Telecommun Syst (2006) 32:71–91
DOI 10.1007/s11235-006-8203-1

Modeling and analysis of hierarchical cellular networks with bidirectional overflow and take-back strategies under generally distributed cell residence times

Shensheng Tang · Wei Li

Received: 1 April 2005 / Accepted: 20 February 2006

Abstract The rapid growth of wireless services and mobile users drives a great interest in cellular networks with a hierarchical structure. Hierarchical cellular networks (HCNs) can provide high system capacity, efficient channel utilization and inherent load-balancing capability. In this paper, we develop an analytical model and a performance analysis method for a two-layer HCN with bidirectional overflow and take-back strategies. Mobile users are divided into two classes. The call requests (including new and handoff calls) of fast and slow users are preferably assigned to the macrolayer and microlayer, respectively. A call from a fast user or slow user can overflow to its non-preferable layer if there is no channel available. The successful overflow call can be taken back to its preferable layer if a channel becomes available. Since the commonly used exponentially distributed assumption for cell residence time and then the channel occupancy time does not hold for emerging mobile networks, we model various cell residence times by general distributions to adapt to more flexible mobility environments. The channel occupancy times are derived in terms of the Laplace transforms of various cell residence times. The handoff rates, overflow rates and take-back rates of each layer are also derived in terms of the new call arrival rates and related probabilities. The stationary probabilities (and then the performance measures) are determined on the basis of the theory of multi-dimensional loss systems.

Keywords Hierarchical cellular networks · Handoff · Overflow · Take-back · Cell residence time · Channel occupancy time

1. Introduction

The demand of wireless services and mobile users is growing rapidly while the existing bandwidth reserved for wireless communications is limited. Methods to improve radio spectrum efficiency

S. Tang (✉) · W. Li
Department of Electrical Engineering and Computer Science,
University of Toledo, Toledo, OH 43606, USA
e-mail: stang@eng.utoledo.edu

W. Li
e-mail: wli@eecs.utoledo.edu

are urgently needed so that higher network capacity can be achieved. Cell splitting can improve the system capacity because of higher frequency reuse efficiency (or shorter frequency reuse distance). This, however, potentially results in more frequent handoff operations and thus causes a heavy signaling burden to the system, since the mobile user needs to cross more cells during its lifetime, especially for a high-speed user. To avoid the overhead of smaller sized cells, a hierarchical cellular architecture is proposed.

The hierarchical cellular network (HCN) consists of different layers. Different layers have different sizes of cells, and large cells are overlaid on small cells. Low-speed users are normally assigned to smaller cells, while high-speed users are assigned to larger cells. As a result, system capacity increases while handoff load is limited. In a typical two-layer HCN, all the macrocells form the upper layer, called the macrolayer; all the microcells form the lower layer, called the microlayer. In general, the macrolayer and microlayer are the referable layers for fast and slow user requests, respectively, but many further mechanisms like handoff priority access, overflow limitation and reversible capability are used [1]. In [2], all new and handoff calls of fast and slow users are first directed on the microcells; the macrocells are used for overflow calls. In contrast to the speed-insensitive selection mechanism, the pure-speed-sensitive strategy allocates slow and fast calls to microcells and macrocells, respectively, but there is no overflow call between layers [3]. Most strategies in the literature are a mixture of the two simple ones. In [4,5,26], the calls with different classes are served at different layers; the new calls and handoff calls from the lower layer can overflow to the higher layer when there is no available channel on the lower layer. In [6] only the handoff calls from the lower layer can overflow to its higher layer. A common feature of these strategies is the unidirectional overflow without reversible capability, i.e., a call can only overflow from microlayer to macrolayer; once a call overflow to the macrolayer, it cannot be taken back again, even when a channel is available in the microlayer. It is obviously unsuitable that the overflow call from a slow user occupies the precious macrolayer resource (from the viewpoint of frequency reuse efficiency) while the call from a fast users is blocked due to lack of available channel, especially at the time when there is an idle channel available in the microlayer.

To improve the above strategies, some researchers model the HCN by allowing the overflow calls to return to the microlayer. An analytical model of HCN with both overflow and underflow schemes is proposed in [7], where the overflow calls of slow users in the macrocell can be directed to the next microcell (take-back) at the boundary when the system finds idle channels in the next microcell. Performance analysis shows that this strategy reduces the blocking and dropping probabilities of fast users; it also reduces the dropping probability of slow users. The disadvantage is that the system control load and the blocking probability of slow users are slightly increased. Similar analytic models are developed in [8–12]. The introduction of the take-back strategy produces an increase in traffic capacity. The price of this performance improvement is greater system complexity, as the system has to monitor the eventual availability of resources in the microlayer. However, these improved strategies are also unidirectional, i.e., the overflow of fast users from macrolayer to microlayer is not allowed. It is obviously unfair to the calls of fast users. Since fast users and slow users are not treated equally, a call from a slow user can use more channel resource than that from a fast user. To make matters worse, in the unbalanced traffic environment, the call from a fast user has to be blocked even though there are many free channels in the microlayer.

In [13], a bidirectional call-overflow scheme based on mobile velocity is proposed and analyzed by using two one-dimensional Markov processes; the results show that it has a better characteristic in balancing the telecommunications traffic load between macrocells and microcells. All the slow and fast calls can naturally share the common channel resources provided by the two layers, and thus give the best performance compared with unidirectional call-overflow schemes. Although call overflow can produce higher overhead and more handoffs, by simulation there is little difference among three schemes (i.e., bidirectional, unidirectional and no overflow schemes) from the aspect of the number of successful handoffs per call when the traffic is not very high. A drawback of [13] is that it does not consider the successive operations of overflow calls after they have overflowed. Obviously, the successful overflow calls should either work on the non-referable layer until completion, or return to their referable layer if possible. As mentioned, the later case is obviously better since it can enable better resource configuration between different layers, higher system capacity and better quality of service (QoS). This is also shown in [14] by numerical examples through a comparative study among the bidirectional overflow and take-back scheme, no overflow scheme, unidirectional overflow scheme, unidirectional overflow and take-back scheme, and bidirectional overflow without take-back scheme. However, the handoff calls of fast users and slow users in [14] are treated the same as the corresponding new calls (i.e., no prioritization of handoff calls is considered), so no service differentiation can be achieved. Moreover, a common drawback of [13] and [14] is that the channel occupancy times of slow and fast users are assumed to be exponentially distributed, as discussed in many other publications related to the performance evaluation of HCNs.

It is convenient and tractable to assume the exponential distribution of channel occupancy time [1–8,10,12–14]. However, recent studies show that channel occupancy times and interarrival times of cell traffic are no longer exponentially distributed [27–29]. The authors in [27] conclude that channel occupancy times and related time variables are not exponentially distributed from a series of tests for mobile cellular systems, and a lognormal distribution or a mixture of Erlang distributions gives a better statistical fitting to the field data. In [28], the cell residence time, which may have various probabilistic characteristics depending upon different mobility environments, is modeled as the sum of hyper exponential (SOHYP) random variables. In [29], a mobility model where the cell residence time is characterized by a hyper-Erlang distribution is analyzed for PCS networks.

In this paper, a two-layer hierarchical cellular network with bidirectional overflow and take-back strategies under generally distributed cell residence times is considered. The macrolayer and microlayer are respectively the referable layers for fast and slow user requests (including new and handoff calls). A call from a fast user will overflow to the microlayer if there is no channel available. The successful overflow call can be taken back to its macrolayer if a channel becomes available. A similar procedure is applied to a call from a slow user. The cell residence times (and thus channel occupancy times) of fast and slow users in different layers are modeled as general distribution to adapt various mobility environments. The Laplace transform approach is employed to determine the channel occupancy times for different types of calls (i.e., new call, handoff call, overflow call and take-back call of fast and slow users). This approach would also be useful in some other queuing models, for instance, with a buffer to queue the new or handoff requests due to lack of available channels. An analytical model is developed, and some important performance measures such as the new call blocking probability and handoff dropping probability of fast and slow users for each layer and for the network, the carried traffic of each type of user in different layers, and the forced-termination probability in the HCN are evaluated.

The remainder of the paper is organized as follows: Section 2 proposes the system model and related assumptions; Section 3 derives the various arrival rates; Section 4 derives the various channel occupancy times; Section 5 determines the stationary probability in each layer by using the results of Sections 3 and 4; Section 6 derives some interesting performance measures; Section 7 presents the numerical results and discussions; and Section 8 concludes the paper.

2. Model description and assumptions

Consider a two-layer hierarchical cellular network consisting of a set of macrocells and microcells, where each macrocell overlays N microcells. Assume that there are two classes of user mobility: the slow user (e.g., pedestrian) and the fast user (e.g., vehicle). A call from a slow user is first directed to a microcell. If it is blocked, it will overflow to the macrocell overlaying the microcell. A call from a fast user is first directed to a macrocell. If it is blocked, it will overflow to one of the underlying microcells. There are many methods in the literature to handle the layer selection, such as layer selection by user speed [11,15–17], layer selection by residence time [6,18,19], and fuzzy layer selection [20,21]. In this analysis, we assume that the problem of the optimum classification of mobile users has been solved. The successful overflow calls can be taken back to their preferable layers if traffic channels become available. The take-back process can be executed immediately when one of the channels in the preferable layer becomes available, or executed when the overflow call crosses the boundary of a microcell. Here we choose the later case for simply modeling the channel occupancy times of overflow and take-back calls (see details in the next section). We further assume that the user speed does not change greatly during a call session (so that the speed of each type of call is not out of its threshold) and a macrocell exactly covers its underlying N microcells.

