Performance Analysis of the WLAN-First Scheme in Cellular/WLAN Interworking

Wei Song, Hai Jiang, *Student Member, IEEE,* and Weihua Zhuang, *Senior Member, IEEE*

Abstract— In the interworking between a cellular network and wireless local area networks (WLANs), a two-tier overlaying structure exists in the WLAN-covered areas. Due to the heterogeneous underlying quality-of-service (QoS) support, the admission of traffic in these areas has a significant impact on QoS satisfaction and overall resource utilization, especially when multiple services are considered. In this paper, we analyze the performance of a simple admission strategy, referred to as *WLAN-first scheme*, in which incoming voice and data service requests always first try to get admission to the WLAN whenever it is available. It is observed that the overall resource utilization can be maximized when the admission regions for voice and data services in a cell and a WLAN are properly configured.

Index Terms— Cellular/WLAN interworking, resource sharing, call admission control.

I. INTRODUCTION

In the interworking between a cellular network and wireless local area networks (WLANs), there is a two-tier overlaying structure which offers both cellular and WLAN access to a dual-mode mobile station (MS) within the WLAN-covered area. Ubiquitous coverage is provided by the cellular network (higher-tier), while WLANs (lower-tier) are deployed in disjoint hot-spot areas. Then, there comes the admission strategy problem of how to properly admit incoming traffic to the cell or WLAN. Specially, a preferred target network, either a cell or a WLAN, should be first selected based on various decision criteria taking into account factors such as service type and network conditions of the two networks. A service request rejected by its first-choice network can just leave the system or further try to access the other network [1]. In addition, a proper resource sharing policy for multiple services is needed in the integrated network. Due to user mobility and traffic assignment in the overlaying area, the underlying network serving a user may alternate dynamically between the cellular network and WLANs. Therefore, instead of separately allocating the resources in each network, the overall resources of the two networks should be jointly considered for allocation.

There are some related researches on similar problems in two-tier hierarchical cellular networks, in which small-size microcells overlay with large macrocells. Many proposed admission strategies [1,2] are based on user mobility and traffic characteristics. However, these strategies are not effective or efficient for cellular/WLAN integrated networks for

reasons as follows. First, usually only one service type is considered in the strategies. As an essential requirement and important motivation for cellular/WLAN interworking, multi-service support is a challenging issue. Due to the resource sharing nature among multiple services, one service should be provided quality-of-service (QoS) guarantee and protection against the others. Second, the heterogeneous QoS provisioning capability in cellular/WLAN integrated networks further complicates the resource allocation for multiple services. The centralized control and reservation-based resource allocation in cellular networks enable fine-grained QoS provisioning to admitted traffic. In WLANs, the mandatory contention-based random access can only provide coarse-grained QoS in a complete resource sharing manner. Third, in cellular/WLAN interworking, it is not practical to allocate resources based on fast/slow user mobility differentiation [1,2] as in hierarchical cellular networks. This is because WLANs are usually deployed in indoor environments where users are static or only have pedestrian-level low mobility. Hence, resource allocation specifically tuned to cellular/WLAN interworking is needed to achieve desired performance.

A simple and easy-to-implement admission strategy for cellular/WLAN interworking is the *WLAN-first scheme*, where WLANs are always preferred by all services whenever the WLAN access is available, so as to take advantage of the low cost and large bandwidth of WLANs. An incoming service request rejected by a WLAN overflows to the cellular network to request admission if it is a new call, or remains in the cellular network if it is an ongoing call carried by the overlaying cell. Although the WLAN-first scheme is a straightforward approach, there is no in-depth analysis in the open literature on its performance in a practical cellular/WLAN interworking scenario. The analysis is very meaningful to obtain important insights on how various services affect the resource allocation and QoS support in a cellular/WLAN integrated network. This research is to contribute to the performance analysis of the WLAN-first scheme in cellular/WLAN interworking.

The remainder of this paper is organized as follows. Section II gives the system model of this research. Section III discusses the call admission policies and formulates the admission problem. The details of the performance analysis for the WLAN-first scheme are presented in Section IV. Section V discusses the numerical results and obtains important observations for the admission scheme. Finally, Section VI concludes the paper.

II. SYSTEM MODEL

Consider a cellular/WLAN integrated network with one overlaying WLAN in each cell as shown in Fig. 1. The area with only cellular access is referred to as *cellular-only area*,
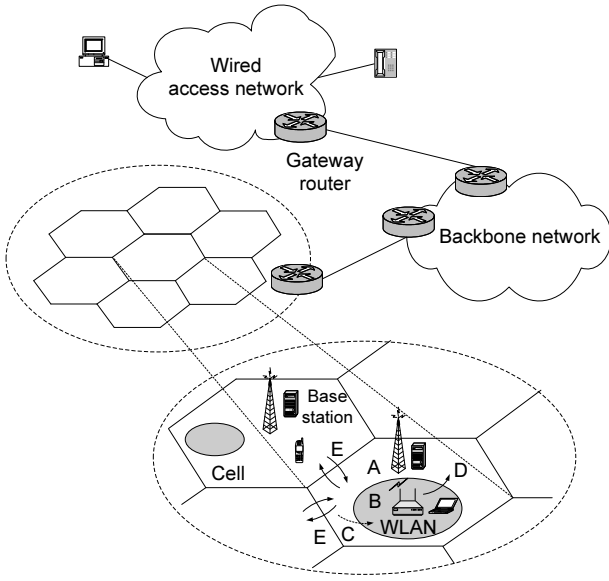
Fig. 1.   New and handoff traffic arrivals in a cellular/WLAN integrated network.

Fig. 2.   Modeling of user residence time in a cell.

while the area covered by both a cell and a WLAN is referred to as *double-coverage area*. Statistical equilibrium is assumed for the whole network. Thus, the analysis is focused on a single cell with an overlaying WLAN, referred to as a *cell cluster*. In this research, we consider real-time voice telephony and interactive data services (such as Web browsing). As WLANs are usually deployed in hot-spot areas, on average, the traffic density in the double-coverage area is higher than that in the cellular-only area. As illustrated in Fig. 1, in addition to new traffic (e.g., "A" and "B"), there are also horizontal handoffs between neighboring cells (e.g., "E") and vertical handoffs between a WLAN and its overlaying cell (e.g., "C" and "D").

As many symbols are used in this paper, we summarize the important ones in Table I. The superscripts "c" and "w" denote cell and WLAN, respectively; subscripts "v" and "d" denote voice and data, respectively; subscripts "n" and "h" denote new and handoff traffic, respectively.

A. Non-uniform Mobility within a Single Cell

WLANs are usually deployed in an indoor environment, where user mobility level is very low and may significantly differ from that of other areas. Hence, a homogeneous mobility model may not be applicable, and it is necessary to differentiate the user mobility characteristics in the double-coverage area from those in the cellular-only area. In the following analysis, a non-uniform model is used to characterize the user mobility within a cell cluster. Let T_r^{co} denote the residence time that a user stays within the cellular-only area before moving to neighboring cells with probability p^{c-c} or to the overlaying WLAN with probability p^{c-w}, and T_r^{dc} the user residence time in the double-coverage area. T_r^{co} and T_r^{dc} are assumed to be exponentially distributed with parameters η^{co} and η^{dc}, respectively. As shown in [3], for an MS with mean velocity of V and uniformly distributed movement direction over $[0, 2\pi]$, the average region boundary cross-over rate η is given by $\eta = V\frac{L}{\pi S}$, where L and S are the boundary
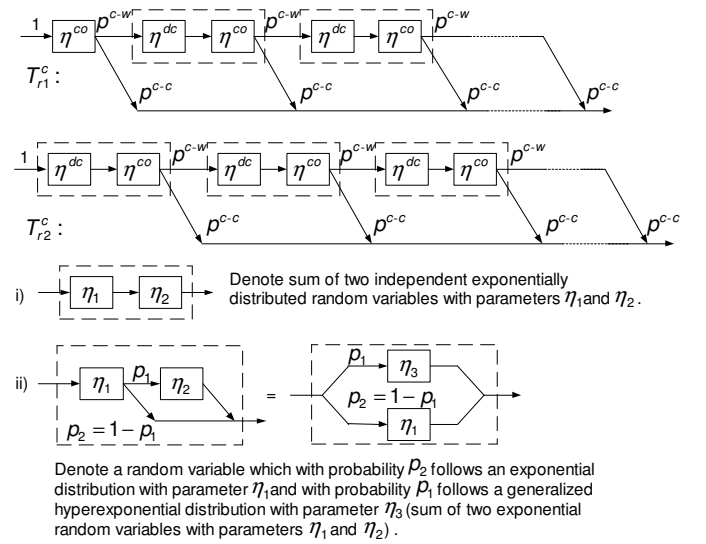
length and area of the region, respectively. Let Δ denote the ratio of the area of the WLAN to that of the cell, and V_{lh} the ratio of the average user moving velocity in the double-coverage area to that in the cellular-only area. Then, we have $\eta^{dc} = V_{lh}\frac{(1-\Delta)}{\Delta}\frac{1}{1+1/\sqrt{\Delta}}\eta^{co} = \frac{V_{lh}(1-\Delta)}{\Delta+\sqrt{\Delta}}\eta^{co}$.

For a new call initiated in the cellular-only area or a handoff call to the cellular-only area, if it is admitted into the cell, let T_{r1}^c denote its residence time within the cell. From the above non-uniform mobility model, it can be seen that T_{r1}^c follows a phase-type distribution shown in Fig. 2. The sum of the exponentially distributed T_r^{co} and T_r^{dc} (the part in the dashed rectangle) follows a generalized hyperexponential distribution with the probability density function (PDF) and moment generating function (MGF) given respectively by

$$f_{T_r^{co}+T_r^{dc}}(t) = \frac{\eta^{dc}}{\eta^{dc}-\eta^{co}}\eta^{co}e^{-\eta^{co}t} + \frac{\eta^{co}}{\eta^{co}-\eta^{dc}}\eta^{dc}e^{-\eta^{dc}t}$$

$$\psi(s) = \mathrm{E}[e^{s(T_r^{co}+T_r^{dc})}] = \frac{\eta^{co}}{\eta^{co}-s}\cdot\frac{\eta^{dc}}{\eta^{dc}-s}. \qquad (1)$$

Hence, the MGF of T_{r1}^c is obtained as

$$\Phi_1(s) = \sum_{i=1}^{\infty}(p^{c-w})^{i-1}p^{c-c}\frac{\eta^{co}}{\eta^{co}-s}[\psi(s)]^{i-1}. \qquad (2)$$

Similarly, for a new call initiated in the double-coverage area and admitted to the cell, its residence time within the cell, denoted by T_{r2}^c, is also modeled by a phase-type distribution shown in Fig. 2. The MGF of T_{r2}^c is

$$\Phi_2(s) = \sum_{i=1}^{\infty}(p^{c-w})^{i-1}p^{c-c}[\psi(s)]^i. \qquad (3)$$

The MGFs of T_{r1}^c and T_{r2}^c are used to obtain channel occupancy time of user traffic admitted to the cell and handoff probabilities from the cell to the overlaying WLAN or neighboring cells.

