

Chapter 2

Resource Allocation for Integrated Voice and Data Services

In the interworking between a cellular network and 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 access selection problem of how to properly admit incoming traffic to the cell or WLAN. Specially, a preferred target network, either a cellular 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 [26]. Due to user mobility and access selection in the overlaying area, the underlying network serving a user may alternate dynamically between the cellular network and WLANs. Due to heterogeneous underlying technologies, the admission choice can have a significant impact on overall resource utilization and QoS satisfaction. In this chapter, we introduce an easy-to-implement resource allocation scheme for voice and data services over the cellular/WLAN integrated network.

2.1 Bandwidth Sharing for Voice and Data Services

Consider real-time voice telephony and interactive data services (such as Web browsing). Each voice call requires a constant bandwidth to meet its strict delay requirement, while data service is adaptive to elastic bandwidth. Each voice call has two voice flows from and to the MS, while each data call has a one-way data flow to the MS. Moreover, we assume the mean arrival rate of new voice (data) calls in the cellular-only area λ_{v1} (λ_{d1}) is proportional to that in the double-coverage area λ_{v2} (λ_{d2}). 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. In addition to new traffic, there are also horizontal handoffs between neighboring cells and vertical handoffs between a WLAN and its overlaying cell.

In the cellular network, with the aid of base stations, the restricted access policy [29] 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 processor sharing (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 [33] 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 both voice and data users in service. Otherwise, the intra-service interference from users of the same service type or inter-service interference from users of the other service type may severely degrade the system performance.

2.1.1 Allocation of WLAN and Cell Bandwidth

To apply joint resource allocation for 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.

Suppose there are n_v^w voice calls and n_d^w data calls admitted in a WLAN. Packets from a voice flow are assumed to arrive with a constant rate, λ_v^P . 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 optional request-to-send and clear-to-send (RTS-CTS) handshaking for channel access, while voice flows use the basic carrier sensing multiple access with collision avoidance (CSMA/CA) mechanism due to the small payload size of voice packets. Following the approach in [43], we can derive the service rates for packets from one voice and data flow, denoted by $\xi_v^w(n_v^w, n_d^w)$ and $\xi_d^w(n_v^w, n_d^w)$, respectively. 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 to be determined experimentally. Given this stability 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.

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 as $R_v^c \cdot N_v^c$, where R_v^c is the bandwidth requirement of a voice call and $N_v^c (\leq \lfloor \frac{C^c}{R_v^c} \rfloor)$ 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

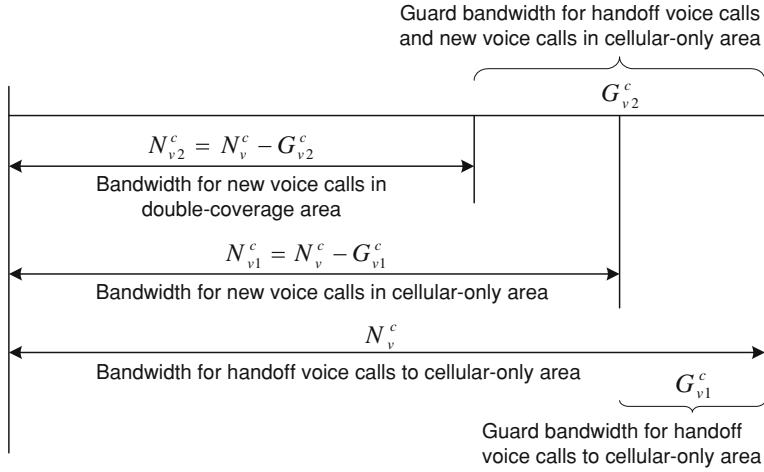


Fig. 2.1 Randomized guard channel policy for voice in the cell

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. 2.1, 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 representing 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 $(C^c - R_v^c \cdot N_v^c)$ 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, data traffic is prioritized similarly to voice based on user location area and new/handoff call differentiation. The data admission region of the cell is given by $(N_d^c, G_{d1}^c, G_{d2}^c)$.

2.1.2 Formulation of Resource Allocation Problem

The WLAN capacity region is derived in terms of (n_v^w, n_d^w) vectors. It is found that the data service rate $\xi_d^w(n_v^w, n_d^w)$ decreases dramatically with more voice calls (i.e., a larger n_v^w), which implies the inefficient voice support of WLANs. To prevent the WLAN from operating in that inefficiency region, instead of directly applying the two-dimensional capacity region as admission criteria, 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. 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 Sect. 2.4 that the configuration of admission regions does significantly affect the overall resource utilization. The optimal configuration assures a maximization of the acceptable traffic load with QoS satisfaction.