Due to the limited resource (channels) in wireless communications, calls with different priorities should be treated differently in resource allocation. A handoff call is usually given higher priority over a new call to maintain lower handoff dropping probability, because dropping an ongoing call is less desirable than blocking a new call from the perception of mobile users. Various handoff priority-based channel allocation schemes have been proposed in the literature, such as cutoff priority scheme [22], queuing priority scheme [23], and new call bounding scheme [24], etc. We choose the new call bounding scheme for each layer of the HCN due to its simplicity: if the number of new calls in a cell (microcell or macrocell) exceeds a threshold when a new call arrives, the new call will be blocked; otherwise it will be admitted. The handoff call is blocked in a current cell only when all channels in the cell are used up. Unlike the case of single-layer cellular networks in [24], the call admission and resource allocation strategy has some successive processes in HCNs, i.e., the blocked call in a layer does not necessarily have to be dropped from the network and overflow to another layer. Furthermore, the successful overflow call can return to its preferable layer when a channel becomes available in that layer.

Each macrocell is allocated C_M channels, of which C_A and C_B channels are the maximum value allowed for new calls and overflow calls, respectively. Each microcell is allocated C_m channels, of which C_a and C_b channels are the maximum value allowed for new calls and overflow calls, respectively. Here the term of "channel" refers to a generic logical channel (e.g., a logical channel could be a double physical channel or several time slots for communication in the up- and down-link directions with a generic multiple access method). Then, the call admission and resource allocation strategies in each macrocell and microcell are as follows.

Macrocell:

- A new call from a fast user is rejected and attempts to overflow to a microcell when there are C_A ongoing new calls of fast users or C_M channels are busy in the macrocell.
- A handoff call from a fast user is rejected and attempts to overflow to a microcell when C_M channels are busy.
- A take-back call from a fast user is rejected (and attempts a handoff to an adjacent microcell) when C_M channels are busy (a take-back call is considered to have the same priority as a handoff call).

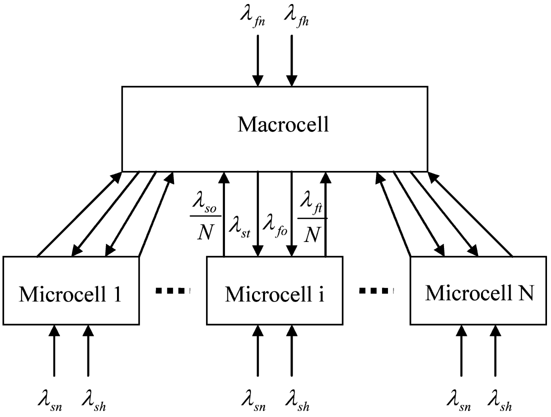


Fig. 1 The two-layer hierarchical cellular network model

- An overflow call from a slow user is rejected when there are C_B ongoing overflow calls or C_M channels are busy in the macrocell.

Microcell:

- A new call from a slow user is rejected and attempts to overflow to a macrocell when there are C_a ongoing new calls of slow users or C_m channels are busy in the microcell.
- A handoff call from a slow user is rejected and attempts overflow to a macrocell when C_m channels are busy.
- A take-back call from a slow user is rejected (and continues to exploit the macrocell resource) when C_m channels are busy.
- An overflow call from a fast user is rejected when there are C_b ongoing overflow calls or C_m channels are busy in the microcell.

A simple block diagram of the two-layer HCN model is shown in Fig. 1.

In the development of the model we introduce the following assumptions:

- The arrival process is a Poisson distribution for new calls and handoff calls. The mean new call arrival rates of fast and slow users to the corresponding cells are λ_{fn} and λ_{sn}, respectively. The mean handoff call arrival rates of fast and slow users to the corresponding cells are λ_{fh} and λ_{sh}, respectively. (The mean handoff call arrival rates will be derived by balancing the incoming and outgoing handoff rates, as explained later).
- The arrival process is a Poisson distribution for overflow and take-back traffic. The mean arrival rates of overflow and take-back calls to a macrocell (or microcell) are λ_{so} and λ_{ft} (or λ_{fo} and λ_{st}), respectively. (The mean arrival rates of overflow and take-back calls will be derived later).
- The cell residence times of each type of user in different layers are of general distributions. Let random variables R_{sm} and R_{fM} denote the cell residence times of slow and fast users in a microcell and macrocell with mean $1/r_{sm}$ and $1/r_{fM}$, and R^r_{sm} and R^r_{fM} denote the corresponding residual cell residence times, respectively. Let random variables R_{fm} and R_{sM} denote the cell residence times of fast and slow users in a microcell and macrocell with mean $1/r_{fm}$ and $1/r_{sM}$, and R^r_{fm} and R^r_{sM} denote the corresponding residual cell residence times, respectively.
- The unencumbered call holding times of fast and slow users are of negative exponential distributions. Let random variables H_M and H_m denote the call holding times of fast and slow users with mean $1/h_M$ and $1/h_m$, and H^r_M and H^r_m denote the corresponding residual call holding times, respectively.
- All macrocells and microcells are stochastically identical in each layer.

Strictly speaking, the handoff and overflow arrival processes may not be exactly Poisson processes because some correlations exist among the call types and the nature of overflow processes. Particularly, overflow traffic is typically rather bursty and usually modeled as an interrupted Poisson process (IPP) [5], a Markov modulated Poisson process (MMPP) [25], or a renewal process [9]. However, for convenience and tractability, the overflow process is also simply approximated as a Poisson process in some literatures, such as [13,14, 26]. The Poisson assumption is often adopted in the analysis of queuing networks, and it is reported that the performance is little degraded under this assumption [9].

3. The various arrival rates

In this section, we derive the arrival rates of handoff calls, overflow calls and take-back calls in each layer. The notations of different arrival rates and time variables are listed in nomenclature.

Consider the macrocell:

The total arrival rate to the macrocell is the sum of the arrival rates of new calls λ_{fn}, handoff calls λ_{fh} and take-back calls λ_{ft} of fast users, as well as the overflow calls of slow users λ_{so}. Note that handoff requests of fast users out of a macrocell may include the following three cases: (a) new calls not blocked attempt to handoff; (b) handoff calls not dropped attempt to handoff; (c) the successful take-back calls of fast users attempt to handoff. By balancing the incoming and outgoing handoff rates in steady state, we have

$$\lambda_{fh} = \lambda_{fn}\big(1 - P^n_{fM}\big)p_{f1} + \lambda_{fh}\big(1 - P^h_{fM}\big)p_{f2} + \lambda_{ft}(1 - P_{ft})\, p_{f3}, \tag{1}$$

where $p_{f1} = P(R^r_{fM} < H_M)$ denotes the probability that a new call from a fast user initiated from a macrocell will continue its session to a neighboring macrocell, $p_{f2} = P(R_{fM} < H^r_M)$ denotes the probability that a handoff call from a fast user will continue its session to a neighboring macrocell, $p_{f3} = P(R^r_{fM} < H^r_M)$ denotes the probability that a take-back call from a fast user will continue its session to a neighboring macrocell. Note that a take-back call has already dwelled some period time in the macrocell through the underlying microcell before it goes back to macrolayer. p_{f1} can be calculated by the Laplace transform approach:

$$p_{f1} = \int_0^\infty \int_t^\infty f_{H_M}(\tau) f_{R^r_{fM}}(t) d\tau dt = \frac{r_{fM}[1 - f^*_{R_{fM}}(h_M)]}{h_M}.$$

The last equation was obtained by the fact that $f_{R^r_{fM}}(t) = r_{fM}(1 - F_{R_{fM}}(t))$, where $f_{R^r_{fM}}(t)$ is the probability density function (pdf) of R^r_{fM}, $F_{R_{fM}}(t)$ is the cumulative distribution function (cdf) of R_{fM}. Similarly, p_{f2} and p_{f3} can be determined as

$$p_{f2} = \int_0^\infty \int_t^\infty f_{H^r_M}(\tau) f_{R_{fM}}(t) d\tau dt = f^*_{R_{fM}}(h_M), \quad p_{f3} = p_{f1} = \frac{r_{fM}\big[1 - f^*_{R_{fM}}\big(h_M\big)\big]}{h_M},$$

where $f_{H^r_M}(t) = f_{H_M}(t)$ is used in above equations due to the memoryless property of the exponential distribution.

