b] = \lambda_v \sum_{n_v=0}^{N_1-1}\sum_{n_d=N-n_v}^{N}\pi(n_v, n_d)$$

$$+\lambda_{vh}\sum_{n_v=N_1}^{N_2-1}\sum_{n_d=N-n_v}^{N}\pi(n_v, n_d).$$

By Little's law, the mean waiting time of the preempted data calls in the buffer is

$$E[W_b] = \frac{E[L_b]}{E[\lambda_b]}. \tag{22}$$

Therefore, the transmission delay of the successfully completed data calls in the system can be calculated as

$$E[T_D] = q_{nonpre}E[D] + q_{back}(E[D] + E[W_b]), \tag{23}$$

where q_{nonpre} and q_{back} are respectively the percentage of the data calls without suffering preemption and the percentage of the preempted data calls eventually connected back to the system, and are given in equations (18) and (19). Since the second term of equation (23) is the additional transmission delay caused by the PP mechanism, we define a performance measure of the transmission delay effect factor, ς, to evaluate the effect of the PP mechanism on the transmission delay.

$$\varsigma = \frac{q_{back}(E[D] + E[W_b])}{E[T_D]}. \tag{24}$$

D. Handoff rates

Since the data traffic is preempted randomly by voice traffic, the preempted data traffic obviously includes new data calls and handoff data calls. From a statistical viewpoint, the total blocking or dropping probability of data calls including those lost from the victim buffer, say P^D_{new} (or $P^D_{handoff}$), should be

$$\begin{cases} P^D_{new} = P_{dn} + (1 - P_{dn})\frac{\lambda_{dn}}{\lambda_{dn}+\lambda_{dh}}P_{pre}, \\ \\ P^D_{handoff} = P_{dh} + (1 - P_{dh})\frac{\lambda_{dh}}{\lambda_{dn}+\lambda_{dh}}P_{pre}. \end{cases} \tag{25}$$

Since the probability that an accepted voice call (respectively data call) will attempt a handoff is $a_v = r_v/(h_v + r_v)$ (respectively $a_d = r_d/(h_d + r_d)$), the rates of the handoff voice calls and handoff data calls departing from a cell, say λ_{vho} and λ_{dho}, are respectively

$$\lambda_{vho} = a_v[(1 - P_{vn})\lambda_{vn} + (1 - P_{vh})\lambda_{vh}] \text{ and}$$

$$\lambda_{dho} = a_d[(1 - P^D_{new})\lambda_{dn} + (1 - P^D_{handoff})\lambda_{dh}],$$

where λ_{vh} and λ_{dh} are (consistant with the previous notation) the handoff arrival rates to a cell for voice calls and data calls, respectively. We know that the handoff arrival rate to a cell consists of the sum of all the handoff departure rates from its neighboring cells. For a homogeneous mobile network, the handoff departure rate from each cell is supposed to be statistically identical, and users leave a cell to each of its neighboring cell with equal probability. Thus, the handoff arrival rate to a cell is equal to handoff departure rate out of the cell, ie., $\lambda_{vh} = \lambda_{vho}$, and $\lambda_{dh} = \lambda_{dho}$.

From the above analysis, we obtain

$$\begin{cases} \lambda_{vh} = \frac{r_v(1-P_vn)\lambda_{vn}}{h_v+r_vP_{vh}}, \\ \\ \lambda_{dh} = \frac{r_d(1-P^D_{new})\lambda_{dn}}{h_d+r_dP^D_{handoff}}. \end{cases} \tag{26}$$

We observe that the handoff rates obtained by equation (26) have the following relationship: (a) λ_{vh} is a function of P_{vn}

and P_{vh}, λ_{dh} is a function of P^D_{new} and $P^D_{handoff}$; (b) P_{vn}, P_{vh}, P^D_{new} and $P^D_{handoff}$ are functions (or through P_{dn}, P_{dh} and P_{pre}) of λ_{vh} and/or λ_{dh}, respectively. Therefore, it is not easy to calculate λ_{dh} and λ_{vh} independently. We extend the iterative technique proposed in [29] to calculate λ_{vh} and λ_{dh} jointly, and then calculate different performance measures. The following is an algorithm example for computing handoff rates λ_{vh} and λ_{dh}, and performance measures P_{vn}, P_{vh}, P_{dn}, P_{dh}, η, P_{pre}, $E[T_D]$, and ς.

```
Input: $N$, $N_1$, $N_2$, $N_3$, $\lambda_{vn}$, $\lambda_{dn}$, $h_v$, $r_v$, $h_d$, $r_d$.
Output: $\lambda_{vh}$, $\lambda_{dh}$, $P_{vn}$, $P_{vh}$, $P_{dn}$, $P_{dh}$, $\eta$, $P_{pre}$, $E[T_D]$, $\varsigma$.
Step 1: Predefine thresholds $\delta_v$ and $\delta_d$ for voice and data traffic, and select an initial value for $\lambda_{vh}$ and $\lambda_{dh}$, respectively.
Step 2: Compute vector $\boldsymbol{\pi_n}$ by equations (8), (9) and (10).
Step 3: Compute $P_{vn}$, $P_{vh}$, $P_{dn}$, $P_{dh}$ and $P_{pre}$ by equations (11), (12), (13), (14) and (17).
Step 4: Compute $P^D_{new}$ and $P^D_{handoff}$ by equation (25).
Step 5: Compute the updated handoff rate $\lambda'_{vh}$ and $\lambda'_{dh}$ by equation (26).
Step 6: If $|\lambda'_{vh} - \lambda_{vh}| < \delta_v$ and $|\lambda'_{dh} - \lambda_{dh}| < \delta_d$, then set $\lambda_{vh} = \lambda'_{vh}$, $\lambda_{dh} = \lambda'_{dh}$ and go to step 7. Otherwise, set $\lambda_{vh} = \lambda'_{vh}$, $\lambda_{dh} = \lambda'_{dh}$ and go to step 2.
Step 7: Compute $P_{vn}$, $P_{vh}$, $P_{dn}$, $P_{dh}$, $\eta$, $P_{pre}$, $E[T_D]$, and $\varsigma$ by equations (11), (12), (13), (14), (15), (17), (23) and (24).
```

E. Busy period time in the cell

Here we consider the busy period time of voice traffic in the system. The busy period time of voice traffic, denoted by B_v, is the time duration starting from the epoch that a cell without voice call is connected with a voice call to the epoch that there is no voice call again for the first time.

