Call Admission Control in Mobile Cellular Networks: A Comprehensive Survey

Majid Ghaderi and Raouf Boutaba
School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1, Canada
Tel: +519 885 5412
Fax: +519 885 1208
{mghaderi,rboutaba}@uwaterloo.ca

Abstract

Call admission control is a key element in the provision of guaranteed quality of service in wireless networks. The design of call admission control algorithms for mobile cellular networks is especially challenging given the limited and highly variable resources, and the mobility of users encountered in such networks. This article provides a survey of admission control schemes for cellular networks and the research in this area. Our goal is to provide a broad classification and thorough discussion of existing call admission control schemes. We classify these schemes based on factors such as deterministic/stochastic guarantees, distributed/local control and adaptivity to traffic conditions. In addition to this, we present some modeling and analysis basics to help in better understanding the performance and efficiency of admission control schemes in cellular networks. We describe several admission control schemes and compare them in terms of performance and complexity. Handoff prioritization is the common characteristic of these schemes. We survey different approaches proposed for achieving handoff prioritization with a focus on reservation schemes. Moreover, optimal and near-optimal reservation schemes are presented and discussed. Also, we overview other important schemes such as those designed for multi-service networks and hierarchical systems as well as complete knowledge schemes and those using pricing for call admission control. Finally, the paper concludes on the state of current research and points out some of the key issues that need to be addressed in the context of call admission control for future cellular networks.

Call Admission Control in Mobile Cellular Networks: A Comprehensive Survey

I. Introduction

Starting in 1921 in the United States, police department experimental mobile radios began operating just above the present AM radio broadcast band. On June 17, 1946 in Saint Louis, AT&T and Southwestern Bell introduced the first American commercial mobile telephone service (typically in automobiles). Installed high above Southwestern Bell's headquarters, a centrally located antenna paged mobiles and provided radio-telephone traffic on the downlink. In the mid-1960s, the Bell System introduced the Improved Mobile Telephone Services (IMTS), which markedly improved the mobile telephone systems. As early as 1947, it was realized that small cells with frequency reuse could increase traffic capacity substantially and the basic *cellular* concept was developed. However, the technology did not exist. In the late 1960s and early 1970s, the cellular concept was conceived and was then used to improve the system capacity and frequency efficiency.

Each cell in a cellular network is equipped with a base station and with a number of radio channels assigned according to the transmission power constraints and availability of spectrum. A channel can be a frequency, a time slot or a code sequence. Any terminal residing in a cell can communicate through a radio link with the base station located in the cell, which communicates with the Mobile Switching Center (MSC), which is in turn connected to the Public Switched Telephone Networks (PSTN) as shown in Fig. 1. When a user initiates or receives a call, the user may roam around the area covered by the network. If the mobile user moves from one cell to another, and the call from/to the user has not finished, the network has to *handoff* the call from one cell to another at the cell boundary crossing without user's awareness of handoff and without much degradation of the service quality.

With the development of digital technologies and microprocessing computing power in the late 1980's and up to today, enormous interest emerged in digital cellular systems, which promised higher capacity and higher quality of services at reduced costs. Historically, mobile cellular communications have undertaken four evolution stages or generations, which are shown in

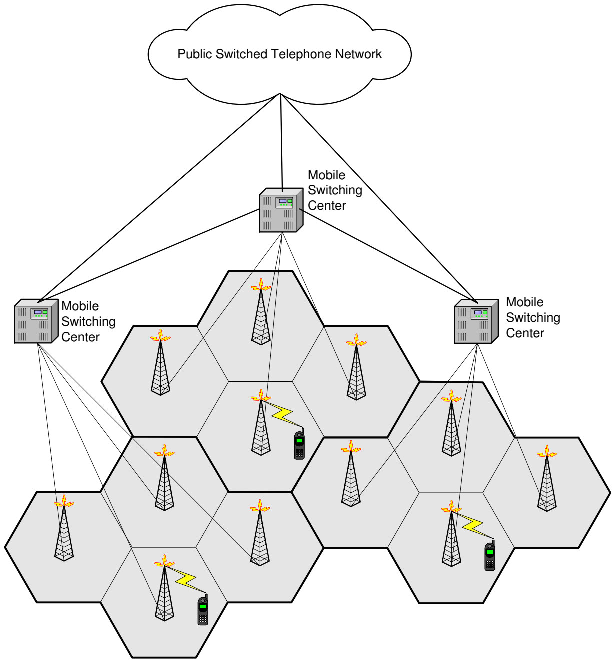


Fig. 1. A cellular system with hexagonal cells.

Table (I) taken from [1]. Analog cellular systems belong to the first generation where the major service provided is voice. Second generation cellular systems use digital technologies to provide better quality of service including voice and limited data with higher system capacity and lower cost. Third generation cellular networks offer multimedia transmission, global roaming across a homogeneous wireless network, and bit rates ranging from 384 Kbps to several Mbps. Worldwide migration to 3G is expected to continue through 2005 [2]. Meanwhile, researchers and vendors are expressing a growing interest in 4G wireless networks that support global roaming across heterogeneous wireless and mobile networks, for example, from a cellular network to a satellite-based network to a high-bandwidth wireless LAN [2]–[4].

Quality of Service (QoS) provisioning in wireless networks is a challenging problem due to the scarcity of wireless resources, i.e. radio channels, and the mobility of users. Call admission control (CAC) is a fundamental mechanism used for QoS provisioning in a network. It restricts the access to the network based on resource availability in order to prevent network congestion and service degradation for already supported users. A new call request is accepted if there are enough idle resources to meet the QoS requirements of the new call without violating the QoS

TABLE I

EVOLUTION OF MOBILE COMMUNICATION SYSTEMS.

Property	1G	2G	2.5G&3G	4G
Starting Time	1985	1992	2002	210-2012
Representative Standard	AMPS	GSM	IMT-2000	UWB
Radio Frequency (Hz)	400M-800M	800M-900M	1800M-2400M	2G-8G
Bandwidth(bps)	2.4K-3K	9.6K-14.4K	384K-2M	20M-100M
Multiple Access Technique	FDMA	TDMA, CDMA	WCDMA	OFDM
Switching Basis	Circuit	Circuit	Circuit,Packet	Packet
Cellular Coverage	Large area	Medium area	Small area	Mini area
Service Type	Voice	Voice, limited data	Voice, data, limited multimedia	Multimedia

for already accepted calls. With respect to the layered network architecture, different quality of service parameters are involved at different layers. At physical layer, bit-level QoS parameters such as bit energy-to-noise density describe the quality of service a mobile user receives. In packet-based communication systems, packet-level QoS parameters such as packet loss, delay and jitter characterize the perceived quality of service. However, most of the existing research on call admission control in cellular networks have focused on an abstract representation of the network in which only call-level QoS parameters, namely, call blocking and dropping probabilities are considered.

The paper is organized as follows. Section II is an overview of some basic concepts which are required for following the rest of the paper. Section III presents the basic modeling and analysis techniques in cellular networks. Section IV identifies three different call admission control problems based on the call-level QoS metrics and gives an overview of call admission control in cellular networks. As the most general approach to admission control, handoff prioritization techniques are reviewed in section V. We then discuss dynamic reservation schemes in section VI and discuss two broad categories of existing admission control techniques, namely, local and distributed admission control. Section VIII covers other important schemes such as those for multi-services networks and hierarchical systems, complete knowledge schemes and the use of pricing for call admission control. Finally, section IX concludes this survey.

II. Basic Concepts

A. Call Dropping and Handoff Failure

When a mobile terminal (mobile user) requests service, it may either be granted or denied service. This denial of service is known as call blocking, and its probability as *call blocking probability* (p_b). An active terminal in a cellular network may move from one cell to another. The continuity of service to the mobile terminal in the new cell requires a successful handoff from the previous cell to the new cell. A handoff is successful if the required resources are available and allocated for the mobile terminal. The probability of a handoff failure is called *handoff failure probability* (p_f). During the life of a call, a mobile user may cross several cell boundaries and hence may require several successful handoffs. Failure to get a successful handoff at any cell in the path forces the network to discontinue service to the user. This is known as call dropping or forced termination of the call and the probability of such an event is known as *call dropping probability* (p_d). In general, dropping a call in progress is considered to have a more negative impact from the user's perspective than blocking a newly requested call.

According to the above definition, the call dropping probability, p_d, and handoff failure probability, p_f, are different parameters. While the handoff failure probability is an important parameter for network management, the probability of call dropping (forced termination) may be more relevant to mobile users and service providers. Despite this fact, most research papers focus on the handoff failure probability because calculating p_f is more convenient.

If H is the number of handoffs throughout the duration of a call then

$$p_d = 1 - (1 - p_f)^H , \tag{1}$$

where H itself is a random variable. Therefore, in average

$$p_d = 1 - \sum_{h=0}^{\infty} (1 - p_f)^h \Pr(H = h) \, . \tag{2}$$

Finally, given the call blocking and dropping probabilities p_b and p_d, the *call completion probability* (p_c) is given by

$$p_c = (1 - p_b)(1 - p_d) \, . \tag{3}$$

Intuitively, call completion probability shows the percentage of those calls successfully completed in the network.

B. Channel Assignment Schemes

Channels are managed at each cell by channel assignment schemes based on co-channel reuse constraints. Under such constraints, three classes of channel assignment schemes have been widely investigated [5]–[7]:

1) Fixed channel assignment (FCA)
2) Dynamic channel assignment (DCA)
3) Hybrid channel assignment (HCA)

In FCA schemes, a set of channels is permanently assigned to each base station. A new call can only be served if there is a free channel available in the cell. Due to non-uniform traffic distribution among cells, FCA schemes suffer from low channel utilization. DCA was proposed to overcome this problem at the expense of increased complexity and signaling overhead. In DCA, all channels are kept in a central pool to be shared among the calls in all cells. A channel is eligible for use in any cell provided the co-channel reuse constraint is satisfied. Although DCA provides flexibility, it has less efficiency than FCA under high load conditions [7]. To overcome this drawback, hybrid allocation techniques, which are a combination of FCA and DCA, were proposed. In HCA each cell has a static set of channels and can dynamically borrow additional channels. For comprehensive survey on channel assignment schemes, the reader is referred to [5]. In this paper, we are interested in that networks where channel assignment is fixed.

C. Handoff Schemes

The handoff schemes can be classified according to the way the new channel is set up and the method with which the call is handed off from the old base station to the new one. At call-level, there are two classes of handoff schemes, namely hard handoff and soft handoff [8], [9].

1) *Hard handoff:* In hard handoff, the old radio link is broken before the new radio link is established and a mobile terminal communicates at most with one base station at a time. The mobile terminal changes the communication channel to the new base station with the possibility of a short interruption of the call in progress. If the old radio link is disconnected before the network completes the transfer, the call is forced to terminate. Thus, even if idle channels are available in the new cell, a handoff call may fail if the network response time for link transfer is too long [10]. Second generation mobile communication systems based on GSM fall in this category.

2) *Soft handoff:* In soft handoff, a mobile terminal may communicate with the network using multiple radio links through different base stations at the same time. The handoff process is initiated in the overlapping area between cells some short time before the actual handoff takes place. When the new channel is successfully assigned to the mobile terminal, the old channel is released. Thus, the handoff procedure is not sensitive to link transfer time [8], [10]. The second and third generation CDMA-based mobile communication systems fall in this category.

Soft handoff decreases call dropping at the expense of additional overhead (two busy channels for a single call) and complexity (transmitting through two channels simultaneously) [10]. Two key issues in designing soft handoff schemes are the handoff initiation time and the size of the active set of base stations the mobile is communicating with simultaneously [9]. This study focuses on cellular networks implementing hard handoff schemes.

D. Performance Criteria

In this subsection, we identify some commonly used performance criteria for comparing CAC schemes. Although others exist, we will focus on the following criteria in this survey:

1) *Efficiency:* Efficiency refers to the achieved utilization level of network capacity given a specific set of QoS requirements. Scheme A is more efficient than scheme B if the network resource utilization with scheme A is higher than that with scheme B for the same QoS parameters and network configuration.
2) *Complexity:* Shows the computational complexity of a CAC scheme for a given network configuration, mobility patterns, and traffic parameters. Scheme A is more complex than scheme B if admission decision making of A involves more complex computations than scheme B.
3) *Overhead:* Refers to the signalling overhead induced by a CAC scheme on the fixed interconnection network among base stations. Some CAC schemes require some information exchange with neighboring cells through the fixed interconnection network.
4) *Adaptivity:* Defined as the ability of a CAC scheme to react to changing network conditions. Those CAC schemes which are not adaptive lead to poor resource utilization. In this paper we only consider adaptivity to traffic load changes. Typically, CAC schemes make

admission decisions based on some internal control parameters, e.g. reservation threshold, which should be recomputed if the load changes.