The take-back requests of fast users may be from the following three cases: (a) overflowed new calls; (b) overflowed handoff calls; (c) the take-back calls of fast users not succeeded at previous attempts (since the take-back operations can be executed at the boundary of every microcell). Thus, the following equation should hold in steady state:

$$\lambda_{ft} = \lambda_{fn} P^n_{fM}(1 - P_{fo}) p_{f4} + \lambda_{fh} P^h_{fM}(1 - P_{fo}) p_{f5} + \lambda_{ft} P_{ft} p_{f6}, \qquad (2)$$

where $p_{f4} = P(R^r_{fm} < H_M)$ denotes the probability that an overflowed new call from a fast user in a microcell attempts a take-back operation at the boundary of the microcell, $p_{f5} = P(R_{fm} < H^r_M)$ denotes the probability that an overflowed handoff call from a fast user attempts a take-back operation at the boundary of the microcell, $p_{f6} = P(R_{fm} < H^r_M)$ denotes the probability that a take-back call from a fast user (not succeeded at previous attempts) continues a take-back operation at the boundary of the microcell. Using the Laplace transform approach, p_{f4}, p_{f5} and p_{f6} can be easily determined by

$$p_{f4} = \int_0^\infty \int_t^\infty f_{H_M}(\tau) f_{R^r_{fm}}(t) d\tau dt = \frac{r_{fm}\big[1 - f^*_{R_{fm}}(h_M)\big]}{h_M},$$

$$p_{f5} = p_{f6} = \int_0^\infty \int_t^\infty f_{H^r_M}(\tau) f_{R_{fm}}(t) d\tau dt = f^*_{R_{fm}}(h_M).$$

From Eqs. (1) and (2), we can solve

$$\lambda_{fh} = \frac{\lambda_{fn}}{Z_f} \big\{(1 - P^n_{fM})(1 - P_{ft} p_{f6}) p_{f1} + P^n_{fM}(1 - P_{fo})(1 - P_{ft}) p_{f3} p_{f4}\big\}, \qquad (3)$$

$$\lambda_{ft} = \frac{\lambda_{fn}}{Z_f} \Big\{P^n_{fM}[1 - (1 - P^h_{fM}) p_{f2}](1 - P_{fo}) p_{f4} + P^h_{fM}(1 - P^n_{fM})(1 - P_{fo}) p_{f1} p_{f5}\Big\}, \qquad (4)$$

where $Z_f = [1 - (1 - P^h_{fM}) p_{f2}](1 - P_{ft} p_{f6}) - P^h_{fM}(1 - P_{fo})(1 - P_{ft}) p_{f3} p_{f5}$.

The arrival rate of the overflow calls of slow users into the macrocell is easily determined by

$$\lambda_{so} = N\big(\lambda_{sn} P^n_{sm} + \lambda_{sh} P^h_{sm}\big). \qquad (5)$$

Consider the microcell:

The total arrival rate to a microcell is the sum of the arrival rates of new calls λ_{sn}, handoff calls λ_{sh} and take-back calls λ_{st} of slow users, as well as the overflow calls of fast users λ_{fo}. By using similar analysis to the macrocell, we have

$$\lambda_{sh} = \lambda_{sn}\big(1 - P^n_{sm}\big) p_{s1} + \lambda_{sh}\big(1 - P^h_{sm}\big) p_{s2} + \lambda_{st}(1 - P_{st}) p_{s3}, \qquad (6)$$

where $p_{s1} = P(R^r_{sm} < H_m)$ denotes the probability that a new call from a slow user initiated from a microcell will continue its session to a neighboring microcell, $p_{s2} = P(R_{sm} < H^r_m)$ denotes the probability that a handoff call from a slow user will continue its session to a neighboring microcell, $p_{s3} = P(R_{sm} < H^r_m)$ denotes the probability that a take-back call from a slow user will continue its session to a neighboring microcell (note the difference between the microcell and macrocell case). These probabilities can be easily calculated by the Laplace transform approach as:

$$p_{s1} = \int_0^\infty \int_t^\infty f_{H_m}(\tau) f_{R^r_{sm}}(t)d\tau dt = \frac{r_{sm}\left[1 - f^*_{R_{sm}}(h_m)\right]}{h_m},$$

$$p_{s2} = p_{s3} = \int_0^\infty \int_t^\infty f_{H^r_m}(\tau) f_{R_{sm}}(t)d\tau dt = f^*_{R_{sm}}(h_m).$$

The take-back requests of slow users may be from the overflowed new calls, the overflowed handoff calls and the take-back calls of slow users not succeeded at previous attempts. Thus, the following equation should hold in steady state:

$$\lambda_{st} = \lambda_{sn} P^n_{sm}(1 - P_{so})p_{s1} + \lambda_{sh} P^h_{sm}(1 - P_{so})p_{s2} + \lambda_{st} P_{st} p_{s3}. \tag{7}$$

From Eqs. (6) and (7), we can solve

$$\lambda_{sh} = \frac{\lambda_{sn}}{Z_s}\left\{\left(1 - P^n_{sm}\right)(1 - P_{st}p_{s3})p_{s1} + P^n_{sm}(1 - P_{so})(1 - P_{st})p_{s1}p_{s3}\right\}, \tag{8}$$

$$\lambda_{st} = \frac{\lambda_{sn}}{Z_s}\left\{P^n_{sm}\left[1 - \left(1 - P^h_{sm}\right)p_{s2}\right](1 - P_{so})p_{s1} + P^h_{sm}\left(1 - P^n_{sm}\right)(1 - P_{so})p_{s1}p_{s2}\right\}, \tag{9}$$

where $Z_s = [1 - (1 - P^h_{sm})p_{s2}](1 - P_{st}p_{s3}) - P^h_{sm}(1 - P_{so})(1 - P_{st})p^2_{s3}$.

The arrival rate of the overflow calls of fast users into a microcell (assume the overflow calls from one macrocell are distributed into the N microcells on average) is easily determined by

$$\lambda_{fo} = \frac{1}{N}\left(\lambda_{fn} P^n_{fM} + \lambda_{fh} P^h_{fM}\right). \tag{10}$$

The various arrival rates obtained above will be used for solving the stationary probabilities in Section 5.

4. The various channel occupancy times

In this section, we use the Laplace transform approach to analyze the various channel occupancy times. This analytical approach is very popular in wireless mobile networks [9,29,30] where [29,30] use it in PCS networks, and [9] uses it in hierarchical cellular networks.

Consider the microcell:

Let T^n_{sm}, T^h_{sm} and T^t_{sm} be the channel occupancy times for new calls, handoff calls and take-back calls of slow users in a microcell, respectively. Then we have

$$T^n_{sm} = \min\{H_m, R^r_{sm}\},\ T^h_{sm} = \min\{H^r_m, R_{sm}\} \text{ and } T^t_{sm} = \min\{H^r_m, R_{sm}\}.$$

From the Theorem 2 of [30], we obtain respectively their Laplace transforms

$$f^*_{T^n_{sm}}(s) = \frac{h_m}{s + h_m} + \frac{s r_{sm}}{(s + h_m)^2}\left[1 - f^*_{R_{sm}}(s + h_m)\right],$$

$$f^*_{T^h_{sm}}(s) = f^*_{T^t_{sm}}(s) = \frac{h_m}{s+h_m} + \frac{s}{s+h_m} f^*_{R_{sm}}(s+h_m),$$

and their means

$$E\left[T^n_{sm}\right] = \frac{1}{h_m} - \frac{r_{sm}}{h_m^2}\left[1 - f^*_{R_{sm}}(h_m)\right], \tag{11}$$

$$E\left[T^h_{sm}\right] = E\left[T^t_{sm}\right] = \frac{1}{h_m}\left[1 - f^*_{R_{sm}}(h_m)\right]. \tag{12}$$

Let T^o_{fm} be the channel occupancy time for the overflow calls of fast users in a microcell. It consists of the channel occupancy times of the overflow new calls, T^n_{fm}, and the overflow handoff calls, T^h_{fm}, of fast users in a microcell. Using the similar analysis to Section A, we have $T^n_{fm} = \min\{H_M, R^r_{fm}\}$and $T^h_{fm} = \min\{H^r_M, R_{fm}\}$.

Then we obtain the Laplace transforms of the probability density functions (pdfs) of the random variables

$$f^*_{T^n_{fm}}(s) = \frac{h_M}{s+h_M} + \frac{sr_{fm}}{(s+h_M)^2}\left[1 - f^*_{R_{fm}}(s+h_M)\right],$$

$$f^*_{T^h_{fm}}(s) = \frac{h_M}{s+h_M} + \frac{s}{s+h_M} f^*_{R_{fm}}(s+h_M),$$

and their means

$$E\left[T^n_{fm}\right] = \frac{1}{h_M} - \frac{r_{fm}}{h_M^2}\left[1 - f^*_{R_{fm}}(h_M)\right], \quad E\left[T^h_{fm}\right] = \frac{1}{h_M}\left[1 - f^*_{R_{fm}}(h_M)\right].$$

Hence, we obtain the Laplace transform of the pdf of T^o_{fm}

$$f^*_{T^o_{fm}}(s) = q f^*_{T^n_{fm}}(s) + (1-q) f^*_{T^h_{fm}}(s),$$

where q represents the fraction of the overflow new calls to all the overflow requests of fast users and is given by $q = \frac{\lambda_{fn}P^n_{fM}}{\lambda_{fn}P^n_{fM} + \lambda_{fh}P^h_{fM}}$. Then we obtain the mean of T^o_{fm}

$$E\left[T^o_{fm}\right] = \frac{1}{h_M} - \frac{\lambda_{fn}P^n_{fM}}{\lambda_{fn}P^n_{fM} + \lambda_{fh}P^h_{fM}} \frac{r_{fm}}{h_M^2}\left[1 - \left(1 - \frac{h_M\lambda_{fh}P^h_{fM}}{r_{fm}\lambda_{fn}P^n_{fM}}\right) f^*_{R_{fm}}(h_M)\right]. \tag{13}$$

Consider the macrocell:

Let T^n_{fM}, T^h_{fM} and T^t_{fM} be the channel occupancy times for new calls, handoff calls and take-back calls of fast users in a macrocell, respectively. Then we have $T^n_{fM} = \min\{H_M, R^r_{fM}\}$, $T^h_{fM} = \min\{H^r_M, R_{fM}\}$ and $T^t_{fM} = \min\{H^r_M, R^r_{fM}\}$.