Let \tilde{B}_v (\tilde{B}_d), \tilde{D}_v (\tilde{D}_d), and \tilde{T}_d denote the upper bounds for voice (data) call blocking and dropping probabilities and mean data transfer time, respectively. Then, the resource allocation problem can be formulated as follows:

$$\begin{aligned} \max_{(N_v^w, N_d^w)} \quad & \lambda_d, \quad s.t. \\ B_v^w \cdot B_{v2}^c & \leq \tilde{B}_v, \quad B_{v1}^c \leq \tilde{B}_v, \quad D_v^c \leq \tilde{D}_v \\ B_d^w \cdot B_{d2}^c & \leq \tilde{B}_d, \quad B_{d1}^c \leq \tilde{B}_d, \quad D_d^c \leq \tilde{D}_d, \quad E[T_d] \leq \tilde{T}_d \end{aligned} \quad (2.1)$$

where λ_d ($=\lambda_{d1} + \lambda_{d2}$) is the mean data call arrival rate in the cell cluster-covered area, 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 in the WLAN, and $E[T_d]$ is the mean data transfer time. Since we fix the voice call arrival rates for simplicity, the maximization of λ_d implies a maximization of the total acceptable traffic load and resource utilization.

2.2 Voice Performance of WLAN-First Scheme

In this section, we investigate a simple and easy-to-implement access selection strategy, referred to as *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, an in-depth analysis is very meaningful to examine how various services affect the resource allocation and QoS support in a cellular/WLAN integrated network.

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* [33]. 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. The number of voice calls in the WLAN can be described by a birth-death process with respect to k_v^w . In this study, we first consider a simple case that the user residence time T_r^w in the double-coverage area is exponentially distributed, i.e., $a = 1$ in (1.3). Since both voice call duration T_v and user residence time T_r^w are exponentially distributed, the voice channel occupancy time $\min(T_v, T_r^w)$ is exponential with mean $1/(\mu_v + \eta^w)$, 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

$$\pi_v^w(k) = \frac{[(\lambda_{v2} + \lambda_{hv}^{cw})/(\mu_v + \eta^w)]^k / k!}{\sum_{i=0}^{N_v^w} [(\lambda_{v2} + \lambda_{hv}^{cw})/(\mu_v + \eta^w)]^i / i!}, \quad 0 \leq k \leq N_v^w \quad (2.2)$$

where λ_{v2} and λ_{hv}^{cw} 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. 2.2 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. The state-dependent transition rates are conditioned on k_v^w and derived as follows.

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

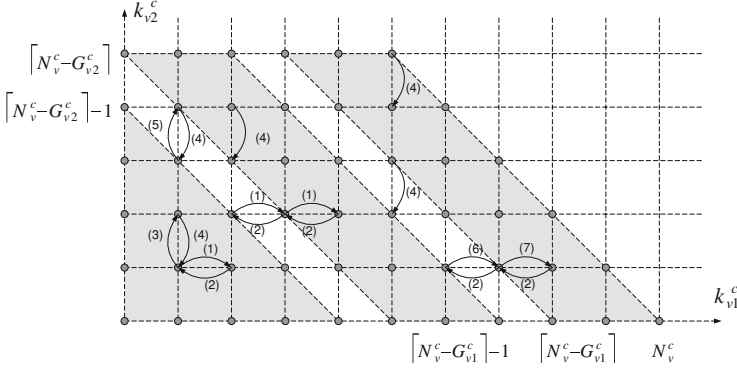
$$P[X > Y] = \int_0^\infty f_Y(y) dy \int_y^\infty \lambda e^{-\lambda x} dx = \int_0^\infty f_Y(y) e^{-\lambda y} dy = \Psi_Y(-\lambda) \quad (2.3)$$

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

$$\begin{aligned} f_Z(z) &= f_X(z)[1 - F_Y(z)] + f_Y(z)[1 - F_X(z)] \\ &= f_X(z) + f_Y(z) - [f_Y(z)F_X(z) + f_X(z)F_Y(z)] \end{aligned} \quad (2.4)$$

where $f_X(\cdot)$ and $F_X(\cdot)$ denote the PDF and CDF of X , respectively. Then, the mean value of Z is

$$\begin{aligned} E[Z] &= \int_0^\infty z f_Z(z) dz = E[X] + E[Y] - \int_0^\infty z [F_X(z)F_Y(z)]' dz \\ &= E[X] - \int_0^\infty f_Y(y) \frac{1}{\lambda} e^{-\lambda y} dy = \frac{1}{\lambda} - \frac{1}{\lambda} \Psi_Y(-\lambda) = \left[\frac{\lambda}{1 - \Psi_Y(-\lambda)} \right]^{-1}. \end{aligned} \quad (2.5)$$



- (1) $\lambda_{v1} + \lambda_{hv}^{cc} + \lambda_{hv}^{wc}$
- (2) $k_{v1}^c \mu_{v1}^c$, if $k_v^w = N_v^w$; $k_{v1}^c (\mu_v + \eta^c)$, if $k_v^w \leq N_v^w - 1$
- (3) λ_{v2} , if $k_v^w = N_v^w$; 0, if $k_v^w \leq N_v^w - 1$
- (4) $k_{v2}^c \mu_{v2}^c$, if $k_v^w = N_v^w$; $k_{v2}^c \frac{\mu_v}{1 - \psi(-\mu_v)}$, if $k_v^w \leq N_v^w - 1$
- (5) $\lambda_{v2} [1 - (G_{v2}^c - \text{floor}(G_{v2}^c))]$, if $k_v^w = N_v^w$; 0, if $k_v^w \leq N_v^w - 1$
- (6) $\lambda_{v1} [1 - (G_{v1}^c - \text{floor}(G_{v1}^c))] + \lambda_{hv}^{cc} + \lambda_{hv}^{wc}$
- (7) $\lambda_{hv}^{cc} + \lambda_{hv}^{wc}$

