Loss Performance Modeling for Hierarchical Heterogeneous Wireless Networks With Speed-Sensitive Call Admission Control

Huang, Qian; Huang, Yue-Cai; Ko, King-Tim; Iversen, Villy Bæk

Published in:
I E E E Transactions on Vehicular Technology

Link to article, DOI:
10.1109/TVT.2011.2142203

Publication date:
2011

Loss Performance Modeling for Hierarchical Heterogeneous Wireless Networks with Speed-Sensitive Call Admission Control

Qian Huang, Yue-Cai Huang, King-Tim Ko and Villy Bæk Iversen

***Abstract*—Hierarchical overlay structure is an alternative solution to integrate the existing and future heterogeneous wireless networks for providing subscribers with better mobile broadband services. Traffic loss performance in such integrated heterogeneous networks is necessary for operator's network dimensioning and planning. This paper investigates the computationally efficient loss performance modeling for multiservice in hierarchical heterogeneous wireless networks. A speed-sensitive call admission control scheme is considered in our model in order to assign overflowed calls to the appropriate tiers. This approach avoids unnecessary and frequent handoff between cells and reduces signaling overheads. An approximation model with guaranteed accuracy and low computational complexity is presented for the loss performance of multiservice traffic. The accuracy of the numerical results is validated by comparing the results from the approximation with the simulations.**

***Index Terms*—Hierarchical networks, heterogeneous wireless networks, mobile traffic management, overflow, multiservice, QoS, performance evaluation.**

I. Introduction

With the rapid development of mobile broadband technologies such as femtocells, WiFi, WiMAX, 3G and LTE, integration of various wireless networks has become necessary to provide mobile users with seamless Internet access through different technologies. A solution that integrates various wireless networks into a hierarchical overlay system based on hierarchical cell structure has been considered [1]. This solution has also been used for the deployment of femtocell networks in 3G, WiMAX and LTE networks [2], [3], [4]. The advantage of this solution is that mobile users in the systems can switch between various wireless networks for more efficient use of network resources. In a WiMAX/femtocell overlay system presented in [2], mobile users can connect to Internet via nearby femtocells, and the calls rejected by femtocell networks due to lack of radio access can overflow to overlaying WiMAX networks. Such kind of call overflow schemes has been used in cellular overlay networks for reducing call blocking probability and improving bandwidth utilization [5], [6]. However, allowing call overflow between the overlay networks results in forced handoff and extra signaling overheads. Additionally, frequent handoff can be incurred by mobile user's movement in those areas covered by small cells such as femtocells and leads to increased signaling overheads and operating costs. A solution to this problem is to take into account user mobility speed in call admission control (CAC) for mobility management as in cellular overlay networks [7], [8], [9], [10], [11]. We refer to this CAC as speed-sensitive CAC. Its basic idea is as follows. When calls are blocked due to cell capacity limit, blocked calls from fast-speed users are redirected to high-tier large cells, e.g. macro-cells; blocked calls from slow-speed users are redirected to low-tier cells, e.g. microcells or femtocells. This approach assigns mobile users to appropriate cells so that frequent call handoff from fast-speed users in small cells can be avoided and signaling overheads can be reduced.

To provide acceptable quality of service (QoS) for mobile users in hierarchical overlay networks, resource allocation between the overlay networks must be carefully designed since the overflow traffic from other networks will compete for bandwidth with local users [12], [13], [14]. As an essential problem to be addressed for optimal resource allocation, computationally efficient methods for traffic loss performance in hierarchical heterogeneous networks are necessary. To obtain the loss performance, a key issue to be solved is the overflow traffic modeling. There is a simple solution that assumes the overflow traffic between the overlay cells to be a Poisson process. This assumption has been used in past loss models for single service (e.g. voice call) in cellular overlay systems [5], [9], [10], [11] and also some multiservice loss models [14], [15], [16]. This simple Poisson assumption, however, ignores the fact that overflow traffic is "bursty" in nature, and has been demonstrated that it leads to erring performance evaluation in multiservice cellular overlay networks [17]. To obtain the accurate performance evaluation, more computationally complicated models have been considered for overflow traffic loss analysis, such as the Markov-modulated Poisson Process (MMPP) [17], [18] and its special two-state case known as Interrupted Poisson Process (IPP) [6]. The major concern with the MMPP model is the high computation effort involved in solving multi-dimensional equilibrium state equations. For a single-tier network with cell capacity C and n traffic flows, the MMPP model has a C^n order of complexity. This computation will become intractable in hierarchical heterogeneous wireless networks with large number of traffic flows and cell capacity.

Our previous work in [19] has developed a so-called Multiservice Overflow Approximation (MOA) model to obtain the multiservice loss performance in a homogeneous micro/macrocell overlay system, where "homogeneous" is defined as having the same offered traffic in each of the microcells. Under this assumption, the loss performance of the overflow traffic out of microcells in the overlaying macrocell

Q. Huang is with the Department of Electrical and Electronic Engineering, The University of Melbourne, Victoria 3010 Australia (huangq@unimelb.edu.au). Y.-C. Huang and K. T. Ko are with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong SAR, China (eektko@cityu.edu.hk, yuechuang2@student.cityu.edu.hk). V. B. Iversen is with the Department of Photonics Engineering, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark (vbi@com.dtu.dk).

can be obtained by using the multi-rate Hayward's approximation [19]. This MOA model, however, is not available for hierarchical heterogeneous wireless systems. In heterogeneous scenarios, the networks located at the same or different tiers can be differentiated in terms of transmission rate, capacity and signal coverage range. These heterogeneous parameters have significant effects on network performance modeling [20]. An effect is the change on the statistical characteristics of channel occupancy times of a traffic flow in different networks. This factor leads to different considerations for traffic loss modeling of heterogeneous overlay networks in this paper, which has not been addressed in our previous work of [19] .

In our current work, we consider heterogeneous overlay networks, where the overflow traffic from different networks has different statistical moments (e.g. mean and variance) which are related to the service time distributions in these networks. In addition, the statistical moments of the overflow traffic offered to a new network are redefined according to the service time distribution in this new network. The latter can be further elaborated by the following example. Consider a two-tier overlay system with the single overflow traffic from tier-1 to tier-2. The service times t_s for a call are determined by a distribution function $F_1(t_s)$ in tier-1 and another distribution function $F_2(t_s)$ in tier-2. From the loss analysis in tier-1, we obtain the mean (m_1) and the variance (v_1) of the overflow traffic from tier-1 in terms of the service time distribution $F_1(t_s)$. For the loss analysis of this overflow traffic in tier-2, we need to recalculate the mean (m_1') and variance (v_1') in terms of the service time distribution $F_2(t_s)$. As $F_1(t_s) \neq F_2(t_s)$, here $m_1' \neq m_1$ and $v_1' \neq v_1$.

In this paper, we present a recursive algorithm to derive the moments of multi-class overflow traffic in hierarchical heterogeneous overlay networks with different service time distributions. With the obtained moments of overflow traffic, we develop the multi-class traffic loss model for a hierarchical heterogeneous overlay system with the speed-sensitive CAC scheme performed on the inter-tier overflow calls. With the proposed loss model, we obtain the numerical solutions of call blocking and dropping probabilities for multi-class traffic in the system. As for the multiservice and multi-QoS traffic in broadband networks, e.g. UMTS, WiMAX, LTE, and 802.11e, a loss model based on bandwidth allocation in an equivalent number of bandwidth units per call or connection can be established. This method enables CAC to be applied to an incoming call and connection. Hence, call blocking and dropping probabilities are still important performance metric of interest in broadband networks.

The remainder of this paper is organized as follows. The system model is described in Section II. The proposed loss model for multi-class traffic in hierarchical heterogeneous overlay systems with the speed-sensitive CAC scheme is presented in Section III. Model validation, numerical results, and discussions are presented in Section IV. We draw our conclusions in Section V.

II. Hierarchical Heterogeneous Overlay System

Consider a two-tier overlay system with heterogeneous wireless networks distinguished from each other in capacity, signal coverage range, statistical characteristics of service time distribution, user mobility, volume, and traffic behavior. The high-tier networks are assumed to have greater signal coverage than those at the low tier; each high-tier network overlays several adjacent low-tier networks.

The aforementioned speed-sensitive CAC scheme is used to manage the overflow traffic between the heterogeneous overlay networks. Initially, if the new calls of fast-speed users in a low-tier network are blocked due to capacity limitation, the blocked new calls are overflowed to a high-tier network for possible service. If the blocked new calls are from slow-speed users in a high-tier network, they are overflowed to a low-tier network. Similar control schemes are used to handle the handoff calls between the neighboring networks at the same tier. If fast-speed users in a low-tier network cannot be handed off to a neighboring network, their handoff calls are overflowed to the networks at the high tier. The failed handoff calls of slow-speed users in the high-tier networks are overflowed to the networks at the low tier. With this scheme, a call is finally dropped when there is no bandwidth available for it in the hierarchical system; call blocking and dropping probabilities can thus be improved.

Additionally, we use the bandwidth reservation scheme [21] to protect the handoff calls. A portion of capacity in each tier network is reserved for handoff calls only; the remaining bandwidth is shared by all arriving calls. For simplicity, the bandwidth required by a class k call, denoted as d_k, is measured by the number of bandwidth units.

Assume that K classes of services are supported in the two-tier system. Fig. 1 shows the traffic flows related to class k service, $1 \leq k \leq K$. The notations in Fig. 1 are defined in Appendix I. The offered traffic from class k service to any network s at tier-l ($l = 1, 2$) can be an aggregation of the three types of traffic: local new call traffic, handoff call traffic from adjacent networks at the same tier, and overflow call traffic from the high/low tier networks. The local new call traffic to each network can be assumed to have a Poisson arrival process. The handoff traffic within the same tier can also be approximated by a Poisson process as in [6], [9], [17], and [21]. However, the overflow traffic is non-Poisson. Therefore, the aggregated traffic to each overlay network from each service class is non-Poisson, and its statistics can be characterized by the first two moments, the mean and the variance.

Define $\lambda_n^{(k,l,s)}$ and $\lambda_{h,in}^{(k,l,s)}$ as the local new call and handoff call arrival rates of class k service to tier-l network s, respectively, with $l = 1, s = s_1$ and $l = 2, s = s_2$. Define $\lambda_{nu}^{(k,1,s_1)}$ and $\lambda_{nd}^{(k,2,s_2)}$ as the overflow call arrival rates from the blocked class k new calls in tier-1 network s_1 and tier-2 network s_2 but overflowed up to tier-2 and down to tier-1 networks, respectively. Similarly define $\lambda_{hu}^{(k,1,s_1)}$ and $\lambda_{hd}^{(k,2,s_2)}$ as the overflow call arrival rates from the blocked class k handoff calls in tier-1 network s_1 and tier-2 network s_2 but overflowed up to tier-2 and down to tier-1 networks, respectively. Hereafter, we use the subscript "n" in the notations to stand for any new call traffic, and the subscript "h" to stand for any handoff traffic.