Similar to the microcell, we obtain the means of T^n_{fM}, T^h_{fM} and T^t_{fM}, respectively.

$$E\left[T^n_{fM}\right] = E\left[T^t_{fM}\right] = \frac{1}{h_M} - \frac{r_{fM}}{h_M^2}\left[1 - f^*_{R_{fM}}(h_M)\right], \tag{14}$$

$$E\left[T_{fM}^{h}\right] = \frac{1}{h_M}\left[1 - f_{R_{fM}}^{*}(h_M)\right]. \tag{15}$$

Let T_{sM}^{o} be the channel occupancy time for the overflow calls of slow users in a macrocell. It consists of the channel occupancy times of the overflow new calls, T_{sM}^{n}, and the overflow handoff calls, T_{sM}^{h}, of slow users in a macrocell. Then we have

$$T_{sM}^{n} = \min\left\{H_m, R_{sM}^{r}\right\} \quad \text{and} \quad T_{sM}^{h} = \min\left\{H_m^{r}, R_{sM}^{r}\right\}.$$

Then we obtain the Laplace transforms of the pdfs of the random variables and their means

$$f_{T_{sM}^{n}}^{*}(s) = f_{T_{sM}^{h}}^{*}(s) = \frac{h_m}{s + h_m} + \frac{s r_{sM}}{(s + h_m)^2}\left[1 - f_{R_{sM}}^{*}(s + h_m)\right],$$

$$E\left[T_{sM}^{n}\right] = E\left[T_{sM}^{h}\right] = \frac{1}{h_m} - \frac{r_{sM}}{h_m^2}\left[1 - f_{R_{sM}}^{*}(h_m)\right].$$

Hence, we obtain the Laplace transform of the pdf of T_{sM}^{o} and its mean

$$f_{T_{sM}^{o}}^{*}(s) = f_{T_{sM}^{n}}^{*}(s) = f_{T_{sM}^{h}}^{*}(s), \quad E[T_{sM}^{o}] = \frac{1}{h_m} - \frac{r_{sM}}{h_m^2}\left[1 - f_{R_{sM}}^{*}(h_m)\right]. \tag{16}$$

The means of various channel occupancy times obtained above will be used for solving the stationary probabilities in Section 5.

5. The stationary probabilities

From the model description, we know that the total arrival process to each layer is a Poisson process and the channel occupancy time is a random variable whose distribution may depend on the type of the call (or customer termed in queueing networks), so we can model the microcell and macrocell as stationary symmetric queues, and describe them as loss systems with multi-servers (channels). This kind of system is known to have an insensitivity property, i.e., the stationary probability depends on the service (channel occupancy time) distribution only through its mean [31]. For a macrocell, let $\pi_M(i, j, k, l)$ denote the joint stationary probability that i is the number of new calls of fast users, j is the number of handoff calls of fast users, k is the number of take-back calls of fast users and l is the number of overflow calls of slow users, respectively. From the theory of loss system [9,31,32], we have

$$\pi_M(i, j, k, l) = G_M^{-1}\frac{\left(\lambda_{fn}E\left[T_{fM}^{n}\right]\right)^i\left(\lambda_{fh}E\left[T_{fM}^{h}\right]\right)^j\left(\lambda_{ft}E\left[T_{fM}^{t}\right]\right)^k\left(\lambda_{so}E\left[T_{sM}^{o}\right]\right)^l}{i!j!k!l!}, \tag{17}$$

$$(i, j, k, l) \in \Psi \equiv \{i \le C_A, \quad j \le C_M, \quad k \le C_M, \quad l \le C_B, \quad i + j + k + l \le C_M\}$$

where G_M is the normalization constant and given by

$$G_M = \sum_{(i,j,k,l)\in\Psi}\frac{\left(\lambda_{fn}E\left[T_{fM}^{n}\right]\right)^i\left(\lambda_{fh}E\left[T_{fM}^{h}\right]\right)^j\left(\lambda_{ft}E\left[T_{fM}^{t}\right]\right)^k\left(\lambda_{so}E\left[T_{sM}^{o}\right]\right)^l}{i!j!k!l!}.$$

Similarly, for a microcell, let $\pi_m(i, j, k, l)$ denote the joint stationary probability that i is the number of new calls of slow users, j is the number of handoff calls of slow users, k is the number of take-back calls of slow users and l is the number of overflow calls of fast users, respectively. Then we have

$$\pi_m(i, j, k, l) = G_m^{-1} \frac{\left(\lambda_{sn} E\left[T_{sm}^n\right]\right)^i \left(\lambda_{sh} E\left[T_{sm}^h\right]\right)^j \left(\lambda_{st} E\left[T_{sm}^t\right]\right)^k \left(\lambda_{fo} E\left[T_{fm}^o\right]\right)^l}{i!j!k!l!}, \tag{18}$$

$$(i, j, k, l) \in \Omega \equiv \left\{ i \leq C_a,\ j \leq C_m,\ k \leq C_m,\ l \leq C_b,\ i + j + k + l \leq C_m \right\}$$

where G_m is the normalization constant and given by

$$G_m = \sum_{(i,j,k,l)\in\Omega} \frac{\left(\lambda_{sn} E\left[T_{sm}^n\right]\right)^i \left(\lambda_{sh} E\left[T_{sm}^h\right]\right)^j \left(\lambda_{st} E\left[T_{sm}^t\right]\right)^k \left(\lambda_{fo} E\left[T_{fm}^o\right]\right)^l}{i!j!k!l!}.$$

6. The performance measures

6.1. The basic performance measures

For the macrocell

: the blocking of a new call from a fast user occurs when the number of channels occupied by new calls is equal to C_A, or there is no available channel in the macrocell (even though the number of channels occupied by new calls is less than C_A). Hence, the new call blocking probability of fast users in a macrocell is

$$P_{fM}^n = \sum_{\substack{i=C_A \ or \\ i+j+k+l=C_M}} \pi_M(i, j, k, l). \tag{19}$$

The handoff failure of a handoff call from a fast user occurs when all the C_M channels in the macrocell are occupied. Hence, the handoff failure probability of fast users in a macrocell is

$$P_{fM}^h = \sum_{i+j+k+l=C_M} \pi_M(i, j, k, l). \tag{20}$$

Obviously, when $C_A = C_M$, we have $P_{fM}^n = P_{fM}^h$ from the definitions, this becomes the non-prioritized handoff scheme in the macrocell.

A take-back request from a fast user fails when all the channels assigned to them in the macrocell are occupied. From Section 2, we know the take-back calls have the same priority as the handoff calls. Hence, the take-back failure probability is $P_{ft} = P_{fM}^h$.

The overflow request from a slow user fails when the number of channels occupied by the overflow calls is equal to C_B, or there is no available channel in the macrocell. Hence, the overflow failure probability of slow users in a macrocell is

$$P_{so} = \sum_{\substack{l=C_B \ or \\ i+j+k+l=C_M}} \pi_M(i, j, k, l). \tag{21}$$

For the microcell:

the blocking of a new call from a slow user occurs when the number of channels occupied by new calls is equal to C_a, or there is no available channel in the microcell. Hence, the new call blocking probability of slow users in a microcell is

$$P^n_{sm} = \sum_{\substack{i=C_a \text{ or} \\ i+j+k+l=C_m}} \pi_m(i, j, k, l). \tag{22}$$

Similarly, the handoff failure probability of slow users in a microcell is

$$P^h_{sm} = \sum_{i+j+k+l=C_m} \pi_m(i, j, k, l). \tag{23}$$

The take-back failure probability of slow users in a microcell is: $P_{st} = P^h_{sm}$. The overflow request from a fast user fails when the number of channels occupied by the overflow calls is equal to C_b, or there is no available channel in the microcell. Hence, the overflow failure probability of fast users in a microcell is

$$P_{fo} = \sum_{\substack{l=C_b \text{ or} \\ i+j+k+l=C_m}} \pi_m(i, j, k, l). \tag{24}$$

For the network:

from previous analysis, we know that a new call will not necessarily be blocked from the HCN when it is rejected in its preferable layer, nor will a handoff call. Now we are interested in the final blocking (or dropping) probability of new calls (or handoff calls) of different types of users in the HCN. By the PASTA (Poisson arrivals see time average) property, the final blocking and dropping probabilities of new calls and handoff calls of fast users in the HCN are respectively

$$P^n_f = P^n_{fM} P_{fo} \quad \text{and} \quad P^h_f = P^h_{fM} P_{fo}. \tag{25}$$