5) *Stability:* Stability is the CAC insensitivity to short term traffic fluctuations. If an adaptive CAC reacts too fast to any load change then it may lead to unstable control. For example during a period of time all connection requests are accepted until a congestion occurs and then all requests are rejected. It is desirable that network control and management avoid such a situation.

Looking at existing CAC schemes, there are many assumptions and parameters involved in each scheme. Therefore, it is extremely difficult to develop a unified framework for evaluating and comparing the performance of CAC schemes using analytical or simulation techniques. For the comparison purposes in this paper, we do not use quantitative values for these criteria instead we use qualitative values. These qualitative values, e.g. “Very High”, “High”, “Moderate” and “Low”, are sufficient for a relative comparison of the CAC schemes investigated in this paper.

III. Cellular Networks Modeling and Analysis

Hong and Rappaport are the first who systematically studied the performance evaluation of cellular networks [11]. Due to the mobility of users and the complex traffic generated by new emerging integrated services, analytical results from classical traffic theory are not applicable to cellular communication systems. Hence, traffic engineering for networks supporting mobile services has added a new dimension in teletraffic theory and requires careful attention. In this section, we present some basic modeling and analysis techniques that will be useful for the remaining of the paper.

A. Assumptions and Definitions

We define the following terms commonly used in the literature to be used throughout this paper.

- *call holding time:* the duration of the requested call connection. This is a random variable which depends on the user behavior (call characteristics).
- *cell residency time:* the amount of time a mobile user spends in a cell. Cell residency is a random variable which depends on the user behavior and system parameters, e.g. cell geometry.

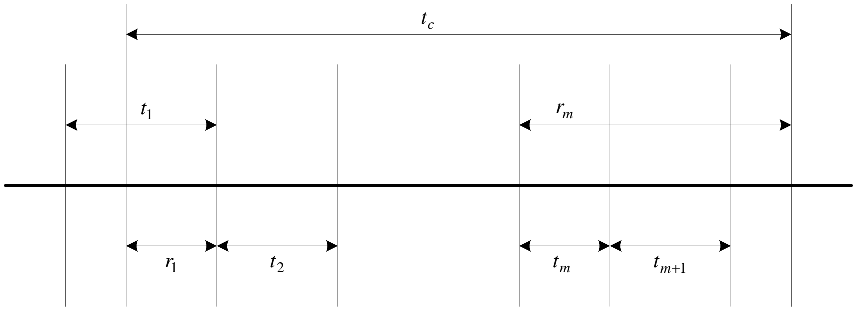


Fig. 2. The time diagram for call holding time and cell residence time.

- *channel holding time:* How long a call which is accepted in a cell and is assigned a channel will use this channel before completion or handoff to another cell. This is a random variable which can be computed from the call holding time and cell residency time and generally is different for new calls and handoff calls.

One of the most important parameters in modeling a cellular network is the channel holding time distribution. Typically, it is assumed that channel holding time is exponentially distributed with the same parameter for both new calls and handoff calls. This is a direct result of the memoryless assumption that call holding time and cell residency times are exponentially distributed [12]. This assumption may not be correct in practice and needs more careful investigation as pointed out in [13]–[17] and references there in.

Fig. 2, taken from [14], shows a time diagram for call holding and cell residency times. Let t_c be the call holding time for a typical new call, t_m be the cell residency time, r_1 be the time between the instant the new call is initiated at and the instant the new call moves out of the cell if the new call is not completed, and $r_m\,(m > 1)$ be the residual life of call holding time when the call finishes the m-th handoff successfully. Let t_{nh} and t_{hh} denote the channel holding times for a new call and a handoff call, respectively. Then from Fig. 2, the new call channel holding time is

$$t_{\mathrm{nh}} = \min\{t_c, r_1\}, \tag{4}$$

and the handoff call channel holding time is

$$t_{\mathrm{hh}} = \min\{r_m, t_m\}. \tag{5}$$

Consequently it can be shown that [18]

$$F_{t_{\mathrm{nh}}}(t) = F_{t_c}(t) + F_{r_1}(t) - F_{t_c}(t)F_{r_1}(t), \tag{6}$$

and

$$F_{t_{\mathrm{hh}}}(t) = F_{r_m}(t) + F_{t_m}(t) - F_{r_m}(t)F_{t_m}(t). \tag{7}$$

where $F_x(t) = \Pr(x \leq t)$ is the probability distribution function of random variable x.

As mentioned earlier, inherited from classical telephony, it is typically assumed that call holding times and cell residency times in mobile cellular networks are exponentially distributed. Assume that call holding times are exponentially distributed with mean $1/\mu$ and cell residency times are also exponentially distributed with mean $1/\eta$. From the memoryless property of exponential distribution, we conclude that r_m has the same distribution as t_c. Similarly, r_1 has the same distribution as t_m. Using (6) and (7), it can be obtained that t_{nh} and t_{hh} are exponentially distributed with the mean $1/(\mu + \eta)$.

B. Cellular Network Modeling and Analysis

Consider a cellular network consisting of M cells. Mobile users move among the cells according to the routing probability matrix $R = [r_{ij}]$. From theoretical point of view, such a network can be modeled as an open queueing network with arbitrary routing where each cell is modeled as a multi-server queue. In this subsection, we focus on such classical modeling techniques based on the Markov chain analysis. Therefore, exponential distributions play a critical role in this analysis. In the following subsection, we shall consider more general cases in which the exponential assumption is relaxed.

The following assumptions are used in this subsection:

- Each cell i has c_i channels.
- The call holding time is exponentially distributed with mean $1/\mu$.
- The new calls arrive into a cell according to a Poisson process. The arrival rate into cell i is λ_i.
- The cell residence times are exponentially distributed. The mean residence time in cell i is $1/\eta_i$.

One important parameter required for the analysis is the handoff arrival process which depends on other system parameters, e.g. cell residence times. Fang et al. [12] showed that with exponential call holding times, handoff arrival process will be Poisson if and only if the cell residence time is exponentially distributed. Let ν_i denote the handoff arrival rate into cell i. We will later show how to compute ν_i for the considered network.

Having this set of assumptions, each cell i in isolation can be modeled as an $M/M/c/c$ queue. Let us define the state of a cell as the total number of active calls in the cell. Let $\pi_i(n)$ denote the steady-state probability of having n calls in cell i. Using the balanced equations (or Erlang-B formula) we find that

$$\pi_i(n) = \frac{\rho^n/n!}{\sum_{n=0}^{c_i} \rho^n/n!}, \quad 1 \leq n \leq c_i \tag{8}$$

where ρ_i denotes the offered load and is expressed as

$$\rho_i = (\lambda_i + \nu_i)/(\mu + \eta)\,. \tag{9}$$

Consequently, the call blocking probability in cell i, $\boldsymbol{P}_b(i)$, is given by

$$\boldsymbol{P}_b(i) = \pi_i(c_i)\,. \tag{10}$$

Since handoff calls are treated in the same way as new calls in the network under investigation, we simply obtain handoff failure probability in cell i, $\boldsymbol{P}_f(i)$, as follows

$$\boldsymbol{P}_f(i) = \boldsymbol{P}_b(i)\,. \tag{11}$$

Let $\pi(n_1, \ldots, n_M)$ denote the steady-state probability of the network being in state $(n_1, \ldots, n_M)$, i.e. n_1 calls in cell 1, n_2 calls in cell 2, and so on . Using the classical queueing theory results [19], it is obtained that

$$\pi(n_1, \ldots, n_M) = \prod_{i=1}^{M} \pi_i(n_i), \quad 0 \leq n_i \leq c_i\,. \tag{12}$$

Now is left to compute the handoff arrival rate in each cell. Thanks to the memoryless property of exponential distribution we have

1) New call channel holding time and handoff channel holding time are exponentially distributed with the same mean value
2) The residual life of a call is exponentially distributed with the same mean value as the call holding time
3) Handoff arrival process is a Poisson process and consequently the joint new and handoff arrival process is Poisson as well

Let us define $\boldsymbol{P}_h(i)$ as the probability that a call currently being served in cell i requires another handoff before completion. Then, $\boldsymbol{P}_h(i)$ can be expressed as

$$\begin{aligned}\boldsymbol{P}_h(i) &= \Pr(t_c > t_i) \\ &= \int_{t=0}^{\infty} \Pr(t_c > t_i | t_i = t)\ \Pr(t_i = t)\, dt = \frac{\eta_i}{\mu + \eta_i}\,.\end{aligned} \tag{13}$$

Then, the rate of handoff out of any cell j is given by

$$(\lambda_j + \nu_j)(1 - \boldsymbol{P}_b(j))\boldsymbol{P}_h(j)\,. \tag{14}$$

Hence, the handoff arrival rate into cell i is given by

$$\nu_i = \sum_{j \neq i} \Big[(\lambda_j + \nu_j)\big(1 - \boldsymbol{P}_b(j)\big)\boldsymbol{P}_h(j)\Big] r_{ji} \tag{15}$$

or in matrix form as follows

$$\boldsymbol{\Lambda}_h = (\boldsymbol{\Lambda}_n + \boldsymbol{\Lambda}_h)(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{\Phi}, \tag{16}$$

where, $\boldsymbol{\Lambda}_n = [\lambda_1, \ldots, \lambda_i]$, $\boldsymbol{\Lambda}_h = [\nu_1, \ldots, \nu_i]$, $\boldsymbol{B} = \mathrm{diag}[\boldsymbol{P}_b(i)]$, $\boldsymbol{I}$ is an $M \times M$ identity matrix, and $\boldsymbol{\Phi}[\phi_{ij}]$ is the handoff rate matrix with $\phi_{ij} = \boldsymbol{P}_h(i) r_{ij}$. A fixed-point iteration [20] can be used to obtain the steady-state handoff arrival rate vector $\boldsymbol{\Lambda}_h$. Fixed-point iteration also known as relaxation method or repeated substitution, is a simple technique for solving the nonlinear equations describing the system. Iteration starts with an initial value for $\boldsymbol{\Lambda}_h$, say $[0, \ldots, 0]$, to obtain a new value for $\boldsymbol{\Lambda}_h$. Then this new value is substituted in (16) to obtain another value. This process continues until $\boldsymbol{\Lambda}_h$ converges with respect to the desired precision.

So far, we have computed the call blocking probability, $\boldsymbol{P}_b(i)$, which is essentially equal to the handoff failure probability, $\boldsymbol{P}_f(i)$, in the network model under investigation. In fact there is no preferential treatment implemented for handoff calls inside the network. Although handoff failure probability is an important measure for network control but call dropping probability is more meaningful for users (refer to section II). In the following discussion we turn our attention to the computation of the network-wide call dropping probability using discrete time Markov chain (DTMC) analysis. Let $\boldsymbol{P}_d(i)$ denote the call dropping probability given that the connection was initiated in cell i. Notice that the call dropping probability in this model is source dependent due to the heterogeneous nature of the network.

User mobility in the considered cellular network can be conveniently represented by a DTMC as follows. Each state i ($1 \leq i \leq M$) of this chain represents the current location (cell index) of the mobile user within the network. In addition to this, there are two absorbing states, one for dropping state (state d) and the other for completion state (state c). Let $\boldsymbol{\Delta}[\delta_{ij}]$ denote the associated transition probability matrix. Then

$$\begin{cases} \delta_{ij} = \phi_{ij}(1 - \boldsymbol{P}_b(j)) & 1 \leq i, j \leq M \\ \delta_{id} = \sum_{j \neq i} \phi_{ij} \boldsymbol{P}_b(j) & 1 \leq i \leq M \\ \delta_{ic} = 1 - \boldsymbol{P}_h(i) & 1 \leq i \leq M \\ \delta_{cc} = \delta_{dd} = 1 \end{cases} \tag{17}$$

This is a transient Markov chain and will finally settle into one of the absorbing states d or c. The transition matrix $\boldsymbol{\Delta}$ has the following canonical form

$$\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{A} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix}, \tag{18}$$

where $\boldsymbol{Q}$ is an $M \times M$ matrix representing the transient states, $\boldsymbol{A}$ is an $M \times 2$ matrix, $\boldsymbol{I}$ is an 2×2 identity matrix and $\mathbf{0}$ is an $2 \times M$ zero matrix.

Let $\boldsymbol{N}$ denote the fundamental matrix [21] of $\boldsymbol{\Delta}$, that is

$$\boldsymbol{N} = (\boldsymbol{I} - \boldsymbol{Q})^{-1}. \tag{19}$$

Let s_{ij} be the probability that a call initiated in cell i will be absorbed in state j ($j = d, c$). Let $\boldsymbol{S}$ be the matrix with entries s_{ij}. Then $\boldsymbol{S}$ is an $M \times 2$ matrix, and

$$\boldsymbol{S} = \boldsymbol{N}\boldsymbol{A}, \tag{20}$$

where $\boldsymbol{N}$ is the fundamental matrix given by (19) and $\boldsymbol{A}$ is as in the canonical form of $\boldsymbol{\Delta}$. Notice that the call completion probability and call dropping probability are then obtained as

$$\boldsymbol{P}_c(i) = s_{ic} \tag{21}$$

$$\boldsymbol{P}_d(i) = s_{id} \tag{22}$$

given that the call was initiated in cell i.

