Chapter 1
An Overview of Call Admission Control in Mobile Cellular Networks

This chapter provides a thorough overview on call admission control techniques commonly employed in mobile cellular networks. It begins with an introduction to cellular technology, and gradually explores various methods and techniques for call admission control undertaken by different research groups. Strategies of call admission control under diversity of network environments have been introduced with special reference to priority of calls, predictive nature of the network and implicitness of the network, call queuing strategy, and channel borrowing schemes. Application of soft computing techniques, including artificial neural nets, genetic algorithm, fuzzy relational approach and particle swarm optimization, in call admission control is illustrated.

1.1 History of Mobile Communication

The origin of radio communications dates back to the 19th century. In 1864 James Clerk Maxwell enunciated the well-known Maxwell Equations for electromagnetic radiation. In 1876 Alexander Graham Bell invented the telephone. In 1887 Heinrich Hertz discovered "hertzian waves" which are now called as radio waves. In 1896 Guillermo Marconi carried out the world's first radio transmission. There had been scope of simplex radio communications particularly for radio reception for common people, and duplex communication among police and investigation departments over the last 50 years. The duplex radio system, however, worked for short range communications, and was far fetched to be considered for realization in a large and global sense.

Wireless communications have changed beyond recognition over the last 15 years. The first widely used cellular mobile phones were analogue and installed initially in cars due to the bulky hardware required at the time. Shortly the Bell telephone company (US) introduced the first cellular public network AMPS (Advanced Mobile Phone Service) in 1978, after hand-held devices weighing more than 1 kg became available. The AMPS had an evolution and merging with NMT (Nordic Mobile Telephone) of Germany, and appeared later in the UK as TACS (Total Access Communication Service).

The development and deployment of second-generation systems took place from the late 1980s to the present day. These were digital rather than analogue

S. Ghosh and A. Konar: CAC in a Mobile Cellular Network, SCI 437, pp. 1–62.
springerlink.com

providing the end user with supposedly better voice quality, whilst providing the operators with considerable improvements to the capacity per unit bandwidth. The other advantage of second-generation systems is their roaming ability. The pan-European standard GSM (Global System for Mobile Communications) allowed international roaming for the first time throughout Europe. Many other countries throughout the world have now adopted GSM. The USA adopted an evolutionary approach to its AMPS system, developing D-AMPS (Digital AMPS). Also, a second standard is also used in the USA (IS-95), which provides an air interface based on CDMA (Code Division Multiple Access).

Wireless standards were also developed for cordless telephone applications, where users had a personal 'base' in their homes connected to a landline. An early standard in the UK and Canada was CT2, with its digital second-generation counterpart DECT (Digital Enhanced Cordless Telephone). In the USA, there are at least two cordless standards: PACS-UB, and IS-136. PACS-UB is primarily intended for a wireless PABX scenario, with multiple ports providing overlapping coverage areas, allowing portables to switch connections frequently between ports.

The Japanese have a similar system called PHS (Personal Handy phone System). The third area attracting considerable interest is Fixed Wireless Access (FWA), or alternatively known as Wireless in the Local Loop (WLL), which is intended to replace the cable from the 'last mile to the home'. FWA will probably prove most successful is in low-density communication scenarios, where cost of cabling is relatively high, and in the developing world where cabling infrastructure is not yet in place.

Third generation systems are currently under research worldwide, and are being designed to support full multimedia access. A worldwide standard may be achieved, the so-called FPLMTS (Future Public Land Mobile Telecommunication Systems). This is currently being worked out in Europe as UMTS (Universal Mobile Telephone Standard). The above scheme is known in USA as W-CDMA (Wideband CDMA). It runs in a 5 MHz bandwidth. Another 3G standard proposed in USA is known as CDMA2000. Since their proposal, there was battle over which standard to follow.

For this, another two standards have emerged. They are EDGE (Enhanced Data rates for GSM evolution) and GPRS (General Packet Radio Service). The above two standards are called 2.5G standard. These are basically 2G standards with some modification. This evolution is still continuing and we have to see where it goes now. Fig. 1.1 shows the evolution of standards.

The basic mechanism of the communication system which will be considered is that a set of entities (users) access a common medium which, in this case, is the radio channel. This concept is depicted schematically in Fig. 1.2. The frequency spectrum or bandwidth that is allocated to a certain system is a limited resource indicated by the rectangular frame. In general, there are many co–existing wireless systems.

In order to avoid interference to and from other systems, a certain level of protection is required. This is indicated by the shaded frame. The aim is to accommodate as many simultaneous users as possible (capacity) within the limited resource. In the example shown in Fig. 1.2, 4 users are considered each of whom requires an equivalent fraction of the total radio resource (illustrated by a circle).

In a digital context, this corresponds to services which require the same information bit–rate. In Fig. 1.2, the size of the circles and hence the required radio

capacity are constant. In real systems the size may be time variant. Consider, for example, a speech service and periods when a speaker is silent, there is no requirement to transmit data and thus the size of the circle would shrink to merely a single point in the space. An ideal multiple access technique supports the time variant request of radio capacity because this means that, at any given time, only those resources are allocated which are actually required. Consequently, situations are avoided where more capacity is allocated than would actually be required.

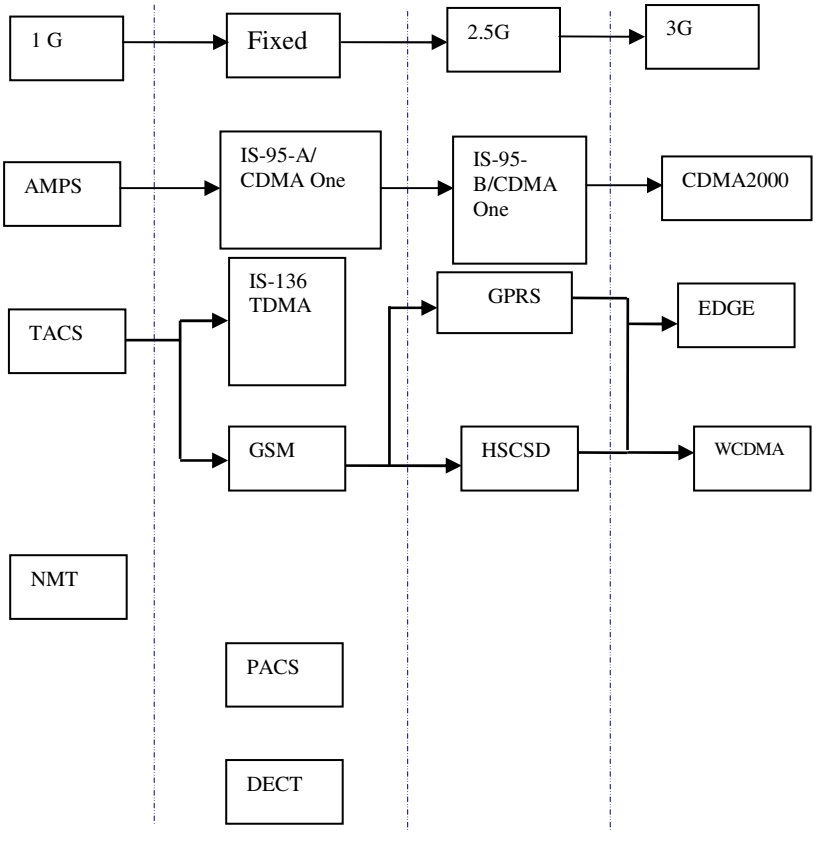


Fig. 1.1 Wireless Standards Evolution to 3rd generation

If the system is not designed carefully, users or mobile stations (MS's) will interfere with each other (gray areas). Therefore each user needs some protection which is equivalent to moving users apart. This measure, however, results in unused radio resources which, considering the immense costs for the radio frequency spectrum is inefficient. Therefore, the aim is to accommodate as many users as possible (minimizing the black colored areas) while keeping the interference at a tolerable level. The separation of users can be done in any dimensions as long as it fulfils the interference requirements. In practice the following dimensions are used:

- Frequency Division Multiple Access (FDMA),
- Time Division Multiple Access (TDMA),
- Space Division Multiple Access (SDMA),
- Code Division Multiple Access (CDMA).

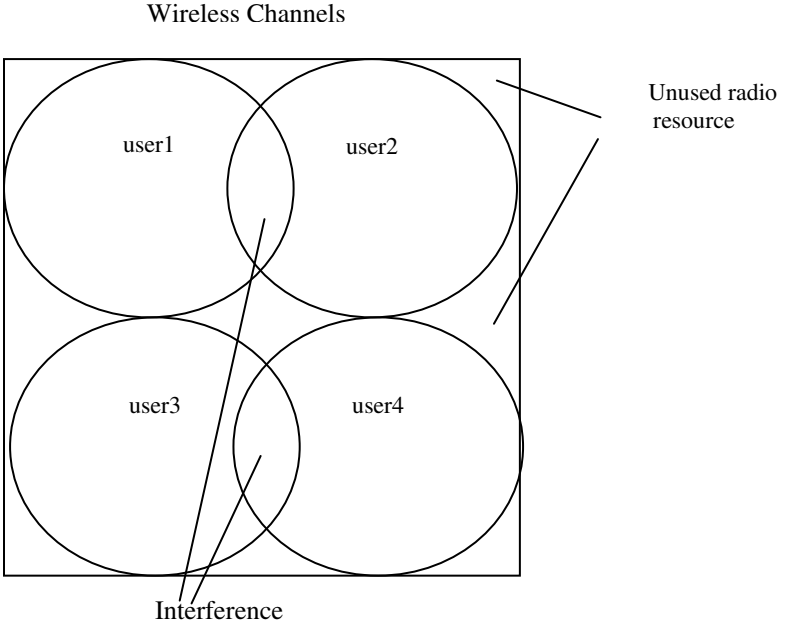


Fig. 1.2 The Principle of Multiuser access

We shall consider the CDMA case elaborately. For certain types of services the aim is to achieve full spatial coverage. In conventional wireless systems a mobile entity is linked to a base station (BS). BS's are connected to a radio network controller that uses additional interfaces that cater for the access to the public switched telephone network (PSTN). The principle structure of a cellular wireless system is shown in Fig. 1.3. The signals on the air–interface experience a distance dependent attenuation. Since the transmit powers are limited, the coverage area of a BS is limited, as well.