The final blocking and dropping probabilities of new calls and handoff calls of slow users in the HCN are respectively

$$P^n_s = P^n_{sm} P_{so} \quad \text{and} \quad P^h_s = P^h_{sm} P_{so}. \tag{26}$$

6.2. Carried traffic distribution

After obtaining the stationary probabilities of each layer, we can easily determine the carried traffic distributions of different types of calls in each layer. Let CT^f_M and CT^s_M denote the carried traffic of fast and slow users in the macrocell, and CT^f_m and CT^s_m denote the carried traffic of fast and slow users in the microcell, respectively. Then, we have

$$CT^f_M = \sum_{(i,j,k,l)\in\Psi} (i + j + k) \cdot \pi_M(i, j, k, l). \tag{27}$$

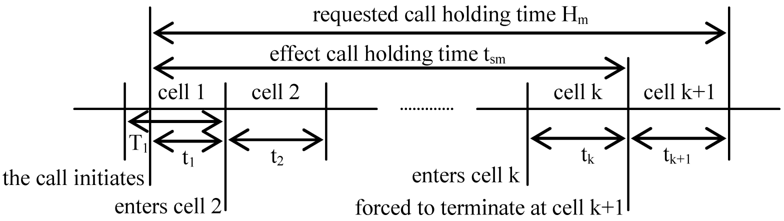


Fig. 2 The timing diagram for a forced-termination call

$$CT^s_M = \sum_{(i,j,k,l)\in\Psi} l \cdot \pi_M(i,j,k,l).. \tag{28}$$

$$CT^f_m = \sum_{(i,j,k,l)\in\Omega} l \cdot \pi_m(i,j,k,l). \tag{29}$$

$$CT^s_m = \sum_{(i,j,k,l)\in\Omega} (i+j+k) \cdot \pi_m(i,j,k,l). \tag{30}$$

The total carried traffic in each macrocell and microcell, CT_M and CT_m, are thus respectively

$$CT_M = CT^f_M + CT^s_M,$$

and

$$CT_m = CT^f_m + CT^s_m.$$

6.3. The forced-termination probabilities

We have already got the following probabilities: P^n_{fM}, P^h_{fM}, P^n_{sm}, P^h_{sm}, P_{fo}, P_{ft}, P_{so} and P_{st}. Now we are interested in the probability of a forced-termination call. We separate the problem into several parts: (a) the forced-termination probability of a call from a slow user in the microlayer, P^{ft}_{sm}; (b) the forced-termination probability of a call from a fast user in the macrolayer, P^{ft}_{fM}; (c) the forced-termination probability of a call from a slow user in the macrolayer, P^{ft}_{sM}; (d) the forced-termination probability of a call from a fast user in the microlayer, P^{ft}_{fm}; and the total forced-termination probability in the HCN, P^{ft}_{total}. (a) Find P^{ft}_{sm}, i.e., the probability that a call from a slow user is forced to terminate from the system at the k-th handoff (note that the call is connected with probability $1 - P^n_{sm}$, and then makes $k-1$ successful handoffs with probability $(1 - P^h_{sm})^{k-1}$, and is forced to terminate at the k-th handoff with probability $P^h_{sm}P_{so}$). Let's consider the timing diagram of a forced-termination call in Fig. 2. T_1 and t_1 are the cell residence time and residual cell residence time of a call from a slow user in microcell 1, respectively; t_i is the cell residence time of a call from a slow user in microcell i $(i \geq 2)$.

According to previous assumptions, T_1, t_2, ..., t_k, ... are independent and identically distributed random variables, which have the same pdf $f_{R_{sm}}(\cdot)$ and cdf $F_{R_{sm}}(\cdot)$ with mean $1/r_{sm}$. Let $f^*_{t_1}(s)$ and $f^*_{t_i}(s)$ be the Laplace transforms of the pdfs of the random variables t_1 and t_i $(i \geq 2)$, respectively. The effective call holding time t_{sm} is the summation of t_i $(i \geq 1)$, i.e., $t_{sm} = t_1 + t_2 + \cdots + t_k$. Let $f_{t_{sm}}(\cdot)$ and $f^*_{t_{sm}}(s)$ be the pdf and the corresponding Laplace transform of the random variable t_{sm}. Then, from the theory of Laplace transform [33], we have

$$f^*_{t_1}(s) = \int_0^\infty r_{sm}[1 - F_{R_{sm}}(\tau)e^{-s\tau}d\tau = \frac{r_{sm}}{s}\big[1 - f^*_{R_{sm}}(s)\big],$$

$$f^*_{t_2}(s) = f^*_{t_3}(s) = \cdots = f^*_{t_k}(s) = f^*_{R_{sm}}(s),$$

and

$$f^*_{t_{sm}}(s) = \prod_{i=1}^{k} f^*_{t_i}(s) = \frac{r_{sm}}{s}\big[1 - f^*_{R_{sm}}(s)\big]\big[f^*_{R_{sm}}(s)\big]^{k-1}. \tag{31}$$

By summing all the probabilities for $1 \le k < \infty$, the probability P^{ft}_{sm} is determined as

$$\begin{aligned}
P^{ft}_{sm} &= \sum_{k=1}^{\infty}\big(1 - P^n_{sm}\big)\big(1-P^h_{sm}\big)^{k-1}P^h_{sm}P_{so}\cdot \Pr(t_1 + t_2 + \cdots + t_k \le H_m)\\
&= \sum_{k=1}^{\infty}\big(1 - P^n_{sm}\big)\big(1-P^h_{sm}\big)^{k-1}P^h_{sm}P_{so}\cdot \int_0^\infty f_{t_{sm}}(t)\Pr\left(H_m \ge t|t_1 + t_2 + \cdots + t_k \le t\right)dt\\
&= \sum_{k=1}^{\infty}\big(1 - P^n_{sm}\big)\big(1-P^h_{sm}\big)^{k-1}P^h_{sm}P_{so}\cdot \int_0^\infty \int_t^\infty f_{t_{sm}}(t)f_{H_m}(\tau)\,d\tau dt\\
&= \sum_{k=1}^{\infty}\big(1 - P^n_{sm}\big)\big(1-P^h_{sm}\big)^{k-1}P^h_{sm}P_{so}\cdot f^*_{t_{sm}}(h_m).
\end{aligned}$$

Substituting Eq. (33) into the above expression, we have

$$\begin{aligned}
P^{ft}_{sm} &= \sum_{k=1}^{\infty}\big(1 - P^n_{sm}\big)\big(1-P^h_{sm}\big)^{k-1}P^h_{sm}P_{so}\cdot \frac{r_{sm}}{h_m}\big[1 - f^*_{R_{sm}}(h_m)\big]\big[f^*_{R_{sm}}(h_m)\big]^{k-1}\\
&= \frac{r_{sm}\big(1 - P^n_{sm}\big)P^h_{sm}P_{so}\big[1 - f^*_{R_{sm}}(h_m)\big]}{h_m\big[1 - \big(1 - P^h_{sm}\big)f^*_{R_{sm}}(h_m)\big]}.
\end{aligned} \tag{32}$$

(b) The probability P^{ft}_{fM} can be determined by following the same method.

$$P^{ft}_{fM} = \frac{r_{sm}\big(1 - P^n_{sm}\big)P^h_{sm}P_{so}[1 - f^*_{R_{sm}}(h_m)]}{h_m\big[1 - \big(1 - P^h_{sm}\big)f^*_{R_{sm}}(h_m)\big]}. \tag{33}$$

(c) Find P^{ft}_{sM}, i.e., the probability that an overflow call from a slow user in the macrolayer is forced to terminate from the system at the k-th macrocell-level overflow. It is noteworthy that one macrocell covers N microcells. For the first time a call from a slow user overflow to a macrocell with probability $1 - P_{so}$, then it attempts a take-back operation at each boundary of a microcell; assume it has failed x_0 take-back operations and reaches the boundary of the current macrocell; we say the call finishes its 1st macrocell-level overflow with probability $A_1 = (1 - P_{so})P^{x_0}_{st}$. Then it attempts a take-back operation with failure and an overflow operation to the target macrocell with succession, and continues to attempt x−1 take-back operations with failure and reaches the boundary of the macrocell; we say the call finishes its 2nd macrocell-level overflow with probability $A_2 = (1 - P_{so})P_{st}^x$. When the call makes $k - 1$ successful macrocell-level overflows with probability A_2^{k-1}, it is forced to terminate from the system at the k-th macrocell-level overflow with probability $P_{st}P_{so}$. By summing all the probabilities for $1 \leq k < \infty$ and following the same method as (a), the probability P_{sM}^{ft} is determined as

$$P_{sM}^{ft} = \frac{r_{sM} A_1 P_{st} P_{so}\big[1 - f^*_{R_{sM}}(h_m)\big]}{h_m\big[1 - A_2 f^*_{R_{sm}}(h_m)\big]} \quad . \tag{34}$$