B. Traffic Model and QoS Requirements

For voice traffic, a constant bandwidth is required for each voice call to meet its strict delay requirement, while data

TABLE I

SUMMARY OF IMPORTANT SYMBOLS

Symbol	Definition
B_{v1}^c (B_{d1}^c)	Blocking probability of a cell for new voice (data) calls in cellular-only area
B_{v2}^c (B_{d2}^c)	Blocking probability of a cell for new voice (data) calls in double-coverage area
B_v^w (B_d^w)	Voice (data) call blocking probability of a WLAN
C^c	Cell bandwidth
D_v^c (D_d^c)	Dropping probability of a cell for handoff voice (data) calls from neiboring cells or overlaying WLAN
f_d	Mean data file size
G_{d1}^c (G_{d2}^c)	Randomized number of guard channels for handoff (new and handoff) data calls in cellular-only area
G_{v1}^c (G_{v2}^c)	Randomized number of guard channels for handoff (new and handoff) voice calls in cellular-only area
H_v^{c-c} (H_d^{c-c})	Handoff probability of voice (data) calls from cell to cell
H_v^{c-w} (H_d^{c-w})	Handoff probability of voice (data) calls from cell to overlaying WLAN
H_v^{w-c} (H_d^{w-c})	Handoff probability of voice (data) calls from WLAN to overlaying cell
N_v^c (N_d^c)	Maximum number of voice (data) calls allowed in a cell
N_v^w (N_d^w)	Maximum number of voice (data) calls allowed in a WLAN
p^{c-c} (p^{c-w})	Probability of a user moving out of current cell to neighboring cells (to overlaying WLAN)
T_d^c (T_d^w)	Data call duration if a data call is served by a cell (WLAN)
T_{r1}^c (T_{r2}^c)	Residence time within a cell of a call initiated in cellular-only (double-coverage) area and admitted into a cell
T_r^{co} (T_r^{dc})	User residence time in cellular-only (double-coverage) area of a cell with mean $(\eta^{co})^{-1}$ $((\eta^{dc})^{-1})$
V_{lh}	Ratio of average user moving velocity in double-coverage area to that in cellular-only area
λ_{d1} (λ_{d2})	Mean new data call arrival rate in cellular-only (double-coverage) area of a cell
λ_{hv}^{c-c} (λ_{hd}^{c-c})	Mean arrival rate of handoff voice (data) calls between neighboring cells
λ_{hv}^{c-w} (λ_{hd}^{c-w})	Mean arrival rate of handoff voice (data) calls from cell to overlaying WLAN
λ_{hv}^{w-c} (λ_{hd}^{w-c})	Mean arrival rate of handoff voice (data) calls from WLAN to overlaying cell
λ_{nv2}^c (λ_{nd2}^c)	Mean arrival rate of new voice (data) calls to a cell from double-coverage area
λ_{v1} (λ_{v2})	Mean new voice call arrival rate in cellular-only (double-coverage) area of a cell
λ_v^p	Mean arrival rate of voice packets from a voice flow
$(\mu_v)^{-1}$	Mean voice call duration
Δ	Ratio of the area of a WLAN to that of a cell
$\xi_v^w(\cdot)$ $(\xi_d^w(\cdot))$	Service rate of packets from one voice (data) flow in a WLAN
$\pi_v^c(\cdot)$ $(\pi_d^c(\cdot))$	Steady-state probability of the number of voice (data) calls in a cell
$\pi_v^w(\cdot)$ $(\pi_d^w(\cdot))$	Steady-state probability of the number of voice (data) calls in a WLAN
$\psi(\cdot)$, $\Phi_1(\cdot)$, $\Phi_2(\cdot)$	Moment generation functions of $T_r^{co}+T_r^{dc}$, T_{r1}^c, and T_{r2}^c, respectively

service is adaptive to elastic bandwidth. To facilitate analysis, all call arrivals are assumed to be Poisson. Voice call duration is exponentially distributed. If the download of a Web page or a data file is viewed as a packet data call, the data call duration (data transfer time) depends on the file size and actual occupied bandwidth. Considering the interactive nature of these data services, the mean data transfer time needs to be bounded within seconds. Because there is no close-form expression for the probability distribution of the data transfer time [4] when a processor sharing (PS) service discipline is applied (to be discussed in the next section), we consider an exponentially distributed data file size, in which case an upper bound for the mean data transfer time can be obtained [5]. The other call-level QoS requirements considered here are call blocking and dropping probabilities.

C. Resource Sharing between Voice and Data Services in a Cell or WLAN

In the cellular network, with the aid of base stations, the restricted access policy [6] can be applied. With this policy, voice is only allowed to occupy certain bandwidth, while the remaining bandwidth is dedicated to data. All the bandwidth unused by current voice traffic is shared equally by existing data calls. That is, a PS service discipline is applied to data traffic, and the total bandwidth occupied by data traffic dynamically varies with voice traffic. This policy is shown to achieve higher utilization than complete sharing and complete partitioning [7] and to offer each service certain QoS protection against the other. In WLANs, with contention-based random access, multiple services are supported in complete sharing. Admission control is necessary to limit the numbers of both voice and data calls in service. Otherwise, the intra-service interference from calls of the same service type or inter-service interference from calls of the other service type may severely degrade the system performance.

III. CALL ADMISSION POLICY

To apply admission control to the integrated network, we need to first analyze the capacity of each network for voice and data services. With centralized control and bandwidth reservation, the cell capacity is relatively easy to analyze, while the contention-based access and complete resource sharing in WLANs complicate the WLAN capacity analysis.

A. WLAN Capacity

In the original IEEE 802.11 standard, a per-node queue with per-node backoff is used for channel access contention and collision resolution. This per-node based principle penalizes heavily-loaded nodes with many flows (e.g., the access point). It is unfair and ineffective to guarantee the various QoS requirements of multimedia traffic. On the other hand, the MACAW [8] uses per-flow queue with per-flow backoff for channel contention. By this means, a node with multiple flows is viewed as multiple virtual nodes, each having one flow. A similar principle is adopted in IEEE 802.11e draft, where in each node there are multiple transmission queues for different access categories. Given the advantage of per-flow contention in multi-service support, we consider per-flow contention-based WLANs and extend the method in [9,10] to analyze the WLAN capacity.

Suppose there are n_v^w voice calls and n_d^w data calls admitted in a WLAN[1]. Packets from a voice flow are assumed to arrive with a constant rate, λ_v^p (frames/slot). For Web browsing, the data file to be transmitted is usually pre-stored in a server. Therefore, it is reasonable to consider that there is always traffic during the lifetime of a data call. Data transmission follows the request to send (RTS)-clear to send (CTS)-DATA-ACK handshaking for channel access, while voice flows do not use the RTS/CTS dialogue due to the small payload size of voice packets. Let $\xi_v^w(n_v^w, n_d^w)$ and $\xi_d^w(n_v^w, n_d^w)$ denote the service rates (frames/slot) for packets from one voice and data flow, respectively.

When a voice flow transmits in a slot, a collision will happen if any other voice or data flow transmits in the same slot. The collision probability is given by

$$p_v = 1 - \left[1 - \frac{\lambda_v^p}{\xi_v^w(n_v^w, n_d^w)}\tau_v\right]^{2n_v^w-1}(1 - \tau_d)^{n_d^w} \quad (4)$$

where τ_v and τ_d are the transmission probability of a voice flow and a data flow in a slot, given by (6) and (7) at the top of next page [11], respectively, with $W = \mathrm{CW}_{\min} + 1$ and $\mathrm{CW}_{\min}$ being the initial backoff window, which is 31 in IEEE 802.11, m the retransmission limit, m' the maximum backoff stage. Similarly, the collision probability for a data flow transmitting in a slot is

$$p_d = 1 - \left[1 - \frac{\lambda_v^p}{\xi_v^w(n_v^w, n_d^w)}\tau_v\right]^{2n_v^w}(1 - \tau_d)^{n_d^w-1}. \quad (5)$$

For data traffic, the time durations of a successful and collided transmission from a data flow are given by $T_{sd} = T_{DIFS}+T_{RTS}+T_{SIFS}+T_{CTS}+T_{SIFS}+T_{D\_DATA}+T_{SIFS}+T_{ACK}$ and $T_{cd} = T_{DIFS} + T_{RTS} + T_{CTS\_TO}$, respectively. The T_{D_DATA}, T_{RTS}, T_{CTS}, T_{ACK}, T_{DIFS}, T_{SIFS}, and T_{CTS_TO} are the transmission time of a DATA frame from a data flow, RTS frame duration, CTS frame duration, transmission time of an ACK frame, distributed coordination function (DCF) interframe space (DIFS), short interframe space (SIFS), and waiting time for a CTS TIMEOUT, respectively.

---

[1]Each voice call in the WLAN has two voice flows from and to the MS, while each data call has a one-way data flow to the MS.

On the other hand, the time needed for a successful transmission from a voice flow is $T_{sv} = T_{DIFS} + T_{V\_DATA} + T_{SIFS} + T_{ACK}$, with T_{V_DATA} being the transmission time of a voice DATA frame. The time for a collided transmission from a voice flow, denoted by T_{cv}, depends on the traffic types of the collided frames. There exist two scenarios: (1) a target voice frame collides with only voice frames, with probability $q_{vv} = (1 - \tau_d)^{n_d^w}\left[1 - (1 - \frac{\lambda_v^p}{\xi_v^w(n_v^w, n_d^w)}\tau_v)^{2n_v^w-1}\right]/p_v$; (2) a target voice frame collides with at least one data frame, with probability $q_{vd} = \left[1 - (1 - \tau_d)^{n_d^w}\right]/p_v$. We then have

$$T_{cv} = (T_{DIFS} + T_{V\_DATA} + T_{ACK\_TO}) \cdot q_{vv} + T_{cd} \cdot q_{vd}$$

where T_{ACK_TO} is the waiting time for an ACK TIMEOUT.