Fig. 2.2 State transition diagram for voice in the cell

For new and handoff voice calls in the cellular-only area, the channel occupancy time is $\min(T_v, T_{r1}^c)$. Based on (1.5) and (2.5), its mean value can be derived by

$$\begin{aligned}
 E[\min(T_v, T_{r1}^c)] &= \frac{1}{\mu_v} - \frac{1}{\mu_v} \sum_{i=1}^{\infty} (p^{cw})^{i-1} p^{cc} \frac{\eta^c}{\eta^c + \mu_v} [\psi(-\mu_v)]^{i-1} \\
 &= \frac{1}{\mu_v} - \frac{1}{\mu_v} p^{cc} \frac{\eta^c}{\eta^c + \mu_v} \frac{1}{1 - p^{cw} \psi(-\mu_v)} \triangleq \frac{1}{\mu_{v1}^c} \quad (2.6)
 \end{aligned}$$

where

$$\psi(-\mu_v) = \frac{\eta^c}{\eta^c + \mu_v} \cdot \frac{\eta^w}{\eta^w + \mu_v}.$$

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

$$\begin{aligned}
E[\min(T_v, T_{r2}^c)] &= \frac{1}{\mu_v} - \frac{1}{\mu_v} \sum_{i=1}^{\infty} (p^{cw})^{i-1} p^{cc} [\psi(-\mu_v)]^i \\
&= \frac{1}{\mu_v} - \frac{1}{\mu_v} p^{cc} \frac{\psi(-\mu_v)}{1 - p^{cw} \psi(-\mu_v)} \triangleq \frac{1}{\mu_{v2}^c}. \quad (2.7)
\end{aligned}$$

To further simplify analysis, we average the transition rates of (k_{v1}^c, k_{v2}^c) over k_v^w depending on whether there is enough free capacity in the WLAN for an arriving voice call. Then, 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 the departure rate from state (k_{v1}^c, k_{v2}^c) to state $(k_{v1}^c, k_{v2}^c - 1)$ ($k_{v2}^c \geq 1$) are respectively approximated by

$$\tilde{\mu}_{v1}^c = B_v^w k_{v1}^c \mu_{v1}^c + (1 - B_v^w) k_{v1}^c (\mu_v + \eta^c) \quad (2.8)$$

$$\tilde{\mu}_{v2}^c = B_v^w k_{v2}^c \mu_{v2}^c + (1 - B_v^w) k_{v2}^c \frac{\mu_v}{1 - \psi(-\mu_v)}. \quad (2.9)$$

As indicated by $\tilde{\mu}_{v1}^c$ and $\tilde{\mu}_{v2}^c$, voice traffic admitted to the cell from the cellular-only area and from the double-coverage area has different mean channel occupancy times 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 [38]. 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. In [55], the state distribution is approximated with a recursive method, 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 [55] to obtain $\pi_v^c(k)$ (i.e., the steady-state probability of k voice calls admitted into the cell) as follows¹:

$$\begin{aligned}
\pi_v^c(k) &= \left(\frac{\lambda_{v1}^c}{\tilde{\mu}_{v1}^c} + \frac{\lambda_{nv2}^c}{\tilde{\mu}_{v2}^c} \right)^k \frac{\pi_v^c(0)}{k!}, \quad 0 \leq k \leq \lfloor N_{v2}^c \rfloor \\
\pi_v^c(k) &= \left(\frac{\lambda_{v1}^c}{\tilde{\mu}_{v1}^c} + \frac{\lambda_{nv2}^c}{\tilde{\mu}_{v2}^c} \right)^{\lfloor N_{v2}^c \rfloor} \rho_{v2} \left(\frac{\lambda_{v1}^c}{\tilde{\mu}_{v1}^c} \right)^{k - \lfloor N_{v2}^c \rfloor - 1} \frac{\pi_v^c(0)}{k!}, \quad \lfloor N_{v2}^c \rfloor + 1 \leq k \leq \lfloor N_{v1}^c \rfloor \\
\pi_v^c(k) &= \left(\frac{\lambda_{v1}^c}{\tilde{\mu}_{v1}^c} + \frac{\lambda_{nv2}^c}{\tilde{\mu}_{v2}^c} \right)^{\lfloor N_{v2}^c \rfloor} \rho_{v2} \left(\frac{\lambda_{v1}^c}{\tilde{\mu}_{v1}^c} \right)^{\lfloor N_{v1}^c \rfloor - \lfloor N_{v2}^c \rfloor - 1} \rho_{v1} \left(\frac{\lambda_{hv}^c}{\tilde{\mu}_{v1}^c} \right)^{k - \lfloor N_{v1}^c \rfloor - 1} \frac{\pi_v^c(0)}{k!}, \\
&\quad \lfloor N_{v1}^c \rfloor + 1 \leq k \leq N_v^c \quad (2.10)
\end{aligned}$$