Due to the radial signal propagation, in theory, a single BS covers a circular area. The area that is covered by a BS is also referred to as a cell. When modeling cellular systems, cells are approximated by hexagons as they can be used to cover a plane without overlap and represent a good approximation of circles.

Since the total available radio resource is limited, the spatial dimension is used to allow wide area coverage. This is achieved by splitting the radio resource into groups. These groups are then assigned to different contiguous cells. This pattern is repeated as often as necessary until the entire area is covered. A single pattern is equivalent to a cluster.

Therefore, a radio resource which is split into i groups directly corresponds to a cell cluster of size i. In this way it is ensured that the same radio resource is only used in cells that are separated by a defined minimum distance. This mechanism is depicted in Fig. 1.4 (A group of radio resource units is indicated by a certain shade). As a consequence the separation distance grows if the cluster size increases.

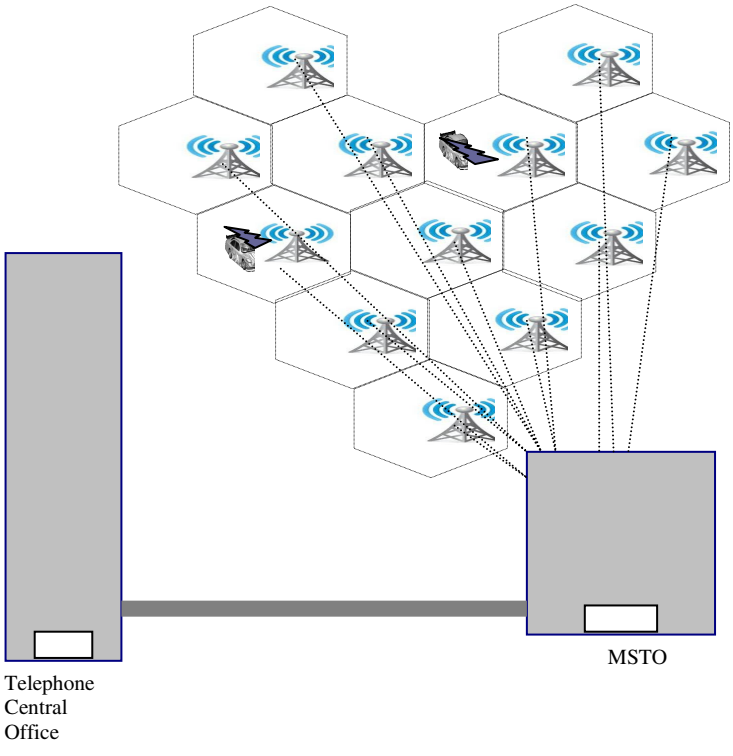


Fig. 1.3 Typical mobile communication systems

Hence, increasing the cluster size acts in favor of low interference. However, an increased cluster size means that the same radio resource is used less often within a given area. As a result, fewer users per unit area can be served. Therefore, there is a trade–off between cluster size and capacity. In an ideal scenario the total available radio resource would be used in every cell whilst the interference was kept at a tolerable level. Herein lies a particular advantage of CDMA over all other multiple access modes since the same frequency carrier can be re–used in every cell [1].

It is clear that this results in increased co–channel interference (CCI) which gradually reduces cell capacity, but the magnitude of the resulting reduction of spectral efficiency is usually less than would be obtained if a fixed frequency re–use distance was applied. The cell capacity, finally, is dependent on many system functions such as power control, handover, etc. which is why capacity in a CDMA

system is described as soft–capacity. However, the fact that in a CDMA system frequency planning can be avoided may not only result in capacity gains, but it eventually makes CDMA a more flexible air interface. Next part gives some basic ideas on CDMA systems.

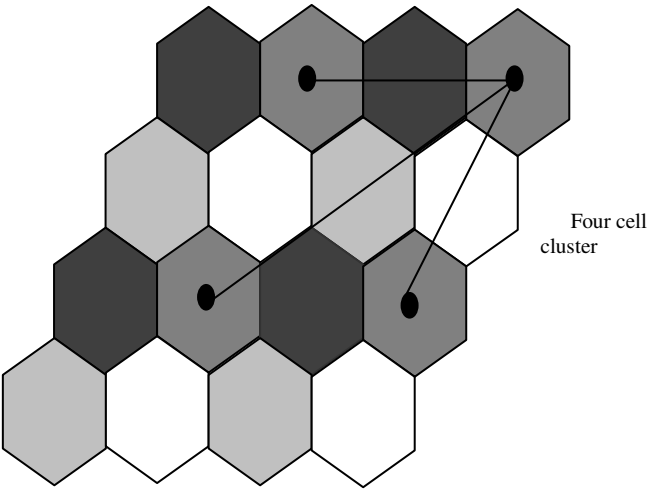


Fig. 1.4 Cellular Concepts

1.2 Cellular CDMA Systems

In CDMA systems, the narrowband message signal is multiplied by a very large bandwidth signal called the spreading signal. The spreading signal is a pseudo noise code sequence that has a chip rate that is orders of magnitudes greater than the data rate of the message. All users in a CDMA system use the same carrier frequency and may transmit simultaneously.

Each user has its own pseudorandom codeword that is approximately orthogonal to all other code words. The receiver performs a time correlation operation to detect only the specific desired codeword. All other code words appear as noise due to de-correlation. For detection of the message signal, the receiver needs to know the codeword used by the transmitter. Each user operates independently with no knowledge of the other users.

In CDMA, the power of multiple users at a receiver determines the noise floor after de-correlation. If the power of each user within a cell is not controlled such that they do not appear equal at the base station receiver, then the near-far problem occurs.

The near-far problem occurs when many mobile users share the same channel. In general, the strongest received mobile signal will capture the demodulator at a base station. In CDMA, stronger received signal levels raise the noise floor at the base station demodulators for the weaker signals, thereby decreasing the

probability that the weaker signals will be received. To combat the near-far problem, power control is used in most CDMA implementations. Power control is provided by each base station in a cellular system and assures that each mobile within the base station coverage area provides the same signal level to the base station receiver.

This solves the problem of a nearby subscriber overpowering the base station receiver and drowning out the signals of far away subscribers. Power control is implemented at the base station by rapidly sampling the radio signal strength indicator (RSSI) levels of each mobile and then sending a power change command over the forward radio link. Despite the use of power control within each cell, out-of-cell mobiles provide interference which is not under the control of the receiving base station.

Spread spectrum techniques build the foundation for CDMA. Therefore, a brief summary of spread spectrum communication is presented in the following. In a spread spectrum system, the frequency bandwidth is greater than the minimum bandwidth required to transmit the desired information. There are different methods as to how the spreading of the spectrum can be accomplished:

Direct sequence (DS) spread spectrum: A signal with a certain information bit rate is modulated on a frequency carrier with a much higher bandwidth than would be required to transmit the information signal. Each user is assigned a unique code sequence which has the property that the individual user's information can be retrieved after dispreading.

Frequency hopping (FH) spread spectrum: The available channel bandwidth is subdivided into a large number of contiguous frequency slots. The transmitted signal occupies one or more of the available frequency slots which are chosen according to a pseudo–random sequence.

Time hopping spread spectrum: A time interval which is much larger than the reciprocal of the information bit rate is subdivided into a large number of timestamps (TS). The information symbols are transmitted in a pseudo–randomly selected TS.

Chirp or pulse–FM modulation system: The frequency carrier is swept over a wide band during a given pulse interval. It is common in all spread spectrum techniques that the available bandwidth, B, is much greater than the bandwidth required transmitting a signal with an information data rate, W. The ratio B/W is the bandwidth spreading factor or processing gain, pg. The processing gain results in a interference suppression which makes spread spectrum systems highly resistant to interference or jamming. This property in particular makes spread spectrum techniques interesting for the application to wireless multiple access communication where a large number of uncoordinated users in the same geographical area access a radio frequency resource of limited bandwidth.

Using the spread spectrum technique, the number of simultaneously active users permitted is proportional to the processing gain [2]. Since the early 1980s,

this has led to the development of the CDMA technology which primarily utilizes the pseudo noise (PN) DS spread spectrum technique [3]. Apart from the PN direct sequencing a second category of CDMA techniques exists. This is described as orthogonal DS–CDMA. In this thesis, PN DS–CDMA systems are considered because orthogonal CDMA systems would require an ideal channel. In addition, CDMA based standards use, at least, a PN code for the final scrambling of the transmitted data.

The wireless communication standards which utilize CDMA techniques, for example, IS–95 and UMTS use a combination of orthogonal codes and PN codes [4], but this is merely aimed to increase the robustness of the system. Since in this thesis PN DS–CDMA techniques are considered, henceforth the expression CDMA will be used to describe this particular multiple access method.