Let $N_B(t)$ and $J_B(t)$ be the number of voice calls and data calls (including those in the buffer) in the system at time t, respectively. We know that $(N_B(t), J_B(t))$ is a two-dimensional absorbing Markov process with absorbing state $\{(0,j)|0 \le j \le N\}$ and the infinitesimal generator matrix

$$Q_B = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ D_1 & E_1 & B_1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & D_{N_2-1} & E_{N_2-1} & B_{N_2-1} \\ 0 & 0 & 0 & \cdots & 0 & D_{N_2} & E_{N_2} \end{bmatrix},$$

where D_i, E_i and B_i are the same as those in equation (7). The busy period time of voice 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,$\boldsymbol{\theta_1}$) to the absorbing state $\{(0,j)|0 \le j \le N\}$, where $\boldsymbol{\theta_1}$= ($\boldsymbol{\pi_1}/(\boldsymbol{\pi_1 e}), 0, \cdots, 0)$. If we denote by

$$T_B = \begin{bmatrix} E_1 & B_1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & D_{N_2-1} & E_{N_2-1} & B_{N_2-1} \\ 0 & 0 & \cdots & 0 & D_{N_2} & E_{N_2} \end{bmatrix},$$

then from Lemma 2.2.2 in [30], we know the distribution of the busy period time, B_v, is

$$P(B_v \le x) = 1- \boldsymbol{\theta_1} exp(T_B x)\mathbf{e} \quad \text{for } x \ge 0. \tag{27}$$

And the mean of busy period time, $E[B_v]$, is

$$E[B_v] = - \boldsymbol{\theta_1}(T_B)^{-1}\mathbf{e}. \tag{28}$$

F. Blocking period time in the cell

Contrasting with the busy period time, the blocking period time studies the duration that calls are blocked in a cell, which was used as a performance measure in voice-centric PCS network in [31]. Here we extend the concept to integrated voice/data environments. First we consider the blocking period time of new voice calls, say Z_{nv}, i.e., the time duration starting from the epoch that a new voice call is blocked when it arrives to the epoch when there is at least one channel available for new voice calls for the first time.

Let $X_{nv}(t)$ and $Y_{nv}(t)$ be the number of voice calls and data calls (including those in the buffer) in the system at time t, respectively. It is easy to prove that $(X_{nv}(t), Y_{nv}(t))$ is a two-dimensional absorbing Markov process with absorbing state $\{(i,j)|0 \le i \le N_1 - 1, 0 \le j \le N\}$ and the infinitesimal generator matrix Q_{nv} given by

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ D_{N_1} & E_{N_1} & B_{N_1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & D_{N_2-1} & E_{N_2-1} & B_{N_2-1} \\ 0 & 0 & 0 & \cdots & 0 & D_{N_2} & E_{N_2} \end{bmatrix},$$

where D_i, E_i and B_i are the same as those in previous equation (7). The blocking period time of new voice calls in the cell is just the absorbing time of the Markov process starting from the initial state (0,$\boldsymbol{\theta_2}$) to the absorbing state, where $\boldsymbol{\theta_2}$= ($\boldsymbol{\pi_{N_1}}/(\boldsymbol{\pi_{N_1} e}), 0, \cdots, 0)$. Based on these arguments, if we denote by

$$T_{nv} = \begin{bmatrix} E_{N_1} & B_{N_1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & D_{N_2-1} & E_{N_2-1} & B_{N_2-1} \\ 0 & 0 & \cdots & 0 & D_{N_2} & E_{N_2} \end{bmatrix},$$

then from Lemma 2.2.2 in [30], we know the distribution of the blocking period time, Z_{nv}, is

$$P(Z_{nv} \le x) = 1- \boldsymbol{\theta_2} exp(T_{nv} x)\mathbf{e} \quad \text{for } x \ge 0. \tag{29}$$

And the mean of blocking period time, $E[Z_{nv}]$, is

$$E[Z_{nv}] = - \boldsymbol{\theta_2}(T_{nv})^{-1}\mathbf{e}. \tag{30}$$

Similarly, the blocking period time of handoff voice calls can easily be obtained.

IV. Computational Issues

In this section, we first present the performance comparisons between the DTBR and CSPP schemes in terms of the system award, handoff voice dropping probability, total channel utilization and transmission delay, then the detailed numerical results of the CSPP scheme, and finally a specific design issue from the CSPP scheme, i.e., the determination of the optimal value for N_1 and N_2. For the sake of fair comparison, similar system configuration is assumed as that in [9]: the total number of channels N of each cell is set to be 30, b_M is set to be 2 and b_m is 1. The data call intensity ρ_d is set to be 7 with mean arrival rate $\lambda_d = 0.007S^{-1}$ and mean service rate $\mu_d = 0.001S^{-1}$, where $h_d = 0.0008S^{-1}$ and $r_d = 0.0002S^{-1}$. For voice call, the mean service rate is assumed to be $\mu_v = 0.0083S^{-1}$ (equivalently, the mean call

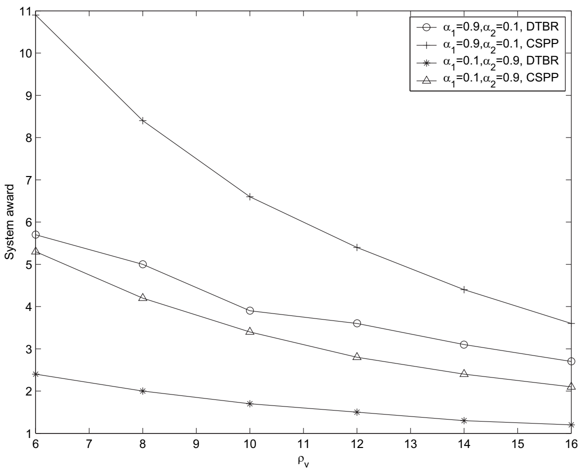

Fig. 2. Comparison of system award.