Suppose that each tier-2 network overlays N_1 tier-1 networks. The aggregated class k traffic offered to tier-1 network

s_1 has the call arrival rates

$$\Lambda_n^{(k,1,s_1)} = \lambda_n^{(k,1,s_1)} + \lambda_{nd}^{(k,2,s_2)}, \quad (1)$$

$$\Lambda_h^{(k,1,s_1)} = \lambda_{h,in}^{(k,1,s_1)} + \lambda_{hd}^{(k,2,s_2)}; \quad (2)$$

and the aggregated class k traffic to tier-2 network s_2 has

$$\Lambda_n^{(k,2,s_2)} = \lambda_n^{(k,2,s_2)} + \sum_{s_1=1}^{N_1} \lambda_{nu}^{(k,1,s_1)}, \quad (3)$$

$$\Lambda_h^{(k,2,s_2)} = \lambda_{h,in}^{(k,2,s_2)} + \sum_{s_1=1}^{N_1} \lambda_{hu}^{(k,1,s_1)}. \quad (4)$$

Let $B_n^{(k,l,s)}$ and $B_h^{(k,l,s)}$ denote the blocking probabilities for class k new calls and handoff calls in tier l network s. The class k handoff call arrival rate to tier l network s, denoted as $\lambda_{h,in}^{(k,l,s)}$, is the summation of the class k handoff call departure rates, denoted as $\lambda_{h,out}^{(k,l,s')}$, from all neighboring networks s' at the same tier l when the system is in the equilibrium state, i.e.

$$\lambda_{h,in}^{(k,l,s)} = \sum_{s' \in \Omega_{l,s}} q(s',s) \lambda_{h,out}^{(k,l,s')}, \quad (5)$$

where $\Omega_{l,s}$ denotes the set of the neighboring networks of tier l network s, for $l = 1, s = s_1$ and $l = 2, s = s_2$; $q(s',s)$ denotes the probability of a call in tier l network s' making a handoff to tier l network s, for $s' \neq s$. In cellular systems with hexagonal-shaped cells, where each cell is surrounded by six neighboring cells and mobile users are uniformly distributed in each cell, we have $q(s',s) = 1/6$.

Let $\nu_n^{(k,l,s)}$, $\nu_v^{(k,l,s)}$ and $\nu_h^{(k,l,s)}$ denote the probabilities of an accepted new call (both local and overflowed), overflow handoff call and local handoff call in tier l network s making a handoff out of its current network, respectively. Here $l = 1, s = s_1$ and $l = 2, s = s_2$. These handoff probabilities for different types of accepted calls in a network can be derived from their service time distributions in the given network. The service time of a call in a network is jointly determined by the call's holding time and sojourn time in the network. Different types of accepted calls in the network follow different service time distributions. For new calls (both local and overflowed), the service time distribution is jointly determined by the call holding time and the residual sojourn time distributions in the given network. For local handoff calls, it is jointly determined by the residual call holding time and the sojourn time distributions. For overflowed handoff calls, it is jointly determined by the residual call holding time and the residual sojourn time distributions.

The departure rates $\lambda_{h,out}^{(k,l,s)}$ for $l = 1, s = s_1$ and $l = 2, s = s_2$ are derived as

$$\begin{aligned} \lambda_{h,out}^{(k,1,s_1)} &= \Lambda_n^{(k,1,s_1)}(1 - B_n^{(k,1,s_1)})\nu_n^{(k,1,s_1)} \\ &+ \lambda_{h,in}^{(k,1,s_1)}(1 - B_h^{(k,1,s_1)})\nu_h^{(k,1,s_1)} \\ &+ \lambda_{hd}^{(k,2,s_2)}(1 - B_h^{(k,1,s_1)})\nu_v^{(k,1,s_1)}, \end{aligned} \quad (6)$$

$$\begin{aligned} \lambda_{h,out}^{(k,2,s_2)} &= \Lambda_n^{(k,2,s_2)}(1 - B_n^{(k,2,s_2)})\nu_n^{(k,2,s_2)} \\ &+ \lambda_{h,in}^{(k,2,s_2)}(1 - B_h^{(k,2,s_2)})\nu_h^{(k,2,s_2)} \\ &+ \sum_{s_1=1}^{N_1} \lambda_{hu}^{(k,1,s_1,)}(1 - B_h^{(k,2,s_2)})\nu_v^{(k,2,s_2)}. \end{aligned} \quad (7)$$

Assume there is no downward overflow traffic by letting $\lambda_{nd}^{(k,2,s_2)} = \lambda_{hd}^{(k,2,s_2)} = 0$. Assume $q(s',s) = 1/6$. The handoff call arrival rates $\lambda_{h,in}^{(k,l,s)}$ in Eqns. (2) and (4) can be resolved by an iterative algorithm [21], [22] aiming to achieve Eqn. (5) under the system equilibrium state. Exemplify tier 1 system.

1. Initially, let $\delta = 10^{-6}$; let $\lambda'^{(k,1,s)}_{h,in} = 1.0$ and $\lambda_{h,out}^{(k,1,s)} = 1.0$ for all s and k in tier 1. Obtain the initial values of $\lambda_{h,in}^{(k,1,s)}$ from Eqn. (5).
2. Without downward overflows, $\Lambda_h^{(k,1,s)} = \lambda_{h,in}^{(k,1,s)}$. It is known that $Z_h^{(k,1,s)} = 1$. From Eqn. (17) and Eqn. (18) obtain $B_n^{(k,1,s)}$ and $B_h^{(k,1,s)}$ by the method described in Section III-B.
3. Obtain $\lambda_{h,out}^{(k,1,s)}$ from Eqn. (6); then obtain $\lambda'^{(k,1,s)}_{h,in} = \sum_{s' \in \Omega_{1,s}} \lambda_{h,out}^{(k,1,s')}$ from Eqn. (5).
4. If $|\lambda'^{(k,1,s)}_{h,in} - \lambda_{h,in}^{(k,1,s)}| > \delta$, let $\lambda_{h,in}^{(k,1,s)} = \lambda'^{(k,1,s)}_{h,in}$, and then repeat step 2 to step 4; else $\lambda_{h,in}^{(k,1,s)} = \lambda'^{(k,1,s)}_{h,in}$.

Similar procedures are used to obtain $\lambda_{h,in}^{(k,2,s)}$.

The key point to solve the problem of loss performance modeling in the overlay system is to determine the first two moments of the aggregated traffic offered to the overlay network, e.g. the call arrival rates $\Lambda_n^{(k,2,s_2)}$ and $\Lambda_h^{(k,2,s_2)}$, and the peakedness[1] $Z_n^{(k,2,s_2)}$ and $Z_h^{(k,2,s_2)}$ of the aggregated new call and handoff call traffic to a tier-2 network s_2. This problem is addressed in the following contents.

III. The Proposed Overflow Loss Model

With the speed-sensitive CAC scheme, bidirectional call overflows, upward and downward, are supported in the hierarchical heterogeneous overlay systems. Blocked calls from fast-speed users are overflowed to the higher-tier networks with larger coverage; blocked calls from slow-speed users are overflowed to the lower-tier networks with smaller coverage. For conciseness, we elaborate our model by assuming that only upward overflow traffic from fast-speed users exists. The same analysis method can be used for downward overflow traffic from slow-speed users.

A. Input traffic modeling

Let $(\Lambda_n^{(k,l,s)},\ Z_n^{(k,l,s)},\ d_k)$ and $(\Lambda_h^{(k,l,s)},\ Z_h^{(k,l,s)},\ d_k)$ represent the traffic from class k new calls and handoff calls input to any network s at tier-l respectively, with the required bandwidth d_k for each class k call.

Using the idea of the multi-rate Hayward's approximation, the loss performance of any network s at tier l with the input traffic $(\Lambda_n^{(k,l,s)},\ Z_n^{(k,l,s)},\ d_k)$ and $(\Lambda_h^{(k,l,s)},\ Z_h^{(k,l,s)},\ d_k)$ can be approximated by the loss performance of an equivalent trunk group with Poisson input $\left(\frac{\Lambda_n^{(k,l,s)}}{Z_n^{(k,l,s)}},\ 1,\ d_k Z_n^{(k,l,s)}\right)$ and $\left(\frac{\Lambda_h^{(k,l,s)}}{Z_h^{(k,l,s)}},\ 1,\ d_k Z_h^{(k,l,s)}\right)$, if the equivalent trunk group has the same service time distribution and the same mean service rates $\mu_n^{(k,l,s)}$ and $\mu_h^{(k,l,s)}$ for class k new calls and handoff calls as the original network s at tier l.

[1] Peakedness is defined as the ratio of variance to mean.

To achieve this equivalent trunk group, the means and variances of the multiservice overflow traffic from tier-1 to tier-2 are required to know. Theoretically, the exact solution of the means and variances of the multiservice overflow traffic can be obtained from the state probability distribution of the aforementioned infinite-server overflow group by solving the related infinite-state Markov chain model. However, it is impractical to solve a multidimensional Markov chain with an infinite number of states. Existing methods solve this problem by assuming a large enough finite-server overflow group in computations [6], [16], [17]. The computations involved are still very extensive for large networks. Here we propose a decomposition method to obtain the mean and peakedness of the multiservice overflow traffic.

We start from tier-1. As it is assumed that the new call and handoff call traffic is Poisson and only upward overflow traffic exists, we have the call arrival rates of the new call and handoff call traffic to tier-1 network s_1 determined by $\Lambda_n^{(k,1,s_1)} = \lambda_n^{(k,1,s_1)}$ and $\Lambda_h^{(k,1,s_1)} = \lambda_{h,in}^{(k,1,s_1)}$, respectively. The offered traffic load intensities are defined as $A_n^{(k,1,s_1)} = \lambda_n^{(k,1,s_1)}/\mu_n^{(k,1,s_1)}$ and $A_h^{(k,1,s_1)} = \lambda_{h,in}^{(k,1,s_1)}/\mu_h^{(k,1,s_1)}$; and the peakedness are given by $Z_n^{(k,1,s_1)} = 1$ and $Z_h^{(k,1,s_1)} = 1$.

Our first step is to decompose the input traffic flows from different classes of new calls and handoff calls to the equivalent trunk group for tier-1 network s_1, by redirecting the input traffic flows to a set of hypothetical groups. This is shown in Fig. 2. For each class of new call or handoff call traffic, there is an independent hypothetical group to accommodate the calls. The hypothetical groups are determined by two constraints. (1) The service time for each class k call in the hypothetical group is identical to that in the equivalent trunk group for tier-1 network s_1, and also identical to that in the original tier-1 network s_1. This gives identical mean service rates $\mu_n^{(k,1,s_1)}$ and $\mu_h^{(k,1,s_1)}$ in the original, the equivalent and the hypothetical groups. (2) The blocking probabilities of class k calls in the hypothetical groups, $\hat{B}_n^{(k,1,s_1)}$ and $\hat{B}_h^{(k,1,s_1)}$, are identical to that in the equivalent trunk group, and also equivalent to that in the original tier-1 network s_1; that is, $\hat{B}_n^{(k,1,s_1)} \approx B_n^{(k,1,s_1)}$ and $\hat{B}_h^{(k,1,s_1)} \approx B_h^{(k,1,s_1)}$.