Based on an analytical method similar to that in [9,10], we further have

$$\frac{1}{\xi_v^w(n_v^w, n_d^w)} = \left[\frac{(2n_v^w - 1)\lambda_v^p}{\xi_v^w(n_v^w, n_d^w)} + 1\right]T_{sv} + n_d^w T_{sd} + \overline{W}_v(p_v)$$
$$+ \frac{1}{k}\left[\left(\frac{(2n_v^w - 1)\lambda_v^p}{\xi_v^w(n_v^w, n_d^w)} + 1\right)\overline{T}_{cv} + n_d^w \overline{T}_{cd}\right] \quad (8)$$

$$\frac{1}{\xi_d^w(n_v^w, n_d^w)} = 2n_v^w \frac{\lambda_v^p}{\xi_v^w(n_v^w, n_d^w)}T_{sv} + n_d^w T_{sd} + \overline{W}_d(p_d)$$
$$+ \frac{1}{k}\left[2n_v^w \frac{\lambda_v^p}{\xi_v^w(n_v^w, n_d^w)}\overline{T}_{cv} + n_d^w \overline{T}_{cd}\right] \quad (9)$$

where k is the average number of (voice or data) flows involved in a collision, $\overline{W}_v(p_v)$ is the average backoff time of a voice flow as a function of p_v, $\overline{W}_d(p_d)$ is the average backoff time of a data flow as a function of p_d, and $\overline{T}_{cv}$ ($\overline{T}_{cd}$) is the average collision time of a frame in a voice (data) flow, which can be obtained from T_{cv} and p_v (T_{cd} and p_d). Note that k is set to 1 in [9] and 2 in [10]. Actually, k can be evaluated more accurately as follows. At any time slot, there are approximately $\bar{n} = 2n_v^w \frac{\lambda_v^p}{\xi_v^w(n_v^w, n_d^w)} + n_d^w$ flows with backlogged traffic, and the transmission probability of a voice or data flow can be approximated by $\bar{\tau} = \frac{\tau_v + \tau_d}{2}$. Denote the probability that i flows transmit in a slot as $p(i) = \binom{\bar{n}}{i}\bar{\tau}^i(1 - \bar{\tau})^{(\bar{n}-i)}$. Then, the average number of (voice and data) flows involved in a collision is given by

$$k = \frac{\sum_{i=2}^{\bar{n}} i \cdot p(i)}{\sum_{i=2}^{\bar{n}} p(i)} = \frac{\bar{n}\bar{\tau} - p(1)}{1 - p(0) - p(1)}. \quad (10)$$

Moreover, to satisfy the real-time requirement of voice traffic, the service rate of a voice flow needs to be greater than the voice packet arrival rate. Thus, the following constraint should be met: $\xi_v^w(n_v^w, n_d^w) > (1 + \delta)\lambda_v^p$, where δ is a design parameter that can be determined experimentally. Based on (4), (5), (8), and (9), and the above constraint, we can get the capacity region, i.e., the feasible set of (n_v^w, n_d^w) vectors, and the corresponding data service rate $\xi_d^w(n_v^w, n_d^w)$ for each (n_v^w, n_d^w) vector in the capacity region.

B. Admission Regions for Voice and Data

Given the cell bandwidth C^c and total voice traffic load, the minimum bandwidth needed to meet the requirements of voice call blocking and dropping probabilities can be obtained

$$\tau_v = \frac{2(1-2p_v)(1-p_v^{m+1})}{W(1-(2p_v)^{m'+1})(1-p_v) + W(1-2p_v)2^{m'}(p_v^{m'+1} - p_v^{m+1}) + (1-2p_v)(1-p_v^{m+1})} \tag{6}$$

$$\tau_d = \frac{2(1-2p_d)(1-p_d^{m+1})}{W(1-(2p_d)^{m'+1})(1-p_d) + W(1-2p_d)2^{m'}(p_d^{m'+1} - p_d^{m+1}) + (1-2p_d)(1-p_d^{m+1})} \tag{7}$$
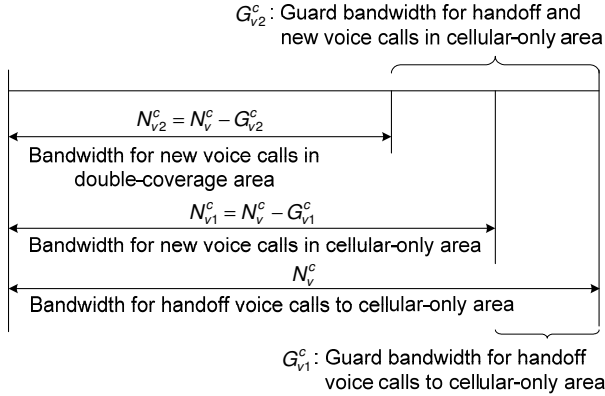

Fig. 3. Randomized guard channel policy for voice in the cell.

as $r_v \cdot N_v^c$, where r_v is the bandwidth requirement of a voice call and N_v^c ($\leq \lfloor \frac{C^c}{r_v} \rfloor$) is the maximum number of voice calls allowed in a cell. Moreover, because only cellular access is available in the cellular-only area, randomized guard channel policy is applied to give the new and handoff traffic in this area a priority to access the cell bandwidth over the traffic in the double-coverage area. Because the call blocking and dropping probabilities are very sensitive to the amount of reserved bandwidth, the guard bandwidth for high-priority voice traffic is randomized instead of an integer number of guard channels. As shown in Fig. 3, the voice admission region of the cell is given by $(N_v^c, G_{v1}^c, G_{v2}^c)$, in which G_{v2}^c ($\leq N_v^c$) is a real number to represent a randomized number of guard channels (guard bandwidth) dedicated to new and handoff voice traffic in the cellular-only area and G_{v1}^c ($\leq G_{v2}^c$) is the guard bandwidth reserved only for handoff voice traffic in this area. On the other hand, the remaining cell bandwidth ($C^c - r_v \cdot N_v^c$) is dedicated to data. All on-going data calls equally share the bandwidth unused by voice, and the data service rate is dynamically adjusted with call arrivals and departures. In addition, in a similar way to voice traffic discussed above, data traffic is also prioritized based on user location area and new/handoff call differentiation. The data admission region of the cell is given by $(N_d^c, G_{d1}^c, G_{d2}^c)$.

On the other hand, in Section III-A, we derive the WLAN capacity region in terms of (n_v^w, n_d^w) vectors. It is found that $\xi_d^w(n_v^w, n_d^w)$ decreases dramatically with a larger n_v^w, which implies the inefficient voice support of WLANs. To avoid the WLAN from operating in that inefficiency region, we limit the maximum number of voice calls and that of data calls admitted in the WLAN by N_v^w and N_d^w, respectively. A new or handoff voice (data) call is admitted to the WLAN when the number of existing voice (data) calls in the WLAN is less than N_v^w (N_d^w). The WLAN admission region (N_v^w, N_d^w) is chosen within the WLAN capacity region to guarantee packet-

level QoS satisfaction. Due to user mobility and the overlaying structure, the QoS performance is jointly determined by the cell and WLAN. Thus, given (N_v^w, N_d^w), based on the QoS requirements, we can derive the admission regions of the cell for voice and data accordingly. It can be seen in Section V that the configuration of admission regions does significantly affect the overall resource utilization. The best configuration assures a maximization of the acceptable traffic load with QoS satisfaction.

C. Admission Control Problem

With the two-tier overlying structure, the incoming traffic in the double-coverage area should be properly admitted to the cell or WLAN. Here, we consider the simple WLAN-first scheme to admit voice and data calls in the double-coverage area. New and handoff calls first try to get admission to the WLAN; blocked new calls overflow to the cell to request admission, while rejected vertical handoff calls remain carried by the cell. Due to different QoS support and resource sharing policies in the underlying networks, the configuration of admission regions of the cell and WLAN can have a significant impact on the overall system performance.

Let B_v^{req} (B_d^{req}), D_v^{req} (D_d^{req}), and T_d^{req} denote the requirements of new voice (data) call blocking and handoff voice (data) call dropping probabilities and mean data transfer time, respectively. Then, the admission control problem can be formulated as follows:

$$\max_{(N_v^w, N_d^w)} \lambda_d, \tag{11}$$

subject to : $\quad B_v^w \cdot B_{v2}^c \leq B_v^{req}, \quad B_{v1}^c \leq B_v^{req}, \quad D_v^c \leq D_v^{req}$

$B_d^w \cdot B_{d2}^c \leq B_d^{req}, \quad B_{d1}^c \leq B_d^{req}, \quad D_d^c \leq D_d^{req}, \quad \mathrm{E}[T_d] \leq T_d^{req}$

where λ_d is the mean data call arrival rate in the cell cluster, B_{v1}^c and B_{v2}^c (B_{d1}^c and B_{d2}^c) are the blocking probabilities of the cell for new voice (data) calls in the cellular-only area and double-coverage area, respectively, D_v^c is the voice handoff dropping probability of the cell, B_v^w (B_d^w) is the probability that a voice (data) call is blocked by the WLAN, and $\mathrm{E}[T_d]$ is the mean data transfer time. Here, we fix the voice call arrival rates for simplicity. Thus, the maximization of λ_d implies a maximization of the total acceptable traffic load and resource utilization.

IV. PERFORMANCE ANALYSIS OF THE WLAN-FIRST SCHEME

To obtain the solution to the problem (11), we use a searching algorithm given in Table II. As seen from steps 4 and 7, the QoS metrics in terms of call blocking/dropping probabilities and mean data transfer time should be evaluated accurately and effectively in each searching round. Due to the coupling

between the cell and WLAN, resource sharing between voice and data, and differentiation of new and handoff traffic in different areas, the analysis is very complex and multiple dimensions are involved. In Section IV-A and Section IV-B, we elaborate our QoS evaluation approach, which enables an efficient searching for the configuration of admission regions by applying proper decomposition and statistical averaging techniques.