As mentioned above, the capacity calculation of a CDMA system is more complex since it is interference limited. Each user contributes to the common noise floor which is usually assumed to be Gaussian [5].Thus, interference is a most important parameter in a CDMA system and capacity analyses focus on calculating interference quantities [5]. Since interference is dependent on many factors, for example, power control, adjacent channel leakage and handover strategies to name only a few, the capacity figures can vary significantly (soft capacity).

CDMA is used in the 2nd generation mobile communication standard IS–95 which gained special interest after it had been claimed that CDMA can achieve a greater spectral efficiency than conventional FDMA and TDMA methods [5]. For example, Viterbi [6] showed that the capacity of a CDMA system can be:

$$\text{Capacity (CDMA)} \cong 1 \text{ Bit/Sector/Hz/Cell;}$$

It is assumed that the voice activity of each user to be 50% and the sectorisation gain to be 4 – 6 dB. This figure was compared to the capacity of GSM (Global System for Mobile communications):

$$\text{Capacity (GSM)} \cong 1/10 \text{ Bit/Sector/Hz/Cell,}$$

where, a frequency re–use factor of 1/4, was assumed. Theoretically, when considering a single cell and an AWGN (additive white Gaussian noise) channel the multiple access schemes CDMA, FDMA and TDMA are equivalent with respect to spectral efficiency [7].

Therefore, the greater spectral efficiency of CDMA systems primarily results from three basic principles:

1. The same channel is used in every cell (channel re–use factor of 10[1]),
2. Interruptions in transmission, e.g., quiet periods of a speaker, when assuming a voice service, are exploited [5].
3. Antenna sectorisation is used.

In general, the net improvement in capacity due to all the above features, of CDMA over digital FDMA or TDMA is on the order of 4 to 6 and over analog FM/FDMA it is nearly a factor of 20 [5].

However, it was demonstrated that the advantages of CDMA systems were slightly overestimated due to two basic hypotheses that usually cannot be fulfilled in a realistic environment [8]:

1. Perfect power control,
2. All MS's are allocated to the most favorable BS, i.e., the BS offering the lowest path loss.

1.3 Radio Resource Allocation Techniques

In a cellular network certain radio resources allocation methods are required to mitigate the detrimental impact of interference (Co-Channel Interface i.e. CCI and Adjacency Channel Interface i.e. ACI).

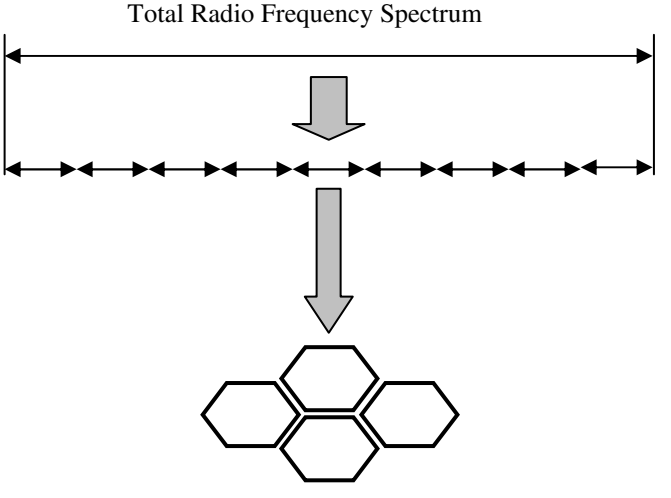


Fig. 1.5 Radio spectrum allocation

Channel Assignment Problem

The total radio spectrum allocated to a particular service producer can be divided into a set of disjoint or non-interfering radio channels (Fig. 1.5). All these channels can be used simultaneously.

The three methods used to divide spectrum into such channels are

i) **Frequency Division (FD):** Here the spectrum is divided into disjoint frequency bands.

ii) **Time Division (TD):** Here the usage of the channel is divided into disjoint time periods called time slots.

iii) **Code Division (CD):** Here the spectrum division is done using different modulation codes.

Furthermore a combination of all three can be also used to achieve desired result. Allocation of the channels among the cells should satisfy traffic demand and the electromagnetic compatibility constrains. The Constrains are categorized as soft and hard constrains. The soft constraints are describe as follows

a. **The co-channel constraint (CCC):** where the same channel cannot be assigned to certain pairs of radio cells simultaneously.
b. **Adjacent channel constraint (ACC):** where channels adjacent in frequency spectrum cannot de assigned to adjacent radio cells simultaneously.
c. **Co-site Constraint (CSC):** where channel assigned in the same radio cell must have minimal separation in frequency between each other.

The hard constrains are

a. Selected Channel should have high ***re-use value***.
b. Selected Channel should produce high ***packing density***.

The major constrains that needed to be considering while establishing a channel assignment algorithm are the Co-channel constrain and the re-use value of the channels.

Table 1 Classification of CAP

Classification criteria	Types /Classes		
Co-channels Separation	Fixed	Dynamic	Hybrid
CIR measurements	Blind	Local CIR measurements	
Control	Centralized	Distributed	

Classification of Channel Allocation Scheme

Channel allocation schemes can be divided into a number of different categories depending on the comparison basis.

Three basic concepts of radio resource allocation are as follows:

- Static or fixed channel assignment (FCA) techniques
- Dynamic channel assignment (DCA) techniques
- Hybrid channel assignment (HCA) techniques

The principles of these methods are described in the following section.

1.3.1 Fixed Channel Assignment Techniques

An FCA method allocates a fixed fraction of all available channels to an individual cell of a cellular environment. The same group of channels is only used in cells that are separated by a minimum distance D. The channel re–use distance D ensures that CCI does not deteriorate the system performance greatly. The cluster size basically determines the system capacity, since it specifies the maximum number of simultaneously active connections that can be supported at any given time. The group size which equals the number of channels per cell, M, can be found from the relation

M= (Available BW) / (channel BW * cluster size).

It can be seen that M is increasing with a decreasing cluster size K, but this also means that the interference is higher which, in turn, reduces the capacity or QoS. This means that a system with a greater number of channels per cell is more efficient than a system with only a few channels. This effect is well known as the trucking gain. Consequently, fixed channel assignment techniques result in poor spectral efficiency. Given that CCI varies with the cell load, there might be traffic scenarios where a lower channel re–use distance can be tolerated in favor of a temporarily higher number of channels available in a single cell (or cluster of cells).

This would require methods which dynamically monitor interference and load situations throughout the network and which carry out channel re–configurations accordingly. In contrast to DCA strategies, FCA techniques are not designed to achieve this flexibility. CDMA systems such as the UTRA–FDD interface of UMTS re–use the same channel in every cell which, in theory, makes FCA or DCA techniques superfluous, but requires special handover techniques (soft–handover). The fixed channel assignment techniques are classified as given in Fig. 1.6.

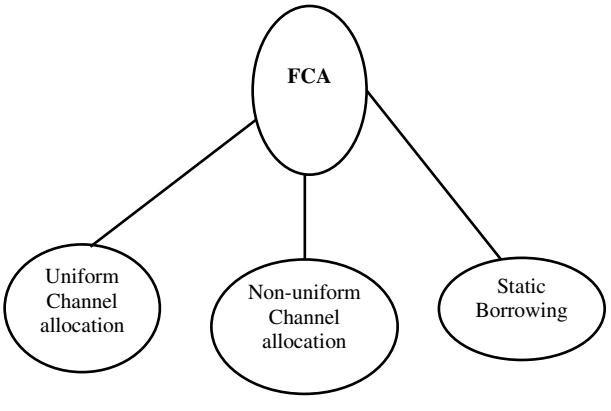


Fig. 1.6 Types of FCA

a) Uniform Channel allocation

The uniform channel allocation is efficient if the traffic distribution of the system is also uniform. Here the overall average blocking probability of the mobile system is the same as the call blocking probability in a cell. A uniform allocation of channels to cells may result in high blocking in some cells and poor channel utilization, since traffic in cellular systems can be non-uniform with temporal and spatial fluctuations.

b) Non-uniform Channel allocation

Here the number of nominal channels allocated to each cell depends on the expected traffic in that cell. So heavily loaded cells are assign more channel than the channels with comparative fewer loads.

c) Static Borrowing

In this scheme, the channels from lightly loaded cells are reassigned to the heavily loaded cells. The channels from the lightly loaded ones can be reassigned only if the distance is more than the minimum distance for reuse. This is known as static borrow since the channels can borrow free channels from its neighboring cells (donors) to accommodate new calls. When a channel is borrowed the other cells are prohibited from using it.

1.3.2 Dynamic Channel Assignment Techniques

In DCA schemes, any base station can use any channel. All channels are kept in a central pool and are assigned dynamically to new calls as they arrive in the system. After each call is completed, the channel is returned to the central pool. It is fairly straightforward to select the most appropriate channel for any call based simply on current allocation and current traffic, with the aim of minimizing the interference.

The advantage with this scheme is that channels can be moved from cells with less demand to cells with heavier demand which is time varying. Dynamic channel assignment is again of two types: a) Centralized DCA b) Distributed DCA as shown in Fig. 1.7.