(d) The probability P_{fm}^{ft} can be determined by using the same analysis as (c).

$$P_{fm}^{ft} = \frac{r_{sM} B_1 P_{ft} P_{fo}\big[1 - f^*_{R_{fm}}(h_M)\big]}{h_M\big[1 - B_2 f^*_{R_{fM}}(h_M)\big]}, \tag{35}$$

where $B_1 = (1 - P_{fo})$ is the probability of 1^{st} microcell-level overflow, $B_2 = P_{ft}(1 - P_{fo})$ is the probability of i-th microcell-level overflow ($i \geq 2$).

e) After obtaining the above four probabilities, the total forced-termination probability in the HCN can be easily calculated by

$$P_{total}^{ft} = \frac{N(\lambda_{sn} + \lambda_{sh} + \lambda_{st})P_{sm}^{ft} + N\lambda_{fo}P_{fm}^{ft} + (\lambda_{fn} + \lambda_{fh} + \lambda_{ft})P_{fM}^{ft} + \lambda_{so}P_{sM}^{ft}}{N(\lambda_{sn} + \lambda_{sh} + \lambda_{st} + \lambda_{fo}) + (\lambda_{fn} + \lambda_{fh} + \lambda_{ft} + \lambda_{so}\lambda_{fo})}. \tag{36}$$

Note that the performance measures (or the joint stationary probabilities) are related to the various arrival rates and channel occupancy times, which also have functional relationships with the performance measures (as shown in Section 6.1 and 6.2). Hence, we need a recursive method to compute those fixed-point equations. Basically, the simple and widely-used iterative technique in the literature, for example [32,34], is enough to compute the performance measures. If the size of C_m or C_M increases, the state space of the system will increase drastically and the computation may be difficult. In this case, the methods proposed in [35,36] may be used. All the numerical results in the next section have convergent values of the performance measures.

7. Numerical results and discussions

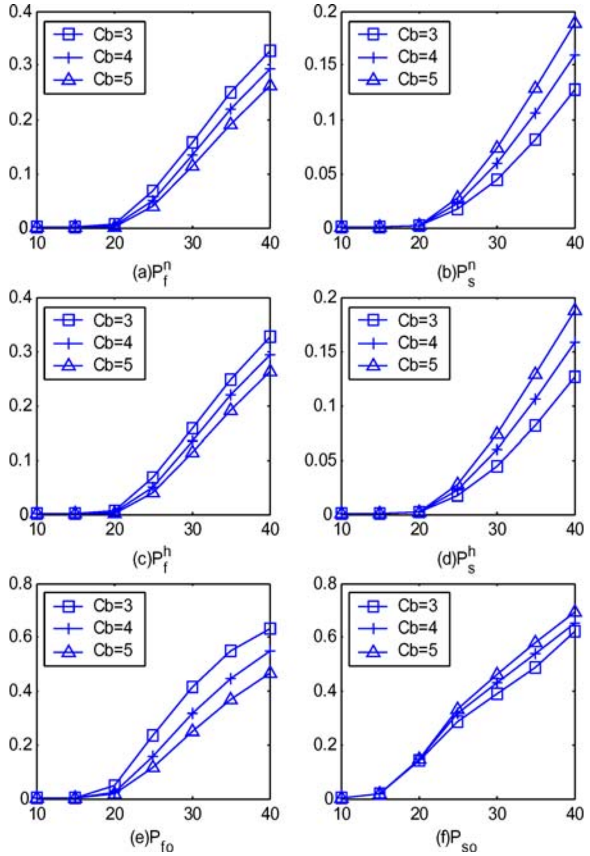
In this section, we present numerical examples to study the performance of the two-layer HCN in terms of the effect of overflow calls, mobility ratio and carried traffic distributions. We choose the hyper-Erlang distribution as the distributions of our cell residence times, since the hyper-Erlang model is shown to provide a universal approximator to any general distribution of nonnegative random variable [30]. The pdf of a hyper-Erlang random variable and its Laplace transform are given as follows:

$$f(t) = \sum_{i=1}^{V} \alpha_i \frac{(v_i\theta_i)^{v_i} t^{v_i - 1}}{(v_i - 1)!} e^{-v_i\theta_i t}, \quad t \geq 0 \quad \text{and} \quad f^*(s) = \sum_{i=1}^{V} \alpha_i \left(\frac{v_i\theta_i}{s + v_i\theta_i}\right)^{v_i},$$

where $\sum_{i=1}^{V} \alpha_i = 1, \alpha_i \geq 0, \theta_i \geq 0$, $N > 0$, $v_i \geq 0$ $(i = 1, 2, \ldots, V)$. Specifically, the parameters are chosen as $V = 2, \alpha_1 = 1 - \alpha_2 = 0.2$, $v_1 = 2$, $v_2 = 2$, $\theta_1 = 0.25$ and $\theta_2 = 0.40$ for slow users in a microcell (mean = 2.8 min) , and $\alpha_1 = 1 - \alpha_2 = 0.1$, $v_1 = 1$, $v_2 = 2$, $\theta_1 = 0.35$

and $\theta_2 = 0.45$ for fast users in a macrocell (mean = 2.286 min), and $\alpha_1 = 1 - \alpha_2 = 0.3$, $\upsilon_1 = 3$, $\upsilon_2 = 2$, $\theta_1 = 0.75$ and $\theta_2 = 0.95$ for fast users in a microcell (mean = 1.137 min), and $\alpha_1 = 1 - \alpha_2 = 0.1$, $\upsilon_1 = 4$, $\upsilon_2 = 2$, $\theta_1 = 0.115$ and $\theta_2 = 0.105$ for slow users in a macrocell (mean = 9.441 min). The call holding time is exponentially distributed with mean 2.857 min for fast users and 2.857 min for slow users. The other system parameters are chosen as follows: $C_M = 30$, $C_A = 20$, $C_B = 10$, $C_m = 10$, $C_a = 8$, $C_b = 5$. Let Λ denote the total arrival rate of new calls and γ denote the mobility ratio, i.e., the fraction of new calls from slow users. Then $\lambda_{sn} = \gamma\Lambda/N$, $\lambda_{fn} = (1-\gamma)\Lambda$, where N is the number of microcells covered by a macrocell ($N = 7$), γ is basically set to be 0.5 except the case of presenting the effect of the mobility ratio. The various handoff rates and take-back rates are recursively calculated from the new call arrival rates and related probabilities.

Figure 3(a)–(f)shows the effect of C_b on different performance measures (note C_B is fixed at this case). As C_b increases, the final new call blocking and handoff dropping probability of fast users, P_f^n and P_f^h, as well as the overflow failure probability of fast users P_{fo} will decrease; while the corresponding measures of slow users, P_s^n, P_s^h and P_{so}, will increase. This is because more calls of fast users can overflow to microcells, and the calls of slow users may thus confront more intense competition from fast users. The effect of C$_B$ on performance measures can also be obtained due to symmetry (not shown in figure). Hence, by adjusting the parameters C$_b$ and C$_B$, we can effectively balance the traffic load with different QoS requirements.



(a)P_f^n (b)P_s^n

(c)P_f^h (d)P_s^h

(e)P_{fo} (f)P_{so}

Fig. 3 The effect of C$_b$ on performance measures. Square: C$_b$ = 3; Plus: C$_b$ = 4; Triangle: C$_b$ = 5. Horizontal axis: total arrival rate of new calls Λ; Vertical axis: performance measures, (a) P_f^n; (b) P_s^n; (c) P_f^h; (d) P_s^h; (e) P_{fo}; (f) P_{so}.

Fig. 4 The effect of mobility ratio on performance measures. Square: $\gamma = 0.4$; Plus: $\gamma = 0.5$; Triangle: $\gamma = 0.6$. Horizontal axis: total arrival rate of new calls Λ; Vertical axis: performance measures, (a) P_f^n; (b) P_s^n; (c) P_f^h; (d) P_s^h; (e) P_{fo}; (f) P_{so}