TABLE II
SEARCHING ALGORITHM FOR ADMISSION REGIONS

1: (n_v^w, n_d^w) = GetWLANCapacity

   %Calculate WLAN capacity region in terms of a feasible set of (n_v^w, n_d^w) vectors (i.e., maximum numbers of voice and data calls simultaneously accommodated) and packet service rates for each (n_v^w, n_d^w) vector in the capacity region

2: $N_{v,max}^w = \max(n_v^w): (n_v^w, n_d^w) \in$ WLAN capacity region

3: **for** $N_v^w = 0, ..., N_{v,max}^w$ **do**

4:     By bisection searching, determine minimum N_v^c and (G_{v1}^c, G_{v2}^c) so that $B_v^w \cdot B_{v2}^c \leq B_v^{req}$, $B_{v1}^c \leq B_v^{req}$, and $D_v^c \leq D_v^{req}$

5:     Determine $N_{d,max}^w$ according to N_v^w and the (n_v^w, n_d^w) vector set

6:     **for** $N_d^w = 0, ..., N_{d,max}^w$ **do**

7:         By bisection searching, determine $(N_d^c, G_{d1}^c, G_{d2}^c)$ and the acceptable mean data call arrival rate λ_d which satisfy $B_d^w \cdot B_{d2}^c \leq B_d^{req}$, $B_{d1}^c \leq B_d^{req}$, $D_d^c \leq D_d^{req}$, and $\mathrm{E}[T_d] \leq T_d^{req}$

8:     **end for**

9: **end for**

10: Output $(N_v^w, N_d^w), (N_v^c, G_{v1}^c, G_{v2}^c)$, and $(N_d^c, G_{d1}^c, G_{d2}^c)$ which maximize the acceptable λ_d with QoS satisfaction

---

A. Analysis for Voice Traffic

Because a voice call duration is of the order of minutes, while a data call is required to finish transmission within seconds, the number of voice calls fluctuates much more slowly than that of data calls. No voice call arrival or departure is assumed during a data call duration. In particular, this limiting behavior for a Markov chain is referred to as *nearly complete decomposability* [4].

1) Voice Call Blocking and Dropping Probabilities: Let $(k_v^w, k_{v1}^c, k_{v2}^c)$ denote the state of voice traffic in a cell cluster, where k_v^w, k_{v1}^c, and k_{v2}^c are the numbers of voice calls admitted to the WLAN, to the cell from the cellular-only area, and to the cell from the double-coverage area, respectively. First, the number of voice calls in the WLAN can be described by a birth-death process with respect to k_v^w. Since both voice call duration T_v and user residence time T_r^{dc} in the double-coverage area are exponentially distributed, the channel occupancy time of voice calls in the WLAN, $\min(T_v, T_r^{dc})$, is exponential with mean $1/(\mu_v + \eta^{dc})$, where $1/\mu_v$ is the mean voice call duration. Then, the steady-state probability of k voice calls in the WLAN is obtained based on an $M/M/K/K$ loss system, given by

$$\pi_v^w(k) = \frac{[(\lambda_{v2} + \lambda_{hv}^{c-w})/(\mu_v + \eta^{dc})]^k/k!}{\sum_{i=0}^{N_v^w} \frac{[(\lambda_{v2}+\lambda_{hv}^{c-w})/(\mu_v+\eta^{dc})]^i}{i!}}, \qquad 0 \leq k \leq N_v^w$$

$$(12)$$

where λ_{v2} and λ_{hv}^{c-w} are the mean arrival rates of new and handoff voice calls to the WLAN, respectively. Thus, the voice call blocking probability in the WLAN is $B_v^w = \pi_v^w(N_v^w)$.

Next, we analyze the voice performance in a cell, which is more complex due to traffic prioritization. We draw in Fig. 4 the state transition diagram of (k_{v1}^c, k_{v2}^c), which is divided into several areas for illustration purpose and in each area only one example transition is shown with respect to k_{v1}^c and k_{v2}^c, respectively. Next, we derive the state-dependent transition rates in Fig. 4, which are given by

$(k_{v1}^c, k_{v2}^c) \rightarrow (k_{v1}^c + 1, k_{v2}^c):$
(1) $\lambda_{v1} + \lambda_{hv}^{c-c} + \lambda_{hv}^{w-c}$, if $k_{v1}^c \leq \lfloor N_v^c - G_{v1}^c \rfloor - 1$
(6) $\lambda_{v1}[1 - (G_{v1}^c - \lfloor G_{v1}^c \rfloor)] + \lambda_{hv}^{c-c} + \lambda_{hv}^{w-c}$,
    if $k_{v1}^c = \lfloor N_v^c - G_{v1}^c \rfloor$
(7) $\lambda_{hv}^{c-c} + \lambda_{hv}^{w-c}$, if $\lceil N_v^c - G_{v1}^c \rceil \leq k_{v1}^c \leq N_v^c - 1$

$$(13)$$

$(k_{v1}^c, k_{v2}^c) \rightarrow (k_{v1}^c, k_{v2}^c + 1):$
(3) λ_{v2}, if $k_{v2}^c \leq \lfloor N_v^c - G_{v2}^c \rfloor - 1$
(5) $\lambda_{v2}[1 - (G_{v2}^c - \lfloor G_{v2}^c \rfloor)]$, if $k_{v2}^c = \lfloor N_v^c - G_{v2}^c \rfloor$

$$(14)$$

$(k_{v1}^c, k_{v2}^c) \rightarrow (k_{v1}^c - 1, k_{v2}^c):$
(2) $k_{v1}^c \mu_v^1$, if $k_{v1}^c \geq 1$, $k_v^w = N_v^w$
(2) $k_{v1}^c \cdot (\mu_v^c + \eta^{co})$, if $k_{v1}^c \geq 1, k_v^w \leq N_v^w - 1$

$$(15)$$

$(k_{v1}^c, k_{v2}^c) \rightarrow (k_{v1}^c, k_{v2}^c - 1):$
(4) $k_{v2}^c \mu_v^2$, if $k_{v2}^c \geq 1$, $k_v^w = N_v^w$
(4) $k_{v2}^c \frac{\mu_v}{1 - \psi(-\mu_v)}$, if $k_{v2}^c \geq 1, k_v^w \leq N_v^w - 1$ (16)

where λ_{v1} is the mean arrival rate of new voice calls in the cellular-only area, λ_{hv}^{w-c} and λ_{hv}^{c-c} are those of handoff voice calls to the cell from neighboring cells and from the WLAN, respectively.
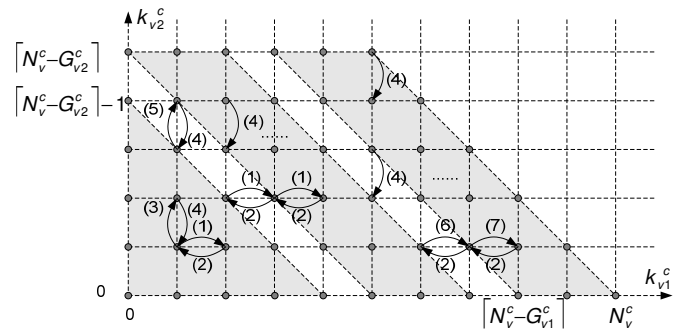

Fig. 4. State transition diagram for voice in the cell.

In general, suppose X and Y are two independent random variables with $X \sim \exp(\lambda)$. Then,

$$\mathrm{P}[X > Y] = \int_0^\infty f_Y(y) \int_y^\infty \lambda e^{-\lambda x} dx dy = \Psi_Y(-\lambda) \quad (17)$$

where $f_Y(\cdot)$ and $\Psi_Y(\cdot)$ are the PDF and MGF of Y, respectively. Letting $Z = \min(X, Y)$, the PDF of Z is given by

$f_Z(z) = f_X(z)(1 - F_Y(z)) + f_Y(z)(1 - F_X(z))$, where $f_X(\cdot)$, $F_X(\cdot)$ and $F_Y(\cdot)$ denote the PDF and cumulative probability function (CDF) of X, and the CDF of Y, respectively. Then, the mean value of Z is

$$\mathrm{E}[Z] = \mathrm{E}[X] - \int_0^\infty f_Y(y)\frac{1}{\lambda}e^{-\lambda y}dy = \frac{1}{\lambda} - \frac{1}{\lambda}\Psi_Y(-\lambda). \quad (18)$$

When there is not enough free capacity in the WLAN for an arriving voice call, the channel occupancy time of new and handoff voice calls admitted to the cell from the cellular-only area is $\min(T_v, T_{r1}^c)$. Based on (2) and (18), its mean value can be derived and is given by

$$\mathrm{E}[\min(T_v, T_{r1}^c)]$$
$$= \frac{1}{\mu_v} - \frac{1}{\mu_v}\sum_{i=1}^\infty (p^{c-w})^{i-1}p^{c-c}\frac{\eta^{co}}{\eta^{co}+\mu_v}(C_p)^{i-1}$$
$$= \frac{1}{\mu_v} - \frac{1}{\mu_v}p^{c-c}\frac{\eta^{co}/(\eta^{co}+\mu_v)}{1-p^{c-w}C_p} \triangleq \frac{1}{\mu_{v1}^c} \quad (19)$$

where

$$C_p = \frac{\eta^{dc}\eta^{co}}{\eta^{dc}-\eta^{co}}\left(\frac{1}{\eta^{co}+\mu_v} - \frac{1}{\eta^{dc}+\mu_v}\right).$$

Similarly, the channel occupancy time of new voice calls admitted to the cell from the double-coverage area is $\min(T_v, T_{r2}^c)$ with mean value

$$\mathrm{E}[\min(T_v, T_{r2}^c)] = \frac{1}{\mu_v} - \frac{1}{\mu_v}\sum_{i=1}^\infty (p^{c-w})^{i-1}p^{c-c}(C_p)^i$$
$$= \frac{1}{\mu_v} - \frac{1}{\mu_v}p^{c-c}\frac{C_p}{1-p^{c-w}C_p} \triangleq \frac{1}{\mu_{v2}^c}. \quad (20)$$