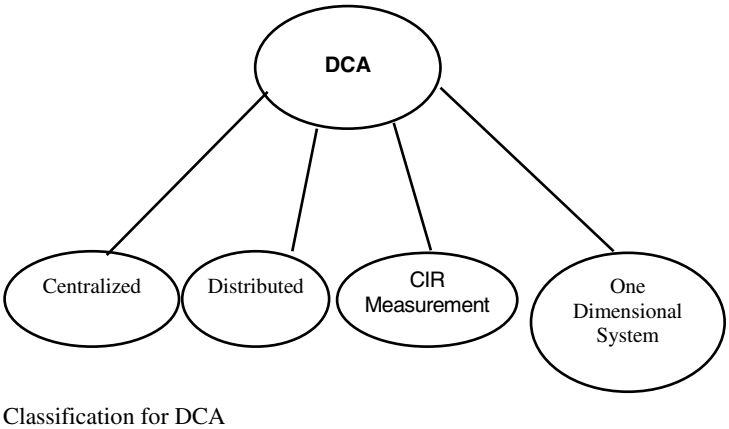


Fig. 1.7 Classification for DCA

a) **Centralized DCA:** The centralized DCA scheme involves a single controller selecting a channel for each cell. Theoretically it provides the best performance at the expense of high-centralized overhead. It is not suitable for high-density micro-cell systems. The disadvantage with this type of system is that if the main system goes down because for some reason then the whole system fails because of no other alternative.

b) **Distributed DCA:** The distributed DCA scheme involves a number of controllers scattered across the network. Here, the channel assignment decision is made by a local instance. Thus, only local information is available. Hence, the complexity is reduced considerably when this type of DCA algorithm is used.

c) **CIR measurement DCA schemes:** All mobile base station pairs are examined in channels in the same order and choose the first available with acceptable CIR

d) **One Dimension Systems:** In this scheme a mobile is assigned a channel that maximizes the minimum of the ClR's of all mobiles being served by the system at that time. A mobile is served only after all mobiles to the left of it have had a chance to be served. This sequential (left to right) order of service is chosen because it appears to be the best way for reusing the channel. The mobile immediately to the right of a given set of mobiles with channels assigned is the one that will cause the most interference at the base station servicing the given set of mobiles, and is also the one which has the most interference from that set of mobiles.

1.3.3 Hybrid Channel Assignment Techniques (HCA)

HCA schemes are the combination of both FCA and DCA techniques. In HCA schemes, the total number of channels available for service is divided into fixed and dynamic sets. The fixed set contains a number of nominal channels that are assigned to cells as in the FCA schemes and, in all cases, are to be preferred for use in their respective cells. All users share the dynamic set in the system to increase flexibility.

In a CDMA system, all users share a common channel, they are differentiated by using codes. For this, channel allocation problem is transformed to the problem of call admission control. Next part gives the idea about call admission problem.

The classification of HCA techniques are shown in Fig. 1.8.

Flexible Channel Allocation (FICA)

This scheme divides the available channels into fixed and flexible sets. Each cell is assigned a set of fixed channels that typically suffices under a light traffic load. The flexible channels are assigned to those cells whose channels have become inadequate under increasing traffic loads.

FICA techniques differ according to the time at which additional channels are assigned. The assignment of these channels among the cells is done in either a

scheduled or predictive manner. In the predictive strategy, the traffic intensity or the blocking probability is constantly measured at every cell site so that the reallocation of the flexible channels can be carried at any point in time.

Fixed and Dynamic Channel Allocation
In this technique the blocking rate depending on traffic intensity. In low traffic intensity the DCA scheme is used; in heavy traffic situations the FCA strategy is used. The transition from one strategy to the other would be done gradually because a sudden transition will cause a lot of blocking.

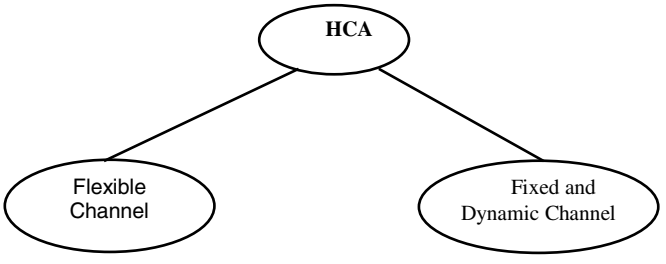


Fig. 1.8 Clasification of HCA

1.4 Call Admission Control

Dynamic channel allocation (DCA), has been extensively studied for FDMA/TDMA cellular systems as a means of increasing capacity and adapting to traffic loading variations. In DS-CDMA cellular systems, however, it is difficult to utilize DCA due to the difficulty of sharing traffic load between cells. Here, the problem of channel allocation can be viewed as call admission control.

Call admission control is one of the most effective methods for optimal resource management. When a call is initiated in one cell, it will request a channel from its home cell. In CDMA systems, assigning a channel means allocating the appropriate power to a requesting mobile. Due to the sharing of spectrum, this induces interference to other users. This kind of situation requires that the interference must be below a certain level to maintain the appropriate level of communication quality. This fact is elaborated in the next paragraph.

In CDMA systems, capacity is limited only by the total level of interference from all connected users. As a result, CDMA utilizes the effect of statistical multiplexing without the complex radio channel allocation or reallocation that is required in frequency-division multiple-access (FDMA) and time-division multiple-access (TDMA) systems. However, in systems based on statistical multiplexing, there exists a tradeoff relationship between the system capacity and the level of communication quality.

Although the so-called graceful degradation based on this tradeoff relationship is one of the most essential features of CDMA systems, the communication quality must be guaranteed to a certain level on average. The number of

simultaneous users occupying a base station (BS) therefore must be limited such that an appropriate level of communication quality can be maintained. Call admission control (CAC) thus plays a very important role in CDMA systems because it directly controls the number of users. CAC must be designed to guarantee both a grade of service (GoS), i.e., the blocking rate, and a quality of service (QoS), i.e., the loss probability for communication quality.

1.4.1 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 the hard and the soft handoff [9-10].

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 [11]. 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 [9], [11]. 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) [11].

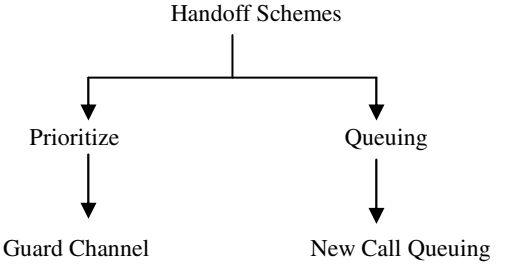


Fig. 1.9 Handoff schemes

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 [10]. This study focuses on cellular networks implementing hard handoff schemes (Fig. 1.9).

a. Prioritizing Schemes

Handoff prioritizing schemes are channel assignment strategies that allocate channels to handoff requests more readily than new calls.

b. Guard Charnels Schemes

In this scheme a number of channels are reserved exclusively for handoff calls in a cell. The remaining channels are used among the new and handoff calls.

c. Queuing Schemes

Here when the power level received by the base station in the current cell reaches a certain threshold, namely the handoff threshold, a called is placed in the queue from the neighbor cell for providing service.

The call remains in the queue until either an available channel in the new cell is found or the power by the base station in the current cell drops below a second threshold, called the receiver threshold.

d. New Call Queuing Schemes

In this method of guard channels and the queuing of new calls was introduced. This method not only minimizes blocking of handoff calls, but also increases total carried traffic.

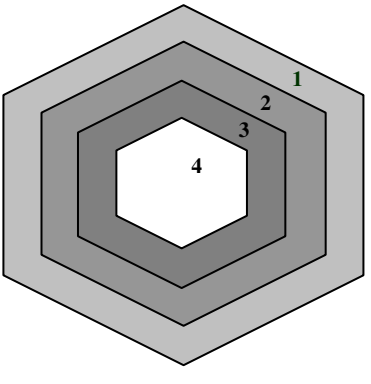


Fig. 1.10 Concentric sub cells

1.4.2 Reused Partitioning

In reused partitioning method each cell in the system is divided into two or more concentric sub cells (zones) as shown in Fig. 1.10. Because the inner zones are closer to the base station located at the center of the cell, the power level required

to achieve a desired CIR in the inner zones can be much lower compared to the outer zones. The Reused Partitioning scheme can be divided as follows as given in Fig. 1.11:

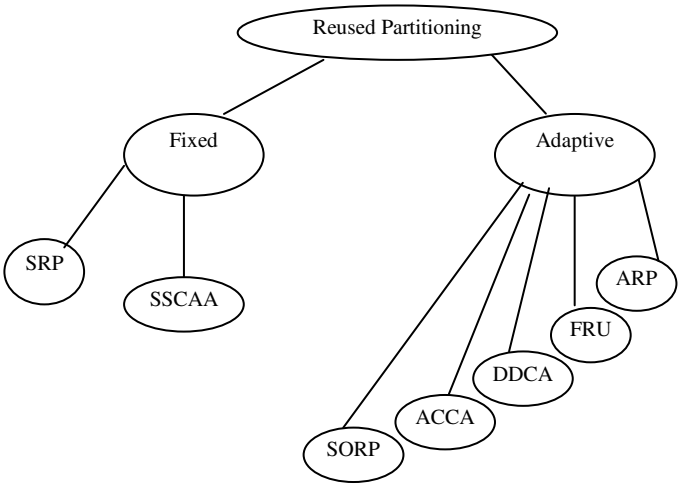


Fig. 1.11 Classification for Reused Partitioning

1.4.2.1 Fixed Reused Partitioning Scheme

In this scheme, available channels are split among several overlaid cell plans with different reuse distances.