¹ 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 [55].

where $\pi_v^c(0) = 1/C_1$ with C_1 being a normalization constant,² and

$$\begin{aligned} N_{v1}^c &= N_v^c - G_{v1}^c, & N_{v2}^c &= N_v^c - G_{v2}^c \\ \lambda_{v1}^c &= \lambda_{v1} + \lambda_{hv}^{wc} + \lambda_{hv}^{cc}, & \lambda_{hv}^c &= \lambda_{hv}^{wc} + \lambda_{hv}^{cc}, & \lambda_{nv2}^c &= B_v^w \lambda_{v2} \\ \rho_{v1} &= \frac{[1 - (G_{v1}^c - \lfloor G_{v1}^c \rfloor)] \lambda_{v1} + \lambda_{hv}^c}{\tilde{\mu}_{v1}^c}, \\ \rho_{v2} &= \frac{\lambda_{v1}^c}{\tilde{\mu}_{v1}^c} + \frac{[1 - (G_{v2}^c - \lfloor G_{v2}^c \rfloor)] \lambda_{nv2}^c}{\tilde{\mu}_{v2}^c}. \end{aligned}$$

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

$$B_{v2}^c = (G_{v2}^c - \lfloor G_{v2}^c \rfloor) \pi_v^c(\lfloor N_{v2}^c \rfloor) + \sum_{i=\lfloor N_{v2}^c \rfloor+1}^{N_v^c} \pi_v^c(i) \quad (2.11)$$

$$B_{v1}^c = (G_{v1}^c - \lfloor G_{v1}^c \rfloor) \pi_v^c(\lfloor N_{v1}^c \rfloor) + \sum_{i=\lfloor N_{v1}^c \rfloor+1}^{N_v^c} \pi_v^c(i) \quad (2.12)$$

$$D_v^c = \pi_v^c(N_v^c). \quad (2.13)$$

2.3 Data Performance of WLAN-First Scheme

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 [33].

2.3.1 Mean Data Transfer Time

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 $(C^c - i R_v^c)$ with a probability $\pi_v^c(i)$, $i = 0, 1, \dots, N_v^c$. Thus, the expected duration of a data call carried by the cell is approximated by [10]

$$E[T_d^c] = \sum_{i=0}^{N_v^c} \pi_v^c(i) \frac{\rho_d^c(i)^{N_d^c+1} [N_d^c \rho_d^c(i) - N_d^c - 1] + \rho_d^c(i)}{\lambda_d^c [1 - \rho_d^c(i)^{N_d^c}] [1 - \rho_d^c(i)]} \quad (2.14)$$

² C_2 , C_3 , and C_4 used in the following are all normalization constants.

where

$$\lambda_d^c = \lambda_{d1} + \lambda_{nd2}^c + \lambda_{hd}^{wc} + \lambda_{hd}^{cc}, \quad \lambda_{nd2}^c = B_d^w \lambda_{d2}, \quad \rho_d^c(i) = \frac{\lambda_d^c f_d}{C^c - i R_v^c}$$

and λ_{nd2}^c is the mean arrival rate of new data calls overflowed to the cell in the double-coverage area, f_d is the mean data file size. Given in (2.14) is actually an upper bound for the mean transfer time of a data call with exponentially distributed size [10]. When data traffic evolves rapidly with respect to voice traffic, i.e., the number of data calls can attain its stationary regime given by an $M/G/1/K - PS$ queue with service capacity $(C^c - i R_v^c)$, 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 a 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

$$\tilde{\pi}_d^w(j) = \sum_{i=0}^{N_v^w} \left[\pi_v^w(i) \tilde{\pi}_d^w(0) \frac{(\lambda_d^w)^j}{\prod_{l=1}^j l \chi_d^w(i, l)} \right], \quad j = 1, 2, \dots, N_d^w \quad (2.15)$$

$$\tilde{\pi}_d^w(0) = 1/C_2, \quad \lambda_d^w = \lambda_{d2} + \lambda_{hd}^{cw}, \quad \chi_d^w(i, l) = \frac{\xi_d^w(i, l)}{f_d}$$

where $\pi_v^w(i)$ is given by (2.2), $\chi_d^w(i, l)$ is the service rate for one data call with i voice calls and l data calls in the WLAN, λ_{d2} and λ_{hd}^{cw} are the mean arrival rates of new and handoff data calls to the WLAN, respectively. Using the Little's law, the expected duration of a data call carried by the WLAN can be obtained as

$$E[T_d^w] = \frac{1}{\lambda_d^w(1 - B_d^w)} \sum_{j=0}^{N_d^w} j \tilde{\pi}_d^w(j) \quad (2.16)$$