According to our admission policy in the WLAN, $B_v^w = \mathrm{P}[k_v^w < N_v^w]$. Depending on the number of existing voice calls in the WLAN (k_v^w), we approximate the departure rate from state (k_{v1}^c, k_{v2}^c) to state $(k_{v1}^c - 1, k_{v2}^c)$ ($k_{v1}^c \geq 1$) and that from state (k_{v1}^c, k_{v2}^c) to state $(k_{v1}^c, k_{v2}^c - 1)$ ($k_{v2}^c \geq 1$) by $k_{v1}^c \cdot \tilde{\mu}_{v1}^c$ and $k_{v2}^c \cdot \tilde{\mu}_{v2}^c$, respectively, where $\tilde{\mu}_{v1}^c$ and $\tilde{\mu}_{v2}^c$ are given by

$$\tilde{\mu}_{v1}^c = B_v^w\mu_{v1}^c + (1 - B_v^w)(\mu_v + \eta^{co}) \quad (21)$$
$$\tilde{\mu}_{v2}^c = B_v^w\mu_{v2}^c + (1 - B_v^w)\frac{\mu_v}{1-\psi(-\mu_v)}. \quad (22)$$

As indicated by (21) and (22), voice traffic admitted to the cell from the cellular-only area and double-coverage area has different mean channel occupancy time approximated by $(\tilde{\mu}_{v1}^c)^{-1}$ and $(\tilde{\mu}_{v2}^c)^{-1}$, respectively. Hence, the cell can be viewed as a multi-service loss system [12]. A product-form state distribution exists and is insensitive to service time distributions, provided that the resource sharing among services is under coordinate convex policies. This requires that transitions between states come in pairs. For loss systems with trunk reservation (e.g., the guard channel policy), the insensitivity property and product-form solution are destroyed due to the one-way transitions at some states. A recursive method is proposed in [13] to approximate the state distribution, which is shown to be accurate for a wide range of traffic intensities and when the service rates (such as $\tilde{\mu}_{v1}^c$ and $\tilde{\mu}_{v2}^c$) do not greatly differ from each other. Moreover, the blocking probabilities

are *almost* insensitive to service time distributions. Hence, we use the recursive approximation in [13] to obtain $\pi_v^c(k)$ (i.e., the steady-state probability of k voice calls admitted into the cell) given by (23) at the top of next page[2], in which

$$N_{v1}^c = N_v^c - G_{v1}^c, \quad N_{v2}^c = N_v^c - G_{v2}^c$$
$$\lambda_{hv}^c = \lambda_{hv}^{w-c} + \lambda_{hv}^{c-c}, \quad \lambda_{v1}^c = \lambda_{v1} + \lambda_{hv}^c, \quad \lambda_{nv2}^c = B_v^w\lambda_{v2}$$
$$\rho_{v1} = \frac{(1-(G_{v1}^c - \lfloor G_{v1}^c\rfloor))\lambda_{v1} + \lambda_{hv}^c}{\tilde{\mu}_{v1}^c}$$
$$\rho_{v2} = \frac{\lambda_{v1}^c}{\tilde{\mu}_{v1}^c} + \frac{(1-(G_{v2}^c - \lfloor G_{v2}^c\rfloor))\lambda_{nv2}^c}{\tilde{\mu}_{v2}^c}.$$

Thus, the voice call blocking and dropping probabilities of the cell are given by

$$B_{v2}^c = (G_{v2}^c - \lfloor G_{v2}^c\rfloor)\,\pi_v^c(\lfloor N_{v2}^c\rfloor) + \sum_{i=\lfloor N_{v2}^c\rfloor+1}^{N_v^c}\pi_v^c(i) \quad (24)$$

$$B_{v1}^c = (G_{v1}^c - \lfloor G_{v1}^c\rfloor)\,\pi_v^c(\lfloor N_{v1}^c\rfloor) + \sum_{i=\lfloor N_{v1}^c\rfloor+1}^{N_v^c}\pi_v^c(i) \quad (25)$$

$$D_v^c = \pi_v^c(N_v^c). \quad (26)$$

2) Mean Arrival Rates of Handoff Voice Calls: It can be seen from (23)-(26) that handoff arrival rates are needed to obtain the voice call blocking and dropping probabilities. The handoff arrival rates are related to the handoff probabilities, which are the probabilities that at least one more handoff is required before the call completion. The handoff probability of voice calls in the cellular-only area to neighboring cells, denoted by H_v^{c-c}, can be obtained according to (17) as $H_v^{c-c} = p^{c-c}\mathrm{P}[T_v > T_{r1}^c] = p^{c-c}\Phi_1(-\mu_v)$. Similarly, the handoff probability of voice calls in the cellular-only area to the overlaying WLAN, denoted by H_v^{c-w}, is given by $H_v^{c-w} = p^{c-w}\Phi_1(-\mu_v)$.

With an exponentially distributed user residence time in the double-coverage area, the handoff probability of voice calls from the WLAN to the overlaying cell is $H_v^{w-c} = \eta^{dc}/(\eta^{dc}+\mu_v)$. Therefore, the handoff traffic from the WLAN to the overlaying cell has a mean arrival rate λ_{hv}^{w-c} given by

$$\lambda_{hv}^{w-c} = H_v^{w-c}(\lambda_{v2} + \lambda_{hv}^{c-w})(1 - B_v^w). \quad (27)$$

Similarly, the mean arrival rates of handoff traffic between neighboring cells and from the cell to the overlaying WLAN can be respectively obtained as

$$\lambda_{hv}^{c-c} = H_v^{c-c}\Big[\lambda_{v1}(1 - B_{v1}^c) + (\lambda_{hv}^{w-c} + \lambda_{hv}^{c-c})(1 - D_v^c)$$
$$+ \lambda_{nv2}^c(1 - B_{v2}^c)H_v^{w-c}\Big] \quad (28)$$
$$\lambda_{hv}^{c-w} = H_v^{c-w}\Big[\lambda_{v1}(1 - B_{v1}^c) + (\lambda_{hv}^{w-c} + \lambda_{hv}^{c-c})(1 - D_v^c)$$
$$+ \lambda_{nv2}^c(1 - B_{v2}^c)H_v^{w-c}\Big]. \quad (29)$$

Thus, voice call blocking and dropping probabilities can be obtained recursively from (12), (23), and (27)-(29).

---

[2] The expression is given under the condition that $\lfloor G_{v1}^c\rfloor \leq \lfloor G_{v2}^c\rfloor - 1$ and $\lfloor G_{v1}^c\rfloor \geq 1$. When $\lfloor G_{v1}^c\rfloor = 0$ or $\lfloor G_{v1}^c\rfloor = \lfloor G_{v2}^c\rfloor$, the expression can be adjusted accordingly based on the recursive method in [13].

$$\pi_v^c(k) = \begin{cases} (\frac{\lambda_{v1}^c}{\tilde{\mu}_{v1}^c} + \frac{\lambda_{nv2}^c}{\tilde{\mu}_{v2}^c})^k \frac{\pi_v^c(0)}{k!}, & 0 \le k \le \lfloor N_{v2}^c \rfloor \\ (\frac{\lambda_{v1}^c}{\tilde{\mu}_{v1}^c} + \frac{\lambda_{nv2}^c}{\tilde{\mu}_{v2}^c})^{\lfloor N_{v2}^c \rfloor} \rho_{v2} (\frac{\lambda_{v1}^c}{\tilde{\mu}_{v1}^c})^{(k-\lfloor N_{v2}^c \rfloor - 1)} \frac{\pi_v^c(0)}{k!}, & \lfloor N_{v2}^c \rfloor + 1 \le k \le \lfloor N_{v1}^c \rfloor \\ (\frac{\lambda_{v1}^c}{\tilde{\mu}_{v1}^c} + \frac{\lambda_{nv2}^c}{\tilde{\mu}_{v2}^c})^{\lfloor N_{v2}^c \rfloor} \rho_{v2} (\frac{\lambda_{v1}^c}{\tilde{\mu}_{v1}^c})^{(\lfloor N_{v1}^c \rfloor - \lfloor N_{v2}^c \rfloor - 1)} \rho_{v1} (\frac{\lambda_{hv}^c}{\tilde{\mu}_{v1}^c})^{(k-\lfloor N_{v1}^c \rfloor - 1)} \frac{\pi_v^c(0)}{k!}, & \lfloor N_{v1}^c \rfloor + 1 \le k \le N_v^c \end{cases} \quad (23)$$

B. Analysis for Data Traffic

Under the limiting case that the time scale of voice calls is much larger than that of data calls, the analysis for data traffic can be approximately decoupled from that of voice [4].

1) Mean Data Transfer Time and Handoff Arrival Rates: First, we consider the performance of data service in the cell. Since all the bandwidth unused by voice traffic is shared equally by existing data calls, a cell behaves like an $M/G/1/K - PS$ queue, whose service capacity is $(C^c - i \cdot r_v)$ with a probability $\pi_v^c(i)$, $i = 0, 1, ..., N_v^c$. Thus, the expected duration of a data call carried by the cell is approximated by [5]

$$\mathrm{E}[T_d^c] = \sum_{i=0}^{N_v^c} \pi_v^c(i) \frac{\rho_d^c(i)^{(N_d^c+1)} \left[N_d^c \rho_d^c(i) - N_d^c - 1 \right] + \rho_d^c(i)}{\lambda_d^c \left[1 - \rho_d^c(i)^{N_d^c} \right] \left[1 - \rho_d^c(i) \right]} \tag{30}$$

where

$$\lambda_d^c = \lambda_{d1} + \lambda_{nd2}^c + \lambda_{hd}^{w-c} + \lambda_{hd}^{c-c}, \quad \rho_d^c(i) = \frac{\lambda_d^c \cdot f_d}{C^c - i \cdot r_v}$$

where λ_{d1} is the mean arrival rate of new data calls in the cellular-only area, with λ_{d2} being that of the double-coverage area, the part overflowed to the cell has a mean rate $\lambda_{nd2}^c = B_d^w \lambda_{d2}$, λ_{hd}^{w-c} and λ_{hd}^{c-c} are the mean arrival rates of handoff data calls to the cell from neighboring cells and from the WLAN, respectively, and f_d is the mean data file size. Given in (30) is actually an upper bound for the mean transfer time of a data call with exponentially distributed size [5]. When data traffic evolves rapidly with respect to voice traffic, the number of data calls can attain its stationary regime given by the $M/G/1/K - PS$ queue. In this case, the upper bound can be used to approximate the mean data transfer time.