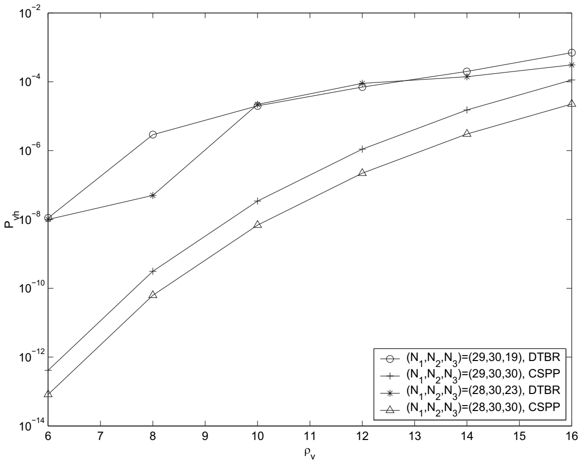

Fig. 3. Comparison of handoff dropping probability.

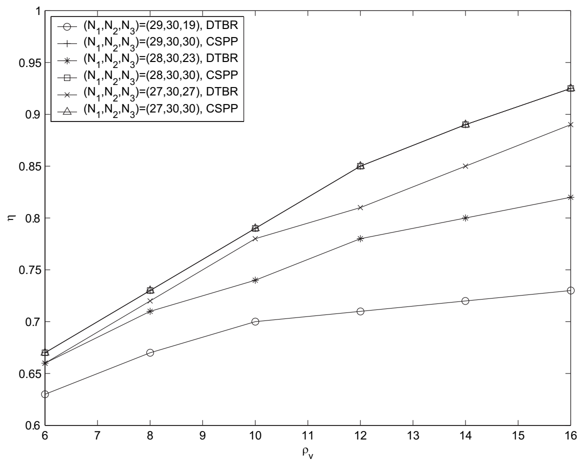

Fig. 4. Comparison of total channel utilization.

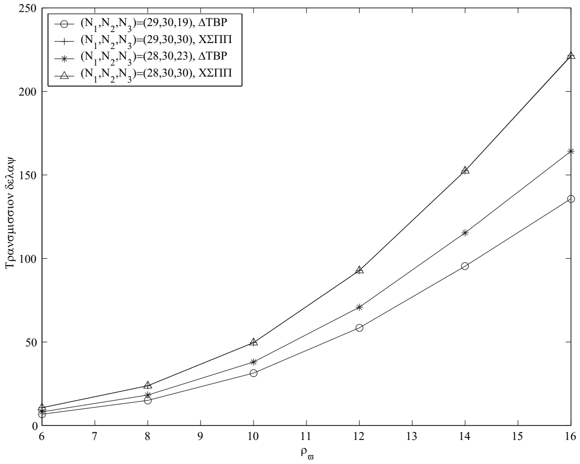

Fig. 5. Comparison of transmission delay.

holding time is about 2 minutes). The voice call intensity ρ_v can vary from 6 to 16. Also, we set the handoff rate same as that in [9], i.e., $\lambda_{vh} = 0.2\lambda_v$ and $\lambda_{dh} = 0.2\lambda_d$.

A. Comparison of DTBR and CSPP schemes

Now we present the system award according to the formula proposed in [9] (The total channel utilization and different probabilities of the CSPP scheme under the same configuration are presented separately in Section IV.B):

$$Award = -(\alpha_1 log_{10}^{P_{vn}} + \alpha_2 log_{10}^{P_d}) + \eta,$$

where α_1 and α_2 are weighting factors and $\alpha_1 + \alpha_2 = 1$. In [9], there is no service distinction between handoff data and new data calls, so we set $N_3 = N$ to remove the data service distinction in CSPP scheme (that is, $P_{dn} = P_{dh}$), and the mixed data blocking probability P_d in our scheme is $P_{dn} + (1 - P_{dn})P_{pre}$. In [9], the handoff voice calls can use total C channels of a cell ($C = 30$), the threshold of new voice calls K_1 (respectively, data calls K_2) is set to be 29, 28 and 27 (respectively, 19, 23 and 27) in its Figs. 4 and 5, and we use the same parameter values in the following Figs. 2-5, i.e., $N_1 = 29/28/27$, $N_2 = 30$. One may argue that it is unfair with the different bandwidth resource for data traffic in the two schemes. However, this is attributed to the different characteristics between the two schemes. In the DTBR scheme, more sharing bandwidth for data and voice traffic can only lead to poor performance of voice calls due to the intrinsic nature of contention in complete sharing approach. Contrastively, in the CSPP scheme, no contention exists between data and voice traffic. The voice traffic can ignore the existence of the data traffic due to the PP mechanism. Moreover, the total channels and the thresholds of voice calls are the same in the comparison. Therefore, the capability of data traffic to share all the bandwidth resource exactly reflects the advantage of the CSPP scheme.