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

2.3.2 Blocking and Dropping Probabilities of Data Calls

Consider the state that there are i voice calls and j data calls in a cell. For 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 $(i, j - 1)$ ($j \geq 1$) by

$$\tilde{\mu}_{d1}^c(i, j) = B_d^w j \mu_{d1}^c(i, j) + (1 - B_d^w) j [\nu_d^c(i, j) + \eta^c] \quad (2.17)$$

where

$$\nu_d^c(i, j) = \frac{C^c - i R_v^c}{j f_d} \quad (2.18)$$

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 (2.5) as

$$\mu_{d1}^c(i, j) = \frac{j \nu_d^c(i, j)}{1 - \Phi_1(-\nu_d^c(i, j))}. \quad (2.19)$$

Similarly, for data calls admitted to the cell from the double-coverage area, the departure rate from state (i, j) to state $(i, j - 1)$ ($j \geq 1$) is

$$\tilde{\mu}_{d2}^c(i, j) = B_d^w \cdot j \mu_{d2}^c(i, j) + (1 - B_d^w) \cdot \frac{j \nu_d^c(i, j)}{1 - \psi(-\nu_d^c(i, j))} \quad (2.20)$$

where

$$\mu_{d2}^c(i, j) = \frac{j \nu_d^c(i, j)}{1 - \Phi_2(-\nu_d^c(i, j))}. \quad (2.21)$$

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 following average departure rate to simplify analysis

$$\tilde{\mu}_d^c(i, j) = p_{d1}^c(j) \tilde{\mu}_{d1}^c(i, j) + p_{d2}^c(j) \tilde{\mu}_{d2}^c(i, j) \quad (2.22)$$

where $p_{d1}^c(\cdot)$ 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

$$p_{d1}^c(j) = \lambda_{d1}^c(j) / \lambda_d^c(j), \quad p_{d2}^c(j) = \lambda_{d2}^c(j) / \lambda_d^c(j), \quad \lambda_d^c(j) = \lambda_{d1}^c(j) + \lambda_{d2}^c(j)$$

$$\lambda_{d1}^c(j) = \begin{cases} \lambda_{d1} + \lambda_{hd}^{cc} + \lambda_{hd}^{wc}, & j \leq \lfloor N_{d1}^c \rfloor, \quad N_{d1}^c = N_d^c - G_{d1}^c \\ \lambda_{d1}(N_{d1}^c - \lfloor N_{d1}^c \rfloor) + \lambda_{hd}^{cc} + \lambda_{hd}^{wc}, & j = \lfloor N_{d1}^c \rfloor + 1 \\ \lambda_{hd}^{cc} + \lambda_{hd}^{wc}, & \lfloor N_{d1}^c \rfloor + 2 \leq j \leq N_d^c \end{cases}$$

$$\lambda_{d2}^c(j) = \begin{cases} \lambda_{hd2}^c, & j \leq \lfloor N_{d2}^c \rfloor, \quad N_{d2}^c = N_d^c - G_{d2}^c \\ \lambda_{hd2}^c(N_{d2}^c - \lfloor N_{d2}^c \rfloor), & j = \lfloor N_{d2}^c \rfloor + 1 \\ 0, & \lfloor N_{d2}^c \rfloor + 2 \leq j \leq N_d^c. \end{cases}$$

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 [20] for data with

$$\phi(j) = \tilde{\mu}_d^c(i, j), \gamma(l, j) = \delta(l, j) = \frac{1}{j}, \quad l = 1, 2, \dots, j, \quad j = 1, 2, \dots, N_d^c \quad (2.23)$$

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 $1, 2, \dots, j$; $\gamma(l, j)$ 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 queueing 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 (2.23) and has a service capacity $(C^c - iR_v^c)$ with a probability $\pi_v^c(i)$, $i = 0, 1, \dots, N_v^c$.

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 [20]. Then, the steady-state probability of j ($j = 1, 2, \dots, N_d^c$) data calls in the cell is approximately given by

$$\pi_d^c(j) = \sum_{i=0}^{N_v^c} \left[\pi_v^c(i) \pi_d^c(0) \prod_{l=1}^j \frac{\lambda_d^c(l)}{\phi(l)} \right] = \sum_{i=0}^{N_v^c} \left[\pi_v^c(i) \pi_d^c(0) \prod_{l=1}^j \frac{\lambda_d^c(l)}{\tilde{\mu}_d^c(i, l)} \right] \quad (2.24)$$

where $\pi_d^c(0) = 1/C_3$. 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 (2.11)–(2.13) 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 $j[\chi_d^w(i, j) + \eta^w]$. Similar to the derivation of $\pi_d^c(j)$, the steady-state probability of j data calls carried by the WLAN is given by

$$\pi_d^w(j) = \sum_{i=0}^{N_v^w} \left[\pi_v^w(i) \pi_d^w(0) \prod_{l=1}^j \frac{\lambda_d^w}{l[\chi_d^w(i, l) + \eta^w]} \right], \quad \pi_d^w(0) = 1/C_4, \quad j = 1, 2, \dots, N_d^w. \quad (2.25)$$