Because a data call may be carried by different cells and/or WLANs during its lifetime, its overall performance depends on both networks. Next, we analyze the expected data call duration when a data call is carried by the WLAN. In the WLAN, the data service rate is state-dependent due to the complete resource sharing between voice and data traffic. The probability of j data calls carried by the WLAN is given by

$$\tilde{\pi}_d^w(j) = \sum_{i=0}^{N_v^w} \left[\pi_v^w(i) \tilde{\pi}_d^w(0) \frac{(\lambda_d^w)^j}{\prod_{l=1}^j l \cdot \chi_d^w(i,l)} \right], j = 1, 2, ..., N_d^w$$

where

$$\lambda_d^w = \lambda_{d2} + \lambda_{hd}^{c-w}, \quad \chi_d^w(i,l) = \frac{\xi_d^w(i,l)}{f_d}$$

λ_{hd}^{c-w} is the mean arrival rate of handoff data calls to the WLAN, $\chi_d^w(i,l)$ is the service rate for one data call with i voice calls and l data calls in the WLAN, $\xi_d^w(i)$ is given by

(9), $\pi_v^w(i)$ is given by (12). Using the Little's law, the expected duration of a data call carried by the WLAN is obtained as

$$\mathrm{E}[T_d^w] = \frac{1}{\lambda_d^w(1 - B_d^w)} \sum_{j=0}^{N_d^w} j \tilde{\pi}_d^w(j) \tag{31}$$

where B_d^w is the data call blocking probability of the WLAN.

As seen from (30)-(31), the mean data transfer time depends on the mean arrival rates and blocking/dropping probabilities of data calls, which are inter-dependent and need to be evaluated recursively. The derivation of data call blocking and dropping probabilities is given in Section IV-B.2. The mean arrival rates of handoff data calls can be obtained similarly to those of handoff voice calls except for a different call duration. Due to the PS service manner for data traffic, it is very complex to derive and apply a precise statistic model for the data call duration. For tractability, the duration of data calls when carried by a cell, T_d^c, is assumed to be exponential with expectation $\mathrm{E}[T_d^c]$. The data call hand-off probability from the cellular-only area to neighboring cells, H_d^{c-c}, can be obtained according to (17) as $H_d^{c-c} = p^{c-c} \Phi_1(-1/\mathrm{E}[T_d^c])$. The data call handoff probability from the cellular-only area to the overlaying WLAN H_d^{c-w} is given by $H_d^{c-w} = p^{c-w} \Phi_1(-1/\mathrm{E}[T_d^c])$. Then, λ_{hd}^{c-c} and λ_{hd}^{c-w} can be derived similarly to (28) and (29). Moreover, under the assumption that T_d^w is exponential, the data call handoff probability from the WLAN to the overlaying cell is given by $H_d^{w-c} = \frac{1/\mathrm{E}[T_d^w]}{1/\mathrm{E}[T_d^w] + \eta^{dc}}$. Thus, the mean rate of handoff data calls out of the WLAN is $\lambda_{hd}^{w-c} = \lambda_d^w(1 - B_d^w)H_d^{w-c}$.

2) Blocking and Dropping Probabilities of Data Calls: Consider the state that there are i voice calls and j data carried by a cell. For new and handoff data calls admitted to the cell from the cellular-only area, by averaging over the WLAN state, we approximate the departure rate from state (i,j) to state $(i, j-1)$ $(j \ge 1)$ by $j \cdot \tilde{\mu}_{d1}^c(i,j)$, where

$$\tilde{\mu}_{d1}^c(i,j) = B_d^w \mu_{d1}^c(i,j) + (1 - B_d^w) \left[\nu_d^c(i,j) + \eta^{co} \right] \tag{32}$$

$\nu_d^c(i,j) = \frac{C^c - i \cdot r_v}{j \cdot f_d}$ and $\mu_{d1}^c(i,j)$ is the inverse of mean cell bandwidth occupancy time of data calls when there is not enough free capacity in the WLAN, which can be obtained from (18) as

$$\mu_{d1}^c(i,j) = \frac{\frac{C^c - i \cdot r_v}{j \cdot f_d}}{1 - \Phi_1(-\frac{C^c - i \cdot r_v}{j \cdot f_d})} = \frac{\nu_d^c(i,j)}{1 - \Phi_1(-\nu_d^c(i,j))}. \tag{33}$$

Similarly, for data calls admitted to the cell from the double-coverage area, the departure rate from state (i,j) to state $(i, j-1)$ $(j \ge 1)$ is $j \cdot \tilde{\mu}_{d2}^c(i,j)$, where

$$\tilde{\mu}_{d2}^c(i,j) = B_d^w \mu_{d2}^c(i,j) + (1 - B_d^w) \frac{\nu_d^c(i,j)}{1 - \psi(-\nu_d^c(i,j))} \tag{34}$$

$$\mu_{d2}^c(i,j) = \frac{\frac{C^c - i \cdot r_v}{j \cdot f_d}}{1 - \Phi_2(-\frac{C^c - i \cdot r_v}{j \cdot f_d})} = \frac{\nu_d^c(i,j)}{1 - \Phi_2(-\nu_d^c(i,j))}. \quad (35)$$

Considering the two-tier overlaying structure in cellular/WLAN interworking, new and handoff data calls in the cellular-only area are prioritized by bandwidth reservation with the randomized guard channel policy. In this case, we use the average departure rate $j \cdot \tilde{\mu}_d^c(i,j)$ to simplify analysis, where

$$\tilde{\mu}_d^c(i,j) = p_{d1}^c(j)\tilde{\mu}_{d1}^c(i,j) + p_{d2}^c(j)\tilde{\mu}_{d2}^c(i,j) \quad (36)$$

$p_{d1}^c(\cdot)$ and $p_{d2}^c(\cdot)$ are respectively the fractions of traffic requesting admission to the cell from the cellular-only area and from the double-coverage area, given by

$$p_{d1}^c(j) = \frac{\lambda_{d1}^c(j)}{\lambda_{d1}^c(j) + \lambda_{d2}^c(j)}, \qquad p_{d2}^c(j) = \frac{\lambda_{d2}^c(j)}{\lambda_{d1}^c(j) + \lambda_{d2}^c(j)}$$

$$\lambda_{d1}^c(j) = \begin{cases} \lambda_{d1} + \lambda_{hd}^{c-c} + \lambda_{hd}^{w-c}, & j \le \lfloor N_{d1}^c \rfloor - 1 \\ \lambda_{d1}\gamma_{d1}^c + \lambda_{hd}^{c-c} + \lambda_{hd}^{w-c}, & j = \lfloor N_{d1}^c \rfloor \\ \lambda_{hd}^{c-c} + \lambda_{hd}^{w-c}, & \lfloor N_{d1}^c \rfloor + 1 \le j \le N_d^c \end{cases}$$

$$\lambda_{d2}^c(j) = \begin{cases} \lambda_{nd2}^c, & j \le \lfloor N_{d2}^c \rfloor - 1 \\ \lambda_{nd2}^c\gamma_{d2}^c, & j = \lfloor N_{d2}^c \rfloor \\ 0, & \lfloor N_{d2}^c \rfloor + 1 \le j \le N_d^c \end{cases}$$

$$N_{d1}^c = N_d^c - G_{d1}^c, \quad N_{d2}^c = N_d^c - G_{d2}^c$$
$$\gamma_{d1}^c = N_{d1}^c - \lfloor N_{d1}^c \rfloor, \quad \gamma_{d2}^c = N_{d2}^c - \lfloor N_{d2}^c \rfloor.$$

Under the assumption of nearly complete decomposition of data traffic from voice, when there are i voice calls carried by the cell, the cell operates like a symmetric queue [14] for data with

$$\phi(j) = j \cdot \tilde{\mu}_d^c(i,j), \qquad \gamma(l,j) = \delta(l,j) = \frac{1}{j} \quad (37)$$
$$l = 1, 2, ..., j, \qquad j = 1, 2, ..., N_d^c$$

where $\phi(j)$ ($\phi(j) > 0$ if $j > 0$) is the total service rate when there are j customers (data calls) in the queue in positions $1, 2, ..., j$; $\gamma(l,j)$ is the fraction of the service rate directed to the customer in position l ($\sum_{l=1}^{j}\gamma(l,j) = 1$); $\delta(l,j+1) = \gamma(l,j+1)$ (symmetric condition) is the probability that an arriving customer moves into position l. A data call carried by the cell may depart due to a handoff to another cell or WLAN. This departure is independent of the queuing position of the data call and behaves like a multi-server loss system without waiting room. In addition, a data call may also depart from the cell due to call completion. Since all the bandwidth unused by current voice calls is shared equally by existing data calls in a PS manner, a fair share of the total service rate is dedicated to each data call irrelevant to its queueing position. Thus, a data call completion or arrival affects the amount of resources allocated to each data call, but each data call still keeps a fair share. Therefore, $\delta(l,j)$ and $\gamma(l,j)$ are independent of the queueing positions (i.e., l) of data calls and satisfy the symmetric condition. As a result, for data service, the cell can be modeled by a symmetric queue, which operates in a manner given by (37) and has a service capacity ($C^c - i \cdot r_v$) with a probability $\pi_v^c(i)$, $i = 0, 1, ..., N_v^c$.

For symmetric queues such as processor-sharing queues and multi-server queues without waiting room (i.e., loss systems), a product-form stationary queue occupancy distribution exists and is applicable to arbitrarily distributed service requirements [14]. Then, the steady-state probability of j data calls in the cell is approximately given by

$$\pi_d^c(j) = \sum_{i=0}^{N_v^c}\left[\pi_v^c(i)\pi_d^c(0)\prod_{l=1}^{j}\frac{\lambda_{d1}^c(l) + \lambda_{d2}^c(l)}{\phi(l)}\right] \quad (38)$$
$$= \sum_{i=0}^{N_v^c}\left[\pi_v^c(i)\pi_d^c(0)\prod_{l=1}^{j}\frac{\lambda_{d1}^c(l) + \lambda_{d2}^c(l)}{l \cdot \tilde{\mu}_d^c(i,l)}\right], \quad j = 1, 2, ..., N_d^c.$$

Then, the data call blocking and dropping probabilities in the cell can be obtained by replacing N_v^c, G_{v1}^c, G_{v2}^c, N_{v1}^c, N_{v2}^c and π_v^c in (24)-(26) with N_d^c, G_{d1}^c, G_{d2}^c, N_{d1}^c, N_{d2}^c and π_d^c, respectively.