The foregoing decomposition process allows us to approximate the overflow traffic from the equivalent trunk group for tier-1 network s_1 by the overflow traffic from the corresponding hypothetical group. The moments of each overflow traffic from the original tier-1 network s_1, $m_{nu}^{(k,1,s_1)}$, $z_{nu}^{(k,1,s_1)}$, $m_{hu}^{(k,1,s_1)}$, and $z_{hu}^{(k,1,s_1)}$ are then derived from the moments of the overflow traffic from the hypothetical group, also the equivalent trunk group, $\hat{m}_{nu}^{(k,1,s_1)}$, $\hat{z}_{nu}^{(k,1,s_1)}$, $\hat{m}_{hu}^{(k,1,s_1)}$ and $\hat{z}_{hu}^{(k,1,s_1)}$, as the equivalent trunk group is a loss system partially equivalent to the original tier-1 network s_1 (see Appendix II) and their moments of overflow traffic are correlated based on Eqn. (25) and Eqn. (26).

Next, the capacity of each hypothetical group shown in Fig. 2 is to be determined. Denote $\beta_n^{(k,1,s_1)}$ and $\beta_h^{(k,1,s_1)}$ as the capacity of the hypothetical groups that accommodate, respectively, class k new calls and handoff calls offered to the equivalent trunk group for tier-1 network s_1. Denote $\hat{B}_n^{(k,1,s_1)}$ as the call blocking probability of class k new calls in the hypothetical group; it is obtained by the Erlang-B formula

$$E(\beta_n^{(k,1,s_1)}; A_n^{(k,1,s_1)}/Z_n^{(k,1,s_1)}, 1, d_k Z_n^{(k,1,s_1)}). \quad (8)$$

Similarly, we obtain the call blocking probability of class k handoff calls in the hypothetical group, denoted as $\hat{B}_h^{(k,1,s_1)}$. It has the same derivation as Eqn. (8), except that all the notation subscripts "n" in Eqn. (8) are replaced by "h" to identify the handoff calls. For conciseness, in the following contents we only present the equations of the loss model for the new calls. The equations of the loss analysis for the handoff calls are identified by replacing the notation subscripts "n" in the equations for the new calls by "h".

Based on the above constraint (2) on the hypothetical groups $\hat{B}_n^{(k,1,s_1)} \approx B_n^{(k,1,s_1)}$, the value of $\beta_n^{(k,1,s_1)}$ is obtained by solving the Erlang-B formula Eqn. (8) in an iterative manner. The same approach is used to obtain the value of $\beta_h^{(k,1,s_1)}$.

The hypothetical group for class k new calls $(\beta_n^{(k,1,s_1)};\ \frac{A_n^{(k,1,s_1)}}{Z_n^{(k,1,s_1)}},\ 1,\ d_k Z_n^{(k,1,s_1)})$ is completely equivalent to a trunk group defined as $(\frac{\beta_n^{(k,1,s_1)}}{d_k Z_n^{(k,1,s_1)}};\ \frac{A_n^{(k,1,s_1)}}{Z_n^{(k,1,s_1)}},\ 1,\ 1)$, with the same call blocking probability and the same moments of the overflow traffic. Thus, the mean $\hat{m}_{nu}^{(k,1,s_1)}$ and peakedness $\hat{z}_{nu}^{(k,1,s_1)}$ of the overflow traffic from hypothetical group $(\beta_n^{(k,1,s_1)};\ \frac{A_n^{(k,1,s_1)}}{Z_n^{(k,1,s_1)}},\ 1,\ d_k Z_n^{(k,1,s_1)})$ can be obtained from trunk group $(\frac{\beta_n^{(k,1,s_1)}}{d_k Z_n^{(k,1,s_1)}};\ \frac{A_n^{(k,1,s_1)}}{Z_n^{(k,1,s_1)}},\ 1,\ 1)$; they are

$$\hat{m}_{nu}^{(k,1,s_1)} = \hat{B}_n^{(k,1,s_1)} \cdot \frac{A_n^{(k,1,s_1)}}{Z_n^{(k,1,s_1)}}\ , \quad (9)$$

$$\hat{z}_{nu}^{(k,1,s_1)} = 1 - \hat{m}_{nu}^{(k,1,s_1)} + \frac{\frac{A_n^{(k,1,s_1)}}{Z_n^{(k,1,s_1)}}}{\frac{\beta_n^{(k,1,s_1)}}{d_k Z_n^{(k,1,s_1)}} + 1 - \frac{A_n^{(k,1,s_1)}}{Z_n^{(k,1,s_1)}} + \hat{m}_{nu}^{(k,1,s_1)}}. \quad (10)$$

The peakedness in Eqn. (10) is derived from Riordan's equation [23]. The same approach is used to obtain the mean $\hat{m}_{hu}^{(k,1,s_1)}$ and the peakedness $\hat{z}_{hu}^{(k,1,s_1)}$ of the overflow traffic from the hypothetical group $(\beta_h^{(k,1,s_1)};\ \frac{A_h^{(k,1,s_1)}}{Z_h^{(k,1,s_1)}},\ 1,\ d_k Z_h^{(k,1,s_1)})$ for class k handoff calls.

As the hypothetical groups give the same call blocking probability as the equivalent trunk group obtained from the multi-rate Hayward's approximation, the means of class k new call overflow traffic from the equivalent trunk group for tier-1 network s_1 can be directly obtained by Eqn. (9). However, the peakedness of class k new call overflow traffic from the equivalent trunk group is *approximated* by Eqn. (10), because the correlation between different overflow traffic from the same equivalent trunk group is compromised by the foregoing decomposition. Here we need to clarify that the mean and the peakedness obtained by Eqn. (9) and Eqn. (10) for class k new call overflow traffic from the hypothetical trunk groups are related to the mean call service time $1/\mu_n^{(k,1,s_1)}$ for class k new calls in tier-1 network s_1.

In the considered heterogeneous overlay system, the networks located at either the same or different tiers can have different statistical characterizations, besides different capacities

and coverage. The service time for a mobile call in a network is jointly determined by the call holding time and sojourn time in the network. Due to user mobility (varying velocity, random trajectory) and irregular cell coverage, a mobile call may have different distributions of sojourn time in different networks or the same distribution but different means. This will lead to different distributions of call service times for a mobile call in different networks, or the same distribution but different mean call service times. For simplicity, we assume that mobile calls from each service class have the same call holding time distribution and the same sojourn time distribution but different means in different networks; i.e. for a given class of calls, they will have the same service time distribution but different mean service times in different networks.

Assume that tier-2 network s_2 covers N_1 tier-1 networks. By the foregoing analysis, we obtain the moments of the overflow traffic from the tier-1 networks according to the mean call service times $1/\mu_n^{(k,1,s_1)}$ for new calls and $1/\mu_h^{(k,1,s_1)}$ for handoff calls in these tier-1 networks. As we further consider the loss performance of these overflow calls in tier-2 network s_2, the moments of the overflow traffic offered to tier-2 should be redefined according to the mean call service times $1/\mu_n^{(k,2,s_2)}$ and $1/\mu_h^{(k,2,s_2)}$ for new calls and handoff calls in tier-2 network s_2. A similar problem was studied in [24], where a recursive algorithm was used to derive the approximate moments of single-service overflow traffic in its primary and secondary trunk groups with different mean service times. Here, we use the similar recursive algorithm to derive the means and the variances of the multiservice overflow traffic from tier-1 networks to tier-2, according to the different mean call service times in tier-1 and tier-2.

Define $\epsilon_n^{(k,s_2,s_1)}$ as the ratio of mean call service time in tier-1 network s_1 to that in tier-2 network s_2 for class k new calls, i.e. $\epsilon_n^{(k,s_2,s_1)} = \mu_n^{(k,2,s_2)}/\mu_n^{(k,1,s_1)}$. Similarly, define $\epsilon_h^{(k,s_2,s_1)} = \mu_h^{(k,2,s_2)}/\mu_h^{(k,1,s_1)}$ for class k handoff calls.

Let $\tilde{m}_{nu}(\mu_n^{(k,2,s_2)})$ and $\tilde{v}_{nu}(\mu_n^{(k,2,s_2)})$ denote the redefined mean and variance of the overflowed new call traffic from the hypothetical group for tier-1 network s_1 to tier-2 network s_2 according to the mean call service time $1/\mu_n^{(k,2,s_2)}$ in tier-2 network s_2. Using the recursive method shown in Appendix III, $\tilde{m}_{nu}(\mu_n^{(k,2,s_2)})$ and $\tilde{v}_{nu}(\mu_n^{(k,2,s_2)})$ are derived by Eqn. (32) and Eqn. (34); they are respectively equivalent to the mean and the variance of class k overflowed new call traffic from the equivalent trunk group for tier-1 network s_1 shown in Fig. 2.

Let $\tilde{z}_{nu}(\mu_n^{(k,2,s_2)})$ denote the redefined peakedness of the overflowed new call traffic from the equivalent trunk group for tier-1 network s_1 to tier-2 network s_2, according to the mean call service time $1/\mu_n^{(k,2,s_2)}$ in tier-2 network s_2, $\tilde{z}_{nu}(\mu_n^{(k,2,s_2)}) = \tilde{v}_{nu}(\mu_n^{(k,2,s_2)})/\tilde{m}_{nu}(\mu_n^{(k,2,s_2)})$. From Eqn. (32) and Eqn. (34), $\tilde{z}_{nu}(\mu_n^{(k,2,s_2)})$ is obtained as

$$\tilde{z}_{nu}(\mu_n^{(k,2,s_2)}) = 1 - \tilde{m}_{nu}(\mu_n^{(k,2,s_2)}) + \frac{\frac{A_n^{(k,1,s_1)}}{Z_n^{(k,1,s_1)}} \cdot m'_{n1}(\lceil \frac{\beta_n^{(k,1,s_1)}}{d_k Z_n^{(k,1,s_1)}} \rceil)}{\epsilon_n^{(k,s_2,s_1)} \cdot \tilde{m}_{nu}(\mu_n^{(k,2,s_2)})}. \quad (11)$$

The recursive method presented in Appendix III is also used to derive the redefined mean, variance and peakedness of the overflowed handoff call traffic from the equivalent trunk group for tier-1 network s_1 to tier-2 network s_2, according to the mean call service time $1/\mu_h^{(k,2,s_2)}$ in tier-2 network s_2.

In particular, for homogeneous hierarchical networks where $\epsilon_n^{(k,s_2,s_1)} = 1$, Eqn. (11) is simplified to Eqn. (10), and Eqn. (32) is simplified to Eqn. (9). This shows that the statistical characterization of overflow traffic is consistent in homogeneous hierarchical overlay networks.