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

Table 2.1 System parameters

Parameter	Value	Parameter	Value
C^w	11 Mbit/s	C^c	2 Mbit/s
λ_{v1}	0.12 calls/s	λ_{v2}	0.18 calls/s
$(\mu_v)^{-1}$	140 s	R_v^c	12.2 kbit/s
$\tilde{B}_v(\tilde{B}_d)$	0.01	$\tilde{D}_v(\tilde{D}_d)$	0.001
f_d	64 KB	\tilde{T}_d	4 s
Δ	0.1	V_{lh}	0.6
$(\eta^c)^{-1}$	10 min	$(\eta^w)^{-1}$	14 min
p^{cc}	0.76	p^{cw}	0.24

As seen from (2.14), (2.16), (2.24), and (2.25), the mean data transfer time depends on the mean arrival rates and blocking and dropping probabilities of data calls, which are inter-dependent and need to be evaluated recursively.

2.4 Numerical Results and Discussion

Given in Table 2.1 are the system parameters, which are selected based on popularly deployed cellular networks (e.g., cdma2000) and WLAN standards (e.g., IEEE 802.11b). Applying the QoS evaluation approach in a search algorithm given in Table 2.2, we can obtain the voice and data allocation parameters. The best configuration should maximize the traffic load acceptable to a given cell/WLAN cluster. In the following, we discuss some important observations obtained from the numerical results of the searching process.

2.4.1 Accuracy Validation

We use a discrete event-driven simulator written in C/C++ language to verify the accuracy of our analysis. 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.

Figures 2.3 and 2.4 illustrate the call-level QoS performance within the derived admission regions. As shown in Fig. 2.3, the simulation results of voice call blocking and dropping probabilities are very close to the analytical results 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

Table 2.2 Search algorithm for allocation parameters

1:	Derive cell capacity region of vectors (n_v^c, n_d^c) to satisfy $\frac{E_b}{N_0}$ requirements
2:	Derive WLAN capacity region of vectors (n_v^w, n_d^w) to meet stability constraints
3:	$N_{v,max}^w = \max(n_v^w)$: $(n_v^w, n_d^w) \in \text{WLAN capacity region}$
4:	$N_{v,max}^c = \max(n_v^c)$: $(n_v^c, n_d^c) \in \text{cell capacity region}$
5:	for $N_v^w = 0, \dots, N_{v,max}^w$ do // Evaluation for voice traffic
6:	By bisection search, determine $(N_v^c, G_{v1}^c, G_{v2}^c)$ s.t.
7:	$B_v^w B_{v2}^c \leq \tilde{B}_v$, $B_{v1}^c \leq \tilde{B}_v$, and $D_v^c \leq \tilde{D}_v$
8:	$N_{d,max}^w = \max(n_d^w)$ with $n_v^w = N_v^w$
9:	for $N_d^w = 0, \dots, N_{d,max}^w$ do // Evaluation for data traffic
10:	Initialize $\lambda_{d,min}$ and $\lambda_{d,max}$
11:	$\lambda_d \leftarrow (\lambda_{d,min} + \lambda_{d,max})/2$ // Update mean arrival rate of data calls λ_d
12:	By bisection search, determine $(N_d^c, G_{d1}^c, G_{d2}^c)$ and acceptable λ_d s.t.
13:	$B_d^w B_{d2}^c \leq \tilde{B}_d$, $B_{d1}^c \leq \tilde{B}_d$, $D_d^c \leq \tilde{D}_d$, $E[T_d^c] \leq \tilde{T}_d$, and $E[T_d^w] \leq \tilde{T}_d$
14:	if solutions for $(N_d^c, G_{d1}^c, G_{d2}^c)$ exist then
15:	$\lambda_{d,min} \leftarrow \lambda_d$; $\lambda_d \leftarrow (\lambda_{d,min} + \lambda_{d,max})/2$
16:	else
17:	$\lambda_{d,max} \leftarrow \lambda_d$; $\lambda_d \leftarrow (\lambda_{d,min} + \lambda_{d,max})/2$
18:	end if
19:	if acceptable λ_d converges then
20:	break
21:	end if
22:	Record maximum acceptable λ_d
23:	end for
24:	end for
25:	Output (N_v^w, N_d^w) , $(N_v^c, G_{v1}^c, G_{v2}^c)$, and $(N_d^c, G_{d1}^c, G_{d2}^c)$ that maximize acceptable λ_d

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. 2.4, the mean data transfer time ($E[T_d]$) 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. It is found that this increase results in QoS violation to the mean data transfer time. This indicates that the relative analytical error of the mean data transfer time is restricted within 1–3 %, and the upper bound of $E[T_d]$ for the derivation of admission regions is tight in the case of integrated voice/data services.