To compute the average (system-wide) call dropping and call completion probabilities, let $\boldsymbol{W} = [w_1, \ldots, w_M]$ be the initial probability distribution of initiated calls, then (as in [22])

$$w_i = \frac{\lambda_i(1 - \boldsymbol{P}_b(i))}{\sum_{j=1}^{M} \lambda_j(1 - \boldsymbol{P}_b(j))}. \tag{23}$$

Therefore, the average call dropping probability is given by

$$p_d = [\boldsymbol{W}\boldsymbol{S}]_d \tag{24}$$

$$p_c = [\boldsymbol{W}\boldsymbol{S}]_c \, . \tag{25}$$

For a simple case, consider a homogeneous network in which all cells have the same capacity and experience the same arrival and handoff rate (users are uniformly distributed). Then all the cells show the same performance parameters, in particular, blocking, dropping and handoff probabilities (denoted by p_b, p_d and p_h) are the same. Hong and Rappaport derived the following result using a direct approach [11] based on the number of possible handoffs,

$$p_d = \sum_{H=0}^{\infty} (p_h)^H (1-p_f)^{H-1} p_f = \frac{p_h p_f}{1-p_h(1-p_f)} \tag{26}$$

where H is the number of successful handoffs that a call makes before being dropped.

C. Call and Channel Holding Times Characterization

Inherited from the fixed telephony analysis, it is commonly assumed that call holding time and cell residence times in cellular networks are exponentially distributed. Although exponential distributions are not accurate in practice but the models based on the exponential assumption are typically tractable and do provide mean value analysis which indicates the system performance trend [23]. In this subsection we first investigate some of the results reported from field data analysis and detailed simulations regarding the call holding time and cell residency times. Then we turn our attention to some proposed models which are able to capture the observed statistical characteristics to some extent. A good model must be general enough to provide a good approximation of the field data, and must also be simple enough to enable us to obtain analytically tractable results for performance evaluation [24].

Using real measurements, Jedrzycki and Leung [15] showed that a lognormal distribution is a more accurate model for cell residency time. Based on simulations, Guerin [17] showed that for some cases the channel occupancy time distribution is quite close to exponential distribution but for the low rate of change of direction the channel occupancy time distribution shows rather poor agreement with the exponential distribution. Using detailed simulations based on cell geometries, Zonoozi and Dassanayake [16] concluded that the cell residency time is well described by a generalized gamma distribution but channel holding time remains exponential.

Gamma distribution is usually a good candidate for fitting a probability distribution to measured data. It can match the first two moments of the measured data and other distributions like exponential and Erlang are its special cases.

Typically, there is an interest in describing the call holding and cell residency times by a mixture of exponential distributions. The usefulness of this approach is that they may be broken down into stages and phases consisting of various exponential distributions and consequently are conveniently described by Markov chains [19].

Rappaport [25] used Erlang-k distributions to model holding times in a cellular network. Let $\{X_i\}_{i=1}^{k}$ denote a set of iid random variables with exponential distribution. Then $X = \sum_{i=1}^{k} X_i$ defines a random variable with Erlang-k distribution. Hyper-exponential distributions have been used in [13]. Let $\{X_i\}_{i=1}^{M}$ denote a set of exponentially distributed random variables with mean μ_i for X_i. Then $X = \sum_{i=1}^{M} \alpha_i X_i$ defines a random variable with hyper-exponential distribution where $\alpha_i \geq 0$ and $\sum_{i=1}^{M} \alpha_i = 1$. The sum of hyper-exponential (SOHYP) distributions was proposed by Orlik and Rappaport [13], [26] for modeling the holding times. The random variable $\sum_{i=1}^{N} X_i$ defines a SOHYP random variable where X_is have hyper-exponential distribution. They showed the generality of SOHYP models by showing that the coefficient of variance (the ratios of square root of variance to mean) can be adjusted to be less than, equal to or greater than unity.

Along the same approach, Fang et al. [24], [27] have investigated the so-called hyper-Erlang distribution which is less complicated than SOHYP distribution. Let $\{X_i\}_{i=1}^{M}$ denote a set of random variables with Erlang-k distribution. Then $X = \sum_{i=1}^{M} \alpha_i X_i$ defines a random variable with hyper-Erlang distribution where $\alpha_i \geq 0$ and $\sum_{i=1}^{M} \alpha_i = 1$. It can be shown that the set of all hyper-Erlang distributions is convex and can approximate any nonnegative random variable [24]. Particularly, hyper-Erlang distributions can be tuned to have coefficient of variance less than, equal to or greater than unity. Fang [24] claimed that hyper-Erlangs can even be tuned to approximate heavy-tailed distributions leading to long-range dependency and self-similarity [28]–[31]. Note that, hyper-Erlang includes exponential, Erlang and hyper-exponential as special cases.

Two shortcomings of mixed exponential models as pointed out by Rajaratnam and Takawira [32] are that they suffer from state space explosion and/or they represent handoff traffic as state-dependent mean arrival rate thus ignoring the higher moments of the handoff arrival process.

Instead, they proposed a model based on the application of gamma distribution for call and channel holding times characterization.

D. Handoff Arrival Process

Chlebus and Ludwin [33] reexamined the validity of Poisson arrivals for handoff traffic in a classical cellular network where everything is exponentially distributed. They concluded that handoff traffic is indeed Poisson in a nonblocking environment. However, they claimed that in a blocking environment handoff traffic is smooth. A smooth process is the one whose coefficient of variance is less than one. Similarly, Rajaratnam and Takawira [34] empirically showed that handoff traffic is a smooth process under exponential channel holding times. Using a solid mathematical framework, Fang et al. [12] proved that for exponential call holding times the merged traffic from new calls and handoff calls is Poisson if and only if the cell residence times are exponentially distributed.

Assume that the cellular network under investigation is uniform. Recall the new call channel holding time $t_{\rm nh}$ and handoff call channel holding time as given by (4) and (5). Let λ and ν denote the arrival rates for new calls and handoff calls, respectively. Let $t_{\rm ch}$ denote the channel holding time whether the call is a new call or a handoff call, thus

$$t_{\rm ch} = \frac{\lambda}{\lambda+\nu} t_{\rm nh} + \frac{\nu}{\lambda+\nu} t_{\rm hh}. \tag{27}$$

Referring to Fig. 2, let $f_c(t)$, $f(t)$, $f_r(t)$, $f_{\rm nh}(t)$, $f_{\rm hh}(t)$ and $f_{\rm ch}(t)$ denote, respectively, the probability density functions of t_c, t_m, r, $t_{\rm nh}$, $t_{\rm hh}$ and $t_{\rm ch}$ with their corresponding Laplace transforms $f_c^*(t)$, $f^*(t)$, $f_r^*(t)$, $f_{\rm nh}^*(t)$, $f_{\rm hh}^*(t)$ and $f_{\rm ch}^*(t)$, respectively. In [12], for a homogeneous network with exponentially distributed call holding times, the following results are obtained.

(i) The Laplace transform of the probability density function of the new call channel holding time is given by

$$f_{\rm nh}^*(s) = \frac{\mu}{s+\mu} + \frac{\eta s}{(s+\mu)^2}[1 - f^*(s+\mu)] \tag{28}$$

and the expected new call channel holding time is

$$E[t_{\rm nh}] = \frac{1}{\mu} - \frac{\eta}{\mu^2}[1 - f^*(\mu)]. \tag{29}$$

(ii) The Laplace transform of the probability density function of the handoff call channel holding time is given by

$$f^*_{\text{hh}}(s) = \frac{\mu}{s+\mu} + \frac{s}{s+\mu} f^*(s+\mu), \tag{30}$$

and the expected handoff call channel holding time is

$$E[t_{\text{hh}}] = \frac{1}{\mu}(1 - f^*(\mu)). \tag{31}$$

(iii) The Laplace transform of the probability density function of the channel holding time is given by

$$f^*_{\text{ch}} = \frac{\lambda}{\lambda+\nu} f^*_{\text{nh}} + \frac{\nu}{\lambda+\nu} f^*_{\text{hh}}, \tag{32}$$

and the expected channel holding time is

$$E[t_{\text{ch}}] = \frac{1}{\mu} - \frac{\lambda\eta}{(\lambda+\nu)\mu^2}\left[1 - \left(1 - \frac{\nu\mu}{\lambda\eta}\right) f^*(\mu)\right]. \tag{33}$$

(iv) The handoff call arrival rate ν is given by

$$\nu = -\eta(1-p_b)\lambda \sum_{p\in\sigma_c} \text{Res}_{s=p} \frac{1 - f^*(s)}{s^2[1-(1-p_f)f^*(s)]} f^*_c(-s), \tag{34}$$

where σ_c is the set of poles of $f^*_c(-s)$ on the right complex plane, $\text{Res}_{s=p}$ is the residue at a pole $s = p$, p_b and p_f are the new call blocking and handoff failure probabilities, respectively.

Since all the given Laplace transforms are in terms of rational functions, one can easily use partial fraction expansion to find the inverse Laplace transforms. Interested readers are referred to [35] for a combined analytical/simulation model with general mobility and call assumptions. For analytical results with generally distributed call holding and cell residency times refer to [36].

IV. Call Admission Control

Call admission control (CAC) is a technique to provide QoS in a network by restricting the access to network resources. Simply stated, an admission control mechanism accepts a new call request provided there are adequate free resources to meet the QoS requirements of the new call request without violating the committed QoS of already accepted calls. There is a tradeoff between the QoS level perceived by the user (in terms of the call dropping probability) and the utilization of scarce wireless resources. In fact, CAC can be described as an optimization problem as we see later in section VII.

We assume that available bandwidth in each cell is channelized and focus on call-level QoS measures. Therefore, the call blocking probability (p_b) and the call dropping probability (p_d) are the relevant QoS parameters in this paper. Three CAC related problems can be identified based on these two QoS parameters [37]:

1) **MINO:** Minimizing a linear objective function of the two probabilities (p_b and p_d).
2) **MINB:** For a given number of channels, minimizing the new call blocking probability subject to a hard constraint on the handoff dropping probability.
3) **MINC:** Minimizing the number of channels subject to hard constraints on the new and handoff calls blocking/dropping probabilities.

As mentioned before, channels could be frequencies, time slots or codes depending on the radio technology used. Each base station is assigned a set of channels and this assignment can be static or dynamic as described in section II.

MINO tries to minimize penalties associated with blocking new and handoff calls. Thus, MINO appeals to the network provider since minimizing penalties results in maximizing the net revenue. MINB places a hard constraint on handoff call blocking thereby guaranteeing a particular level of service to already admitted users while trying to maximize the net revenue. MINC is more of a network design problem where resources need to be allocated apriori based on, for example, traffic and mobility characteristics [37].

Since dropping a call in progress is more annoying than blocking a new call request, handoff calls are typically given higher priority than new calls in access to the wireless resources. This preferential treatment of handoffs increases the blocking of new calls and hence degrades the bandwidth utilization [38]. The most popular approach to prioritize handoff calls over new calls is by reserving a portion of available bandwidth in each cell to be used exclusively for handoffs.

In general there are two categories of CAC schemes in cellular networks:

1) *Deterministic CAC*: QoS parameters are guaranteed with 100% confidence [39], [40]. Typically, these schemes require extensive knowledge of the system parameters such as user mobility which is not practical, or sacrifice the scarce radio resources to satisfy the deterministic QoS bounds.
2) *Stochastic CAC*: QoS parameters are guaranteed with some probabilistic confidence [11],

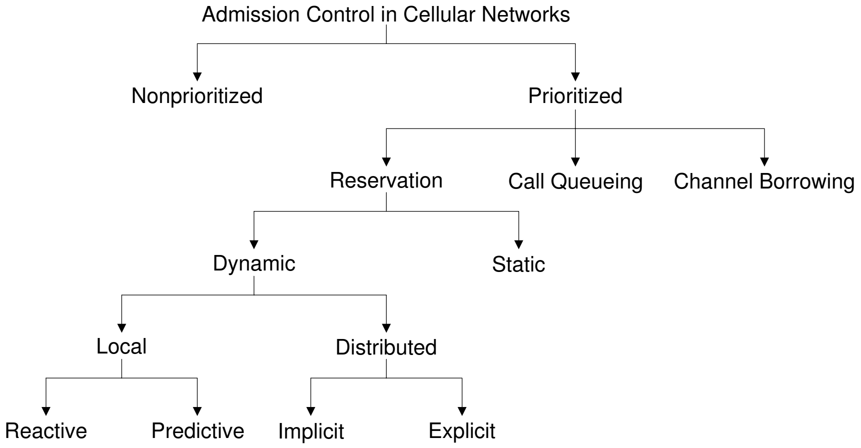


Fig. 3. Stochastic call admission control schemes in cellular networks.