1. **Simple Reuse Partitioning (SRP)**
   Simple RUP can be implemented by dividing the spectrum allocation into two or more groups of mutually exclusive channels. Channel assignment within the ith group is then determined by the reuse factor N_i for that group. Mobile units with the best received signal quality will be assigned to the group of channels with the smallest reuse value factor value, while those with the poorest received signal quality will be assigned to the group of channels with the largest reuse factor value.

2. **Simple Sorting Channel Assignment Algorithm (SSCAA)**
   Here each cell is divided into a number of concentric zones and assigned a number of channels. For each mobile in the cell, the base station measures the level of SIR and places the measurements in a descending order. Then it assigns channels to the set of at most mobiles with the largest values of SIR, where M is the number of available channels in the entire cell. The mobile in the set with the smallest value of SIR is assigned a channel from the outer cell zone.

1.4.2.2 Adaptive Channel Allocation Reuse Partitioning Schemes

1. **Autonomous Reuse Partitioning(ARP)**
   In this scheme all the channels are viewed in the same order by all base stations, and the first channel that satisfies the threshold condition is allocated to the mobile attempting the call. Thus, each channel is reused at a minimum distance with respect to the strength of the received desired signal.

2. **Flexible Reuse (FRU)**
   In this scheme whenever a call requests service, the channel with the smallest CIR margin among those available is selected. If there is no available channel, the call is blocked.

3. **Self-organized Reuse Partitioning Scheme (SORP)**
   In this method, each base station has a table in which average power measurements for each channel in its cell and the surrounding cells are stored. When a call arrives, the base station measures the received power of the calling mobile station and selects a channel, which shows the average power closest to the measured power.

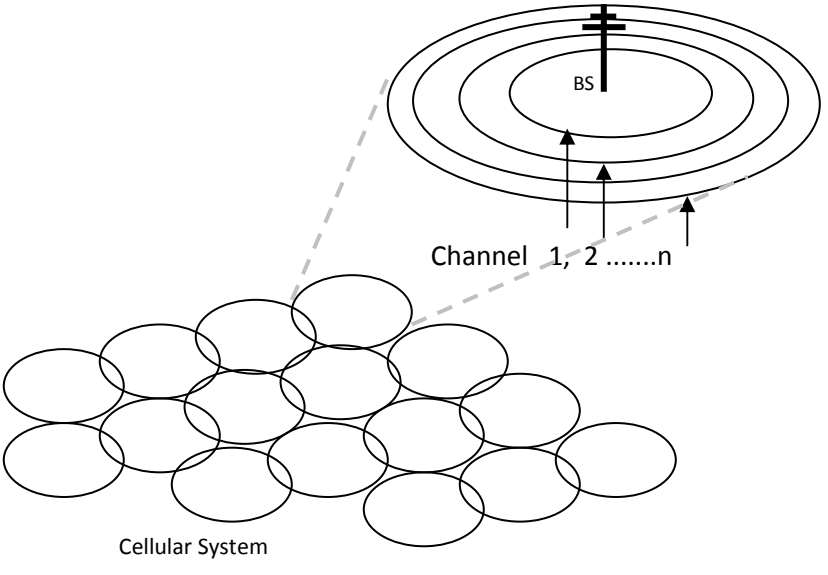


Fig. 1.12 Principle of the all-channel concentric allocation

4. **All-Channel Concentric Allocation (ACCA)**
   All radio channels of a system are allocated nominally in the same manner for each cell. Each cell is divided into N concentric regions; each region has its own channel allocation. Here, each channel is assigned a mobile belonging to the concentric region in which that channel is allocated, and has

a specific desired signal level corresponding to the channel location. Therefore, each channel has its own reuse distance determined from the desired signal level (Fig. 1.12).

5. **Distributed Control Channel Allocation (DCCA)**
   In this scheme all cells are identical, and channels are viewed in the same order, starting with channel number one, by all the base stations in the network. It consists of an omni-directional central station connected to six symmetrically oriented substations.

The substations are simple transceivers, and can be switched on and off under the control of the main station. When the traffic density of the cell is low, all the substations are off and the only operating station is the main station, at the center of the cell covering the entire cell area. Gradually, as call traffic increases, forced call blocking will occur due to an unacceptable level of co-channel interference o r the unavailability of resources. In this case, the main base station switches on the nearest substation to the mobile unit demanding access.

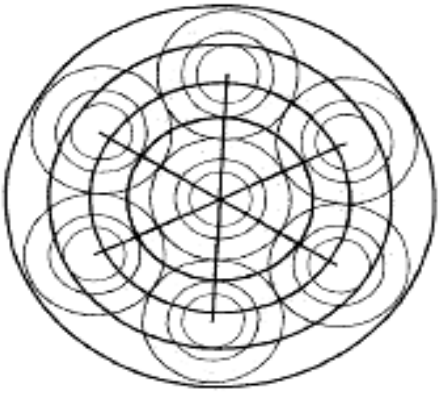


Fig. 1.13 DCCA structure

1.4.3 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 of scheme B for the same QoS parameters and the 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 signaling 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. ***Adaptively:*** 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.
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 congestion occurs and then all requests are rejected. It is desirable that network control and management avoid such a situation.

1.5 Call Admission Control Schemes

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. We assume that available bandwidth in each cell is channelized and focus on call-level QoS measures. Therefore, the call blocking probability (pb) and the call dropping probability (pd) are the relevant QoS parameters in this chapter. Three CAC related problems can be identified based on these two QoS parameters [12]:

1. *MINO:* Minimizing a linear objective function of the two probabilities (pb and pd).
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.

1. **Deterministic CAC:** QoS parameters are guaranteed with 100% confidence [14], [15]. Typically, these schemes require extensive knowledge of the system parameters such as user mobility that 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 [16], by relaxing QoS guarantees, these schemes can achieve a higher utilization than deterministic approaches. Most of the CAC schemes that are investigated in this paper fall in the stochastic category. Fig. 1.14 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.

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 *a priori* based on, for example, traffic and mobility characteristics [12].

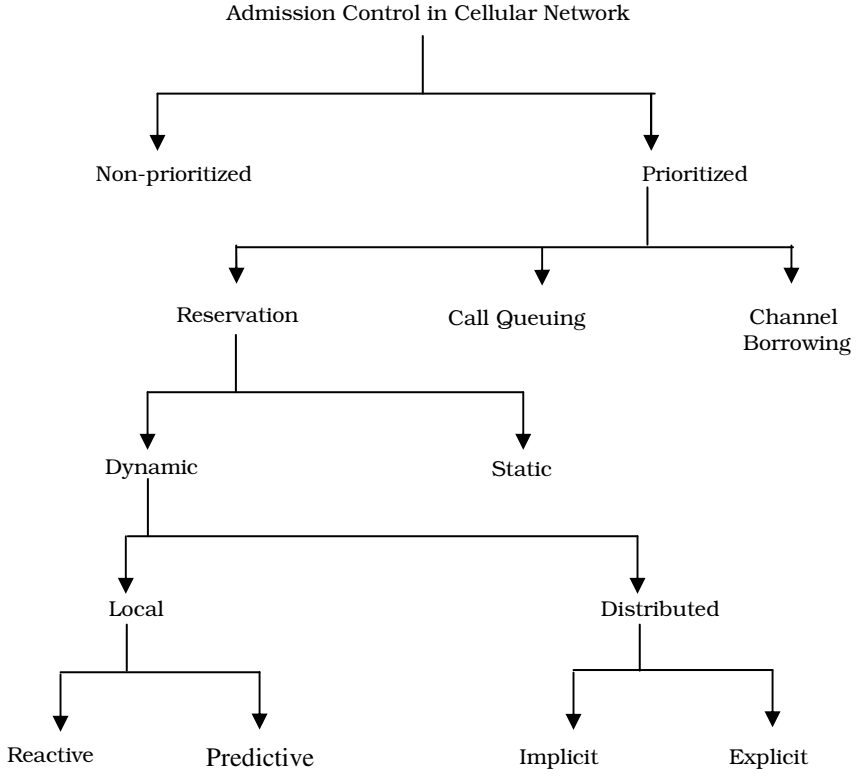


Fig. 1.14 Stochastic call admission control schemes

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 [13]. 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:

1.5.1 Prioritization Schemes

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

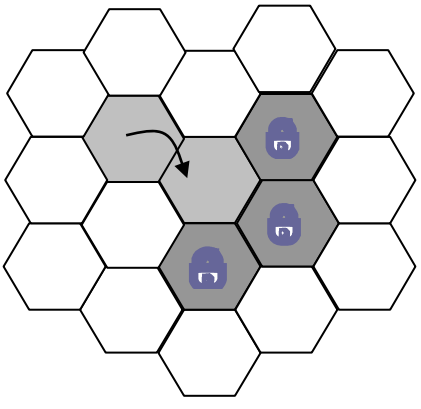


Fig. 1.15 Channel locking Scheme

1.5.1.1 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 [17], [18], [19].

A cell can borrow a channel, 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 [20]. The number of such cells depends on the cell layout and the initial channel allocation. For example, for a hexagonal planar layout with reuse distance of one cell, a borrowed channel is locked in three neighboring cells (see Fig. 1.15).

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 was provided by Katzela and Naghshinehin [17].

1.5.1.2 Call Queuing Schemes

Queuing of handoff requests in absence of channel availability 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.

Queuing can be done for any combination of new and handoff calls. The queue itself can be finite [21] or infinite [16]. Although finite queue systems are more realistic, systems with infinite queue are more convenient for analysis. Fig. 1.16 depicts a classification of call queuing schemes.