2.4.2 Variation with WLAN Data Traffic

Figure 2.5 illustrates 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. As observed in Fig. 2.5a, the data traffic load increases with N_d^w when N_d^w is relatively small. This can be explained as follows. Due to the coupling between the

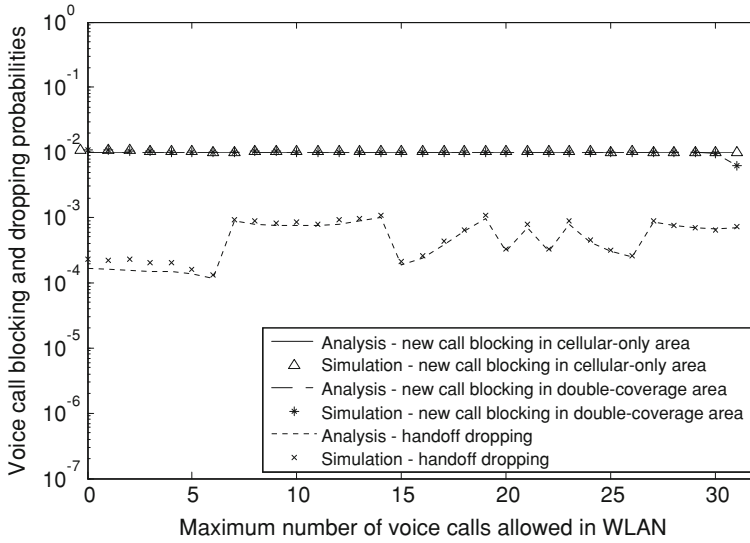


Fig. 2.3 Analytical and simulation results of voice call blocking/dropping probabilities

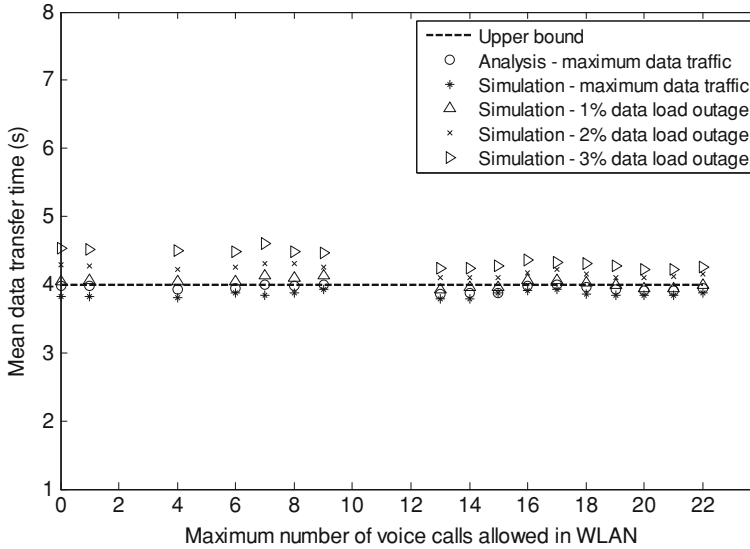


Fig. 2.4 Analytical and simulation results of mean data transfer time

cell and its overlaying WLAN, the less data calls allowed in the WLAN, the more data calls that need to be accommodated by the cell. With the PS sharing for data traffic, the less data calls admitted, the faster they will leave the system as a larger bandwidth is available for each data call. Due to user mobility, both the times that