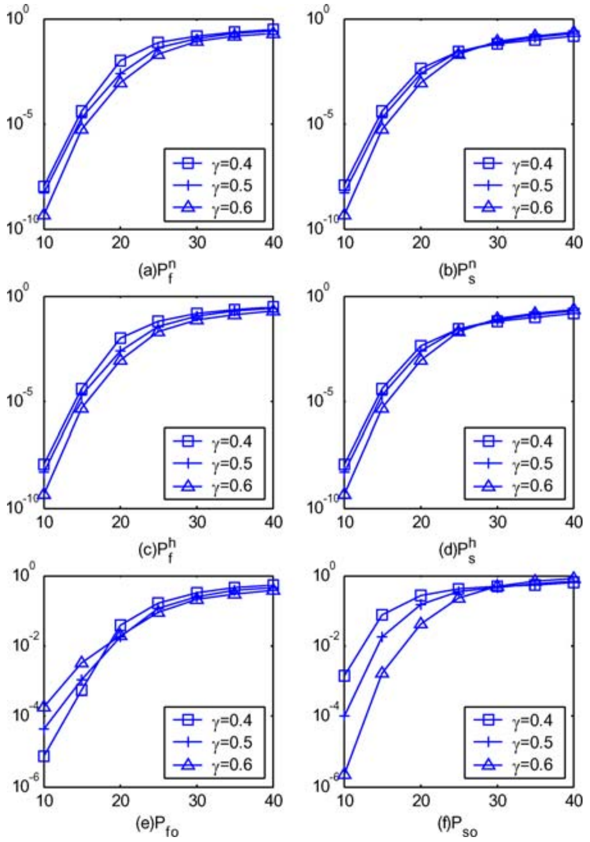


Figure 4(a)–(f)shows the effect of mobility ratio γ on different performance measures (note C_b and C_B are fixed at this case). As γ increases (equivalently, λ_{sn} increases and λ_{fn} decreases), from Fig. 4(a) and (c), it seems that the final new call blocking and handoff dropping probability of fast users, P_f^n and P_f^h, will decrease. However, this may not always true. It is noteworthy that the corresponding measures of slow users, P_s^n and P_s^h, as well as the overflow failure probabilities P_{fo} and P_{so}, vary irregularly (from Fig. 4(b), (d)–(f)). The reason can be explained as follows. When γ increases, two opposite results will be caused at each layer. On one hand, the calls of slow users will increase and those of fast users will decrease, the performance of microcell will be degraded and that of macrocell will be upgraded; On the other hand, the more overflow calls of slow users will degrade the performance of macrocell, and the less overflow calls of fast users will upgrade the performance of microcell. Hence, there are tradeoffs with respect to γ and C_b in the microcell and with respect to γ and C_B in the macrocell.

Figure 5(a)–(d)shows the carried traffic distribution of different user types at different layers. We can observe that most of the traffic carried by the macrocell is of the fast user type and most of the traffic carried by the microcell is of the slow user type. The capability of such an autonomous mobility management maintains with respect to the change of mobility ratio or number of channels assigned to overflow calls.

Figure 6(a)–(f)shows how the forced-termination probabilities depend on the change of mobility ratio γ and total arrival rate of new calls Λ (note C_b and C_B are fixed at this case). As Λ increases, all the forced-termination probabilities increase. This is intuitive, since the number

Fig. 5 The carried traffic distribution of different user types at different layers. Square: fast users in a macrocell; Plus: slow users in a macrocell; Triangle: slow users in a microcell; Asterisk: fast users in a microcell. Horizontal axis: total arrival rate of new calls Λ; Vertical axis: carried traffic distribution, (a) $\gamma = 0.4$, $C_b = 5$; (b) $\gamma = 0.5$, $C_b = 5$; (c) $\gamma = 0.6$, $C_b = 5$; (d) $\gamma = 0.5$, $C_b = 3$.

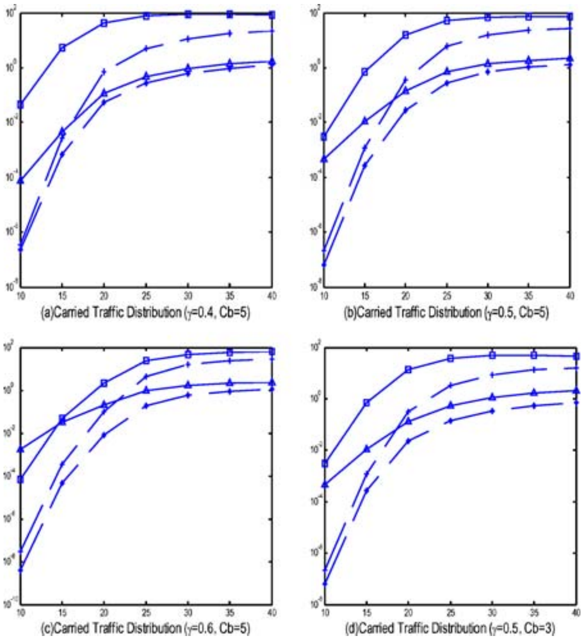
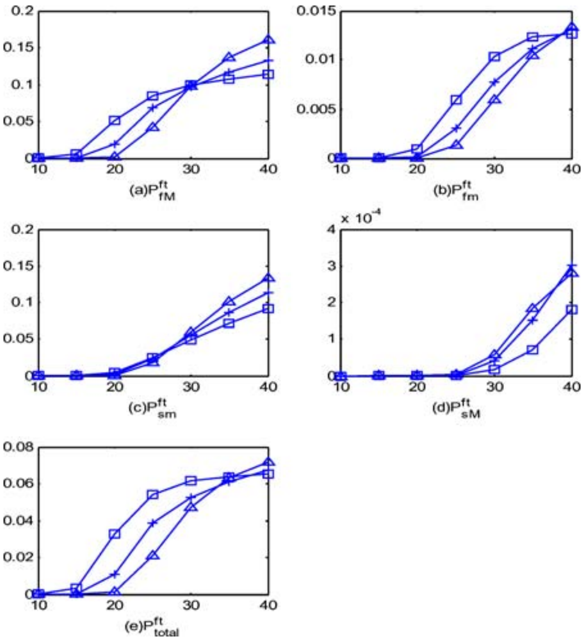


Fig. 6 The forced-termination probabilities. Square: γ =0.4; Plus: γ =0.5; Triangle: γ =0.6. Horizontal axis: total arrival rate of new calls Λ; Vertical axis: different forced-termination probabilities, (a)P_{fM}^{ft}; (b)P_{fm}^{ft}; (c)P_{sm}^{ft}; (d)P_{sM}^{ft}; (e) P_{total}^{ft}



of channels is finite. As γ increases (equivalently, λ_{sn} increases and λ_{fn} decreases), similar to Fig. 4, the forced-termination probabilities vary irregularly; there are tradeoffs with respect to γ and C_b in the microlayer and with respect to γ and C_B in the macrolayer. We can explain this by using Fig. 6 (a) as an example. When γ increases, on one hand, the calls of fast users will decrease, leading to the decrease of P_{fM}^{ft}; on the other hand, the calls of slow users will increase and more overflow calls of slow users will enter the macrolayer to contend with fast users, leading to the increase of P_{fM}^{ft}. Hence, there is a tradeoff between γ and C_B. It seems that there are no obvious results in Fig. 6. However, if we compare Fig. 6(a) and (b), Fig. 6(c) and (d), respectively, it is obvious that P_{fm}^{ft} is far smaller than P_{fM}^{ft}, and P_{sM}^{ft} is even negligible by comparing with P_{sm}^{ft}. This validates that the bidirectional overflow and take-back strategy proposed in this paper does not cause unacceptable forced-termination calls, which is consistent with the simulation result in [13] (i.e., little difference among the bidirectional, unidirectional and no overflow schemes from the aspect of the number of successful handoffs per call when the traffic is not very high). Furthermore, the advantages of this strategy are quite obvious, such as enabling better resource configuration between different layers, higher system capacity, and better QoS.

8. Conclusions

We have developed an analytical model and a performance analysis method for a hierarchical cellular network with bidirectional overflow and take-back strategies under generally distributed cell residence times. Mobile users are divided into two classes. The call requests (including new and handoff calls) of fast and slow users are preferably assigned to the macrolayer and microlayer, respectively. A call from a quickly moving user (a fast user) will overflow to microlayer if there is no macro-channel available. The successful overflow call can be taken back to macrolayer at the boundary of a microcell if a channel becomes available. The similar procedure is applied to a call from a slowly moving user. Since the commonly used exponentially distributed assumption for cell residence time and then the channel occupancy time does not hold for emerging mobile networks, we have modeled various cell residence times by general distributions, which have more flexibility in modeling different mobility environments with various probability characteristics. The channel occupancy times are derived in terms of the Laplace transforms of various cell residence times. The various arrival rates such as handoff rates, overflow rates and take-back rates of each layer, as well as the related performance measures are also derived by a number of fixed-point equations. Finally, since the various arrival rates and the channel occupancy times, as well as the stationary probabilities and the performance measures are analyzed separately according to macrocell case and microcell case, the developed model and analysis method can be easily extended to multi-layer HCNs without much computational effort.

Nomenclature

$\lambda_{sn}(\lambda_{sh})$	New (handoff) call arrival rate of slow users in a microcell.
$\lambda_{fn}(\lambda_{fh})$	New (handoff) call arrival rate of fast users in a macrocell.
$\lambda_{so}(\lambda_{fo})$	Overflow rate of slow (fast) users to a macrocell (microcell) from the N microcells (the macrocell).
$\lambda_{st}(\lambda_{ft})$	Take-back rate of slow (fast) users to a microcell (macrocell) from the macrocell (the N microcells).
$H_m(H_M)$	Requested call holding time of slow (fast) users.
$H_m^r(H_M^r)$	Residual call holding time of slow (fast) users from an arbitrary epoch.
$R_{sm}(R_{fM})$	Cell residence time of slow (fast) users in a microcell (macrocell).
$R_{sm}^r(R_{fM}^r)$	Residual cell residence time of slow (fast) users in a microcell (macrocell) from an arbitrary epoch.
$R_{fm}(R_{sM})$	Cell residence time of fast (slow) users in a microcell (macrocell).

$R^r_{fm}(R^r_{sM})$ Residual cell residence time of fast (slow) users in a microcell (macrocell) from an arbitrary epoch.