[37], [41]. By relaxing QoS guarantees, these schemes can achieve a higher utilization than deterministic approaches.

Most of the CAC schemes which are investigated in this paper fall in the stochastic category. Fig. 3 depicts a classification of stochastic CAC schemes proposed for cellular networks. In the rest of this paper, we discuss each category in detail. In some cases, we will further expand this basic classification.

V. Prioritization Schemes

In this section we discuss different handoff prioritization schemes, focusing on reservation schemes. Channel borrowing, call queueing and reservation are studied as the most common techniques.

A. Channel Borrowing Schemes

In a channel borrowing scheme, a cell (an acceptor) that has used all its assigned channels can borrow free channels from its neighboring cells (donors) to accommodate handoffs [5], [42], [43]. A channel can be borrowed by a cell if the borrowed channel does not interfere with existing calls. When a channel is borrowed, several other cells are prohibited from using it. This is called channel locking and has a great impact on the performance of channel borrowing schemes [44]. The number of such cells depends on the cell layout and the initial channel allocation. For

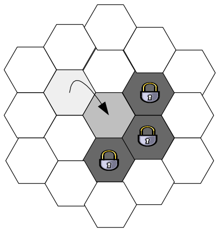


Fig. 4. Channel locking.

example, for a hexagonal planar layout with reuse distance of one cell, a borrowed channel is locked in three neighboring cells (see Fig. 4).

The proposed channel borrowing schemes differ in the way a free channel is selected from a donor cell to be borrowed by an acceptor cell. A complete survey on channel borrowing schemes is provided by Katzela and Naghshinehin [5].

B. Call Queueing Schemes

Queueing of handoff requests, when there is no channel available, can reduce the dropping probability at the expense of higher new call blocking. If the handoff attempt finds all the channels in the target cell occupied it can be queued. If any channel is released it is assigned to the next handoff waiting in the queue. Queueing can be done for any combination of new and handoff calls. The queue itself can be finite [45] or infinite [11]. Although finite queue systems are more realistic, systems with infinite queue are more convenient for analysis. Fig. 5 depicts a classification of call queueing schemes.

Hong and Rappaport [11] analyzed the performance of the simple *guard channel* scheme (see section V-C) with queueing of handoffs where handoff call attempts can be queued for the time duration in which a mobile dwells in the handoff area between cells. They used the FIFO queueing strategy and showed that queueing improves the performance of the pure guard channel scheme, i.e. p_d is lower for this scheme while there is essentially no difference for p_b.

The tolerable waiting time in queues is an important parameter. The performance of queueing schemes is affected by the reneging of queued new calls due to caller impatience and the dropping of queued handoff calls as they move out of the handoff area before the handoff is accomplished successfully. Chang et al. [45] analyzed a priority-based queueing scheme in which handoff

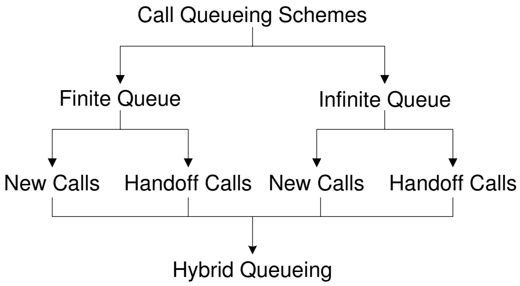


Fig. 5. Call queueing schemes.

calls waiting in queue have priority over new calls waiting in queue to gain access to available channels. They simply assumed that those calls waiting in queue can not handoff to another cell. Recently, Li and Chao [46] investigated a general modeling framework which can capture call queueing as well. They proved that the steady-state distribution of the equivalent queueing model has a product form solution. Queueing schemes have been mainly proposed for circuit-switched voice traffic. Their generalization to multiple classes of traffic is a challenging problem [47]. Lin and Lin [48] analyzed several channel allocation schemes including queueing of new and handoff calls. They concluded that the scheme with new and handoff calls queueing has the best performance.

C. Reservation Schemes

The notion of guard channels was introduced in the mid 80s as a call admission control mechanism to give priority to handoff calls over new calls. In this policy, a set of channels called the guard channels are permanently reserved for handoff calls. Hong and Rappaport [11] showed that this scheme reduces handoff dropping probability significantly compared to the nonprioritized case. They found that p_d decreases by a significantly larger order of magnitude compared to the increase of p_b when more priority is given to handoff calls by increasing the number of handoff channels.

Consider a cellular network with C channels in a given cell. The guard channel scheme (GC) reserves a subset of these channels, say $C - T$, for handoff calls. Whenever the channel occupancy exceeds a certain threshold T, GC rejects new calls until the channel occupancy goes below the threshold. Assume that the arrival process of new and handoff calls is Poisson with

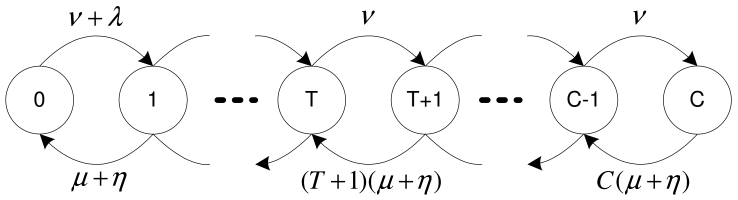


Fig. 6. State transition diagram of the guard channel scheme.

rate λ and ν, respectively. The call holding time and cell residency for both types of call is exponentially distributed with mean $1/\mu$ and $1/\eta$, respectively. Let $\rho = (\lambda + \nu)/(\mu + \eta)$ denote the traffic intensity. Further assume that the cellular network is homogeneous, thus a single cell in isolation is a representative for the network.

Define the state of a cell by the number of occupied channels in the cell. Therefore, the cell channel occupancy can be modeled by a continuous time Markov chain with C states. The state transition diagram of a cell with C channels and $C-T$ guard channels is shown in Fig. 6. Given this, it is straight forward to derive the steady-state probability P_n, that n channels are busy

$$P_n = \begin{cases} (\frac{\rho^n}{n!})P_0, & 0 \leq n \leq T \\ \rho^T(\frac{\nu^{n-T}}{n!})P_0, & T \leq n \leq C \end{cases} \tag{35}$$

where

$$P_0 = \left[\sum_{n=0}^{T} \frac{\rho^n}{n!} + \rho^T \sum_{n=T+1}^{C} \frac{\nu^{n-T}}{n!}\right]^{-1} \tag{36}$$

and then $p_b = \sum_{n=T+1}^{C} P_n$ and $p_f = P_C$.

However, Fang and Zhang [49] showed that when the mean cell residency times for new calls and handoff calls are significantly different (as is the case for non-exponential channel holding times), the traditional one-dimensional Markov chain model may not be suitable and a two-dimensional Markov model must be applied which is more complicated.

A critical parameter in this basic scheme is the optimal number of guard channels. In fact, there is a tradeoff between minimizing p_d and minimizing p_b. If the number of guard channels is conservatively chosen then admission control fails to satisfy the specified p_d. A static reservation typically results in poor resource utilization. To deal with this problem, several dynamic reservation schemes [41], [50]–[53] were proposed in which the optimal number of guard channels

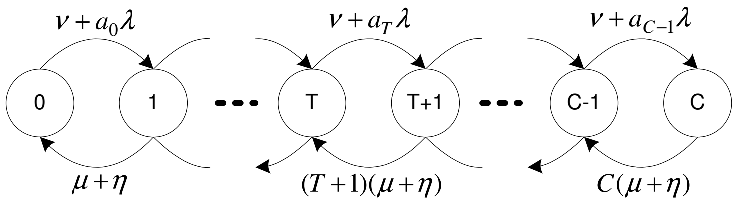


Fig. 7. State transition diagram of the fractional guard channel scheme.

is adjusted dynamically based on the observed traffic load and dropping rate in a control time window. If the observed dropping rate is above the guaranteed p_d then the number of reserved channels is increased. On the other hand, if the current dropping rate is far below the target p_d then the number of reserved channels is decreased. The next section investigates dynamic reservation schemes.

A different variation of the basic GC scheme is known as *fractional guard channel* (FGC) [37]. Whenever the channel occupancy exceeds the threshold T, the GC policy is to reject new calls until the channel occupancy goes below the threshold. In the fractional GC policy, new calls are accepted with a certain probability that depends on the current channel occupancy. Thus we have a randomization parameter which determines the probability of acceptance of a new call. Note that both GC and FGC policies accept handoff calls as long as there are some free channels. One advantage of FGC over GC is that it distributes the newly accepted calls evenly over time which leads to a more stable control [54].

The behavior of FGC in a cell with C channels is depicted in Fig. 7. Note that in state n, the acceptance ratio is a_n. Using balance equations, the steady-state probability of having n channels busy is given by

$$P_n = \frac{\prod_{i=0}^{n-1}(\nu + a_i\lambda)}{(\mu+\eta)^n} P_0, \quad 1 \leq n \leq C \tag{37}$$

where

$$P_0 = \left[1 + \sum_{n=1}^{C} \frac{\prod_{i=0}^{n-1}(\nu + a_i\lambda)}{(\mu+\eta)^n} \right]^{-1} . \tag{38}$$

Therefore, $p_b = \sum_{n=0}^{C}(1-a_n)P_n$ and $p_f = P_C$ where $a_C = 0$. Note that, GC is a special case of FGC where $a_i = 1$ for $0 \leq i \leq T-1$, and $a_i = 0$ for $T \leq i \leq C$.

It has been shown in [38] that due to advance reservation in reservation schemes the efficiency of cellular systems has an upper bound even if no constraint is specified on the call blocking probability. This upper bound is related to call and mobility characteristics through the mean number of handoffs per call. Moreover, the achievable efficiency decreases with decreasing cell size and with increasing call holding time [38].

VI. Dynamic Reservation Schemes

There are two approaches in dynamic reservation schemes: local and distributed (collaborative) depending on whether they use local information or gather information from neighbors to adjust the reservation threshold. In local schemes, each cell estimates the state of the network using local information only, while in distributed schemes each cell gathers network state information in collaboration with its neighboring cells.

A. Local Schemes

We categorize local admission control schemes into *reactive* and *predictive* schemes. By reactive approaches we refer to those admission policies that adjust their decision parameters, i.e. threshold and reservation level, as a result of an event such as call arrival, completion or rejection. Predictive approaches refer to those policies that predict future events and adjust their parameters in advance to prevent undesirable QoS degradations.

1) Reactive Approaches: The well-known guard channel (cell threshold, cut-off priority or trunk reservation) scheme (GC) is the first one in this category. GC has a reservation threshold and when the number of occupied channels reaches this threshold, no new call requests are accepted. One natural extension of this basic scheme is to use more than one threshold (e.g. two thresholds [50]) in order to have more control of the number of accepted calls. It has been shown [55] that the simple guard channel scheme performs remarkably well, often better than more complex schemes during periods in which the load does not differ from the expected level. For a discussion on different reservation strategies refer to [56] by Epstein and Schwartz.

2) Predictive Approaches: Local admission control schemes are very simple but they suffer from the lack of global information about the changes in network traffic. On the other hand, distributed admission control schemes have access to global traffic information at the expense of increased computational complexity and signaling overhead induced by information exchange between cells. To overcome the complexity and overhead associated with distributed schemes and benefit from the simplicity of local admission schemes, predictive admission control schemes were proposed. These schemes try to estimate the global state of the network by using some modeling/prediction technique based on information available locally.

Two different approaches can be distinguished in this category:

(i) *Structural (parameter-based) modeling :*

The changing traffic parameters such as call arrival and departure rates are locally estimated. Assume that the control mechanism periodically measures the arrival rate. Our goal is to compute the expected arrival rate from such online measurements. Typically, a simple exponentially weighted moving average (EWMA) is used for this purpose. Let $\hat{\lambda}(i)$ and $\lambda(i)$ denote the estimated and measured new call arrival rate at the beginning of control period i, respectively. Using EWMA technique, we have

$$\hat{\lambda}(i+1) = \epsilon\hat{\lambda}(i) + (1-\epsilon)\lambda(i), \tag{39}$$

where ϵ is the smoothing coefficient which must be properly selected. In general, a small value of ϵ (thus, a large value of $1-\epsilon$) can keep track of the changes more accurately, but is perhaps too heavily influenced by temporary fluctuations. On the other hand, a large value of ϵ is more stable but could be too slow in adapting to real traffic changes. This technique can be used to estimate the mean cell residency and call holding times as well. Then based on these parameters, a traffic model which can describe the channel occupancy in each cell is derived. Typically, several assumptions are made about traffic parameters in this approach which are necessary to have a tractable problem (for example see [11], [37], [41], [54]).

It is clear that the EWMA in (39) is a special case of the so-called *auto regressive moving average* (ARMA) model [57] in time series analysis. There is virtually no restriction on using more complicated (and perhaps more accurate) estimation techniques.