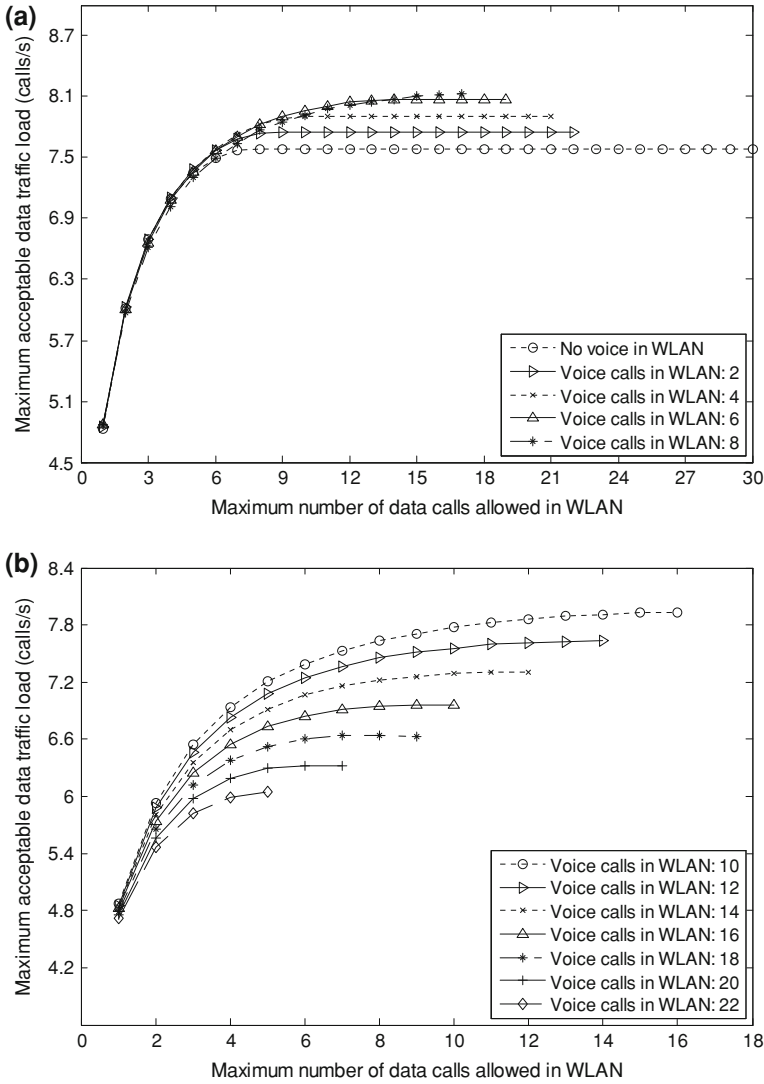


Fig. 2.5 Maximum acceptable data traffic load (mean data call arrival rate λ_d) versus maximum number of data calls allowed in the WLAN (N_d^w) under QoS constraints (blocking probabilities ≤ 0.01 , dropping probabilities ≤ 0.001 , and data transfer time ≤ 4 s)

a data call is carried by WLANs and by cells contribute to the total transfer time of a data call. 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) traffic load that can be supported.

As illustrated in Fig. 2.5b, 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 maximum 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 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 approximates the WLAN capacity for voice.

2.4.3 Variation with WLAN Voice Traffic

For each curve in Fig. 2.5, there is a maximum data call arrival rate achieved with a certain value of N_d^w . From these curves, we can obtain Fig. 2.6, which shows the relationship between the 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 acceptable data traffic load. With this configuration, N_v^w is less than the WLAN capacity for voice service (in this example, the maximum number of voice calls that can be carried with the total WLAN bandwidth is 28). 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, 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 the penalty incurred by voice support in WLANs overwhelms the advantage of the load balancing effect, the maximum acceptable data traffic load begins to decrease.

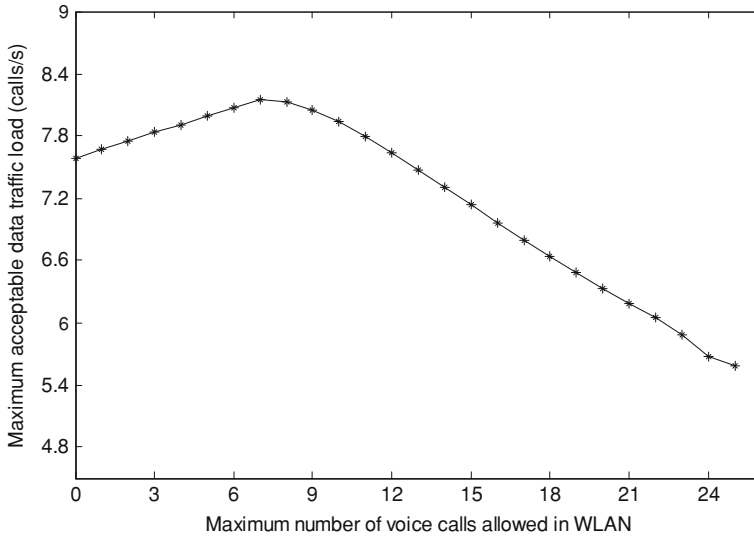


Fig. 2.6 Maximum acceptable data traffic load (mean data call arrival rate λ_d) versus maximum number of voice calls allowed in the WLAN (N_v^w) under QoS constraints (blocking probabilities ≤ 0.01 , dropping probabilities ≤ 0.001 , and data transfer time ≤ 4 s)

2.5 Summary

In this chapter, we introduce the *WLAN-first* resource allocation scheme for the cellular/WLAN intergraded network. In this simple and easy-to-implement scheme, 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. The main features and observations of this work are as follows:

- In this resource allocation scheme, new and handoff calls in different areas are prioritized with limited fractional guard channel policies. To compensate for the limited QoS differentiation capability of WLANs, restricted access mechanism is applied in the cell for resource sharing between voice and data services.
- An analytical model based on two-dimensional Markov processes is developed to evaluate QoS metrics in terms of call blocking/dropping probabilities and mean data transfer time. The model properly captures the location-dependent user mobility within the cell/WLAN cluster.
- Numerical results demonstrate that the QoS performance is closely related to the admission regions for voice and data services in the cell and WLAN. The best admission regions can be determined by applying the QoS evaluation in a search algorithm. Because data traffic accepts elastic bandwidth and better exploits the low mobility and large bandwidth in the double-coverage area, the maximum number of voice calls allowed in a WLAN should be restricted not to occupy all WLAN capacity.

Interworking of Wireless LANs and Cellular Networks

Song, W.; Zhuang, W.

2012, VIII, 67 p. 24 illus., Softcover

ISBN: 978-1-4614-4378-0