Fig. 2 presents the system award for the two schemes when ($\alpha_1 = 0.9$, $\alpha_2 = 0.1$) and ($\alpha_1 = 0.1$, $\alpha_2 = 0.9$). It can be observed that the system award is identically decreased with the increase of voice traffic intensity, but the CSPP scheme achieves much larger system award in both cases. Fig. 3 compares the handoff voice dropping probability of the two schemes. Obviously, the P_{vh} in CSPP scheme is much lower than that in DTBR scheme. Fig. 4 compares the total channel utilization of the two schemes. We can observe that the channel utilization in both schemes is increased with the increase of voice traffic intensity, but CSPP scheme gets higher channel utilization than DTBR scheme and the channel utilization in CSPP scheme is not sensitive to the change of parameters N_1, N_2 and N_3, which is a particular advantage

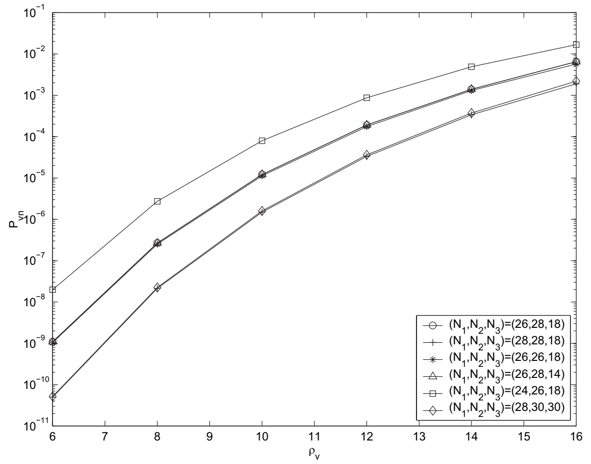

Fig. 6. New voice blocking probability P_{vn}.

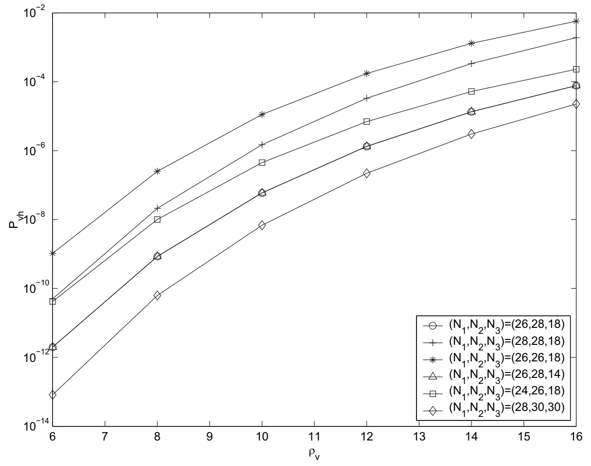

Fig. 7. Handoff voice dropping probability P_{vh}.

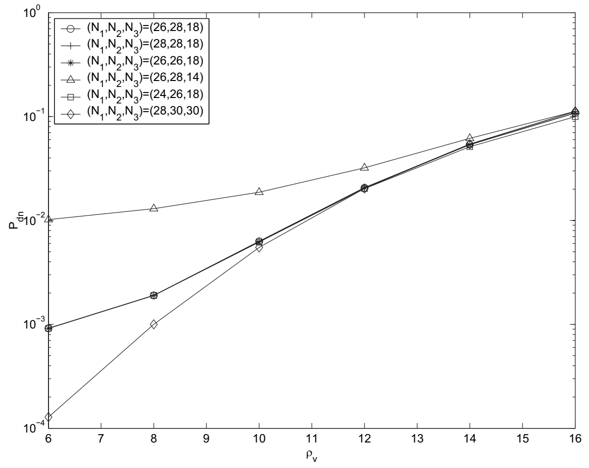

Fig. 8. New data call blocking probability P_{dn}.

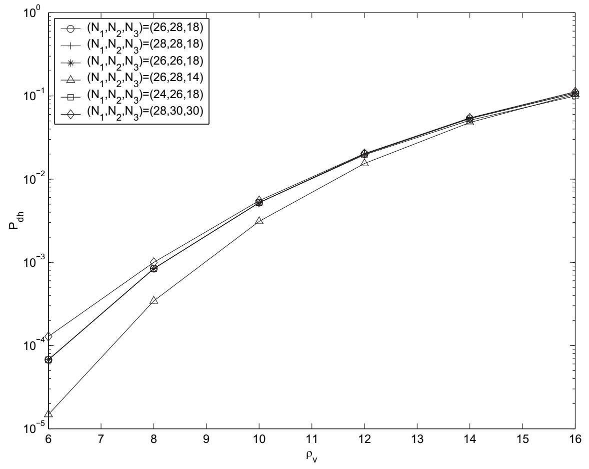

Fig. 9. Handoff data dropping probability P_{dh}.

of CSPP scheme and can be used to satisfy different QoS requirements by adjusting parameters N_1, N_2 and N_3 while meantime maintaining high channel utilization.

The good performances of system award, handoff voice dropping probability and total channel utilization are achieved at the expense of the increase of transmission delay of data traffic, especially under high voice traffic intensity. This can be seen from Fig. 5. Fortunately, the data traffic is usually delay-tolerant and loss-sensitive [9]. And the goal of the CSPP scheme is to achieve good system performance at the expense of some degradation of data traffic (e.g., being squeezed or preempted, and the increase of transmission delay). Moreover, the feature of loss-sensitive of data traffic is well satisfied in the CSPP scheme. This can be shown by comparing the P_{dn}, P_{dh} and P_{pre} in Figs. 8, 9 and 10 in the next subsection: P_{pre} is negligible as compared with P_{dn} and P_{dh}.

B. Performance of CSPP scheme

Figs. 6 to 13 study the system performance of CSPP scheme through different probabilities, channel utilization and transmission delay.