(ii) *Black-Box (measurement-based) modeling:*

Instead of looking at the individual components of traffic, this approach directly looks at the actual traffic. In other words, it tries to model the aggregated traffic without relying on the underlying arrival and departure processes. This approach has been proposed for multimedia systems where most of the assumptions of structural modeling are not valid [58]. The main advantage of this scheme is that it does not make any assumption about the distribution of new call arrival, handoff arrival, channel holding time and bandwidth requirements.

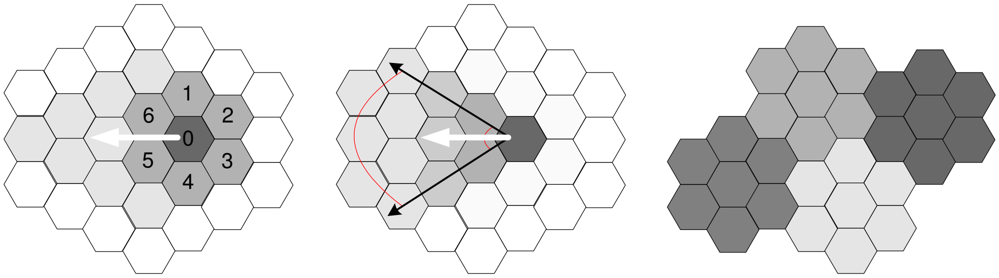
One of the key issues in this approach is to predict traffic in the next control time interval based on the online measurements of traffic characteristics. The goal is to forecast future traffic variations as precisely as possible, based on the measured traffic history. Traffic prediction requires accurate traffic models which can capture the statistical characteristics of actual traffic. Inaccurate models may overestimate or underestimate network traffic.

Recently, there has been a significant change in the understanding of network traffic. It has been found in numerous studies that data traffic in high-speed networks exhibits self-similarity [28]–[30] that can not be captured by classical models, hence self-similar models have been developed. Among these self-similar models, fractional ARIMA [59], [60] and fractional Brownian motion [61], [62] have been widely used for network traffic modeling and prediction.

Considering that future wireless networks will offer the same services to mobile users as their wireline counterparts, it is highly possible that traffic in these networks will also exhibit self-similarity (as reported for wireless data traffic by Jiang et al. [31]). Hence, simple modeling and prediction techniques may not be accurate. Admission control based on self-similar traffic models has been already investigated for wireline networks [63], [64]. Similar approaches may be applicable to cellular communications.

B. Distributed Schemes

The fundamental idea behind all distributed schemes [41], [51]–[54], [65], [66] is that every mobile terminal with an active wireless connection exerts an influence upon the cells in the vicinity of its current location and along its direction of travel [51]. A group of cells which are geographically or logically close together form a *cluster*, as shown in Fig. 8. Either each mobile terminal has its own cluster independent of other terminals or all the terminals in a cell share the same cluster. Typically, the admission decision for a connection request is made in cooperation with other cells of the cluster associated to the mobile terminal asking for admission. In Fig. 8(a) a cluster is defined assuming that a terminal affects all the cells in the vicinity of its current location and along its trajectory, while in Fig. 8(b) it is assumed that those cells that form a sector in the direction of mobile terminal's trajectory are most likely to be affected (visited) by the terminal. And, Fig. 8(c) shows a static cluster which is fixed regardless of the terminal



(a) Shadow cluster [51]. (b) Most likely cluster [66]. (c) Virtual connection tree [67].

Fig. 8. Three examples of cluster definition.

mobility.

Each user currently in the system may either remain in the cell it is in or move to a neighboring cell, hence it can be modeled using a binomial random variable. We approximate the joint behavior of binomial distributions with a normal distribution and hence, the number of active calls in a cell at any time follows a Gaussian distribution. Also, we neglect the possibility of users having moved a distance of two or more cells and of a user arriving/completing a call during a time interval of length T.

Now, consider a hexagonal cellular system similar to those depicted in Fig. 8. Assume that at time $t = t_0$ a new call has arrived. New calls are admitted into the system provided that the predicted handoff failure probability of any user in the home and neighboring cells at time $t = t_0 + T$ is below the target threshold P_{QoS}. Let $n_i(t)$ denote the number of active calls in cell i at time t. Assuming that handoff failure in each cell can be approximated by the overload probability, it is obtained that

$$p_f = \Pr(n(t_0 + T) > c) \tag{40}$$

Therefore the handoff failure in cell i is given by

$$\boldsymbol{P}_f(i) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{c_i - \mathrm{E}[n_i(t_0+T)]}{\sqrt{2\,\mathrm{Var}[n_i(t_0+T)]}}\right) \tag{41}$$

where c_i is the capacity of cell i and $\mathrm{erfc}(x)$ is the complementary error function defined as

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^{\infty} e^{-t^2}\,dt \tag{42}$$

and the expected and variance of the number of calls at time $t_0 + T$ in cell i is given by

$$\mathrm{E}[n_i(t_0 + T)] = n_i(t_0)p_s + p_h \textstyle\sum_{j=1}^{6} n_j(t_0), \tag{43}$$

$$\mathrm{Var}[n_i(t_0 + T)] = n_i(t_0)v_s + v_h \textstyle\sum_{j=1}^{6} n_j(t_0)\,. \tag{44}$$

Where, p_s is the probability of staying in the current cell and p_h is the probability of handing off to another cell during the time period T, which are given by

$$p_s = e^{-(\mu+h)T}, \quad p_h = \frac{1}{6}(1 - e^{-hT})\,. \tag{45}$$

Similarly, v_s and v_h are, respectively, the variances of binomial processes of stay and handoff with parameters p_s and p_h, which are expressed as

$$v_s = (1 - p_s)p_s, \quad v_h = (1 - p_h)p_h\,. \tag{46}$$

The idea of distributed admission control was originally proposed by Naghshineh and Schwartz [41]. They proposed a collaborative admission control known as *distributed call admission control* (DCAC). DCAC periodically gathers some information, namely the number of active calls, from the adjacent cells of the local cell to make the admission decision in combination with the local information. The analysis we presented earlier is slightly different from the original DCAC and is based on the work by Epstein and Schwartz [53]. DCAC is very restrictive in the sense that it takes into consideration information from direct neighbors only and assumes at most one handoff during the control period.

It has been shown that DCAC is not stable and violates the required dropping probability as the load increases [54]. Levin et al. [51] proposed a more complicated version of the original DCAC based on the *shadow cluster* concept, which uses dynamic clusters for each user based on its mobility pattern instead of restricting itself (as DCAC) to direct neighbors only. A practical limitation of the shadow cluster scheme in addition to its complexity and overhead is that it requires a precise knowledge of the mobile trajectory. The so-called *active mobile probabilities* and their characterization are very crucial to the CAC algorithm. Active mobile probabilities for each user give the projected probability of being active in a particular cell at a particular instance of time.

Wu et al. [54] proposed a dynamic, distributed and stable CAC scheme called SDCA which extends the basic DCAC [41] in several ways such as using a diffusion equation to describe

TABLE II

CLUSTER TYPE VS. CAC PERFORMANCE.

Cluster type	CAC efficiency	CAC complexity
Static	Moderate	Moderate
Dynamic	High	High

the evolution of the time-dependent occupancy distribution in a cell instead of the widely used Gaussian approximation. SDCA is a distributed version of the fractional guard channel in that it computes an acceptance ratio a_i for each cell i to be used for the current control period.

Consider the single-call transition probability $f_{ik}(t)$ that an ongoing call in cell i at the beginning of the control period ($t = 0$) is located in cell k at time t. This is in fact very similar to the active mobile probabilities introduced in [51]. For an effective control enforcing dropping probabilities in the order of 10^{-4} to 10^{-2}, essentially all calls handoff successfully. Wu et al. showed that for a uniform network with hexagonal cells, the probability of having n handoffs by time t, $q_n(t)$, takes the simple form

$$q_n(t) = \frac{1}{n!}\left(\frac{\eta t}{6}\right)^n e^{-(\mu+\eta)t} \,. \tag{47}$$

Hence $f_{ik}(t)$ is obtained by summing over all possible paths between i and k. For example $f_{ii}(t)$ can be expressed as

$$f_{ii}(t) = q_0(t) + 6q_2(t) + 12q_3(t) + \cdots \,. \tag{48}$$

Similar equations can be easily derived for $f_{ik}(t)$ [54]. Using these time-dependent transition probabilities Wu et al.computed the time-dependent mean and variance of the channel occupancy distribution, $P_{n_i}(t)$, in cell i at time t. By using a diffusion approximation [68], the authors were able to find the time-dependent handoff failure, $P_{f_i}(t)$, for each cell i. Hence, the average handoff failure probability over a control period of length T is found as

$$\tilde{P}_{f_i} = \frac{1}{T}\int_0^T P_{f_i}(t)\,dt \,. \tag{49}$$

Finally, the acceptance ratio a_i can be obtained by numerically solving the following equation [69]:

$$\tilde{P}_{f_i} = P_{\text{QoS}}, \quad 0 \le a_i \le 1 \,. \tag{50}$$

C. *Classification of Distributed Schemes*

Distributed CACs can be classified according to two factors:

1) Cluster definition
2) Information exchange and processing

A cluster can be either static or dynamic. In the static approach, the size and shape of the cluster is the same regardless of the network situation. In the dynamic approach however, shape and/or size of the cluster change according to the congestion level and traffic characteristics. The virtual connection tree of [67] is an example of a static cluster while the shadow cluster introduced in [51] is a dynamic cluster. A shadow cluster is defined for each individual mobile terminal based on its mobility information, e.g. trajectory, and changes as the terminal moves. It has been shown that it is not worth involving several cells in the admission control process when the network is not congested [70]. Table II shows a tradeoff between the cluster type and the corresponding CAC performance. Typically, dynamic clusters have a better performance at the expense of increased complexity.

In general, distributed CACs can be categorized into *implicit* or *explicit* based on the involvement of cells in the decision making process:

1) *Implicit Approach:* In this approach, all the necessary information is gathered from the neighboring cells, but the processing is local. The virtual connection tree concept introduced in [67] is an example of an implicitly distributed scheme. In this scheme each connection tree consists of a specific set of base stations where each tree has a network controller. The network controller is responsible for keeping track of the users and resources. Despite the fact that information is gathered from a set of neighboring cells, the final decision is made locally in the network controller.
2) *Explicit Approach:* In this approach, not only information is gathered from the neighboring cells, but also the neighboring cells are involved in the decision making process. The shadow cluster concept introduced in [51] is an example of an explicitly distributed scheme. In this scheme a cluster of cells, the shadow cluster, is associated with each mobile terminal in a cell. Upon admitting a new call, all the cells in the corresponding cluster calculate a preliminary response which after processing by the original cell will form the final decision.

TABLE III

COMPARISON OF DYNAMIC CAC SCHEMES.

CAC scheme		Efficiency	Overhead	Complexity	Adaptivity
Local	Reactive	Low	Low	Low	Moderate
	Predictive	Moderate	Low	Moderate	Moderate
Distributed	Implicit	High	Very High	High	High
	Explicit	High	High	Very High	High

TABLE IV

COMPARISON OF DISTRIBUTED CAC SCHEMES.

CAC scheme	Efficiency	Complexity	Stability
Basic distributed	Moderate	Moderate	Moderate
Shadow cluster	High	High	Moderate
Stable dynamic	Very High	High	High

Although it is theoretically possible to involve all the network cells in the admission control process, it is expensive and sometimes useless in practice. To consider the effect of all the cells, analytical approaches involve huge matrix exponentiations. In [54] and [22] two different approximation techniques have been proposed to compute these effects with a lower computational complexity.

Table III shows a comparison of different dynamic CAC schemes. In general, there is a tradeoff between the efficiency and the complexity of local and distributed schemes. Table IV compares three major distributed CAC schemes. In this table, *basic distributed* was proposed by Naghshineh and Schwartz [41], *shadow cluster* refers to the work of Levin et al. [51] and *stable dynamic* is due to Wu et al. [54].

VII. OPTIMAL CONTROL

Recall that a call admission policy is the set of decisions that indicate when a new call will be allocated a channel and when and existing call will be denied a handoff from one cell to another. In this section we investigate the optimal and near-optimal admission policies proposed for three admission problems defined in section IV, namely, MINO, MINB and MINC. Although optimal policies are more desirable, near-optimal policies are more useful in practice due to the

TABLE V

COMPARISON OF OPTIMAL CAC SCHEMES.

CAC scheme		Efficiency	Complexity
Optimal	Single service	High	High
	Multiple services	High	Very High
Near-Optimal	Single service	Moderate	Low
	Multiple services	Moderate	Moderate

complexity of optimal policies which usually leads to an intractable solution. Table V shows a comparison of optimal and near-optimal schemes.