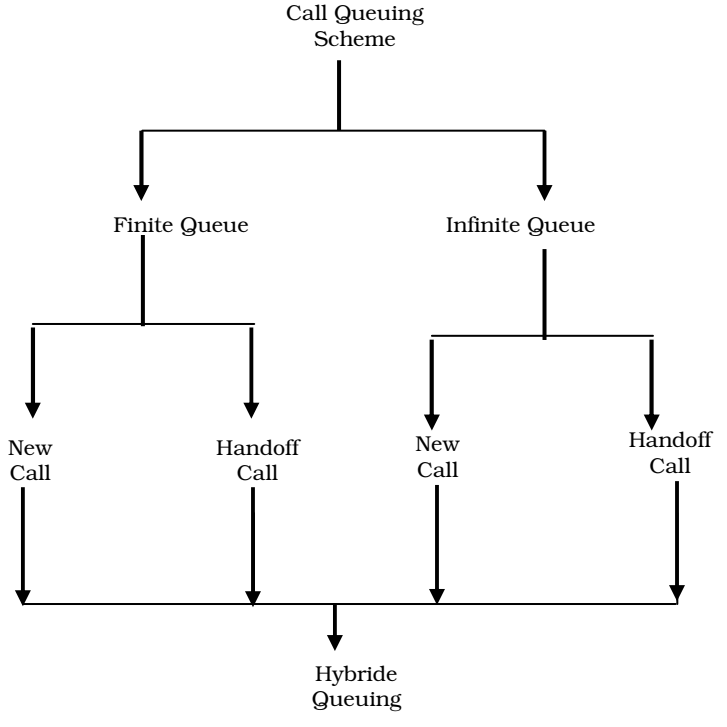


Fig. 1.16 Call queuing schemes.

Hong and Rappaport [16] analyzed the performance of the simple guard channel scheme with queuing 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 queuing strategy and showed that queuing improves the performance of the pure guard channel scheme, i.e., probability of call drop (pd) is lower for this scheme while there is essentially no difference for probability of call block (pb).

The tolerable waiting time in queues is an important parameter. The rearranging 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 affect the performance of queuing schemes.

Chang et al. [21] analyzed a priority-based queuing scheme in which handoff 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 cannot handoff to another cell.

Recently, Li and Chao [22] investigated a general modeling framework that can capture call queuing as well. They proved that the steady-state distribution of the equivalent queuing model has a product form solution. Queuing schemes have been mainly proposed for circuit-switched voice traffic. Their generalization to multiple classes of traffic is a challenging problem [23]. Lin and Lin [24] analyzed several channel allocation schemes including queuing of new and handoff calls. They concluded that the scheme with new and handoff calls queuing has the best performance.

1.5.1.3 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 [16] showed that this scheme reduces handoff-dropping probability significantly compared to the non-prioritized case. They found that pd decreases by a significantly larger order of magnitude compared to the increase of pb 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 rate λ and ν, respectively. The call holding time and cell residency for both types of call is exponentially distributed with mean 1/μ and 1/η, 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, a continuous time Markov chain with C states can model the cell channel occupancy. The state transition diagram of a cell with C channels and C−T guard channels is shown in Fig. 1.8. Given this, it is straightforward to derive the steady-state probability P_n, that n channels are busy

$$P_n = \begin{cases} \left(\frac{\rho^n}{n!}\right) P_0 & 0 \le n \le T \\ \rho^T \left(\frac{\nu^{n-T}}{n!}\right) P_0 & T \le n \le C \end{cases} \tag{1.1}$$

Where

$$P_0 = \left[\sum_{n=0}^{T} \frac{\rho^n}{n!} + \rho^T \sum_{n=T+1}^{C} \frac{\nu^{n-T}}{n!} \right]^{-1} \tag{1.2}$$

$$\text{And then } p_b = \sum_{n=T+1}^{C} P_n \quad \text{and} \quad p_f = P_C \tag{1.3}$$

However, Fang and Zhang [25] 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 pd and minimizing pb. If the number of guard channels is conservatively chosen then admission control fails to satisfy the specified pd. A static reservation typically results in poor resource utilization. To deal with this problem, several dynamic reservation schemes [17], [26–29] were proposed in which the optimal number of guard channels is adjusted dynamically based on the observed traffic load and dropping rate in a control time window.

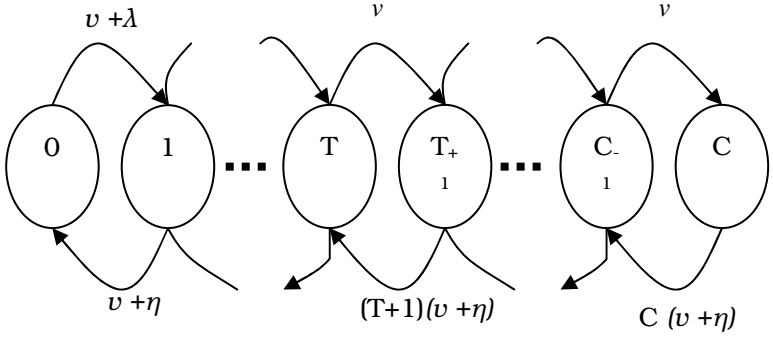


Fig. 1.17 State transition diagram of the guard channel scheme

If the observed dropping rate is above the guaranteed pd then the number of reserved channels is increased. On the other hand, if the current dropping rate is far below the target pd 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) [12]. 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 [30] where,

$$P_n = \frac{\prod_{i=0}^{n-1}(\upsilon + a_i\lambda)}{(\mu+\eta)^n} P_0 \qquad 1 \le n \le C \tag{1.4}$$

$$P_0 = \left[1 + \sum_{n=1}^{C}\left(\frac{\prod_{i=0}^{n-1}(\upsilon + a_i\lambda)}{(\mu+\eta)^n}\right)\right]^{-1} \tag{1.5}$$

Therefore $p_b = \sum_{n=0}^{C}(1-a_n)P_n$ and $p_f = P_C$ where $a_c = 0$.

It has been shown in [13] 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 [13]

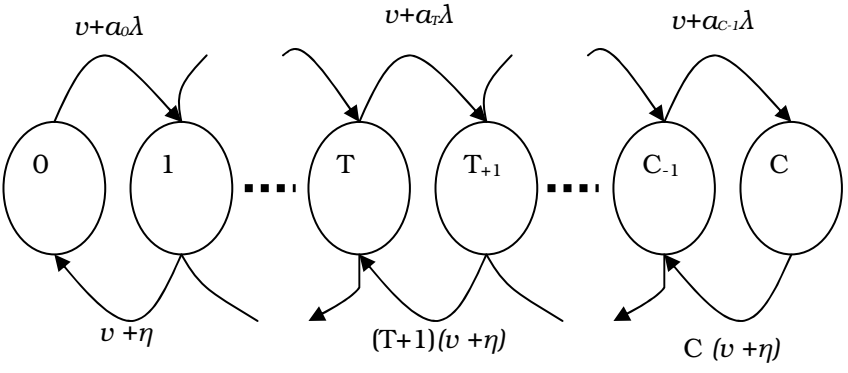


Fig. 1.18 State transition diagram of the fractional guard channel scheme

1.5.1.3.1 Dynamic Reservation Scheme

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.

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 [26]) in order to have more control of the number of accepted calls. It has been shown [31] 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 [32] 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:

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

$$\hat{\lambda}(i+1) = \varepsilon\, \hat{\lambda}(i) + (1-\varepsilon)\lambda(i) \tag{1.6}$$

where ε is the smoothing coefficient which must be properly selected. In general, a small value of ε (thus, a large value of 1- ε) 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 that 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.

It is clear that the EWMA in [14] is a special case of the so-called *auto regressive moving average* (ARMA) model [33] in time series analysis. There is virtually no restriction on using more complicated (and perhaps more accurate) estimation techniques.

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 [34]. 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 that 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 [39–41] that cannot be captured by classical models, hence self-similar models have been developed. Among these self-similar models, fractional ARIMA [35], [36] and fractional Brownian motion [37], [38] 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 wire line counterparts, it is highly possible that traffic in these networks will also exhibit self-similarity (as reported for wireless data traffic by Ziang et al. [34]). Hence, simple modeling and prediction techniques may not be accurate. Admission control based on self-similar traffic models has been already investigated for wire line networks [42], [43]. Similar approaches may be applicable to cellular communications.

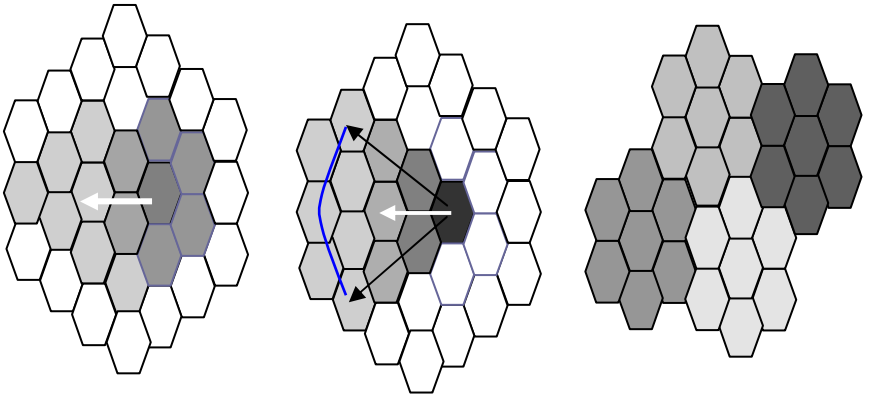
1.5.1.3.2 Distributed Schemes

The fundamental idea behind all distributed schemes [27-30], [44-46] 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 [27]. A group of cells, which are geographically or logically close together, form a *cluster*, as shown in Fig. 1.19. 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. 1.19(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. 1.19(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. Fig. 1.19(c) shows a static cluster that is fixed regardless of the terminal 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.