Finally, from Eqn. (25), Eqn. (26), and Eqn. (11), the peakedness of the overflow traffic offered to tier-2 network s_2 from class k overflowed new calls of tier-1 network s_1, denoted as $z_{nu}^{(k,1,s_1)}$, is obtained as

$$z_{nu}^{(k,1,s_1)} = \tilde{z}_{nu}(\mu_n^{(k,2,s_2)}) Z_n^{(k,1,s_1)}. \quad (12)$$

Based on Eqn. (25) and Eqn. (32), the mean of the overflow traffic offered to tier-2 network s_2 from class k overflowed new calls of tier-1 network s_1, denoted as $m_{nu}^{(k,1,s_1)}$, is obtained as:

$$m_{nu}^{(k,1,s_1)} = \tilde{m}_{nu}(\mu_n^{(k,2,s_2)}) Z_n^{(k,1,s_1)} = \frac{\hat{B}_n^{(k,1,s_1)} A_n^{(k,1,s_1)}}{\epsilon_n^{(k,s_2,s_1)}}. \quad (13)$$

The $m_{nu}^{(k,1,s_1)}$ obtained also represents the load intensity of class k overflowed new call traffic from tier-1 network s_1 to tier-2 network s_2. Then, the average call arrival rate of the overflowed new calls from s_1 to s_2 can be obtained according to the mean call service time in tier-2 network s_2 as

$$\lambda_{nu}^{(k,1,s_1)} = m_{nu}^{(k,1,s_1)} \mu_n^{(k,2,s_2)}. \quad (14)$$

The overflow traffic flows from different tier-1 networks are independent of each other; the upward overflow traffic offered to tier-2 network s_2 is therefore the superposition of the overflow traffic from the covered N_1 networks at tier-1. The input traffic to tier-2 network s_2 is the aggregation of the local Poisson traffic (new and handoff call traffic) and the aggregated overflow traffic from tier-1; it is non-Poisson. We have the mean $A_n^{(k,2,s_2)}$ and the variance $V_n^{(k,2,s_2)}$ of the input traffic to tier-2 network s_2 from class k new calls obtained as

$$A_n^{(k,2,s_2)} = \frac{\lambda_n^{(k,2,s_2)}}{\mu_n^{(k,2,s_2)}} + \sum_{s_1=1}^{N_l} m_{nu}^{(k,1,s_1)}, \quad (15)$$

$$V_n^{(k,2,s_2)} = \frac{\lambda_n^{(k,2,s_2)}}{\mu_n^{(k,2,s_2)}} + \sum_{s_1=1}^{N_l} m_{nu}^{(k,1,s_1)} z_{nu}^{(k,1,s_1)}, \quad (16)$$

and the peakedness obtained as $Z_n^{(k,2,s_2)} = \frac{V_n^{(k,2,s_2)}}{A_n^{(k,2,s_2)}}$.

By Eqn. (15) and Eqn. (16), we have determined the input traffic model of the upward overflow traffic offered to a high tier network. The characterizations for the input traffic, when handoff call traffic and downward overflow traffic are involved, can be derived by the same method.

B. Call-level loss performance analysis

Let $(A_n^{(k,l,s)}, Z_n^{(k,l,s)}, d_k)$ represent the input new call traffic of class k to tier l network s. The values of $A_n^{(k,l,s)}$ and $Z_n^{(k,l,s)}$ for $l = 2$ are derived by Eqn. (15) and Eqn. (16). The input handoff call traffic of class k to tier l network s, represented as $(A_h^{(k,l,s)}, Z_h^{(k,l,s)}, d_k)$ is determined by the same method.

By the model proposed in Section III, we can simplify the problem of non-Poisson traffic loss analysis in tier l network s to a problem of Poisson traffic loss analysis in the equivalent trunk group which is offered with Poisson traffic $\left(\frac{A_n^{(k,l,s)}}{Z_n^{(k,l,s)}}, 1, d_k Z_n^{(k,l,s)}\right)$ from new calls and $\left(\frac{A_h^{(k,l,s)}}{Z_h^{(k,l,s)}}, 1, d_k Z_h^{(k,l,s)}\right)$ from handoff calls, for $1 \leq k \leq K$, under the condition that each call in the equivalent group has the same service time as that in tier l network s. Here, the mean service time is $1/\mu_n^{(k,l,s)}$ for class k new calls and $1/\mu_h^{(k,l,s)}$ for class k handoff calls.

The loss performance of the equivalent trunk group can be obtained by existing loss calculation methods for multiservice networks with Poisson traffic. In the case of multiservice sharing in a network with no bandwidth reservation, the call blocking probabilities can be derived from a product-form solution of the multiservice link occupancy distribution obtained by Kaufman-Roberts' recursion [25], [26] or the convolution algorithm [27]. Such a product-form solution is not obtainable in multiservice networks with bandwidth reservation, where the reversibility of the system state transition process is disrupted by bandwidth reservation [28]. For a completely shared multiservice link, the link occupancy distribution is derived from a unique state space. However, for a multiservice link with bandwidth reservation, multiple composite state spaces are mapped to the same link occupancy distribution; the different composite state spaces are determined by the arrival order of different service classes at a given link occupancy state. This point can be elaborated by the following example. Assuming a link with four channels to accommodate two classes of calls; each class 1 call occupies one channel, and each class 2 call occupies two channels. Bandwidth reservation is used to protect class 2 calls by reserving two channels in the link for their use. In this example, there exist two composite state spaces for the link occupancy distribution. Let n denote the link occupancy state, which represents the number of occupied channels in the link, $n = 0, 1, 2, ..., 4$. Let n_1 and n_2 denote the number of channels in the link occupied by class 1 and class 2 calls, respectively. At a given link occupancy state n, $n = 0, 1, 2, ..., 4$, we have one composite state (n_1, n_2) for the case that a class 1 call is already in this link when a class 2 call arrives, and the other composite state (n_2, n_1) for the case that a class 2 call is already in this link when a class 1 call arrives; here $n_1 + n_2 = n$ for $n = 0, 1, 2, ..., 4$. The state space for the composite state (n_1, n_2) is defined as

$$\underbrace{(0,0)}_{n=0}; \underbrace{(1,0)}_{n=1}; \underbrace{(2,0),(0,2)}_{n=2}; \underbrace{(1,2)}_{n=3}; \underbrace{(0,4),(2,2)}_{n=4};$$

the state space for the composite state (n_2, n_1) is defined as

$$\underbrace{(0,0)}_{n=0}; \underbrace{(0,1)}_{n=1}; \underbrace{(0,2),(2,0)}_{n=2}; \underbrace{n/a}_{n=3}; \underbrace{(4,0)}_{n=4}.$$

The difference between the composite state space of (n_2, n_1) and that of (n_1, n_2) is due to the bandwidth reservation for class 2 calls. As two channels are reserved for class 2 calls, a class 1 call will be rejected when the occupied channels are equal to or more than two channels. Hence, in the state space of (n_2, n_1), the composition for $n = 3$ is not reachable (n/a), and for $n = 4$ only the composition $(4, 0)$ is reachable.

Now consider K service classes sharing a link with bandwidth reservation. Based on the arrival order of different service classes at a given link occupancy state, there exist $K!$ composite state spaces mapping to the link occupancy distribution of the K service classes sharing the link. From each composite state space, we can obtain a corresponding link occupancy distribution for K service classes in the link by using the convolution algorithm of [27]. The final link occupancy distribution for K service classes in the link is obtained with a weighted summation of all link occupancy distributions obtained from the $K!$ composite state spaces. The weight for a given composite state space is determined by the offered traffic load proportion from each service class in the link at the given composite state [28]. That is, the weight reflects the impact of a given composite state space on the final link occupancy distribution. The approximate blocking probability of each class of calls is then calculated based on the approximate final link occupancy distribution that has been obtained. We name this approximation method the permutational convolution algorithm (PCA). The accuracy of the PCA approximation is verified by extensive comparisons with the exact solutions from the multi-dimensional Markov chain. More details of the PCA are presented in [28].

Here we use PCA to calculate the call blocking probability of each class of approximate Poisson traffic in the equivalent trunk group for tier l network s. Let $C_{l,s}$ denote the capacity of tier l network s in bandwidth units (BUs). Let tr_k be the reservation threshold in number of BUs reserved for class k ($1 \leq k \leq K$) handoff calls (both local and overflowed) in each tier network. The maximum number of BUs which can be assigned to class k new calls in tier l network s is $C_{l,s} - tr_k$; for class k handoff calls it is $C_{l,s}$.

The final link occupancy distribution for the K service classes sharing tier l network s, denoted as $\mathbf{Q}_t^{(l,s)}$, is obtained as the weighted summation of the link occupancy distributions obtained from different composite state spaces by PCA. Let $q_t^{(l,s)}(x)$, $x = 0, 1, 2, ..., C_{l,s}$, denote the element of $\mathbf{Q}_t^{(l,s)}$, representing the probability in the link occupancy state that there are x channels occupied in the link by the K classes of calls. Then the call blocking probabilities for class k new and handoff calls in tier l network s are calculated by

$$B_n^{(k,l,s)} = \sum_{x=C_{l,s}-tr_k-d_k Z_n^{(k,l,s)}+1}^{C_{l,s}} q_t^{(l,s)}(x), \tag{17}$$

$$B_h^{(k,l,s)} = \sum_{x=C_{l,s}-d_k Z_h^{(k,l,s)}+1}^{C_{l,s}} q_t^{(l,s)}(x). \tag{18}$$

The results for non-integer $d_k Z_n^{(k,l,s)}$ and $d_k Z_h^{(k,l,s)}$ are obtained by interpolation algorithms.

Consider a L-tier hierarchical overlay system. Let s_l denote a network at tier-l, for $1 \leq l \leq L$. For class k new calls of tier l network s, the call blocking probability in the L-tier overlay system is the probability that the calls are blocked by tier l network s and also rejected when attempting to overflow to

high tier networks, i.e. tier l' network $s_{l'}$, tier $l'+1$ network $s_{l'+1}$,..., and tier L network s_L. Thus, the final call blocking probability $B_c^{(k,l,s)}$ for class k new calls from tier l network s in the L-tier hierarchical overlay system the is obtained as

$$B_c^{(k,l,s)} = \prod_{l'=l}^{L} B_n^{(k,l',s_{l'})}, \tag{19}$$

where $B_n^{(k,l',s_{l'})}$ denotes the call blocking probability of class k service in tier l' network $s_{l'}$.

The probability that a class k call accepted in tier l network s is dropped during handoff due to a capacity limit is defined as the call dropping probability of class k service; it is denoted as $P_d^{(k,l,s)}$ $(1 \leq l \leq L)$. As the hierarchical heterogeneous overlay system consists of networks with different statistical characteristics at either the same or different tiers, it is hard to obtain a close-form equation for the call dropping probability. However, we can derive the call dropping probability from the state transition diagram shown in Fig. 3. The derivation method is described in Appendix IV. Suppose a two-tier overlay system ($L = 2$). Assume that tier-1 network s is covered by tier-2 network s' and the overflow calls from tier-1 network s are offered to tier-2 network s'. The call dropping probability for class k calls from tier-1 network s in the two-tier overlay system is obtained as

$$P_d^{(k,1,s)} = p_f^{(k,1)} B_h^{(k,2,s')} + p_f^{(k,1)}(1 - B_h^{(k,2,s')}) p_f^{(k,2)}. \tag{20}$$

In Eqn. (20), $p_f^{(k,l)}$ denotes the probability that a class k call accepted in a network at tier l makes another handoff within tier l but is rejected due to capacity limits in the neighboring networks at tier l. For conciseness, we refer to $p_f^{(k,l)}$ as the intra-tier handoff failure probability of class k calls at tier l; this is also derived from the state transition diagram shown in Fig. 3. Also see Eqn. (35) in Appendix IV.