Decision theoretic approaches based on *Markov decision process* (MDP) [71] have been extensively studied to find the optimal CAC policy using standard optimization techniques [72]. However, for simple cases such as the one of an isolated cell in a voice system, simple Markov chains have been applied successfully [37]. A Markov decision process is just like a Markov chain, except that the transition matrix depends on the action taken by the decision maker (CAC) at each time step. The CAC receives a reward, which depends on the action and the state. The goal is to find a policy which specifies which action to take in each state, so as to maximize some function (e.g. the mean or expected sum) of the sequence of rewards. A problem formulated as an MDP can be solved iteratively [73]. This is called policy iteration, and is guaranteed to converge to the unique optimal policy. The best theoretical upper bound on the number of iterations needed by policy iteration is exponential in the number of states. However, by formulating the problem as a linear programming problem, it can be proved that one can find the optimal policy in polynomial time.

A. Optimal CAC Schemes

1) Single Service Case: Ramjee et al. [37] showed that the well-known GC policy is optimal for the MINO problem and a restricted version of the FGC policy is optimal for the MINB and MINC problems. In their work, channel occupancy is described by a Markov chain similar to the one in section VI. Although admission policies derived from the MDP formulation of the CAC [74], [75] are optimal for the MINO problem, it has been shown that a dynamic guard channel scheme is more realistic and at the same time approaches the optimal solution [75], [76].

2) Multiple Services Case: Introducing multiple services changes the system behavior dramatically. In contrast to single service systems, GC is no longer optimal for the MINO problem. While the optimal admission policy for single service (voice) systems is computationally complex, for multiple services (multimedia) systems it is even more complicated and expensive. In this situation, a *semi-Markov decision process* (SMDP) has been applied successfully. Optimal policies are reported for multimedia traffic in [72], [77]–[80]. In particular, Choi et al. [81] presented a centralized CAC based on SMDP, Kwon et al. [77] and Yoon et al. [82] proposed distributed CAC schemes based on SMDP, all for non-adaptive multimedia applications. Xiao et al. [78] developed an optimal scheme using SMDP for adaptive multimedia applications. Adaptive multimedia applications can change their bit-rate to adapt to network resource availability.

B. Near-Optimal CAC Schemes

As mentioned before, when the state of the system can be modeled as a Markov process, there exist methods to calculate the optimal call admission policy using a Markov decision process. However, for systems with a large number of states (which grows exponentially with the cell capacity and known as the *curse of dimensionality*) this method is impractical since it requires solving large systems of linear equations. Therefore, methods which can calculate a near-optimal policy are proposed in the literature. In particular, near-optimal approaches based on Markov decision processes [83], *genetic algorithms* [84], [85], and *reinforcement learning* [86] have been proposed.

VIII. Other Admission Control Schemes

A. Multiple Services Schemes

Moving from single service systems to multiple services systems raises new challenges. Particularly, wireless resource management and admission control become more crucial for efficient use of wireless resources [39], [47], [53], [87], [88]. Despite the added complexity to control mechanisms, multiple services systems are typically more flexible in terms of resource management. Usually there are some low priority services, e.g. best effort service, which can utilize unused bandwidth. This bandwidth can be released and allocated to higher priority services upon request, e.g. when the system is fully loaded and a high priority handoff arrives. Fig. 9

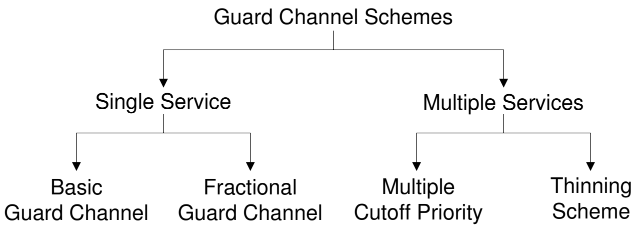


Fig. 9. Single service and multiple services guard channel schemes.

shows a classification of guard channel based CAC schemes in single service and multiple services systems. In the figure, *multiple cutoff priority* [47] and *thinning scheme* [88] are the multiple services counterparts of GC and FGC schemes in single service systems respectively.

In this context, the *thinning scheme* [88] is proposed as a generalization of the basic FGC for multiple classes prioritized traffic. Assume that the wireless network has call requests of r priority levels and each base station has C channels. Let α_{ij} ($i = 0, \ldots, C$ and $j = 1, \ldots, r$) denote the acceptance probabilities of prioritized classes respectively. When the number of busy channels at a base station is i, an arriving type-j call will be admitted with probability α_{ij}. All calls will be blocked when all channels are busy. Call arrivals of priority classes are independent of each other and assumed to be Poisson with rate λ_j for class j. Call durations are exponentially distributed with parameter μ. This system can be characterized by a Markov chain in which the state variable is the number of busy channels in the cell. Let P_n denote the stationary probability at state n, $\rho_j = \lambda_j/\mu$ and $\alpha_k = \sum_{j=1}^{r} \alpha_{kj}\rho_j$. Using balance equations we have

$$P_n = \frac{\prod_{k=0}^{n-1} \alpha_k}{n!} P_0, \tag{51}$$

where

$$P_0 = \left[\sum_{n=0}^{C} \left(\frac{\prod_{k=0}^{n-1} \alpha_k}{n!} \right) \right]^{-1} . \tag{52}$$

Then the blocking probability for class j is given by

$$P_b^j = \sum_{i=T+1}^{C} (1 - \alpha_{ij}) P_i \, . \tag{53}$$

Similarly, a natural extension to the basic GC can be achieved by setting different reservation thresholds for each class of service [47]. Pavlidou [89] analyzed an integrated voice/data cellular system using a two-dimensional Markov chain. Haung et al. [23] analyzed the *movable boundary*

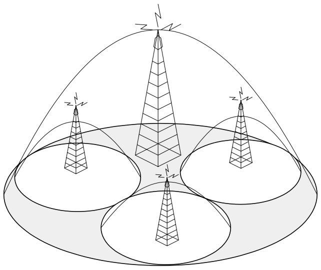


Fig. 10. A hierarchical system with micro/macro cells.

scheme with finite data buffering. In the movable boundary scheme, voice and data traffic each have a dedicated set of the available channels. Once dedicated channels are occupied, voice and data calls will compete for the shared channels. Wu et al. [87], [90] considered a different approach in which voice and data calls first compete for the shared channels and then will use dedicated channels, which can be considered as a natural extension of GC. Interested readers are referred to [91] for a discussion on fixed and movable boundary schemes. A general discussion on bandwidth allocation schemes for voice/data integrated systems can be found in [92].

B. Hierarchical Schemes

As mentioned earlier, micro-/pico-cell systems can improve spectrum efficiency better than macrocell systems because they can provide more spectrum resources per unit coverage area. However, micro-/pico- cell systems are not cost effective in areas with low user population (due to base station cost) and areas with high user mobility (leading to a large number of handoffs). As a consequence, hierarchical architectures [93]–[96] were proposed to take advantage of both macrocell and microcell systems. Fig. 10 shows an example of a hierarchical cellular system.

In this architecture, overlaid microcells cover high-traffic areas to enhance system capacity. Overlaying macrocells cover all of the area to provide general service in low-traffic areas and to provide channels for calls overflowing from the overlaid microcells. In particular, in a hierarchical system with an overflow scheme, it seems more significant to support guard channel for handoff protection and buffers for new and handoff calls in overlaying macrocells than to provide them in microcells [97]. In overflow schemes, when a call is rejected in a micro-cell, it is considered for admission by the macro-cell covering the micro-cell area.

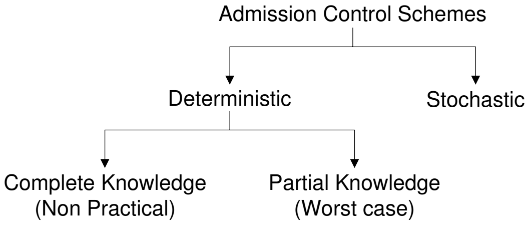


Fig. 11. Call admission control schemes.

Recently Marsan et al. [98] have investigated the performance of a hierarchical system under general call and channel holding time distributions. They used the idea of equivalent flow to break the mixed exponential process into independent exponential processes which can be then solved using classical Markov analysis.

C. Complete Knowledge Schemes

User mobility has an important impact in wireless networks. If the mobility pattern is partially [39], [40] or completely [99] known at the admission time then the optimal decision can be made rather easily.

Many researchers believe that it is not possible in general to have such mobility information at admission time. Even for indoor environments complete knowledge is not available [40]. Nevertheless, such an imaginary perfect knowledge scheme is helpful for benchmarking purposes [99]. Fig. 11 depicts a classification of CAC schemes according to their knowledge about user mobility. Partial knowledge schemes must reserve resources in several cells [39] to provide deterministic guarantees, hence we call them *worse case* schemes.

In addition to CAC schemes assuming deterministic mobility information, there is a large body of research work addressing the probabilistic estimation and prediction of mobility information. Some of them are heuristic-based [52], [66], [100], [101], some others are based on geometrical modeling of user movements [16] and street layouts [102], and some others are based on artificial intelligence techniques [103]. For instance, the distributed CACs introduced in section VI-B are based on probabilistic mobility information.

D. Economic Schemes

Economic models are widely discussed as a means for traffic management and congestion control in providers networks [104]–[106]. Through pricing, the network can send signals to users to change their behavior. It has been shown that for a given wireless network there exists a new call arrival rate which can maximize the total utility of users [106]. Based on this, the admission control mechanism can adjust the price dynamically according to the current network load in order to prevent congestion inside the network.

In terms of economics, utility functions describe user's level of satisfaction with the perceived QoS; the higher the utility, the more satisfied the users. It is sometimes useful to view the utility functions as of money a user is willing to pay for certain QoS. As mentioned earlier (see section IV), call blocking and dropping probabilities are the fundamental call-level QoS parameters in cellular networks. Let us define the QoS metric ϕ as a weighted sum of the call blocking and dropping probabilities as follows

$$\phi = \alpha p_b + \beta p_d, \tag{54}$$

where α and β are constants that denote the penalty associated with blocking a new call or dropping an ongoing call respectively (with $\beta > \alpha$ to reflect the costly call dropping). In section III, we showed that p_b and p_d are functions of new call and handoff call arrival rates λ and ν. Using (34), ν is itself a function of λ. Therefore

$$\phi = f(\lambda), \tag{55}$$

where f is a monotonic and nondecreasing function of λ. Let us define U as the user utility function in terms of the QoS metric ϕ, and let

$$U = g(\phi), \tag{56}$$

where g is a monotonic and nonincreasing function of ϕ. Therefore, the utility function U is maximized at $\phi = 0$. Let λ^* denote the optimal arrival rate for which U is maximized. In [106], it has been shown that the sufficient condition for λ^* is that

$$\left.\frac{dU}{d\lambda}\right|_{\lambda=\lambda^*} = 0\,. \tag{57}$$

Using the optimal arrival rate λ^* obtained using (57), we can characterize a pricing function to achieve the maximum utilization. Let $p(t)$ denote the price charged to users at time t. Define

$H(t)$ as the percentage of users who will accept the price at time t, then

$$\lambda_{\text{in}}(t) = \big(\lambda(t) + \nu(t)\big)H(t), \quad 0 \leq H(t) \leq 1 \tag{58}$$

where $\lambda_{\text{in}}(t)$ is the actual new call arrival rate at time t. $H(t)$ must be designed in such a way that always

$$\lambda_{\text{in}}(t) \leq \lambda^*, \tag{59}$$

and consequently

$$H(t) \leq \min\left\{1, \tfrac{\lambda^*}{\lambda(t)+\nu(t)}\right\}. \tag{60}$$

As mentioned before, pricing can influence the way the users use resources and is usually characterized by demand functions. A simple demand function can be characterized as follows [106],

$$D(t) = e^{-\left(\frac{p(t)}{p_0}-1\right)^2}, \quad p(t) \geq p_0 \tag{61}$$

where p_0 is the normal price. In fact, $D(t)$ denotes the percentage of users that will accept the price $p(t)$. In order to realize control function $H(t)$ we should have $H(t) = D(t)$. Using (60) and (61), the price that should be set at time t to obtain the desired QoS can be expressed as

$$p(t) = p_0\left(1 + \sqrt{\max\left\{0, -\ln\tfrac{\lambda^*}{\lambda(t)+\nu(t)}\right\}}\right). \tag{62}$$

It is worth noting that pricing-based control assumes that network users are sensitive and responsive to price changes. If this is not true for a particular network, e.g. noncommercial networks, then price-based control can not be applied.

IX. Conclusion

Due to the unique characteristics of mobile cellular networks, mainly mobility and limited resources, the wireless resource management problem has received tremendous attention. As a result, a large body of work has been done extending earlier work in fixed networks as well as introducing new techniques. A large portion of this research has been in the area of call admission control. In this paper, we have provided a survey of the major call admission control approaches and related issues for designing efficient schemes. A broad and detailed categorization of the existing CAC schemes was presented. For each category, we explained the main idea and described the proposed approaches for realizing it and identified their distinguishing features.