(a) Shadow cluster (b) Most likely (c) Virtual connection tree

Fig. 1.19 Three cluster definition.

Now, consider a hexagonal cellular system similar to those depicted in Fig. 1.19. 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

$$p_f = \Pr(n(t_0 + T) > c) \tag{1.7}$$

Therefore the handoff failure in cell i is given by

$$p_f(i) = \frac{1}{2} erfc\left(\frac{c_i - E[n_i(t_0 + T)]}{\sqrt{2Var[n_i(t_0 + T)]}} \right) \tag{1.8}$$

where c_i is the capacity of cell i and erfc(x) is the complementary error function defined as

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2} dt \tag{1.9}$$

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

$$E\left[n_i(t_0+T)\right] = n_i(t_0)\, p_s + p_h \sum_{j=1}^{6} n_j(t_0) \tag{1.10}$$

$$Var\left[n_i(t_0+T)\right] = n_i(t_0)\, v_s + v_h \sum_{j=1}^{6} n_j(t_0) \tag{1.11}$$

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

$$p_s = e^{-(\mu+h)T}, \quad p_h = \frac{1}{6}(1-e^{-hT}) \tag{1.12}$$

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

$$v_s = (1-p_s)p_s, \quad v_h = (1-p_h)p_h \tag{1.13}$$

Naghshineh and Schwartz originally proposed the idea of distributed admission control [17]. 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 [29]. 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 [30]. Levin et al. [27] 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. [30] proposed a dynamic, distributed and stable CAC scheme called SDCA which extends the basic DCAC [17] in several ways such as using a diffusion equation to describe 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 cells 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 [27]. For an effective control enforcing dropping probabilities in the order of 10–4 to 10–2, essentially all calls handoff successfully.

Table 2a Cluster Type vs. CAC Performance.

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

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

$$q_n(t) = \frac{1}{n!}\left(\frac{\eta t}{6}\right)^n e^{-(\mu+\eta)t} \tag{1.14}$$

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

$$f_{ii}(t) = q_0(t) + 6q_2(t) + 12q_3(t) + \ldots\ldots \tag{1.15}$$

Similar equations can be easily derived for $f_{ik}(t)$ [30]. Using these time-dependent transition probabilities Wu et al. computed the time-dependent mean and variance of the channel occupancy distribution, $P_{ni}(t)$, in cell i at time t. By using a diffusion approximation [47], the authors were able to find the time-dependent handoff failure, $P_{fi}(t)$, for each cell i. Hence, the average handoff failure probability over a control period of length T is found as

$$\tilde{P}_{f_i} = \frac{1}{T}\int_0^T P_{f_i}(t)\,dt \tag{1.16}$$

Finally, the acceptance ratio *ai* can be obtained by numerically solving the following equation [49]:

$$\tilde{P}_{f_i} = P_{Q \circ S}, \qquad 0 \le a_i \le 1 \tag{1.17}$$

A. Classification of Distributed Schemes

Distributed CAC s 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 [46] is an example of a static cluster while the shadow cluster introduced in [27] 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.

Table 2b Comparison of Dynamic CAC Schemes

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

It has been shown that it is not worth involving several cells in the admission control process when the network is not congested [49]. Table1a 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 CAC s 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 [46] 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 [27] 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 that after processing by the original cell will form the final decision.

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 [30] and [50] two different approximation techniques have been proposed to compute these effects with a lower computational complexity.

Table 2b 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 3 compares three major distributed CAC schemes

In this table, *Naghshineh proposed basic distributed* and Schwartz [17], *shadow cluster* refers to the work of Levin et al. [27] and *stable dynamic* is due to Wu et al. [30].

Table 3 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

1.5.2 Non Prioritized Scheme

There are several schemes which functions without having any priority. The schemes are given as follows.

1.5.2.1 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. Here we investigate the optimal and near-optimal admission policies proposed for three admission problems, namely,

MINO, MINB and MINC. Although optimal policies are more desirable, near-optimal policies are more useful in practice due to the complexity of optimal policies that usually leads to an intractable solution. Table IV shows a comparison of optimal and near-optimal schemes.

Decision theoretic approaches based on *Markov decision process* (MDP) [49] have been extensively studied to find the optimal CAC policy using standard optimization techniques [51].

However, for simple cases such as the one of an isolated cell in a voice system, simple Markov chains have been applied successfully [12]. 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.

Table 4 Comparison of Optimal CAC Schemes

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

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 [51]. 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. [12] 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, a Markov chain similar to the one stated before describes channel occupancy. Although admission policies derived from the MDP formulation of the CAC [54], [55] 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 [55], [56].

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 [52], [57]–[60]. In particular, Choi et al. [61] presented a centralized CAC based on SMDP, Kwon et al. [57] and Yoon et al. [62] proposed distributed CAC schemes based on SMDP, all for non-adaptive multimedia applications. Xiao et al. [58] developed an optimal scheme using SMDP for adaptive multimedia applications. Adaptive multimedia applications can change their bit-rate to adapt to network resource availability.

1.5.2.2 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 [63], genetic algorithms [64], [65], and reinforcement learning [66] have been proposed.

1.6 Other Admission Control Scheme

There are other schemes which are also efficient in handle call admission in a network system.

1.6.1 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 [14], [23], [29], [67], [68]. 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. 1.20 shows a classification of guard channel based CAC schemes in single service and multiple services systems. In the figure, *multiple cutoff priority* [23] and *thinning scheme* [68] are the multiple services counterparts of GC and FGC

schemes in single service systems respectively. In this context, the *thinning scheme* [68] 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 a_{ij} (i = 0, . . ., C and j = 1, . . . , 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 a_{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 λj for class j. Call durations are exponentially distributed with parameter μ. A Markov chain in which the state variable is the number of busy channels in the cell can characterize this system. Let Pn denote the stationary probability at state n, $\rho_j = \frac{\lambda_j}{\mu}$ and $\alpha_k = \sum_{j=1}^{r} \alpha_{k,j}\rho_j$ Using balance equations we have

$$P_n = \frac{\prod_{k=0}^{n-1} \alpha_k}{n!} P_0 \tag{1.18}$$

where,

$$P_0 = \left[\sum_{n=0}^{C} \left(\frac{\prod_{k=0}^{n-1} \alpha_k}{n!} \right) \right]^{-1} \tag{1.19}$$

Then the blocking probability for class *j* is given by

$$P_b{}^j = \sum_{i=T+1}^{C} (1 - \alpha_{ij}) P_i \tag{1.20}$$

Similarly, a natural extension to the basic GC can be achieved by setting different reservation thresholds for each class of service. Pavlidou [89] analyzed an integrated voice/data cellular system using a two-dimensional Markov chain. Haung et al. [87] analyzed the *movable boundary* 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. [67], [70] 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 [71] 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 [72].

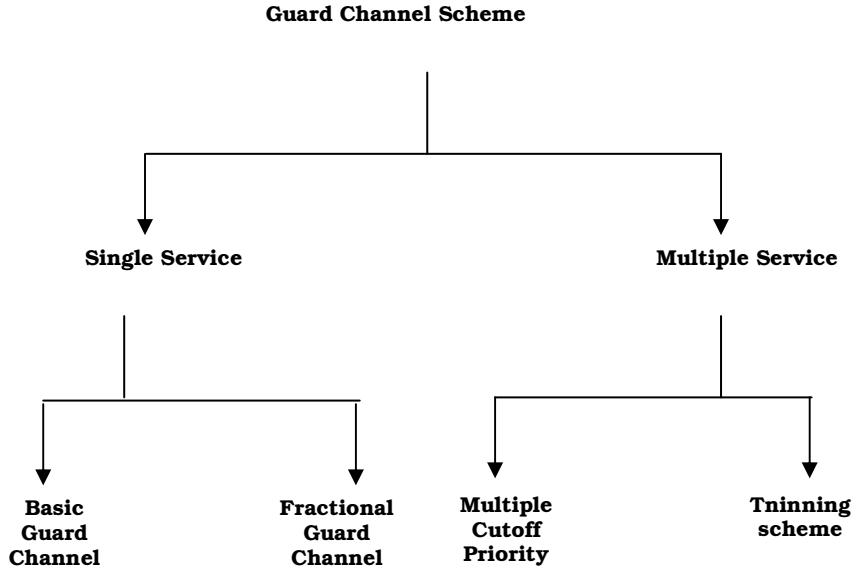


Fig. 1.20 Single service and multiple services guard channel schemes.

1.6.2 Hierarchical Schemes

As mentioned earlier, micro-pico-cell systems can improve spectrum efficiency better than macro cell 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 [73–76] were proposed to take advantage of both macro cell and micro cell systems. Fig. 1.21 shows an example of a hierarchical cellular system. In this architecture, overlaid microcells cover high-traffic areas to enhance system capacity. Overlaying macro cells 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 macro cells than to provide them in microcells [77]. 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.