Fig. 6 shows that P_{vn} decreases with the increase of N_1. The reason is when N_1 increases, new voice traffic can use more channels from the system. We also observe that if N_1 is fixed, P_{vn} just has a little change with the change of N_2 or N_3. For instance, P_{vn} of $(N_1, N_2, N_3) = (28, 30, 30)$ just has a little bit change with that of $(28, 28, 18)$.

Fig. 7 shows that P_{vh} are affected by both N_1 and N_2. P_{vh} decreases with the increase of N_2 when fixing N_1, while increases with the increase of N_1 when fixing N_2. For instance, P_{vh} of $(26, 28, 18)$ is lower than that of $(26, 26, 18)$ and $(28, 28, 18)$. This is because when N_2 increases, handoff voice calls can use more channels from the system. While increasing N_1 and fixing N_2, there must be more new voice calls to contend with handoff voice calls. Moreover, in our system configuration new voice calls have much higher arrival rate than handoff voice calls ($\lambda_{vn} = 0.8\lambda_v$). Anyway, if we just change N_3, such as case $(26, 28, 18)$ and $(26, 28, 14)$, then P_{vh} does not change. This is because the voice traffic ignores the data traffic by the PP mechanism.

Fig. 8 shows that P_{dn} decreases with the increase of N_3, but the effect of N_3 becomes gradually small with the increase of N_3. This is because when N_3 increases, new data traffic can usually use more channels from the system. But at the condition of relative high ρ_v, most channels are quickly occupied by voice traffic, the effect of N_3 on P_{dn} gradually decreases and P_{dn} at different values of N_3 approaches the same. Contrary to P_{dn}, P_{dh} in Fig. 9 increases with the

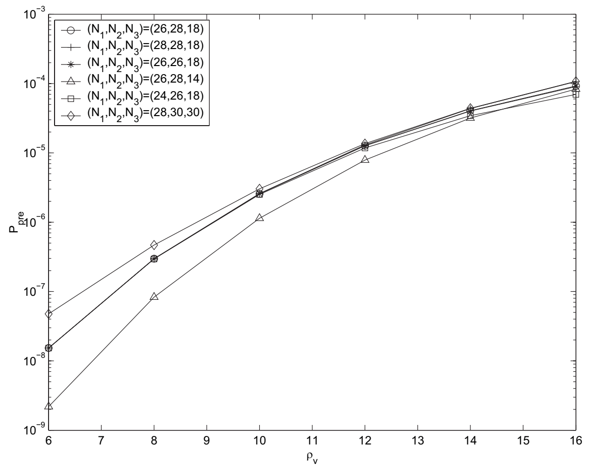

Fig. 10. Loss probability of preempted data traffic P_{pre}.

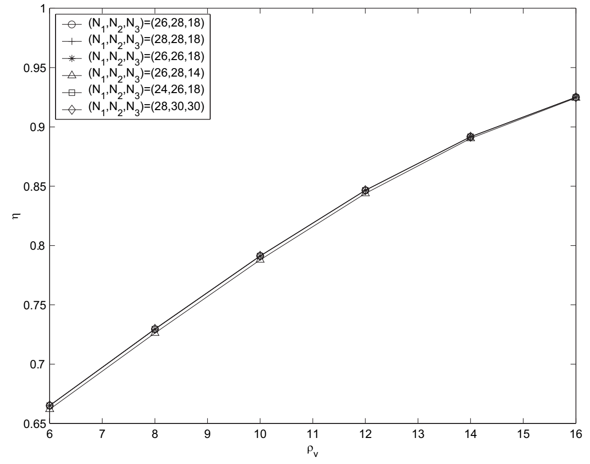

Fig. 11. Total channel utilization.

increase of N_3 at the relative low ρ_v. This is because the increase of N_3 leads to fewer channels (i.e., $N - N_3$) reserved for handoff data calls. But at the condition of relative high ρ_v, P_{dh} at different values of N_3 becomes larger and approaches the same. The reason is the same as that of P_{dn}.

Fig. 10 shows how P_{pre} depends on the changes of parameters N_1, N_2 and N_3. We can see that P_{pre} decreases with the decrease of N_3 when fixing N_1 and N_2, such as $(26, 28, 18)$ and $(26, 28, 14)$. This can be explained as: when N_3 decreases, smaller new data calls are admitted into the system, and new data calls have much higher arrival rate ($\lambda_{dn} = 0.8\lambda_d$) in our configuration, which leads the smaller number of total admitted data calls and therefore the chance of being preempted becomes smaller at the same voice traffic condition. We also found that P_{pre} decreases with the decrease of $N_1 + N_2$ when fixing N_3. For instance, P_{pre} gradually decreases at the order of case $(28, 28, 18)$, $(26, 28, 18)$, $(26, 26, 18)$, and $(24, 26, 18)$ (these are not clearly in Fig. 10, but can be distinguished by amplifying the figure in computer). The reason is that the decrease of voice calls leads to small preemptions. Furthermore, by comparing with P_{dn} in Fig. 8 and P_{dh} in Fig. 9, we can find that P_{pre} is negligible. If we pick up $(26, 28, 14)$ as an example, at $\rho_v = 6$, the numerical scale of P_{dn} is 10^{-2} and P_{dh} is 10^{-5}, while P_{pre} is 10^{-9}. This also validates the effectiveness of using a victim buffer to compensate the injury of data traffic in our scheme. It is noteworthy that the victim buffer is implemented by software and its size can be easily adjusted according to the change of N_2. The main cost is some memory space that requires to be partitioned in the memory management module. Moreover, the resource allocation entity managing the victim buffer and associated CSPP scheme is developed in the base station (rather than the terminal). Thus, the extra cost is trivial as compared with the great benefit the victim buffer brings.