IV. Model Validation And Numerical Results

A. Parameter settings

We evaluate the performance of the proposed approximate model in a two-tier heterogeneous mobile network based on the hierarchical structure, where the networks of large coverage overlay those with small coverage. Here we assume two large cells at the top tier (tier-2) and each large cell overlays two small cells at tier-1. Call handoffs are allowed between cell-1 and cell-2, cell-3 and cell-4 at tier-1, and also between the two tier-2 cells. Call overflow is allowed between overlay networks. It is assumed that each network has a circular coverage and that the capacity is measured in number of BUs. The values of the radius R and the capacity C of the networks are given in Table. I.

We consider two service classes with different bandwidth requirements for each call. A call from a class 1 service is allocated one BU and a call from a class 2 service is allocated two BUs. Bandwidth reservation is used to protect handoff calls. The reservation thresholds are chosen to equalize call blocking probabilities for both classes of new calls. To achieve this goal, two BUs are reserved in each network for handoff calls of class 1 service, and one BU is reserved in each network for handoff calls of class 2 service, i.e. $tr_1 = 2$ and $tr_2 = 1$.

For demonstration purpose, what is of more interest is the distributions of the call holding time and the sojourn time in a cell, rather than the users' mobility pattern or trajectory. Negative exponential distribution has been commonly used to approximate call holding time and sojourn time distributions in mobile cellular networks [17], [21]. Recent investigations show that the negative exponential distribution for sojourn time may lead to a slightly overestimated or underestimated call loss performance when compared with the other specific distributions [11], [16]. On the other hand, the negative exponential distribution has been verified as a good approximation for call duration time compared to other general distributions. In our analytical model, we assume that both call holding time and sojourn time follow the negative exponential distribution. The limitation of this assumption has been evaluated by comparing with the simulation results. In the simulation, we assume that the call holding time follows the negative exponential distribution and the sojourn time follows the generalized Gamma distribution as in [29] and [30].

We choose the mean call holding time equal to 180 seconds for both class 1 and class 2 new calls, and choose the mean sojourn times denoted as τ_s for slow-speed users and τ_f for fast-speed users in different tier cells as shown in Table I.

TABLE II
THE PROPORTION OF DIFFERENT SPEED CALLS IN DIFFERENT CLASS OF SOURCE TRAFFIC.

Mixture pattern	Class 1 service (λ_1)		Class 2 service (λ_2)	
	slow call	fast call	slow call	fast call
Case-1	0	100%	100%	0
Case-2	50%	50%	50%	50%
Case-3	100%	0	0	100%

To demonstrate the impact of mobility speed on traffic overflow, we evaluate the loss performance under different mixtures of slow-speed and fast-speed calls in one class of source traffic. Let λ_1 represent the new call arrival rate from class 1 source traffic to a network. Table II shows the various mixture patterns we consider, where the slow-speed new calls in class 1 service are assigned a call arrival rate $\lambda_1^{(s)}$ equal to 0%, 50%, 100% of λ_1, and the fast-speed new calls in class 1 service are assigned a call arrival rate $\lambda_1^{(f)}$ given by $\lambda_1^{(f)} = \lambda_1 - \lambda_1^{(s)}$.

We use the proposed analytical model to evaluate the call-level loss performance of mobile users with different service classes and at different speeds in the two-tier heterogeneous overlay system. For verification, all of the analytical results obtained with our model are compared to the results obtained with an overflow queuing model simulation using OPNET [31]. In simulation, the confidence intervals for new call blocking probabilities are kept within 2% of the simulation results, and for call dropping probabilities are kept within 5% of the simulation results, both obtained with a 95% level of confidence based on Student's t-distribution.

TABLE I
PARAMETER SETTINGS OF THE SYSTEM MODEL.

Tier-2 Cell-1: $R = 200$m, $C = 40$ BUs $\tau_s = 1130.97$s, $\tau_f = 22.62$s		Tier-2 Cell-2: $R = 400$m, $C = 60$ BUs $\tau_s = 2261.95$s, $\tau_f = 45.24$s	
Tier-1 Cell-1: $R = 100$m, $C = 15$ BUs $\tau_s = 565.5$s, $\tau_f = 11.3$s	Tier-1 Cell-2: $R = 50$m, $C = 10$ BUs $\tau_s = 282.7$s, $\tau_f = 5.65$s	Tier-1 Cell-3: $R = 250$m, $C = 12$ BUs $\tau_s = 1413.7$s, $\tau_f = 28.3$s	Tier-1 Cell-4: $R = 100$m, $C = 12$ BUs $\tau_s = 565.5$s, $\tau_f = 11.3$s

B. Loss performance evaluation

The loss performance of the two service classes is evaluated in three scenarios: the performance in tier-1, the performance in tier-2 and the performance in the two-tier system. Based on the particular reservation thresholds we have chosen, equalized call blocking probability is obtained for both class 1 and class 2 services. For any class of calls originating in the tier-1 cells, reduced call blocking probabilities are obtained in the two-tier overlay system, because the blocked calls in tier-1 due to capacity limit can overflow to tier-2 cells. The impact of users' mobility speed on the loss performance is evaluated by the call dropping probabilities of slow-speed and fast-speed calls with respect to the three mixture patterns defined in Table II.

In Fig. 4 and Fig. 5, we present the call blocking probabilities of the new calls originating in the tier-1 cells in the considered system. In Fig. 6 and Fig. 7 we present the call dropping probabilities of the two classes of calls originating in tier-1 cell-1 and cell-2. All of the results obtained by the proposed analytical model are verified by comparison with the simulation results in the figures. Our analytical results match very well the simulation results in most cases.

In Fig. 4, the call blocking probabilities of the new calls originating in tier-1 cell-2 under the three scenarios, tier-1, tier-2 and the two-tier overlay system, decrease as the proportion of fast-speed new calls in class 1 service decreases from 100% to zero. This trend is also observed in Fig. 5 for the new calls originating in tier-1 cell-3. This observation is reasonable because Table II shows that the decrease of fast-speed calls in class 1 service corresponds to the decrease of the slow-speed calls in class 2 service. In our example, slow-speed calls of class 2 service not only demand more bandwidth than class 1 calls, but also have longer sojourn times and bandwidth occupancies than class 2 fast-speed calls. The decrease of the slow-speed calls in class 2 service thus gives more available bandwidth units to other calls. As for the increased fast-speed calls in class 2 service, they are overflowed to tier-2 cells if blocked by a capacity limit at tier-1.

A different phenomenon is observed in Fig. 5 for the call blocking probabilities in tier-2 cell-2 (covering tier-1 cell-3 and cell-4) and the two-tier overlay. For small new call arrival rates, the call blocking probability in these two scenarios increases as the proportion of fast-speed calls in class 1 service decreases from 100% to zero. The reason is also related to the proportions of fast-speed calls in the two service classes. Due to the speed-sensitive CAC scheme, blocked fast-speed calls at the lower tier networks can overflow to the higher tier networks for possible service. From Table II, the decrease of fast-speed calls in class 1 service corresponds to the increases of slow-speed calls in class 1 and fast-speed calls in class 2. Compared with the overflows from class 1 service, we have more overflow calls to tier-2 cell-2 from fast-speed calls of class 2. The overflowed fast-speed calls of class 2 compete bandwidth with the original calls in tier-2 cell-2. As the proportion of fast-speed calls in class 2 increases, there is increased bandwidth competition in tier-2 cell-2 and increased call blocking probability. An alternative solution to this problem is to reserve bandwidth for the original calls in the higher tier networks to guarantee a required blocking probability.

The speed-sensitive CAC scheme allows the blocked calls of fast-speed users in the lower tier to be overflowed upward to the higher tier and share the bandwidth resource with the local new calls and handoff calls in the higher tier. The increase of the overflow traffic from fast-speed calls leads to increased call blocking and dropping probabilities in the higher tier. The results shown in Fig. 6 and Fig. 7 demonstrate that fast-speed users have a higher rate of handoff among the neighboring networks, and thus experience a higher possibility of handoff failure than slow-speed users. It is again demonstrated that the decrease of slow-speed calls in class 2 increases the bandwidth available for other calls, therefore the call dropping probabilities for both slow-speed and fast-speed calls of class 2 shown in Fig. 7 are reduced as the proportion of slow-speed calls in class 2 decreases.

Now we evaluate the limitation of the exponential distribution assumption for the sojourn time. In Fig. 8 to Fig. 10 we present the simulation results of the call blocking and dropping probabilities for the calls originating in tier-1 cell-1 in the considered two-tier overlay system, under the assumptions that the sojourn times follow the Gamma distributions with the fixed mean sojourn time and the shape parameter equal to 2, 3, 4. Fig. 8 shows that the new call blocking probabilities obtained by simulation under the three Gamma distribution assumptions are similar to the analytical and simulated results obtained under the exponential assumption. It shows that given the mean sojourn time, the sojourn time distribution has slight influence on the new call blocking probability. Fig. 9 and Fig. 10 show that the sojourn time distribution has much more influence on the call dropping probability of slow-speed users than on the fast-speed users. This shows the limitation of our analytical model using the exponential assumption on the sojourn times for slow-speed users.

For validation of our analytical model, in Fig. 8 to Fig. 10 we present the performance obtained under the assumption that the aggregated overflow traffic of the same service class from tier-1 cells to tier-2 is Poisson traffic, and compare it with the results obtained by our analytical model and by the simulation. The comparison shows that significant underestimated performance evaluation is obtained by the Poisson assumption. It demonstrates that in hierarchical heterogeneous

overlay systems, the aggregated overflow traffic from one tier to another cannot be modeled as a Poisson process since the aggregated overflow call arrivals are not from a sufficiently large number of independent overflow sources [32].

V. Conclusions

By taking the effects of user mobility, bandwidth reservation, cell coverage and varying service time distributions for cells at the same or different tiers into consideration, we have proposed a comprehensive loss model to obtain the numerical solution of multiservice loss performance in hierarchical heterogeneous overlay networks. We have also demonstrated that the use of speed-sensitive call admission control scheme in hierarchical heterogeneous overlay networks helps improve the call-level loss performance.

Acknowledgement

The work described in this paper was supported by a grant from The City University of Hong Kong (Project No. 7002375), and the Department of Electronic Engineering, City University of Hong Kong. We would also like to express our sincere appreciations to the anonymous reviewers for their helpful suggestions.

Appendix I: Notations

$\lambda_n^{(k,l,s)}$, $\lambda_h^{(k,l,s)}$: average call arrival rates of class k local new calls and local handoff calls to tier l network s.

$\lambda_{nu}^{(k,1,s_1)}$, $\lambda_{hu}^{(k,1,s_1)}$: average call arrival rates of class k new and handoff call overflows from tier-1 network s_1 to tier-2.

$z_{nu}^{(k,1,s_1)}$, $z_{hu}^{(k,1,s_1)}$: peakedness of class k overflow traffic of new and handoff calls from tier-1 network s_1 to tier-2.

$\lambda_{nd}^{(k,2,s_2)}$, $\lambda_{hd}^{(k,2,s_2)}$: average call arrival rates of class k new and handoff call overflows from tier-2 network s_2 to tier-1.