On the other hand, the departure rate of data calls with i voice calls and j data calls in the WLAN is $j \cdot (\chi_d^w(i,j) + \eta^{dc})$. Similar to the derivation of $\pi_d^c(j)$, the steady-state probability of j data calls carried by the WLAN is given by

$$\pi_d^w(j) = \sum_{i=0}^{N_v^w}\left[\pi_v^w(i)\pi_d^w(0)\prod_{l=1}^{j}\frac{\lambda_d^w}{l \cdot (\chi_d^w(i,l) + \eta^{dc})}\right] \quad (39)$$
$$j = 1, 2, ..., N_d^w.$$

The data call blocking probability of the WLAN is then obtained as $B_d^w = \pi_d^w(N_d^w)$.

V. NUMERICAL RESULTS AND DISCUSSION

In the following, we discuss some important observations obtained from the numerical results of the searching process. Given in Table III are the analysis parameters, which are selected based on popularly deployed cellular networks (e.g., cdma2000) and WLAN standards (e.g., IEEE 802.11b).

TABLE III
SYSTEM PARAMETERS

Parameter	Value	Parameter	Value
λ_{v1}	0.12 calls/s	λ_{v2}	0.18 calls/s
$(\mu_v)^{-1}$	140 s	r_v	14.4 Kbps
C^c	2 Mbps	B_v^{req}	0.01
B_d^{req}	0.01	D_v^{req} (D_d^{req})	0.001
f_d	64K bytes	T_d^{req}	4 s
Δ	0.1	V_{lh}	0.6
$(\eta^{co})^{-1}$	10 min	$(\eta^{dc})^{-1}$	14 min
p^{c-c}	0.76	p^{c-w}	0.24
m'	5	m	7
W	32	slot (WLAN)	20 μs
T_{SIFS}	10 μs	T_{DIFS}	50 μs
T_{RTS}	13.6 slots	T_{CTS}	12.4 slots
T_{ACK}	10.2 slots	λ_v^p	0.0004 frames/slot
Voice payload	50 bytes/packet	Data payload	1000 bytes/packet

A. Variation of Acceptable Data Traffic Load with N_d^w

Fig. 5(a) and Fig. 5(b) illustrate how the acceptable data traffic load (mean data call arrival rate λ_d) varies with the maximum number of data calls allowed in the WLAN (N_d^w) when the maximum number of voice calls allowed in the WLAN (N_v^w) is fixed to different values. It is observed from Fig. 5(a) that the acceptable data traffic load increases with N_d^w when N_d^w is relatively small. This can be explained as follows. With the coupling between the cell and its overlaying WLAN, both the time that a data call is carried by WLANs and by cells contributes to the total transfer time of a data call. With the PS sharing manner for data traffic, the fewer data calls admitted, the faster they will leave the system as a larger bandwidth is available for each data call. When fewer data calls are allowed in the WLAN by choosing a smaller N_d^w, more data calls need to be accommodated by the cell. Since the cell bandwidth is much lower than the WLAN bandwidth, the increase of data transfer time in cells cannot be compensated by the reduction of data transfer time when a data call is carried by WLANs. Hence, the mean data transfer time is longer (shorter) with a decrease (increase) of N_d^w, which results in a smaller (larger) data traffic load that can be supported.

As illustrated in Fig. 5(b), the increase of data traffic load with N_d^w becomes unnoticeable when N_d^w is large (say, more than 10). Indeed, when more data traffic is assigned to the WLAN, the transfer time of data calls in the cell is reduced. However, the reduction is almost balanced by the increase of data transfer time in the WLAN because a larger number of data calls share the WLAN bandwidth. As a result, the acceptable data traffic load is almost the same with large values of N_d^w. On the other hand, with a very large value of N_v^w (e.g., 20), the acceptable data traffic load even decreases negligibly with an increase of N_d^w. This is due to the severe drop of data service rate when N_v^w approaches the WLAN capacity for voice (i.e., the maximum number of voice calls that can be carried with the total WLAN bandwidth).

B. Variation of Acceptable Data Traffic Load with N_v^w

For each curve in Fig. 5, there is a maximum data call arrival rate achieved with a certain value of N_d^w. From these curves, we can obtain Fig. 6, which shows the relationship between the above maximum acceptable data traffic load and the maximum number of voice calls allowed in the WLAN (N_v^w). It is observed that there exists a value of N_v^w (i.e., 8 in the example) which maximizes the above acceptable data traffic load. With this configuration, N_v^w is less than the WLAN capacity for voice service, which is 28 in this example. That is, voice traffic in the double-coverage area should be restricted not to occupy all the WLAN bandwidth. This results from the cellular/WLAN coupling and voice/data resource sharing. First, since a larger value of N_v^w indicates that more voice traffic in the double-coverage area is assigned to the WLAN and relieved from the cell, more cell bandwidth can be used for data traffic in the cellular-only area, and the overall data transfer time is reduced (load balancing effect). This leads to a larger acceptable data traffic load. Second, when N_v^w is further increased to approach the WLAN capacity,
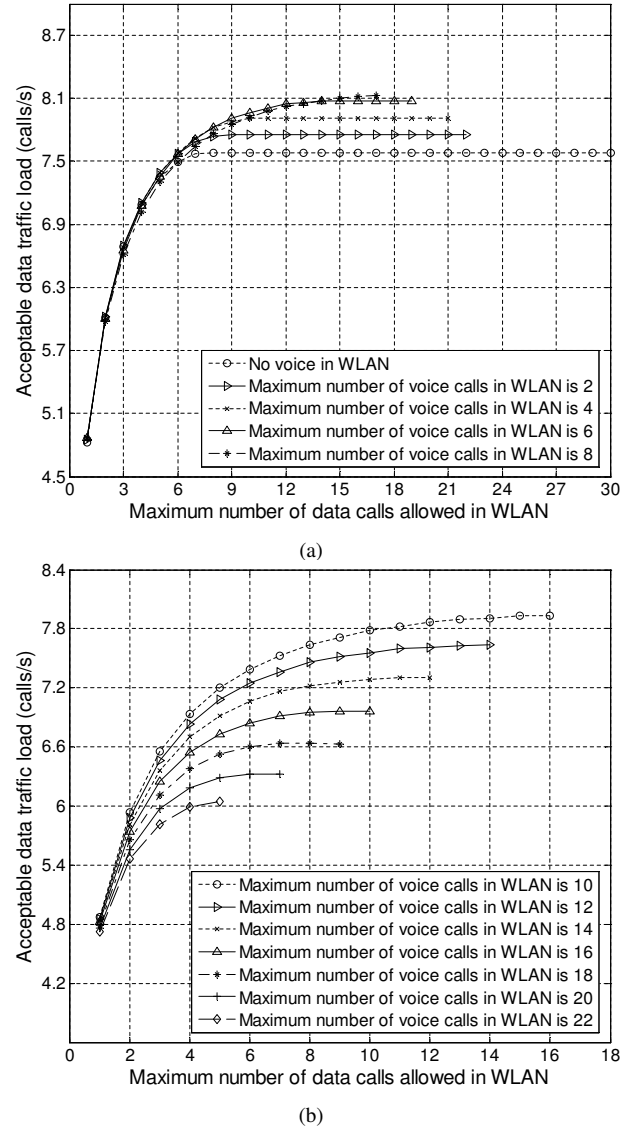
(a)

(b)

Fig. 5. Maximum acceptable data traffic load (mean data call arrival rate λ_d) versus maximum number of data calls allowed in the WLAN (N_d^w) under QoS constraints (blocking probabilities ≤ 0.01, dropping probabilities ≤ 0.001, and mean data transfer time ≤ 4 s).

the acceptable data traffic load decreases. When more voice calls are admitted to the WLAN, the number of data calls that can be simultaneously accommodated by the WLAN decreases and the data service rate drops. As a result, the maximum number of data calls allowed in the cell (N_d^c) needs to be increased so that the overall data call blocking and dropping probabilities meet the corresponding requirements. Due to the much smaller cell bandwidth, an increased traffic load assigned to the cell results in a longer data transfer time. When this penalty incurred by voice support in WLANs overwhelms the advantage of the load balancing effect, the acceptable data traffic load begins to decrease.

C. Validation of Analytical Results by Simulation

We use a discrete event-driven simulator written in C/C++ language to verify the accuracy of our analysis. The simulation model is consistent with the system model used for analy-
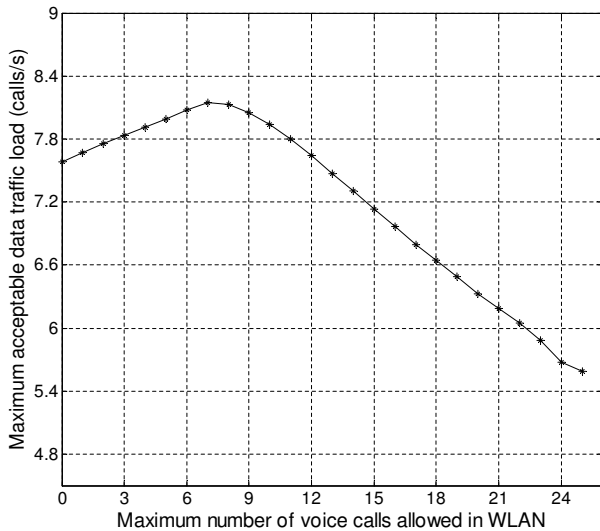
Fig. 6. Maximum acceptable data traffic load (mean data call arrival rate λ_d) versus maximum number of voice calls allowed in the WLAN (N_v^w) under QoS constraints (blocking probabilities ≤ 0.01, dropping probabilities ≤ 0.001, and mean data transfer time ≤ 4 s).