Fig. 11 shows the channel utilization increases with the increase of voice traffic intensity and has almost no change with the change of parameters N_1, N_2 and N_3. This is just the advantage of "completely sharing" plus "preemptive priority" in our CSPP scheme, which can be used to cater for different QoS requirements by adjusting parameters N_1, N_2 and N_3 while meantime not reducing the channel utilization. The reason is, in our scheme data calls can use all the channels of a cell if required, which avoids the case that data calls are blocked while meantime there are still some free channels in the system, and thus increases the total channel utilization.

Fig. 12 shows how the mean transmission delay $E[T_D]$ changes with the change of (N_1, N_2, N_3). It can be observed from the amplified figure that: (i) $E[T_D]$ increases with the increase of ρ_v in all situations. The reason is that more data calls are squeezed or preempted by voice calls. (ii) $E[T_D]$ decreases with the decrease of N_3 when N_1 and N_2 are fixed, see $(26, 28, 18)$ and $(26, 28, 14)$ for example. This is intuitive since at the same condition fewer data calls will share more bandwidth on average. (iii) $E[T_D]$ decreases with the decrease of N_1 and/or N_2 when N_3 is fixed, see $(28, 28, 18)$, $(26, 28, 18)$ and $(24, 26, 18)$ for example. The reason is the same as that in (i).

Fig. 13 shows the effect of PP mechanism on transmission delay by the effect factor ς. As expected, ς will increase with the increase of ρ_v. The reason is that when ρ_v increases, the PP mechanism gradually plays a more important role, leading to the increase of ς. We also observe that the effect of PP mechanism on transmission delay is very small[3] (ς is about 3% even when $\rho_v = 16$), which means that the negative effect of PP mechanism on $E[T_D]$ is trivial as compared with its advantage, such as maintaining good performance for voice traffic, allowing data traffic to share all the system resource and thus maintaining high resource utilization.

C. Optimal design for N_1 and N_2

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 new voice blocking probability P_{vn} and handoff voice dropping probability P_{vh} in the CSPP scheme can be controlled through the design parameters N_1, N_2 (their relationships have been obtained in Figs. 6 and 7 in Section IV.B). On the contrary, the optimal design for N_1 and N_2 can be determined if the

[3] It should be noted that, in Fig. 5, much difference between DTBR and CSPP schemes at high voice traffic intensity is attributed to the fact that only part bandwidth resource can be shared by data traffic in DTBR, while all the bandwidth resource can be shared in CSPP. That is, more data calls can be admitted into the system with CSPP, leading to larger transmission delay.

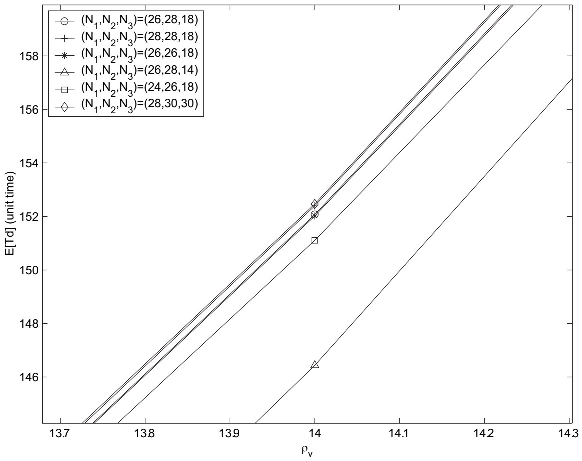

Fig. 12. Mean transmission delay.

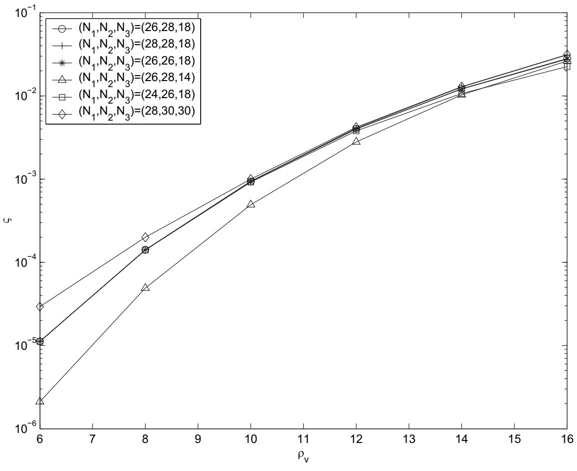

Fig. 13. Transmission delay effect factor ς.

specific QoS requirement $(P_{vn},\, P_{vh})$ is given. At the condition of dynamic traffic load, the QoS requirements are not usually satisfied due to the changes of voice and data traffic intensity. However, the CSPP scheme is a good selection to guarantee the QoSs of voice traffic: a) when the voice traffic intensity is changed, the QoSs can be guaranteed through the adaptive adjustment of N_1 and N_2; b) whether or not the data traffic intensity is changed, the QoS parameters $(P_{vn},\, P_{vh})$ will not be affected due to the PP mechanism (i.e., the voice traffic can ignore the existence of data traffic). The specific algorithm for the optimal design of (N_1, N_2) is as follows:

Step 1: Predefine the QoS threshold (Q_{vn}, Q_{vh})[4] and try some initial value of (N_1, N_2).

Step 2: Compute (P_{vn}, P_{vh}) by the algorithm given in Section III.D.

Step 3: If $P_{vn} > Q_{vn}$ ($P_{vh} > Q_{vh}$), then increase N_1 (N_2) by 1.

Step 4: Repeat Steps 2 and 3 until $P_{vn} < Q_{vn}$, then stop increasing N_1, but continue to increase N_2 and compute P_{vh} until $P_{vh} < Q_{vh}$.