We have compared the various schemes based on some of the most important criteria including efficiency, complexity, overhead, adaptivity and stability. We believe that this article, which is the first comprehensive survey on this subject, can help other researchers in identifying challenges and new research directions in the area of call admission control for cellular networks.

One of the interesting observations stemming from this study and illustrated in [76] is how comparable is the performance of simple reservation-based CAC schemes, e.g. GC, to more complex ones. This is particularly true when the traffic conditions are known a priori [55]. Yet, a large body of research in this area focused on designing more and more sophisticated schemes in the hope of improving the CAC performance. Many assumptions about mobility and traffic characteristics made in CAC related research are often not practical. Therefore, most of the schemes proposed in the literature are difficult to deploy in current and future cellular systems. Furthermore, most of the researchers in the area developed their own simulation environments, making it difficult to reproduce and compare the results. There is a clear lack of implementation and testing of CAC schemes in more realistic situations.

Some of the lessons learned from surveying and analyzing the literature and from which recommendations can be drawn are as follows:

- To use more realistic (non-exponential) mobility and traffic (packet-based) models in designing and analyzing CAC schemes. New mobility models may not necessarily preserve the Markovian property. Meanwhile, new traffic modeling and engineering techniques are aiming at a more accurate description of traffic dynamics not only at call level but at packet level as well. In this perspective, recent findings in traffic analysis such as self-similarity [28]–[31] must be taken into consideration. To avoid complex schemes and eliminate impractical assumptions about traffic and mobility, measurement-based CAC schemes [107]–[109] must be further studied for wireless cellular networks.
- To apply cross layer design [110] in order to improve the performance of CAC schemes and achieve bit-level, packet-level and call-level QoS. In particular, scheduling mechanisms at packet level and control mechanisms at call level can benefit from the information about the state of the wireless channel to achieve a superior performance.
- To design CAC schemes for multiple services networks so as to support emerging multimedia services. Efficient sharing of wireless resources between multiple services is of paramount importance. However, the design and analysis of efficient CAC schemes for such

multiple services networks is much more complicated than that of single service networks.

- To consider heterogeneous networks and interoperability issues in order to achieve global roaming and quality of service end-to-end. A key aspect is seamless handoff among possibly different networks ranging from reliable and managed cellular networks to unreliable and unmanaged wireless LANs. A CAC scheme must be able to communicate with other control components of the network through standard mechanisms to provide end-to-end QoS guarantees. The current trend is towards IP-based architectures and mechanisms for achieving integrated wireless networks [111] and for better resource sharing. This is demonstrated by the increasing research interest in all-IP wireless network architectures [112].

We believe that the most challenging problems to be solved are mobility and wireless channel effects, particularly when considering multiple services networks. Mobility and wireless channel impacts on call-level, packet-level and bit-level system dynamics complicate significantly the modeling of cellular networks traffic which is essential for devising the appropriate CAC schemes. As discussed previously, measurement-based admission control is a promising approach to overcome the complexity of the CAC problem and alleviate some of the impractical assumptions about traffic and mobility.

In conclusion, CAC research remains an exciting area. The state of the art in CAC research suggests that existing CAC schemes cannot handle many of the challenges inherent to future heterogeneous multi services wireless networks. CAC research should continue, but must bear in mind the realistic limits that may be imposed by the inherent nature of the wireless channel, the traffic characteristics and the impact of user mobility. For that to be achieved, the nature of those challenges must be better understood.

References

[1] J.-Z. Sun, J. Sauvola, and D. Howie, “Features in future: 4G visions from a technical perspective,” in *Proc. IEEE GLOBECOM’01*, vol. 6, San Antonio, USA, Nov. 2001, pp. 3533–3537.

[2] U. Varshney and R. Jain, “Issues in emerging 4G wireless networks,” *IEEE Computer*, vol. 34, no. 6, pp. 94–96, June 2001.

[3] S. Y. Hui and K. H. Yeung, “Challenges in the migration to 4G mobile systems,” *IEEE Computer*, vol. 41, no. 12, pp. 54–59, Dec. 2003.

[4] T. Zahariadis and D. Kazakos, “(R)evolution toward 4G mobile communication systems,” *IEEE Wireless Communications Magazine*, vol. 10, no. 4, pp. 6–7, Aug. 2003.

[5] I. Katzela and M. Naghshineh, “Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey,” *IEEE Personal Communications Magazine*, vol. 3, no. 3, pp. 10–31, June 1996.

[6] Y. Iraqi and R. Boutaba, “Resource management issues in future wireless multimedia networks,” *J. High Speed Networking*, vol. 9, no. 3-4, pp. 231–260, 2000.

[7] M. Zhang and T.-S. P. Yum, “Comparisons of channel assignment strategies in cellular mobile telephone systems,” *IEEE Transactions on Vehicular Technology*, vol. 38, no. 4, pp. 211–215, Nov. 1989.

[8] D. Wong and T. J. Lim, “Soft handoffs in CDMA mobile systems,” *IEEE Personal Communications Magazine*, vol. 4, no. 6, pp. 6–17, Dec. 1997.

[9] R. Prakash and V. V. Veeravalli, “Locally optimal soft handoff algorithms,” *IEEE Transactions on Vehicular Technology*, vol. 52, no. 2, pp. 231–260, Mar. 2003.

[10] Y.-B. Lin and A.-C. Pang, “Comparing soft and hard handoffs,” *IEEE Transactions on Vehicular Technology*, vol. 49, no. 3, pp. 792–798, May 2000.

[11] D. Hong and S. S. Rappaport, “Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures,” *IEEE Transactions on Vehicular Technology*, vol. 35, no. 3, pp. 77–92, Aug. 1986, see also: CEAS Tech. Rep. No. 773, College of Engineering and Applied Sciences, State University of New York, June 1999.

[12] Y. Fang, I. Chlamtac, and Y.-B. Lin, “Channel occupancy times and handoff rate for mobile computing and PCS networks,” *IEEE Transactions on Computers*, vol. 47, no. 6, pp. 679–692, June 1998.

[13] P. V. Orlik and S. S. Rappaport, “A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions,” *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 788–803, June 1998.

[14] Y. Fang, I. Chlamtac, and Y.-B. Lin, “Call performance for a PCS network,” *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 8, pp. 1568–1581, Oct. 1997.

[15] C. Jedrzycki and V. C. M. Leung, “Probability distribution of channel holding time in cellular telephone systems,” in *Proc. IEEE VTC’96*, vol. 1, Atlanta, GA, May 1996, pp. 247–251.

[16] M. M. Zonoozi and P. Dassanayake, “User mobility modeling and characterization of mobility patterns,” *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 7, pp. 1239–1252, Sept. 1997.

[17] R. Guerin, “Channel occupancy time distribution in a cellular radio system,” *IEEE Transactions on Vehicular Technology*, vol. 35, no. 3, pp. 89–99, 1987.

[18] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1965.

[19] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, 3rd ed. John Wiley & Sons, Inc., 1998.

[20] F. P. Kelly, “Fixed point models of loss networks,” *Australian Mathematical Society*, vol. 31, pp. 204–218, 1989.

[21] C. M. Grinstead and J. L. Snell, *Introduction to Probability*, 2nd ed. American Mathematical Society, 1997.

[22] K. Mitchell and K. Sohraby, “An analysis of the effects of mobility on bandwidth allocation strategies in multi-class cellular wireless networks,” in *Proc. IEEE INFOCOM’01*, vol. 2, Anchorage, USA, Apr. 2001, pp. 1005–1011.

[23] Y.-R. Haung, Y.-B. Lin, and J.-M. Ho, “Performance analysis for voice/data integration on a finite-buffer mobile system,” *IEEE Transactions on Vehicular Technology*, vol. 49, no. 2, pp. 367–378, Mar. 2000.

[24] Y. Fang, “Hyper-Erlang distribution model and its application in wireless mobile networks,” *Wireless Networks*, vol. 7, no. 3, pp. 211–219, May 2001.

[25] S. S. Rappaport, “Blocking, hand-off and traffic performance for cellular communications with mixed platforms,” in *Proc. Inst. Elect. Eng.*, vol. 140, no. 5, 1998, pp. 389–401.

[26] P. V. Orlik and S. S. Rappaport, “Traffic performance and mobility modeling of cellular communications with mixed platforms and highly variable mobilities,” in *Proc. of the IEEE*, vol. 86, July 1998, pp. 1464–1479.

[27] Y. Fang and I. Chlamtac, “A new mobility model and its application in the channel holding time characterization in pcs networks,” in *Proc. IEEE INFOCOM’99*, vol. 1, New York, USA, Mar. 1999, pp. 20–27.

[28] W. E. Leland, M. Taque, W. Willinger, and D. Wilson, “On the self-similar nature of Ethernet traffic (extended version),” *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[29] M. E. Crovella and A. Bestavros, “Self-similarity in world wide web traffic: Evidence and possible causes,” *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

[30] J. Beran *et al.*, “Long-range dependence in variable-bit-rate video traffic,” *IEEE Transactions on Communications*, vol. 43, no. 2, pp. 1566–1579, Feb. 1995.

[31] M. Jiang, M. Nikolic, S. Hardly, and L. Trajkovic, “Impact of self-similarity on wireless data network performance,” in *Proc. IEEE ICC’01*, Helsinki, Finland, June 2001, pp. 477–481.

[32] M. Rajaratnam and F. Takawira, “Handoff traffic characterization in cellular networks under nonclassical arrivals and service time distributions,” *IEEE Transactions on Vehicular Technology*, vol. 50, no. 4, pp. 954–970, July 2001.

[33] E. Chlebus and W. Ludwin, “Is handoff traffic really Poissonian?” in *Proc. IEEE ICUPC’95*, Tokyo, Japan, Nov. 1995, pp. 348–353.

[34] M. Rajaratnam and F. Takawira, “Nonclassical traffic modeling and performance analysis of cellular mobile networks with and without channel reservation,” *IEEE Transactions on Vehicular Technology*, vol. 49, no. 3, pp. 817–834, May 2000.

[35] H. Zeng, Y. Fang, and I. Chlamtac, “Call blocking performance study for PCS networks under more realistic mobility assumptions,” *Telecommunication Systems*, vol. 19, no. 2, pp. 125–146, Feb. 2002.

[36] Y. Fang and I. Chlamtac, “Analytical generalized results for handoff probability in wireless networks,” *IEEE Transactions on Communications*, vol. 50, no. 3, pp. 396–399, Mar. 2002.

[37] R. Ramjee, D. Towsley, and R. Nagarajan, “On optimal call admission control in cellular networks,” *Wireless Networks*, vol. 3, no. 1, pp. 29–41, Mar. 1997.

[38] A. G. Valko and A. T. Campbell, “An efficiency limit of cellular mobile systems,” *Computer Communications Journal*, vol. 23, no. 5-6, pp. 441–451, Mar. 2000.

[39] A. K. Talukdar, B. Badrinath, and A. Acharya, “Integrated services packet networks with mobile hosts: Architecture and performance,” *Wireless Networks*, vol. 5, no. 2, pp. 111–124, 1999.

[40] S. Lu and V. Bharghavan, “Adaptive resource management algorithms for indoor mobile computing environments,” in *Proc. ACM SIGCOMM’96*, Palo Alto, USA, Aug. 1996, pp. 231–242.

[41] M. Naghshineh and M. Schwartz, “Distributed call admission control in mobile/wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 4, pp. 711–717, May 1996.

[42] C.-J. Chang, P.-C. Huang, and T.-T. Su, “A channel borrowing scheme in a cellular radio system with guard channels and finite queues,” in *Proc. IEEE ICC’96*, vol. 2, Dallas, USA, June 1996, pp. 1168–1172.

[43] X. Wu and K. L. Yeung, “Efficient channel borrowing strategy for multimedia wireless networks,” in *Proc. IEEE GLOBECOM’98*, vol. 1, Sydney, Australia, Nov. 1998, pp. 126–131.

[44] T.-P. Chu and S. S. Rappaport, "Generalized fixed channel assignment in microcellular communication systems," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 3, pp. 713–721, Aug. 1994.

[45] C.-J. Chang, T.-T. Su, and Y.-Y. Chiang, "Analysis of a cutoff priority cellular radio system with finite queueing and reneging/dropping," *IEEE/ACM Transactions on Networking*, vol. 2, no. 2, pp. 166–175, Apr. 1994.

[46] W. Li and X. Chao, "Modeling and performance evaluation of a cellular mobile network," *IEEE/ACM Transactions on Networking*, vol. 12, no. 1, pp. 131–145, Feb. 2004.

[47] B. Li, S. Chanson, and C. Lin, "Analysis of a hybrid cutoff priority scheme for multiple classes of traffic in multimedia wireless networks," *Wireless Networks*, vol. 4, no. 4, pp. 279–290, July 1998.