Fig. 7. Average service rate for packets from one data flow $\xi_d^w(n_v^w, n_d^w)$ with n_d^w fixed to 10 and n_v^w varying within the WLAN capacity region.

sis. Contention-based WLAN MAC protocol is simulated for packet transmissions in the WLAN. More than 10^7 voice and data call arrivals, departures and handoffs are generated in each simulation round to collect statistics on call blocking/dropping probabilities and data call transfer time. The results of multiple simulation rounds are averaged to remove randomness effect. The statistics are collected after the simulated system attains the equilibrium state. For real-time voice traffic, to guarantee the latency requirement, by applying admission control, we restrict the maximum numbers of voice and data flows contending for channel access in the WLAN. As an example, Table IV shows the average voice packet delay when n_d^w is fixed to 10 and n_v^w varies in the analytically derived WLAN capacity region. We can see that, within the capacity region, the voice packet delay is bounded. Fig. 7 further compares the analytical and simulation results of service rates for packets from each data flow. It is observed that the analytical results agree well with the simulation results, which validates the accuracy of our analytical model.

TABLE IV

AVERAGE VOICE PACKET DELAY IN THE WLAN.

n_v^w	1	2	3	4	5
Voice delay (ms)	34.4861	34.9917	38.9335	34.7121	33.8862
n_v^w	6	7	8	9	10
Voice delay (ms)	30.9186	28.7078	24.3805	30.0224	25.5337
n_v^w	11	12	13	14	15
Voice delay (ms)	25.0802	31.6095	28.7314	28.7526	27.8619

In addition, Fig. 8 and Fig. 9 illustrate the call-level QoS performance with different N_v^w. As shown in Fig. 8, the simulation results (denoted by "Simu.") of voice call blocking and dropping probabilities are very close to the analytical results (denoted by "Anal.") and tightly bounded by the corresponding
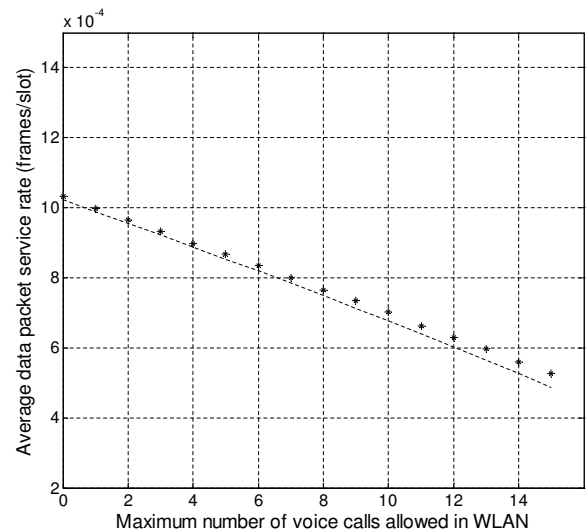
requirements. The performance fluctuation of handoff dropping probability is because the maximum numbers of calls allowed in the cell and WLAN are both integer variables. As we apply randomized guard channel policy to increase the granularity of bandwidth reservation, the fluctuation is actually smaller than that of traditional guard channel policy. On the other hand, as shown in Fig. 9, the mean data transfer time ($\mathrm{E}[T_d]$) is also well bounded and agrees well with the analytical results. To verify whether the user QoS is tightly bounded, we increase the maximum data call arrival rate λ_d obtained analytically by 1%, 2%, and 3%, respectively. This increase results in violation to the originally bounded mean data transfer time, which indicates that the upper bound of $\mathrm{E}[T_d]$ used for the derivation of admission regions is tight in this case of integrated voice/data services.
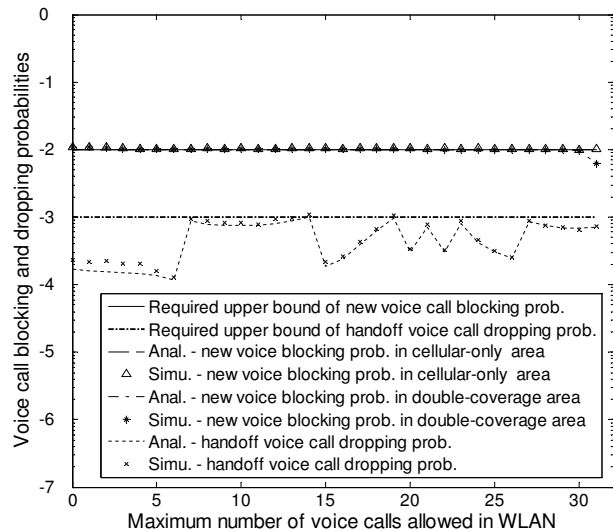

Fig. 8. Analytical and simulation results of voice call blocking and dropping probabilities.
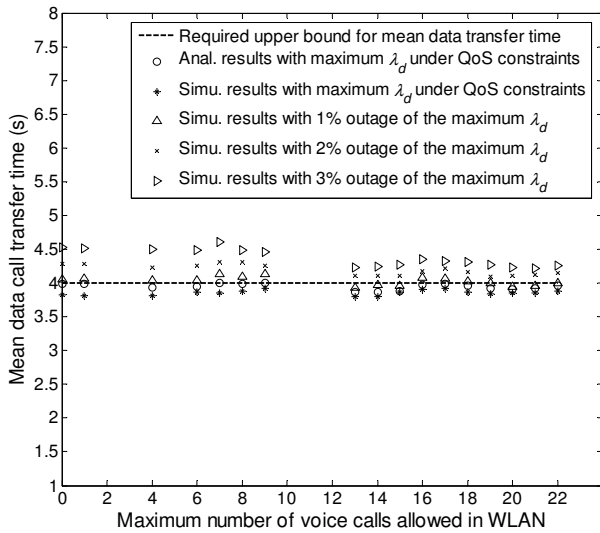
Fig. 9.   Analytical and simulation results of mean data transfer time.

VI. Conclusions

In this paper, we analyze the performance of the WLAN-first scheme in cellular/WLAN interworking. The analytical results are validated by computer simulation. It is observed that the QoS performance is closely related to the admission regions for voice and data services in the cellular network and WLANs. In the best configuration, the maximum number of voice calls allowed in a WLAN is less than the WLAN capacity for voice service. That is, voice traffic in the double-coverage area should be restricted not to occupy all the WLAN bandwidth. Indeed, because data traffic is adaptive to elastic bandwidth, it can take a good advantage of the low mobility and large bandwidth in double-coverage areas. Consequently, the total resources in a cellular/WLAN integrated network should be properly allocated to voice and data services by ensuring appropriate admission control, so as to maximize the overall resource utilization.

Acknowledgment

References

[1] K. Maheshwari and A. Kumar, "Performance analysis of microcellization for supporting two mobility classes in cellular wireless networks," *IEEE J. Select. Areas Commun.*, vol. 49, no. 2, pp. 321–333, Mar. 2000.

[2] T. Klein and S.-J. Han, "Assignment strategies for mobile data users in hierarchical overlay networks: performance of optimal and adaptive strategies," *IEEE J. Select. Areas Commun.*, vol. 22, no. 5, pp. 849–861, June 2004.

[3] B. Jabbari, "Teletraffic aspects of evolving and next-generation wireless communication networks," *IEEE Pers. Commun.*, vol. 3, no. 6, pp. 4–9, Dec. 1996.

[4] R. Núñez Queija, "Processor-sharing models for integrated-services networks," Ph.D. dissertation, Eindhoven Univ. of Tech., Jan. 2000.

[5] F. Delcoigne, A. Proutière, and G. Régnié, "Modeling integration of streaming and data traffic," *Performance Evaluation*, vol. 55, no. 3-4, pp. 185–209, Feb. 2004.

[6] M. Naghshineh and A. Acampora, "QoS provisioning in micro-cellular networks supporting multiple classes of traffic," *Wireless Networks*, vol. 2, no. 3, pp. 195–203, Aug. 1996.

[7] R. Núñez Queija, J. L. van den Berg, and M. R. H. Mandjes, "Performance evaluation of strategies for integration of elastic and stream traffic," in *Proc. 16th Int'l Teletraffic Congress*, 1999, pp. 1039–1050.

[8] V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, "MACAW: a media access protocol for wireless LAN's," in *Proc. ACM SIGCOMM'94*, Oct. 1994, pp. 212–225.

[9] O. Tickoo and B. Sikdar, "A queueing model for finite load IEEE 802.11 random access MAC," in *Proc. IEEE ICC'04*, June 2004, pp. 175–179.

[10] L. Cai, X. Shen, J. W. Mark, L. Cai, and Y. Xiao, "A queueing model for finite load IEEE 802.11 random access MAC," in *Proc. QShine'05*, Aug. 2005.

[11] H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma, "Performance of reliable transport protocol over IEEE 802.11 wireless LAN: analysis and enhancement," in *Proc. IEEE INFOCOM'02*, pp. 599–607.

[12] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*.   Springer-Verlag, 1995.

[13] P. Tran-Gia and F. Hübner, "An analysis of trunk reservation and grade of service balancing mechanisms in multiservice broadband networks," in *Proc. IFIP Workshop TC6'03*, pp. 83–97.

[14] F. P. Kelly, *Reversibility and Stochastic Networks*.   Wiley, 1979.

Wei Song received the B.S. degree in electrical engineering from Hebei University, Baoding, China, in 1998 and the M.S. degree in computer science from Beijing University of Posts and Telecommunications, Beijing, China, in 2001. She is currently working toward the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo, Canada. Her current research interests include resource allocation and quality-of-service (QoS) provisioning for the integrated cellular networks and wireless local area networks (LANs).

Hai Jiang (S'04) received the B.S. degree in 1995 and the M.S. degree in 1998, both in electronics engineering, from Peking University, Beijing, China. He is currently working toward the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include radio resource management, cellular/WLAN interworking, and cross-layer design for wireless multimedia communications.

Weihua Zhuang (M'93-SM'01) received the B.Sc. and M.Sc. degrees from Dalian Maritime University, Liaoning, China, and the Ph.D. degree from the University of New Brunswick, Fredericton, NB, Canada, all in electrical engineering. Since October 1993, she has been with the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada, where she is a Professor. She is a co-author of the textbook *Wireless Communications and Networking* (Prentice Hall, 2003). Her current research interests include multimedia wireless communications, wireless networks, and radio positioning. Dr. Zhuang is a licensed Professional Engineer in the Province of Ontario, Canada. She received the Premier's Research Excellence Award (PREA) in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions. She is an Associate Editor of *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, and *EURASIP Journal on Wireless Communications and Networking*.