Step 5: If $P_{vn} \ll Q_{vn}$ ($P_{vh} \ll Q_{vh}$), then decrease N_1 (N_2) by 1[5]*.*

TABLE I

THE DETERMINATION OF OPTIMAL VALUE FOR $(N_1,\, N_2)$

$(N_1,\, N_2)$	$\rho_v = 8$	$\rho_v = 10$	$\rho_v = 12$
(Q_{vn}, Q_{vh})=$(10^{-2}, 10^{-6})$	(15, 19)	(18, 22)	(21, 25)
(Q_{vn}, Q_{vh})=$(10^{-2}, 10^{-4})$	(15, 17)	(18, 20)	(21, 23)
(Q_{vn}, Q_{vh})=$(10^{-4}, 10^{-6})$	(22, 24)	(25, 27)	(27, 29)

Step 6: Repeat Steps 2 and 5 until $P_{vh} < Q_{vh}$, then stop decreasing N_2, but continue to decrease N_1 and compute P_{vn} until $P_{vn} < Q_{vn}$.

The above steps are for the initial optimal design of (N_1, N_2), and Steps 2 to 6 can be used for adaptive adjustment of (N_1, N_2) under dynamic load. Next, we do some numerical simulations to validate the above algorithm and determine the optimal value of (N_1, N_2). The parameter configuration is the same as the first paragraph of this section, and the results are shown in Table 1. We can observe that the optimal value of (N_1, N_2) has been changed under different traffic loads. For example, at (Q_{vn}, Q_{vh}) = $(10^{-2}, 10^{-6})$, we have (N_1, N_2) = (15, 19) in ρ_v = 8, and (21, 25) in ρ_v = 12. We also observe that the different optimal value of (N_1, N_2) can be achieved if the different QoS requirement (Q_{vn}, Q_{vh}) is given. For example, at ρ_v = 10, we have (N_1, N_2) = (18, 22) in (Q_{vn}, Q_{vh}) = $(10^{-2}, 10^{-6})$, and (25, 27) in (Q_{vn}, Q_{vh}) = $(10^{-4}, 10^{-6})$.

V. CONCLUSIONS

We have proposed a novel adaptive bandwidth allocation scheme, called CSPP scheme, for integrated voice/data mobile networks. It distinguishes the service performance among different traffic classes, and uses the complete sharing approach (within all the bandwidth resource of the system) and degradation/compensation mechanism to further improve the service performance and the bandwidth utilization efficiency. To implement this, the CSPP scheme considers a PP mechanism to maintain high performance of voice traffic while the injury of data traffic is compensated through a victim buffer. The analytical model of the CSPP scheme is analyzed by a two dimensional Markov process. The steady state probability vector is obtained recursively by the matrix-analytic method, and many important performance measures are then determined, such as the new call blocking probabilities and handoff dropping probabilities, respectively, for voice and data traffic, the loss probability of preempted data traffic, the total channel utilization of the system, the transmission delay of data traffic, the busy period time and blocking period time in the cell.

From the performance analysis and comparisons as well as numerical results, we can conclude that the CSPP scheme can achieve complete service differentiation for different traffic classes and very high and stable channel utilization at various traffic loads. By adjusting the defined constraint factors, we can not only maintain higher priority for handoff voice over new voice calls and voice calls over data calls but also provide

[4] In general, Q_{vn} is less than Q_{vh} due to the higher priority of handoff voice traffic.

[5] The reason for decreasing (N_1, N_2) is to protect the performance of data traffic as much as possible, since a very low P_{vn} or P_{vh} (far less than Q_{vn} or Q_{vh}) is also a waste of resource. Note that in the CSPP scheme, the $N - N_2$ bandwidth units are exclusively used by data traffic.

necessary protection for data calls in hot-spot data traffic situations.

Acknowledgment

The authors would like to thank the three anonymous reviewers for their valuable comments and suggestions, which enhanced the quality of the paper. The work of Wei Li was supported in part by the National Science Foundation under Grant CCF-0515263.

References

[1] 3GPP TS 23.107 V6.3.0, "3GPP; Technical Specification Group Services and System Aspects; Quality of Service (QoS) concepts and architecture," June 2005.

[2] O. Alani and M. Al-Akaidi, "Call admission control for multitier networks with integrated voice and data services," *London Communications Symposium (LCS)*, 2003.

[3] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey," *IEEE Pers. Commun.*, vol.3, no.3, pp.10-31, June 1996.

[4] Y.-R. Haung, Y.-B. Lin, and J. M. Ho, "Performance analysis for voice/data integration on a finite-buffer mobile systems," *IEEE Trans. Veh. Technol.*, vol. 49, no. 2, pp. 367-378, Feb. 2000.

[5] H. Heredia-Ureta, F. A. Cruz-Pérez, and L. Ortigoza-Guerrero, "Capacity optimization in multiservice mobile wireless networks with multiple fractional channel reservation," *IEEE Trans. Veh. Technol.*, vol. 52, no. 6, pp. 1519-1539, Nov. 2003.

[6] C. W. Leong *et al.*, "Call admission control for integrated on/off voice and best-effort data services in mobile cellular communications," *IEEE Trans. Commun.*, vol. 52, no. 5, pp. 778-790, May 2004.

[7] G. Schembra, "A resource management strategy for multimedia adaptive-rate traffic in a wireless network with TDMA access," *IEEE Trans. Wireless Commun.*, vol. 4, no. 1, pp. 65-78, Jan. 2005.

[8] L. Ortigoza-Guerrero, F. A. Cruz-Pérez, and H. Heredia-Ureta, "Call level performance analysis for multiservices wireless cellular networks with adaptive resource allocation strategies," *IEEE Trans. Veh. Technol.*, vol. 54, no. 4, pp. 1455-1472, July 2005.

[9] L.-Z. Li *et al.*, "Call level performance analysis for multi-services wireless cellular networks," *IEEE ICC'2003*, Anchorage, USA, May 11-15, 2003.

[10] B. Li *et al.*, "On handoff performance for an integrated voice/data cellular system," *Wireless Networks*, vol. 9, No. 4, pp. 393-402, 2003.