[48] P. Lin and Y.-B. Lin, "Channel allocation for GPRS," *IEEE Transactions on Vehicular Technology*, vol. 50, no. 2, pp. 375–384, Mar. 2001.

[49] Y. Fang and Y. Zhang, "Call admission control schemes and performance analysis in wireless mobile networks," *IEEE Transactions on Vehicular Technology*, vol. 51, no. 2, pp. 371–382, Mar. 2002.

[50] J. R. Moorman and J. W. Lockwood, "Wireless call admission control using threshold access sharing," in *Proc. IEEE GLOBECOM'01*, vol. 6, San Antonio, USA, Nov. 2001, pp. 3698–3703.

[51] D. Levine, I. Akyildiz, and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 1–12, Feb. 1997.

[52] S. Choi and K. G. Shin, "Predictive and adaptive bandwidth reservation for handoffs in QoS-sensitive cellular networks," in *Proc. ACM SIGCOMM'98*, vol. 27, Vancouver, Canada, Oct. 1998, pp. 155–166.

[53] B. M. Epstein and M. Schwartz, "Predictive QoS-based admission control for multiclass traffic in cellular wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 523–534, Mar. 2000.

[54] S. Wu, K. Y. M. Wong, and B. Li, "A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 2, pp. 257–271, Apr. 2002.

[55] J. M. Peha and A. Sutivong, "Admission control algorithms for cellular systems," *Wireless Networks*, vol. 7, no. 2, pp. 117–125, Mar. 2001.

[56] B. Epstein and M. Schwartz, "Reservation strategies for multi-media traffic in a wireless environment," in *Proc. IEEE VTC'95*, vol. 1, Chicago, USA, July 1995, pp. 165–169.

[57] G. E. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, 2nd ed. Holden-Day, 1976.

[58] T. Zhang, E. Berg, J. Chennikara, P. Agrawal, J. C. Chen, and T. Kodama, "Local predictive resource reservation for handoff in multimedia wireless IP networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 1931–1941, Oct. 2001.

[59] J. R. M. Hosking, "Fractional differencing," *Biometrika*, vol. 83, no. 1, pp. 165–176, Apr. 1981.

[60] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed. Springer-Verlag, 1991.

[61] G. Gripenberg and I. Norros, "On the prediction of fractional brownian motion," *Journal of Applied Probability*, vol. 33, pp. 400–410, 1996.

[62] I. Norros, "On the use of fractional brownian motion in the theory of connectionless networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 953–962, Aug. 1995.

[63] Y. Shu, Z. Jin, J. Wang, and O. W. Yang, "Prediction-based admission control using FARIMA models," in *Proc. IEEE ICC'00*, vol. 3, New Orleans, USA, June 2000, pp. 1325–1329.

[64] Y. Shu *et al.*, "Traffic prediction using FARIMA models," in *Proc. IEEE ICC'99*, vol. 2, Vancouver, Canada, June 1999, pp. 891–895.

[65] C. Oliveira, J. B. Kim, and T. Suda, "An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 6, pp. 858–874, Aug. 1998.

[66] A. Aljadhai and T. F. Znati, "Predictive mobility support for QoS provisioning in mobile wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 1915–1930, Oct. 2001.

[67] A. Acampora and M. Naghshineh, "An architecture and methodology for mobile-executed handoff in cellular ATM networks," *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 8, pp. 1365–1375, Oct. 1994.

[68] S. M. Ross, *Stochastic Process*, 2nd ed. American Mathematical Society, 1997.

[69] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1992.

[70] Y. Iraqi and R. Boutaba, "When is it worth involving several cells in the call admission control process for multimedia cellular networks?" in *Proc. IEEE ICC'01*, vol. 2, Helsinki, Finland, June 2001, pp. 336–340.

[71] M. L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 1994.

[72] Z. Haas, J. Y. Halpern, L. Li, and S. B. Wicker, "A decision-theoretic approach to resource allocation in wireless multimedia networks," in *Proc. ACM 4th Workshop Discrete Alg. Mobile Comput. Commun.*, Boston, USA, Aug. 2000, pp. 86–95.

[73] H. C. Tijms, *Stochastic Modeling and Analysis: A Computational Approach*. John Wiley & Sons, 1989.

[74] M. Saquib and R. Yates, "Optimal call admission to a mobile cellular network," in *Proc. IEEE VTC'95*, vol. 1, Chicago, USA, July 1995, pp. 190–194.

[75] D. Chen, S. B. Hee, and K. S. Trivedi, "Optimal call admission control policy for wireless communication networks," in *Proc. International Conference on Information, Communication and Signal processing ICICS'01*, Singapore, Dec. 2001.

[76] Q. Gao and A. Acampora, "Performance comparisons of admission control strategies for future wireless networks," in *Proc. IEEE WCNC'02*, vol. 1, Orlando, USA, Mar. 2002, pp. 317–321.

[77] T. Kwon, Y. Choi, and M. Naghshineh, "Optimal distributed call admission control for multimedia services in mobile cellular networks," in *Proc. Mobile Multimedia Ccommunication MoMuC'98*, Berlin, Germany, Oct. 1998.

[78] Y. Xiao, C. L. P. Chen, and Y. Wang, "An optimal distributed call admission control for adaptive multimedia in wireless/mobile networks," in *Proc. IEEE MASCOTS'00*, San Francisco, USA, Aug. 2000, pp. 477–482.

[79] T. Kwon, Y. Choi, and M. Naghshineh, "Call admission control for adaptive multimedia in wireless/mobile networks," in *Proc. ACM WOWMOM'98*, Dallas, USA, Oct. 1998, pp. 111–116.

[80] T. Kwon, I. Park, Y. Choi, and S. Das, "Bandwidth adaptation algorithms with multi-objectives for adaptive multimedia services in wireless/mobile networks," in *Proc. ACM WOWMOM'99*, Seattle, USA, Aug. 1999, pp. 51–59.

[81] J. Choi, T. Kwon, Y. Choi, and M. Naghshineh, "Call admission control for multimedia service in mobile cellular networks: A markov decision approach," in *Proc. IEEE ISCC'00*, Antibes, France, July 2000, pp. 594–599.

[82] I.-S. Yoon and B. G. Lee, "A distributed dynamic call admission control that supports mobility of wireless multimedia users," in *Proc. IEEE ICC'99*, Vancouver, Canada, June 1999, pp. 1442–1446.

[83] T. Kwon, J. Choi, Y. Choi, and S. Das, "Near optimal bandwidth adaptation algorithm for adaptive multimedia services in wireless/mobile networks," in *Proc. IEEE VTC'99*, vol. 2, Amsterdam, Netherlands, Sept. 1999, pp. 874–878.

[84] A. Yener and C. Rose, "Near optimal call admission policies for cellular networks using genetic algorithms," in *Proc. IEEE Wireless'94*, Calgary, Canada, July 1994, pp. 398–410.

[85] Y. Xiao, C. L. P. Chen, and Y. Wang, "A near optimal call admission control with genetic algorithm for multimedia services in wireless/mobile networks," in *Proc. IEEE NAECON'00*, Dayton, USA, Oct. 2000, pp. 787–792.

[86] E.-S. El-Alfy, Y.-D. Yao, and H. Heffes, “A learning approach for call admission control with prioritized handoff in mobile multimedia networks,” in *Proc. IEEE VTC’01*, vol. 2, Rhodes, Greece, May 2001, pp. 972–976.

[87] B. Li, L. Li, B. Li, and X.-R. Cao, “On handoff performance for an integrated voice/data cellular system,” *Wireless Networks*, vol. 9, no. 4, pp. 393–402, July 2003.

[88] Y. Fang, “Thinning schemes for call admission control in wireless networks,” *IEEE Transactions on Computers*, vol. 52, no. 5, pp. 686–688, May 2003.

[89] F.-N. Pavlidou, “Two-dimensional traffic models for cellular mobile systems,” *IEEE Transactions on Communications*, vol. 42, no. 2/3/4, pp. 1505–1511, 1994.

[90] H. Wu, L. Li, B. Li, L. Yin, I. Chlamtac, and B. Li, “On handoff performance for an integrated voice/data cellular system,” in *Proc. IEEE PIMRC’02*, vol. 5, Lisboa, Portugal, Sept. 2002, pp. 2180–2184.

[91] J. E. Wieselthier and A. Ephremides, “Fixed- and movable-boundary channel-access schemes for integrated voice/data wireless networks,” *IEEE Transactions on Communications*, vol. 43, no. 1, pp. 64–74, Jan. 1995.

[92] M. C. Young and Y.-R. Haung, “Bandwidth assignment paradigms for broadband integrated voice/data networks,” *Computer Communications Journal*, vol. 21, no. 3, pp. 243–253, 1998.

[93] C.-L. I, L. J. Greenstein, and R. D. Gitlin, “A microcell/macrocell architecture for low and high mobility wireless users,” *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 6, pp. 885–891, Aug. 1993.

[94] S. S. Rappaport and L.-R. Hu, “Microcellular communication systems with hierarchical macrocell overlays: Traffic performance models and analysis,” in *Proc. of the IEEE*, vol. 82, Sept. 1994, pp. 1383–1397.

[95] L.-R. Hu and S. S. Rappaport, “Personal communication systems using multiple hierarchical cellular overlays,” *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 2, pp. 406–415, Feb. 1995.

[96] K. L. Yeung and S. Nanda, “Channel management in microcell/macrocell cellular radio systems,” *IEEE Transactions on Vehicular Technology*, vol. 45, no. 4, pp. 601–612, Nov. 1996.

[97] C. Chang, C. J. Chang, and K.-R. Lo, “Analysis of a hierarchical cellular system with reneging and dropping for waiting new and handoff calls,” *IEEE Transactions on Vehicular Technology*, vol. 48, no. 4, pp. 1080–1091, July 1999.

[98] M. A. Marsan, G. Ginella, R. Maglione, and M. Meo, “Performance analysis of hierarchical cellular networks with generally distributed times and dwell times,” *IEEE Transactions on Wireless Communications*, vol. 3, no. 1, pp. 248–257, Jan. 2004.

[99] R. Jain and E. W. Knightly, “A framework for design and evaluation of admission control algorithms in multi-service mobile networks,” in *Proc. IEEE INFOCOM’99*, vol. 3, New York, USA, Mar. 1999, pp. 1027–1035.

[100] F. Yu and V. C. Leung, “Mobility-based predictive call admission control and bandwidth reservation in wireless cellular networks,” in *Proc. IEEE INFOCOM’01*, vol. 1, Anchorage, USA, Apr. 2001, pp. 518–526.

[101] S. Lim, G. Cao, and C. Das, “An admission control scheme for QoS-sensitive cellular networks,” in *Proc. IEEE WCNC’02*, vol. 1, Orlando, USA, Mar. 2002, pp. 296–300.

[102] W.-S. Soh and H. S. Kim, “Qos provisioning in cellular networks based on mobility prediction techniques,” *IEEE Communications Magazine*, vol. 41, no. 1, pp. 86–92, Jan. 2003.

[103] X. Shen, J. W. Mark, and J. Ye, “User mobility profile prediction: An adaptive fuzzy inference approach,” *Wireless Networks*, vol. 6, no. 5, pp. 363–374, 2000.

[104] C. Evci and B. Fino, “Spectrum management, pricing, and efficiency control in broadband wireless communications,” in *Proc. of the IEEE*, vol. 89, Jan. 2001, pp. 105–115.

[105] T. Heikkinen, “Congestion based pricing in a dynamic wireless network,” in *Proc. IEEE VTC’01*, vol. 2, Rhodes, Greece, May 2001, pp. 1073–1076.

[106] J. Hou, J. Yang, and S. Papavassiliou, “Integration of pricing with call admission control for wireless networks,” in *Proc. IEEE VTC’01*, vol. 3, Atlantic City, USA, Oct. 2001, pp. 1344–1348.

[107] S. Jamin *et al.*, “A measurement-based admission control algorithm for integrated services packet networks,” *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 524–540, Feb. 1997.

[108] J. Kim and A. Jamalipour, “Traffic management and QoS provisioning in future wireless IP networks,” *IEEE Pers. Commun.*, vol. 8, no. 5, pp. 46–55, Oct. 2001.

[109] A. Jamalipour and J. Kim, “Measurement-based admission control scheme with priority and service classes for application in wireless IP networks,” *J. Communication Systems*, vol. 16, no. 6, pp. 535–551, May 2003.

[110] G. Carneiro, J. Ruela, and M. Ricardo, “Cross layer design in 4G wireless terminals,” *IEEE Wireless Communications Magazine*, vol. 11, no. 2, pp. 7–13, Apr. 2004.

[111] IEEE Wireless Communications Magazine, Oct. 2003, Special Issue on Merging IP and Wireless Networks.

[112] IEEE Journal on Selected Areas in Communications, May 2004, Special Issue on All-IP Wireless Networks.