$z_{nd}^{(k,2,s_2)}$, $z_{hd}^{(k,2,s_2)}$: peakedness of overflow traffic from class k new and handoff calls in tier-2 network s_2 to tier-1.

$\Lambda_n^{(k,l,s)}$, $\Lambda_h^{(k,l,s)}$: aggregated average call arrival rates of class k new and handoff call traffic to tier l network s.

$Z_n^{(k,l,s)}$, $Z_h^{(k,l,s)}$: peakedness of aggregated class k new call and handoff call overflow traffic to tier l network s.

$\nu_n^{(k,l,s)}$, $\nu_v^{(k,l,s)}$, $\nu_h^{(k,l,s)}$: probabilities of class k accepted new calls (both local and overflowed), accepted overflow handoff calls and accepted local handoff calls, respectively, making a handoff out of tier l network s.

$B_n^{(k,l,s)}$, $B_h^{(k,l,s)}$: blocking probability for class k new calls and handoff calls in tier l network s.

$1/\mu_n^{(k,l,s)}$, $1/\mu_h^{(k,l,s)}$: mean service times for class k new call and handoff call in tier l network s, respectively.

d_k: the number of BUs allocated to a class k call.

tr_k: reserved capacity for class k handoff calls in each tier network in numbers of BUs.

$C_{l,s}$: capacity of tier l network s in numbers of BUs.

Appendix II: Partially Equivalent Loss Systems

The Hayward's approximation [33] is derived under the condition that the non-Poisson traffic with intensity A and peakedness Z offered to a N-server trunk group system, i.e. $(N; A, Z)$, is equivalent to a system comprising Z independent and identical subgroups; each subgroup is assigned N/Z servers and the calls from the non-Poisson source traffic (A, Z) are evenly scheduled into each subgroup to ensure that the traffic offered to each subgroup is Poisson with intensity A/Z [33]. Let the resultant subgroup be represented as $(\frac{N}{Z};\ \frac{A}{Z}, 1)$. The two systems, $(N; A, Z)$ and $(\frac{N}{Z}; \frac{A}{Z}, 1)$ have equivalent call blocking probabilities but different statistical moments for their overflow traffic. We refer to these two systems as partially equivalent loss systems.

The accurate moments of the overflow traffic from any subgroup can be derived from the distribution of the number of overflow calls accepted in a "fictitious" overflow group with infinite number of servers. Let $p_x(i, j)$ denote the probability of the state in which there are i new calls accepted in subgroup x and j overflow calls in its infinite-server overflow group, for $0 \le i \le N/Z$, $0 \le j \le \infty$. As the subgroups are identical and synchronized, and each of them is offered by Poisson traffic with the same intensity A/Z, we have the same state probability for any two subgroups denoted as x and x', $1 \le x, x' \le Z$, that is

$$p_x(i, j) = p_{x'}(i, j). \tag{21}$$

Let m and v denote the mean and the variance of the overflow traffic from the subgroup x to its infinite-server overflow group. They are determined as

$$m = \sum_{i=0}^{N/Z} \sum_{j=0}^{\infty} j p_x(i, j), \tag{22}$$

$$v = \sum_{i=0}^{N/Z} \sum_{j=0}^{\infty} j^2 p_x(i, j) - m^2. \tag{23}$$

The overflow traffic from the system $(N; A, Z)$ is the superposition of the overflow traffic from all Z subgroups. Let $p(i', j')$ denote the state probability of the system $(N; A, Z)$ that there are i' new calls in the system $(N; A, Z)$ and j' overflow calls in its infinite-server overflow group, $0 \le i' \le N, 0 \le j' \le \infty$. As the calls from the non-Poisson source traffic (A, Z) are evenly scheduled to each subgroup, the state probability $p(i', j')$ is as same as the state probability $p_x(i, j)$ for $i' = Z \cdot i$, $j' = Z \cdot j$, i.e.

$$p_x(i, j) = p(i', j') = p(Z \cdot i, Z \cdot j). \tag{24}$$

Thus, the mean of the overflow traffic from the system $(N; A, Z)$, denoted as M_o, is derived as

$$M_o = \sum_{i'=0}^{N} \sum_{j'=0}^{\infty} j' p(i', j') = Z \sum_{i=0}^{N/Z} \sum_{j=0}^{\infty} j p_x(i, j) = Z \cdot m \tag{25}$$

and the variance V_o is derived as

$$V_o = \sum_{i'=0}^{N} \sum_{j'=0}^{\infty} j'^2 p(i', j') - M_o^2 = Z^2 \cdot v. \tag{26}$$

Then the peakedness of the overflow traffic from the system $(N; A, Z)$, denoted as Z_o, is obtained by $Z_o = \frac{V_o}{M_o} = Z \cdot \frac{v}{m}$.

Appendix III: Derivation of Moments of Overflow Traffic in Heterogeneous Scenarios

Let the new calls that cannot be admitted to the hypothetical group for the new call traffic in tier-1 network s_1 overflow to a secondary trunk group with infinite number of servers. The hypothetical group is shown in Fig. 2.

Let $p_n(x_1, x_2)$ denote the probability of the equilibrium state that there are x_1 class k new calls accepted in the hypothetical group and x_2 overflowed new calls accepted in the infinite-server secondary trunk group, with $0 \leq x_1 \leq \lceil \frac{\beta_n^{(k,1,s_1)}}{d_k Z_n^{(k,1,s_1)}} \rceil$ and $x_2 > 0$. We have the r-th moment m_{nr} ($r \geq 0$) and the conditional r-th moment $m'_{nr}(x_1)$ for the following recursion, the overflowed new call traffic from the hypothetical group defined as the probability distributions of the number of overflow calls accepted in the infinite-server secondary trunk group, that is

$$m_{nr} = \sum_{x_2=0}^{\infty} \sum_{x_1=0}^{\lceil \frac{\beta_n^{(k,1,s_1)}}{d_k Z_n^{(k,1,s_1)}} \rceil} (x_2)^r \cdot p_n(x_1, x_2), \quad (27)$$

$$m'_{nr}(x_1) = \sum_{x_2=0}^{\infty} (x_2)^r \cdot p_n(x_1, x_2). \quad (28)$$

As shown in [24], the conditional first moment $m'_{n1}(x_1)$ of the overflowed new call traffic can be derived by

$$m'_{n1}(x_1) = \frac{1}{x_1} \left(\frac{A_n^{(k,1,s_1)}}{Z_n^{(k,1,s_1)}} - x_1 - 1 + \epsilon_n^{(k,s_2,s_1)} \right) \cdot m'_{n1}(x_1 - 1) - \frac{A_n^{(k,1,s_1)} \cdot m'_{n1}(x_1 - 2)}{Z_n^{(k,1,s_1)} \cdot x_1}, \quad \text{for } 1 \leq x_1 \leq \lceil \frac{\beta_n^{(k,1,s_1)}}{d_k Z_n^{(k,1,s_1)}} \rceil; \quad (29)$$

and the second moment m_{n2} is derived as

$$m_{n2} = \frac{1}{\epsilon_n^{(k,s_2,s_1)}} \cdot \frac{A_n^{(k,1,s_1)}}{Z_n^{(k,1,s_1)}} \cdot \left(m'_{n1}(x_1) + m'_{n0}(x_1) \right), \quad \text{for } x_1 = \lceil \frac{\beta_n^{(k,1,s_1)}}{d_k Z_n^{(k,1,s_1)}} \rceil. \quad (30)$$

In Eqn. (30), $m'_{n0}(x_1) = \sum_{x_2=0}^{\infty} p_n(x_1, x_2)$; for $x_1 = \lceil \frac{\beta_n^{(k,1,s_1)}}{d_k Z_n^{(k,1,s_1)}} \rceil$, $m'_{n0}(x_1)$ is just equal to the new call blocking probability for class k service in its hypothetical group, i.e. $\hat{B}_n^{(k,1,s_1)}$, and we have $\hat{B}_n^{(k,1,s_1)} \approx B_n^{(k,1,s_1)}$ for class k service in the equivalent trunk group for tier-1 network s_1.

The variance of the new call overflow from the hypothetical group for tier-1 network s_1 to tier-2 network s_2 is defined as

$$\tilde{v}_{nu}(\mu_n^{(k,2,s_2)}) = m_{n2} - (m_{n1})^2, \quad (31)$$

Let $\tilde{m}_{nu}(\mu_n^{(k,2,s_2)})$ denote the redefined mean of the overflowed new call traffic from the hypothetical group for tier-1 network s_1 to tier-2 network s_2 according to the mean call service time $1/\mu_n^{(k,2,s_2)}$ in tier-2 network s_2. It is obtained as

$$\tilde{m}_{nu}(\mu_n^{(k,2,s_2)}) = \hat{B}_n^{(k,1,s_1)} \cdot \frac{A_n^{(k,1,s_1)}}{Z_n^{(k,1,s_1)}} \cdot \frac{1}{\epsilon_n^{(k,s_2,s_1)}}. \quad (32)$$

Based on the definition of m_{n1} in Eqn. (27), we have

$$m_{n1} = \tilde{m}_{nu}(\mu_n^{(k,2,s_2)}). \quad (33)$$

From Eqn. (29), Eqn. (30), and Eqn. (31), we derive the variance of the class k overflowed new call traffic from the hypothetical group for tier-1 network s_1 to tier-2 network s_2, according to the mean call service time $1/\mu_n^{(k,2,s_2)}$:

$$\tilde{v}_{nu}(\mu_n^{(k,2,s_2)}) = \frac{1}{\epsilon_n^{(k,s_2,s_1)}} \cdot \frac{A_n^{(k,1,s_1)}}{Z_n^{(k,1,s_1)}} \cdot m'_{n1}(\lceil \frac{\beta_n^{(k,1,s_1)}}{d_k Z_n^{(k,1,s_1)}} \rceil) + \tilde{m}_{nu}(\mu_n^{(k,2,s_2)}) - \left(\tilde{m}_{nu}(\mu_n^{(k,2,s_2)}) \right)^2. \quad (34)$$

Appendix IV: Derivation of Call Dropping Probabilities

For either fast or slow-speed calls admitted to tier l network j in a L-tier hierarchical system, $1 \leq l \leq L$, if their handoff attempts to the neighbors of tier l network j are rejected, the handoff calls can be overflowed to the overlaying tier $l+1$ or overlaid tier $l-1$ networks for possible services based on the speed-sensitive CAC. If the overflowed handoff calls are rejected again, further upward or downward overflow can also be tried until all available tiers have been checked. An on-going call is dropped in such systems only if it fails in all handoff attempts. Here we exemplify the fast-speed calls and derive their call dropping probabilities in the considered hierarchical heterogeneous overlay system. Similar method is also valid for the call dropping probability of slow-speed calls.