[11] F. Valois and V. Veque, "Preemption policy for multitier cellular network," *5th IEEE Workshop on Mobile Multimedia Communication (MoMuC'98)*, Berlin, pp. 75-78, Oct. 1998.

[12] B. Li *et al.*, "A preemptive priority handoff scheme in integrated voice and data cellular mobile systems," *IEICE Trans. Commun.*, vol. E82-B, no. 10 Oct. 1999.

[13] S. Tang and W. Li, "Performance Analysis of a Channel Allocation Scheme with Preemptive Priority for Integrated Voice/Data Mobile Networks," in *Proc. 24th IEEE International Performance Computing and Communications Conference (IPCCC'05)*, pp. 417-422, 2005.

[14] H. Wu *et al.*, "On handoff performance for an voice/data integrated cellular system, part II: data buffer case," *IEEE PIMRC'2002*, Sept. 2002.

[15] S. Racz, M. Telek, and G. Fodor, "Call level performance analysis of 3rd generation mobile core networks," *IEEE ICC'2001*, pp.456-461, 2001.

[16] P. Lin, "Channel allocation for GPRS with buffering mechanisms," *Wireless Networks*, Vol. 9, pp. 431-441, 2003.

[17] J.-Y. Jeng and Y.-B. Lin, "Equal resource sharing scheduling for PCS data services," *ACM Wireless Networks*, vol. 5, pp. 41-45, 1999.

[18] H. Heredia-Ureta, F.A. Cruz-Perez, and L. Ortigoza-Guerrero, "Performance analysis of adaptive resource allocation strategies with service time dependence on the allocated bandwidth," *IEEE WCNC 2003*, vol. 3, pp. 1850-1855, 2003.

[19] G. Fodor, S. Racz, and M. Telek, "On providing blocking probability- and throughput guarantees in a multi-service environment," *Int. J. Commun. Syst.* 15, pp. 257-285, 2002.

[20] M. Mouly and M.-B. Pautet, The GSM System for Mobile Communications, *Telecom Publishing*, June 1992.

[21] 3GPP TS 23.205 V7.0.0, "3GPP; Technical Specification Group Core Network; High Speed Circuit Switched Data (HSCSD); Stage 2," Dec. 2004.

[22] J. R. Garcia, J. Melero, and T. Halonen, *GSM, GPRS and EDGE Performance: Evolution Toward 3G/UMTS, 2nd edition*. John Wiley & Sons Inc., 2003.

[23] V. Raghunathan *et al.*, "Energy Aware Wireless Systems with Adaptive Power-Fidelity Tradeoffs," *IEEE Trans. VLSI Systems*, vol. 13, no. 2, pp. 1-15, Feb. 2005.

[24] H. Holma and A. Toskala, WCDMA For UMTS, *John Wiley & Sons Inc.*, 2000.

[25] 3GPP TS 23.205 V7.0.0, "3GPP; Technical Specification Group Core Network; Bearer-independent circuit-switched core network; Stage 2," Dec. 2005.

[26] 3GPP TS 22.101 V7.4.0, "3GPP; Technical Specification Group Services and System Aspects; Service principles," Dec. 2005.

[27] D. P. Gaver, P. A. Jacobs, and G. Latouche, "Finite birth-and death models in randomly changing environments," *Adv. Applied Prob.*, vol. 16, pp. 715-731, 1984.

[28] W. Li and A. S. Alfa, "A PCS network with correlated arrival process and splitted-rate channels," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1318-1325, July 1999.

[29] Y.-B. Lin, "Performance modeling for mobile telephone networks," *IEEE Network Mag.*, Vol. 11, pp. 63-68, Nov./Dec. 1997.

[30] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*. Baltimore, MD: Johns Hopkins Univ. Press, 1981.

[31] A. S. Alfa and W. Li, "A Homogeneous PCS Network with Markov call arrival process and phase type cell residence time," *ACM Wireless Networks*, vol. 8, pp. 597-605, 2002.

Shensheng Tang received his B.S. degree from the Tianjin University and M.S. degree from the China Academy of Telecommunications Technology (CATT), both in Electronic Engineering. He is presently pursuing his Ph.D. 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 product-development of electronic instruments and mobile communication equipments as well as the research related to 3G TD-SCDMA and WCDMA radio technologies. Before studying at The University of Toledo, he was the Head and Project Manager of 3G SA (services & system aspects) Standardization Group of CATT, focusing on R&D and standardization of TD-SCDMA and WCDMA technologies. His current research interests focus on stochastic modeling, queuing theory and its applications, wireless communications and networking including wireless LANs, mobile ad hoc networks, wireless sensor networks, 3G and future generations of wireless cellular networks.

Wei Li is currently an Associate Professor in the Department of Electrical Engineering and Computer Science at the University of Toledo, USA. He received his Ph.D. degree from the Chinese Academy of Sciences in 1994. Dr. Li's research interests are in the routing protocols and security in wireless internet and mobile Ad Hoc Networks; adaptation, design and implementation of dynamic models for wireless and mobile networks; radio resource allocations, channel schemes and handoff strategies in wireless multimedia networks; bio-molecular networks, information systems, mobile and high-performance computing; queueing networks, reliability networks, decision analysis and their applications in communications 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, and for International Journal of High Performance Computing and Networking. He is also serving or has served as a Co-Chair/TPC member/Session Chairs for some IEEE 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. Dr. Li is listed in Who's Who in the World (November 2005), Who's Who in America (October 2005), Who's Who among Executives and Professionals (2004/2005 Honors Edition), Who's Who in Science and Engineering (December 2004) and Who's Who in Engineering Academia (2002). He was once a recipient of Hong Kong Wang Kuan Cheng Research Award in 2003 and US Air Force Summer Faculty Fellowship in 2005.