$T^n_{sm}(T^h_{sm})$ Channel occupancy time of new (handoff) calls of slow users in a microcell.

$T^n_{fM}(T^h_{fM})$ Channel occupancy time of new (handoff) calls of fast users in a macrocell.

$T^o_{sM}(T^o_{fm})$ Channel occupancy time of overflow calls of slow (fast) users in a macrocell (microcell).

$T^t_{sm}(T^t_{fM})$ Channel occupancy time of take-back calls of slow (fast) users in a microcell (macrocell).

$P^n_{sm}(P^n_{fM})$ Blocking probability of new calls of slow (fast) users in a microcell (macrocell).

$P^h_{sm}(P^h_{fM})$ Handoff failure probability of handoff calls of slow (fast) users in a microcell (macrocell).

$P_{so}(P_{fo})$ Overflow failure probability of slow (fast) users to a macrocell (microcell).

$P_{st}(P_{ft})$ Take-back failure probability of slow (fast) users from macrocell to microcell (microcell to macrocell).

$P^n_s(P^n_f)$ Blocking probability of new calls of slow (fast) users in the HCN.

$P^h_s(P^h_f)$ Dropping probability of handoff calls of slow (fast) users in the HCN.

$CT^f_M(CT^s_M)$ The carried traffic of fast (slow) users in the macrocell.

$CT^f_m(CT^s_m)$ The carried traffic of fast (slow) users in the microcell.

$P^{ft}_{sm}(P^{ft}_{sM})$ Forced-termination probability of a call from a slow user in the microlayer (macrolayer).

$P^{ft}_{fM}(P^{ft}_{fm})$ Forced-termination probability of a call from a fast user in the macrolayer (microlayer).

P^{ft}_{total} The total forced-termination probability in the HCN.

X Arbitrary random variable X has mean $E[X]$, probability density function $f_X(t)$, cumulative distribution function $F_X(t)$ and corresponding Laplace transform $f^*_X(s)$.

Acknowledgment The authors would like to thank the referees for the careful reading of the original manuscript. Their comments and suggestions have led to a much better presentation of the paper. The work of Wei Li was supported in part by the National Science Foundation under Grant CCF-0515263 and by the National Natural Science Foundation of China under Grant 60474067.

References

1. T.S. Rappaport, *Wireless Communications Principles and Practice. (Prentice Hall*, 1996).
2. X. Lagrange and P. Godlewski, Teletraffic analysis of a hierarchical cellular network, in: *Proc. IEEE Veh. Technol. Conf.* (VTC '95), (1995) pp. 882–886.
3. C.L. I, L.J. Greenstein and R.D. Gitlin, A microcell/macrocell cellular architecture for low and high-mobility wireless user, IEEE J. Selected Areas Comm. 11(1993) 885–891.
4. S.S. Rappaport and L.R. Hu, Microcellular communication systems with reneging and dropping for waiting new and handoff calls, in: *Proc. IEEE*, 82 (1994) 1383–1397.
5. L.R. Hu and S.S. Rappaport, Personal communication systems using multiple hierarchical cellular overlays, IEEE J. Selected Areas Comm. 13 (1995) 406–415.
6. X. Lagrange and P. Godlewski, Performance of a hierarchical cellular network with mobility-dependent handover strategies, Proc. VTC, 3 (1996) 1868–1872.
7. S.H. Wie, J.S. Jang, B.C. Shin and D.H. Cho, Handoff analysis of the hierarchical cellular system, IEEE Trans. Vehic. Tech. 49 (2000) 2027–2036.
8. B. Li, C. K. Wu and A. Fukuda, Performance analysis of flexible hierarchical cellular systems with a bandwidth efficient handoff scheme, IEEE Trans. on Vehic. Tech. 50(4) (2001) 971–980.

9. K. Yeo and C.-H. Jun, Modeling and analysis of hierarchical cellular networks with general distributions of call and cell residence times, IEEE Trans. Vehic. Tech. 51 (2002) 1361–1374.
10. X. Lagrange, Performance of reversible and non reversible hierarchical cellular networks, in: *Proc. 3rd European Personal Mobile Commun,* (Paris, France, 1999) pp. 197–202.
11. V.T. Vakili, A. Aziminejad and M.R. Dehbozorgi, A novel speed-sensitive bidirectional overflow and hand-down resource allocation strategy for hierarchical cellular networks, in: *Proc. 1st International Symposium on Info. and Comm. Tech.,* (Dublin, Ireland. 2003) pp. 126–131.
12. G. Boggia and P. Camarda, Teletraffic analysis of hierarchical cellular communication networks, IEEE Trans. Vehic. Tech. 52(4) (2003) 931–946.
13. W. Shan, P. Fan and Y. Pan, Performance evaluation of a hierarchical cellular system with mobile velocity-based bidirectional call-overflow scheme, IEEE Trans. Parallel Distrib. Syst. 14(1) (2003) 72–83.
14. B. Jabbari and W.F. Fuhrmann, Teletraffic modeling and analysis of flexible hierarchical cellular networks with speed-sensitive handoff strategy, IEEE J. Select. Areas Comm. 15 (1997) 1539–1548.
15. K. Kawabata, T. Nakamura and E. Fukuda, Estimating velocity using diversity reception in: *Proc. IEEE VTC,* vol. 1 (1994) pp. 371–374.
16. K.L. Yeung and S. Nanda, "Channel management in microcell/macrocell cellular radio systems , IEEE Trans. Vehic. Tech. 45(4) (1996) 601–12.
17. W. Huang and V.K. Bhargava, Effects of user mobility on handoff performance in a hierarchical cellular system, in: *Proc. Canadian Conf. Elec. Comp. Eng.* 1 (1995) 551–54.
18. M. Benveniste, Cell selection in two-tier microcellular/macrocellular systems, in: *Proc. IEEE GLOBECOM,* vol. 2 (1995) pp. 1532–36.
19. C. Sung and W. Wong, User speed estimation and dynamic channel allocation in hierarchical cellular system, in: *Proc. IEEE VTC,* vol. 1(1994) pp. 91–95.
20. K. Shum and C. Sung, Fuzzy layer selection method in hierarchical cellular systems , IEEE Trans. Vehic. Tech. 48(6) (1999) 1840–1849.
21. C. Sung and K. Shum, Fuzzy channel assignment and layer selection in hierarchical cellular system with fuzzy control, IEEE Trans. Vehic. Tech. 50(3) (2001) 657–663.
22. D. Hong and S.S. Rappaport, Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures, IEEE Trans. Vehic. Tech. 35(3) (1986) 77–92.
23. C.H. Yoon and C.K. Un, Performance of personal portable radio telephone systems with and without guard channel, IEEE J. Select. Areas Commun. 11(6) (1993) 911–917.
24. Y. Fang and Y. Zhang, Call admission control schemes and performance analysis in wireless mobile networks, IEEE Trans. on Vehic. Tech. 51(2) (2002) 371–382.
25. Y. Zhou and B. Jabbari, Performance modeling and analysis of hierarchical wireless communication networks with overflow and take-back traffic, in: *Proc. IEEE PIMRC,* vol. 3 (1998) pp. 1176–1180.
26. M.A. Marsan, G. Ginella, R. Maglione and M. Meo. Performance analysis of hierarchical cellular networks with generally distributed call holding times and dwell times, IEEE Trans. on Wireless Comms, 3(1) (2004) 248–257.
27. F. Barcelo and J. Jordan, Channel holding time distribution in cellular telephony, in: *Proc. 9th Int. Conf. Wireless Commun. (Wireless '97), 1, (Alta., Canada,* 1997) pp. 125–134.
28. P.V. Orlik and S.S. Rappaport, A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions, IEEE J. Select. Areas Commun. 16 (1998) 788–803.
29. Y. Fang and I. Chlamtac, Teletraffic analysis and mobility modeling of PCS networks, IEEE Trans. Commun., 47 (1999) 1062–1072.
30. Y. Fang, Hyper-Erlang distribution and its applications in wireless and mobile networks, Wireless Networks (WINET) 7(3) (2001) 211–219.
31. F. P. Kelly, *Reversibility and Stochastic Networks*. (New York Wiley, 1979).
32. S. Tang and W. Li, Modeling and evaluation of the 3*g* mobile network with hot-spot wlans Accepted for publication in International Journal of Wireless and Mobile Computing (IJWMC), Inderscience Publishers.
33. E.J. Watson, *Laplace Transforms and Applications*. (Birkhauserk, 1981).
34. Y.B. Lin, S. Mohan and A. Noerpel, Queueing priority channel assignment strategies for handoff and initial access for a PCS network, IEEE Trans. Veh. Tech. 43 (1994) 704–712.
35. P. Gazdzicki, I. Lambadaris and R.R. Mazumdar,Blocking probabilities for large multirate Erlang loss systems, Adv. Appl. Prob., vol. 25(4) (1993) 997–1009.
36. A. Simonian, F. Theberge, J.R. Roberts and R.R. Mazumdar, Asymptotic estimates for blocking probabilities in a large multi-rate loss network, Adv. Appl. Prob. 29(3) (1997) 806–829.