Fig. 3 shows our model for call dropping, in which we consider all possible handoff failures for a new call accepted in tier l network j, $1 \leq l \leq L$. This accepted new call in tier l network j makes its first handoff to its neighbor network $j+1$ in tier l with a probability $\nu_n^{(k,l,j)}$. The first handoff is accepted by tier l network $j+1$ with a probability $1 - B_h^{k,l,j+1}$. After this successful handoff, this call may make another handoff out of tier l network $j+1$ with a probability $\nu_h^{(k,l,j+1)}$. On the other hand, if the first handoff is rejected with a probability $B_h^{k,l,j+1}$, this handoff call is overflowed to its overlaying network at tier $l+1$. If this overflowed handoff call is accepted in a tier $l+1$ network j, with a probability $1 - B_h^{(k,l+1,j)}$, continuous handoffs may occur between the neighboring networks at tier $l+1$. For a class k new call accepted to tier l network j, let $p_f^{(k,l)}$ denote the probability that this call fails in its t_l-th handoff at tier l due to a capacity limit and overflows to its overlaying network at tier $l+1$, $p_f^{(k,l)}$ is derived as

$$p_f^{(k,l)} = \nu_n^{(k,l,j)} B_h^{(k,l,j+1)} + \nu_n^{(k,l,j)} \sum_{t_l=2}^{\infty} \prod_{x=1}^{t_l-1} (1 - B_h^{(k,l,j+x)}) \nu_h^{(k,l,j+x)} B_h^{(k,l,j+t_l)}, \quad (35)$$

The first item in the right side of Eqn. (35) represents handoff failure at the first handoff, the second item represents handoff failures at the subsequent second, third,..., t_l-th handoff, respectively. Replace $\nu_n^{(k,l,j)}$ in Eqn. (35) with $\nu_v^{(k,l,j)}$, we obtain

the probability that an overflowed handoff call accepted in tier l network j fails in another handoff at tier l and overflows to tier $l+1$.

In particular, if all networks in the L-tier hierarchical system are homogeneous and identical, Eqn. (35) can be written as

$$p_f^{(k,l)} = \frac{\nu_n^{(k,l)} B_h^{(k,l)}}{1-(1-B_h^{(k,l)})\nu_h^{(k,l)}}. \quad (36)$$

Moreover, if both call holding time and sojourn time follow exponential distributions, $\nu_n^{(k,l,j)}$, $\nu_v^{(k,l,j)}$ and $\nu_h^{(k,l,j)}$ are derived as $\nu_n^{(k,l,j)} = \nu_v^{(k,l,j)} = \nu_h^{(k,l,j)} = \eta_k^{(l)}/(\xi_k + \eta_k^{(l)})$, here $1/\xi_k$ denotes mean call holding time and $1/\eta_k^{(l)}$ denotes mean sojourn time of a class k call in a network at tier l.

Combining all possible handoff failures for a new call accepted in tier l network j, the call dropping probability for class k new calls accepted in tier l network j can be derived from the state transition model shown in Fig. 3. It is very tedious, but not impossible, to write out the full expression of the call dropping probability for the accepted calls in a L-tier hierarchical system. We here exemplify the case of $L = 3$. The call dropping probability for class k new calls accepted in tier $l = 1$ network is written as

$$\begin{aligned} P_d^{(k,l,j)} &= p_f^{(k,l)} B_h^{(k,l+1,j)} B_h^{(k,l+2,j)} + \\ & p_f^{(k,l)} (1-B_h^{(k,l+1,j)}) p_f^{(k,l+1)} B_h^{(k,l+2,j)} + \\ & p_f^{(k,l)} B_h^{(k,l+1,j)} (1-B_h^{(k,l+2,j)}) p_f^{(k,l+2)} + \\ & p_f^{(k,l)} (1-B_h^{(k,l+1,j)}) p_f^{(k,l+1)} (1-B_h^{(k,l+2,j)}) p_f^{(k,l+2)}. \quad (37) \end{aligned}$$

REFERENCES

[1] X. Wu, B. Mukherjee, and D. Ghosal, "Hierarchical architectures in the third-generation cellular network," *IEEE Wireless Commun. Mag.*, vol. 11, no. 3, pp. 62–71, Jun. 2004.

[2] S.-P. Yeh, S. Talwar, S.-C. Lee, and H. Kim, "Wimax femtocells: a perspective on network architecture, capacity, and coverage," *IEEE Commun. Mag.*, vol. 46, no. 10, pp. 58–65, Oct. 2008.

[3] R. Y. Kim, J. S. Kwak, and K. Etemad, "Wimax femtocells: requirements, challenges, and solutions," *IEEE Commun. Mag.*, vol. 47, no. 9, pp. 84–91, Sep. 2009.

[4] D. Calin, H. Claussen, and H. Uzunalioglu, "On femto deployment architectures and macrocell offloading benefits in joint macro-femto deployments," *IEEE Commun. Mag.*, vol. 48, no. 1, pp. 26–32, Jan. 2010.

[5] S. S. Rappaport and L.-R. Hu, "Microcellular communication systems with hierarchical macrocell overlays: traffic performance models and analysis," *Proc. IEEE*, vol. 82, no. 9, pp. 1383–1397, Sep. 1994.

[6] L.-R. Hu and S. S. Rappaport, "Personal communication systems using multiple hierarchical cellular overlays," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 2, pp. 406–415, Feb. 1995.

[7] C. Yang, C. Tsai, J. Hu, and T. Chung, "On the design of mobility management scheme for 802.16-based network environment," *Computer Networks*, vol. 51, no. 8, pp. 2049–2066, 2007.

[8] D. J. Lee, B. C. Shin, and D. H. Cho, "Speed estimation of mobile station in additive noise and rayleigh fading environments," *Wireless Networks*, no. 8, pp. 541–548, 2002.

[9] B. Jabbari and W. F. Fuhrmann, "Teletraffic modeling and analysis of flexible hierarchical cellular networks with speed-sensitive handoff strategy," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 8, pp. 1539–1548, Oct. 1997.

[10] G. Boggia, P. Camarda, and N. D. Fonzo, "Teletraffic analysis of hierarchical cellular communication networks," *IEEE Trans. Veh. Technol.*, vol. 52, no. 4, pp. 931–946, Jul. 2003.

[11] K. Yeo and C. Jun, "Modeling and analysis of hierarchical cellular networks with general distributions of call and cell residence times," *IEEE Trans. Veh. Technol.*, vol. 51, no. 6, pp. 1361–1374, Nov. 2002.

[12] G. T. Karetsos, S. A. Kyriazakos, E. Groustiotis, F. D. Giandomenico, and I. Mura, "A hierarchical radio resource management framework for integrating wlans in cellular networking environments," *IEEE Wireless Commun. Mag.*, vol. 12, no. 6, pp. 11–17, Jun. 2005.

[13] T. E. Klein and S.-J. Han, "Assignment strategies for mobile data users in hierarchical overlay networks; performance optimal and adaptive strategies," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 5, pp. 849–861, Jun. 2004.

[14] W. Song and W. Zhuang, "Multi-service load sharing for resource management in the cellular/wlan integrated network," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 725–735, Feb. 2009.

[15] B.-J. Chang and R.-H. Hwang, "Performance analysis for hierarchical multirate loss networks," *IEEE/ACM Trans. Netw.*, vol. 12, no. 1, pp. 187–199, Feb. 2004.

[16] M. A. Marsan, G. Ginella, R. Maglione, and M. Meo, "Perforamnce analysis of hierarchical cellular networks with generally distributed call holding times and dwell times," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 248–257, Jan. 2004.

[17] S.-P. Chung and J.-C. Lee, "Performance analysis and overflowed traffic characterization in multiservice hierarchical wireless networks," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 904–918, May 2005.

[18] F. Brochin and E. Pradel, "A call traffic model for integrated services digital networks," in *Proc. GLOBECOM*, Dec. 1992, pp. 1508–1512.

[19] Q. Huang, K. T. Ko, and V. B. Iversen, "Approximation of loss calculation for hierarchical networks with multiservice overflows," *IEEE Trans. Commun.*, vol. 56, no. 3, pp. 466–473, Mar. 2008.

[20] J. Jobin, M. Faloutsos, and S. K. Tripathi, "The effects of heterogeneity in parameters in wireless cellular network modeling," in *Proc. IEEE VTC*, vol. 6, Sep., pp. 4432–4436.

[21] P. V. Orlik and S. S. Rappaport, "On the handoff arrival process in cellular communications," *Wireless Networks*, vol. 7, no. 2, pp. 147–157, 2001.

[22] Y.-R. Haung and J.-M. Ho, "Distributed call admission control for a heterogeneous pcs network," *IEEE Trans. Comput.*, vol. 15, no. 12, 2002.

[23] R. I. Wilkinson, "Theories for toll traffic engineering in the U.S.A." *Bell Syst. Tech. J.*, vol. 35, no. 2, pp. 421–514, 1956.

[24] R. G. Schehrer, "A two moments method for overflow systems with different mean holding times," in *Proc. the 15th International Teletraffic Congress*, Washington, DC, USA, Jun. 1997, pp. 1303–1314.

[25] J. S. Kaufman, "Blocking in a shared resource environment," *IEEE Trans. Commun.*, vol. 29, no. 10, pp. 1474–1481, Oct. 1981.

[26] J. W. Roberts, "A service system with heterogeneous user requirements – application to multi-service telecommunications systems," in *Performance of Data Communications Systems and Their Applications*, G. Pujolle, Ed. New York: North Holland, 1981, pp. 423–431.

[27] V. B. Iversen, "The exact evaluation of multi-service loss system with access control," *Teleteknik*, vol. 31, no. 2, pp. 56–61, 1987.

[28] Q. Huang, K.-T. Ko, and V. B. Iversen, "A new convolution algorithm for loss probability analysis in multiservice networks," *Performance Evaluation*, vol. 68, no. 1, pp. 76–87, Jan. 2011.

[29] M. M. Zonoozi and P. Dassanayake, "User mobility modeling and characterization of mobility patterns," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 7, pp. 1239–1252, Sep. 1997.

[30] I. Akyildiz and W. Wang, "A dynamic location management scheme for next-generation multitier pcs systems," *IEEE Trans. Wireless Commun.*, vol. 1, no. 1, pp. 178–189, Jan. 2002.

[31] OPNET University Program:. [Online]. Available: http://www.opnet.com/services/university/

[32] L. Kleinrock, *Queueing Systems: Volume I: Theory*. Wiley-Interscience: New York, 1975.

[33] A. A. Fredericks, "Congestion in blocking systems – a simple approximation technique," *Bell Syst. Tech. J.*, vol. 59, no. 6, pp. 805–827, 1980.

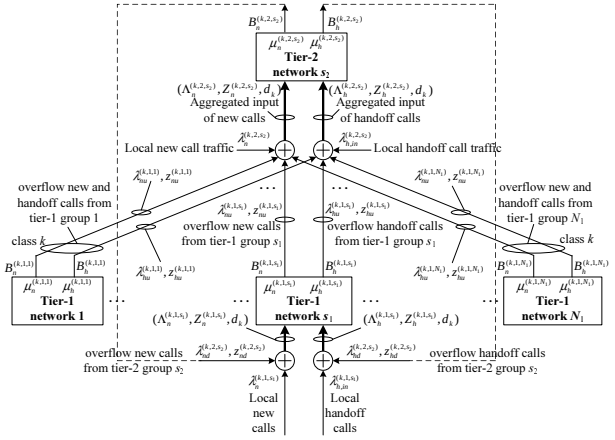

Fig. 1. Class k traffic flows in a two-tier heterogeneous overlay system.

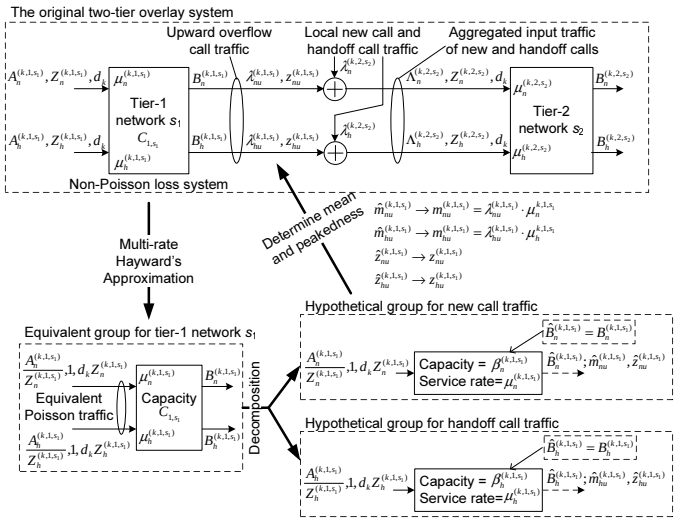

Fig. 2. Loss modeling for class k traffic.

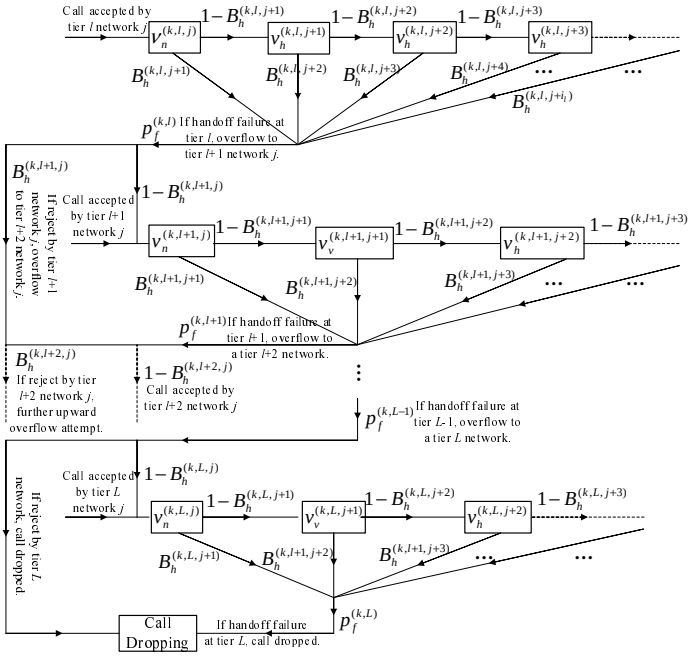

Fig. 3. All possible forced terminations (call dropping) for a call accepted in tier l network j.

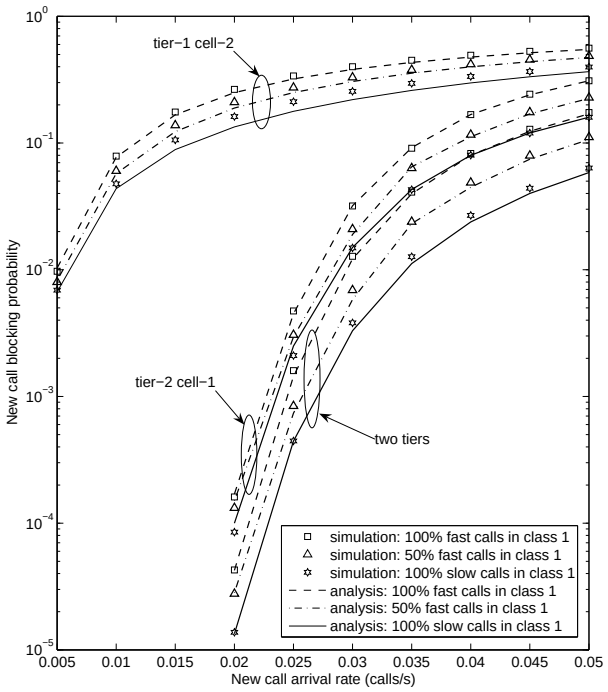

Fig. 4. New call blocking probability for calls from tier-1 cell-2.

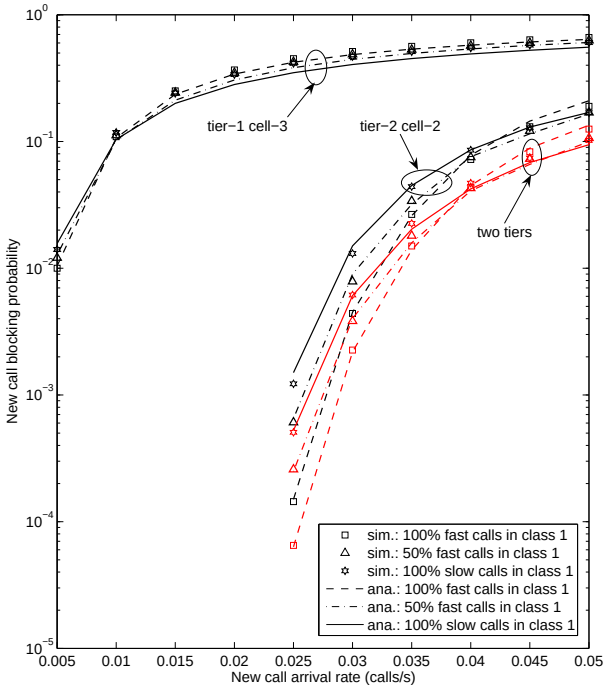

Fig. 5. New call blocking probability for calls from tier-1 cell-3.

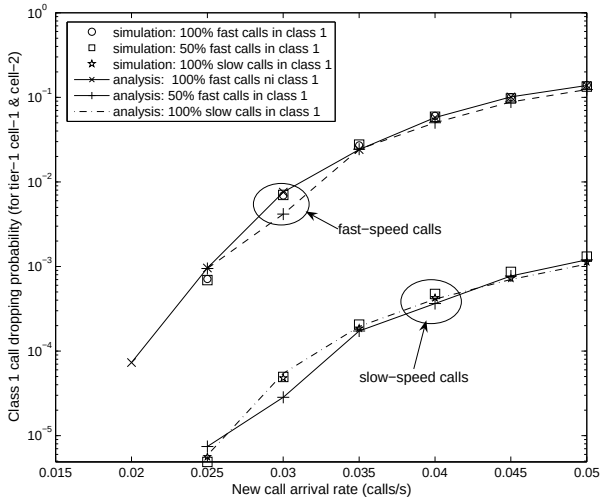

Fig. 6. Call dropping probability for class 1 service in tier-1 cell-1 and cell-2.

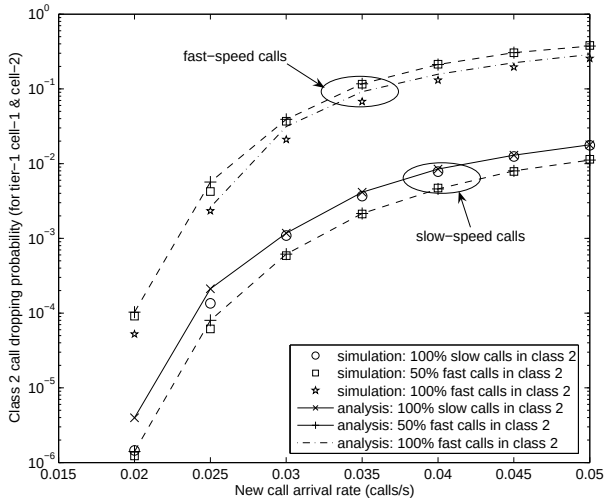

Fig. 7. Call dropping probability for class 2 service in tier-1 cell-1 and cell-2.

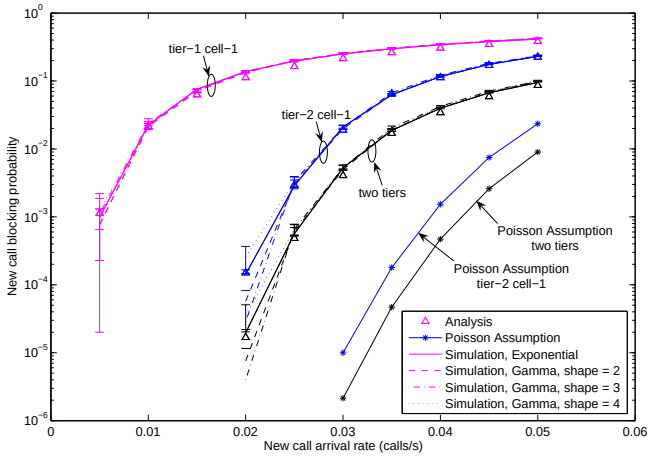

Fig. 8. New call blocking probability for calls from tier-1 cell-1 under mixture pattern case-2.

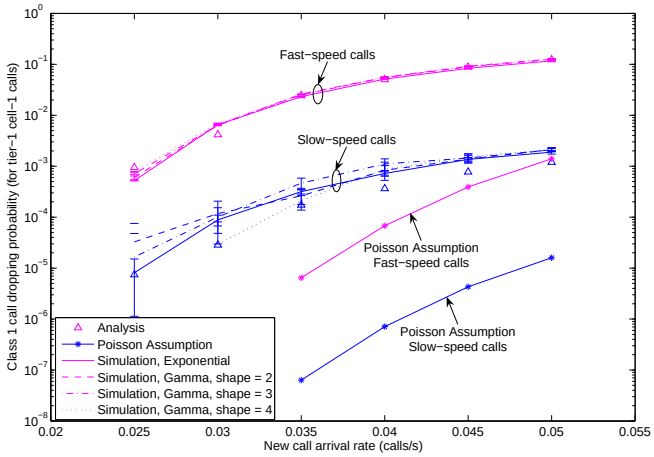

Fig. 9. Call dropping probability for class 1 calls from tier-1 cell-1 under mixture pattern case-2.

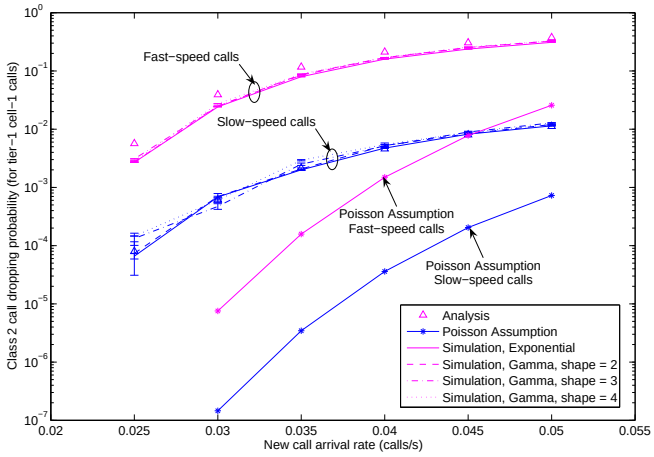

Fig. 10. Call dropping probability for class 2 calls from tier-1 cell-1 under mixture pattern case-2.