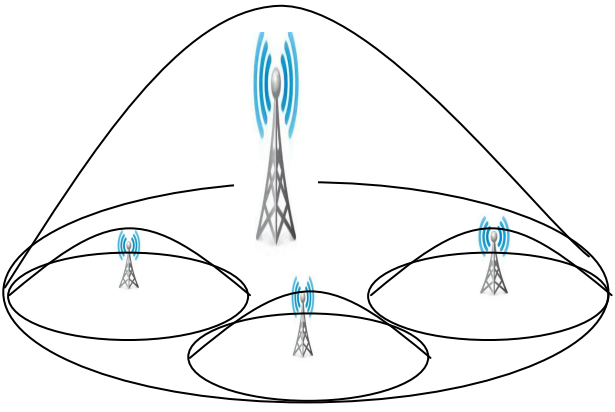


Fig. 1.21 A hierarchical system of micro/macro cells.

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.

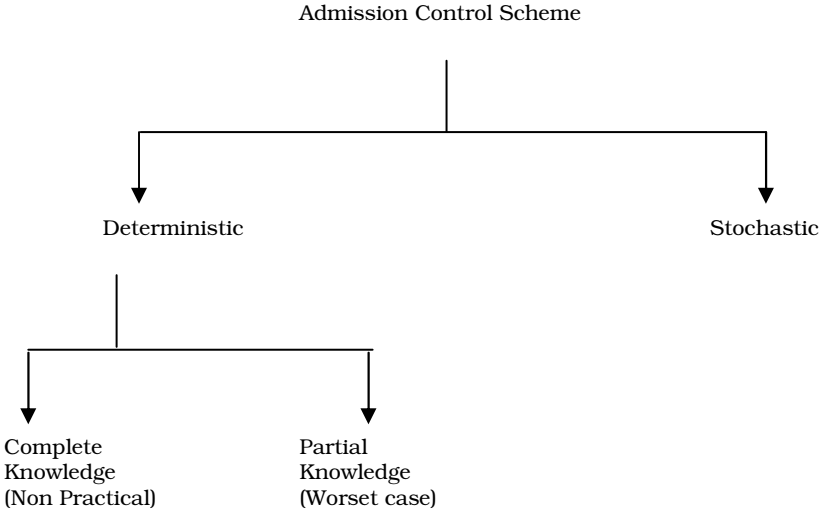


Fig. 1.22 Call admission control schemes.

1.6.3 Complete Knowledge Schemes

User mobility has an important impact in wireless networks. If the mobility pattern is partially [14] or completely [79] 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]. Nevertheless, such an imaginary perfect knowledge scheme is helpful for benchmarking purposes [79]. Fig. 1.22 depicts a classification of CAC schemes according to their knowledge about user mobility. Partial knowledge schemes must reserve resources in several cells [14] 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 [28], [45], [80], [81], some others are based on geometrical modeling of user movements and street layouts [82], and some others are based on artificial intelligence techniques [83]. For instance, the distributed CACs introduced before are based on probabilistic mobility information.

1.6.4 Economic Schemes

Economic models are widely discussed as a means for traffic management and congestion control in provider's networks [84–86]. 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 [86]. 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, 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

$$\phi = \alpha p_b + \beta p_d \tag{1.21}$$

where α and β are constants that denote the penalty associated with blocking a new call or dropping an ongoing call respectively (with $\alpha > \beta$ to reflect the costly call dropping). Earlier we showed that pb and pd are functions of new call and handoff call arrival rates ν and λ. Therefore

$$\phi = f(\lambda) \tag{1.22}$$

where f is a monotonic and non decreasing function of λ. Let us define U as the user utility function in terms of the QoS metric Φ, and let $U = g(\Phi)$,

where g is a monotonic and non-increasing function of Φ. Therefore, the utility function U is maximized at Φ = 0. Let λ* denote the optimal arrival rate for which U is maximized. In [86],it has been shown that the sufficient condition for λ* is that

$$\left.\frac{dU}{d\lambda}\right|_{\lambda=\lambda^*} = 0 \tag{1.23}$$

Using the optimal arrival rate λ* obtained, 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

$$\lambda_{in}(t) = \big(\lambda(t)+\nu(t)\big)H(t), \qquad 0 \le H(t) \le 1 \tag{1.24}$$

where λin(t) is the actual new call arrival rate at time t. H(t) must be designed in such a way that always

$$\lambda_{in}(t) \le \lambda^*, \tag{1.25}$$

and consequently

$$H(t) \le \min\left\{1, \frac{\lambda^*}{\lambda(t)+\nu(t)}\right\} \tag{1.26}$$

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 [86]

$$D(t) = e^{-\left(\frac{p(t)}{p_0}-1\right)^2}, \quad p(t) \ge p_0 \tag{1.27}$$

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). The price that should be set at time t to obtain the desired QoS can be expressed as

$$p(t) = p_0\left(1+\sqrt{\max\left\{0, -\ln\frac{\lambda^*}{\lambda(t)+\nu(t)}\right\}}\right) \tag{1.28}$$

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.

1.7 Call Admission Control Schemes Based on Fuzzy Logic and Evolutionary Algorithms

In this section we describe call admission control schemes using different algorithm.

1.7.1 Call Admission Control Using Fuzzy Logic

Jun Ye et al [88] proposed a call admission control (CAC) scheme using fuzzy logic for the reverse link transmission in wideband code division multiple access (CDMA) cellular communications. The fuzzy CAC scheme first estimates the effective bandwidths of the call request from a mobile station (MS) and its mobility information, and then makes a decision to accept or reject the connection request based on the estimation and system resource availability. Numerical results are given to demonstrate the effectiveness of the proposed fuzzy CAC scheme in terms of new call blocking probability/handoff call dropping probability, outage probability, and resource utilization.

Y. H. Chen et al [89] proposed outage-based fuzzy call admission controller with multi-user detection (OFCAC-MUD) is proposed for wideband code division multiple access (WCDMA) systems. The OFCAC-MUD determines the new call admission based on the uplink signal-to-interference ratios from home and adjacent cells and system outage probabilities. The OFCAC-MUD possesses both the effective reasoning capability of a fuzzy logic system and the aggressive processing ability of MUD. Simulation results reveal that OFCAC-MUD without power control (PC) improves the system capacity by 70.5% as compared to an SIR-based CAC-RAKE with perfect PC. It also enhances the system capacity by 53.9% as compared to an OFCAC-RAKE with perfect PC, by 6.7% as compared to an SIR-based CAC-MUD without PC and by 12.9% as compared to an OFCAC-MUD with perfect PC, given the same outage probability requirements. Moreover, OFCAC-MUD can prevent the violation of outage probability requirements in the hotspot environment, which is hardly achieved by SIR-based CAC.

Chung-Ju et al. [90] proposed a neural fuzzy call admission and rate controller (NFARC) scheme for WCDMA cellular systems providing multirate services. The NFARC scheme can guarantee the quality of service (QoS) requirements and improve the utilization of the system. Simulation results show that the NFARC scheme achieves low forced termination probability and high system capacity even in the bursty traffic conditions. NFARC accepts users more than intelligent call admission controller (ICAC) by an amount of 45.35%.

In the present and next generation wireless networks, cellular system remains the major method of telecommunication infrastructure. Since the characteristic of the resource constraint, call admission control is required to address the limited resource problem in wireless network. The call dropping probability and call blocking probability are the major performance metrics for quality of service (QoS) in wireless network. Chenn-Jung Huang et al[91] proposed an adaptive call admission control and bandwidth reservation scheme using fuzzy logic control concept to reduce the forced termination probability of multimedia handoffs.

The authors adopt particle swarm optimization (PSO) technique to adjust the parameters of the membership functions in the proposed fuzzy logic systems. The simulation results show that the proposed scheme can achieve satisfactory performance when performance metrics are measured in terms of the forced termination probability for the handoffs, the call blocking probability for the new connections and bandwidth utilization.

1.7.2 Call Admission Control Using Genetic Algorithm(GA)

Shyamalie Thilakawardana and Rahim Tafazolli in [92] anticipated that a wide variety of data applications, ranging from WWW browsing to Email, and real time services like packetized voice and videoconference will be supported with varying levels of QoS. Therefore there is a need for packet and service scheduling schemes that effectively provide QoS guarantees and also are simple to implement. This paper describes a novel dynamic admission control and scheduling technique based on genetic algorithms, focusing on static and dynamic parameters of service classes. A performance comparison of this technique on a GPRS system is evaluated against data services and also a traffic mix comprising voice and data.

Sheng-Ling Wang et al [93] introduced an adaptive threshold-based Call Admission Control (CAC) scheme used in wireless/mobile network for multiclass services is proposed. In the scheme, each class's CAC thresholds are solved through establishing a reward penalty model which tries to maximize network's revenue in terms of each class's average new call arrival rate and average handoff call arrival rate, the reward or penalty when network accepts or rejects one class's call etc. To guarantee the real time running of CAC algorithm, an enhanced Genetic Algorithm is designed. Analyses show that the CAC thresholds indeed change adaptively with the average call arrival rate. The performance comparison between the proposed scheme and Mobile IP Reservation (MIR) scheme shows that with the increase of average call arrival rate, the average new Call Blocking Probability (CBP) and the average Handoff Dropping Probability (HDP) within 2000 simulation intervals of the proposed scheme are confined to lower levels, and they show approximate periodical trends of first rise and then decline. While these two performance metrics of MIR always increase. At last, the analysis shows the proposed scheme outperforms MIR in terms of network's revenue.

Shengling Wang et al [94] introduced a dynamic multi-threshold CAC scheme is proposed to serve multi-class service in a wireless/mobile network. The thresholds are renewed at the beginning of each time interval to react to the changing mobility rate and network load. To find suitable thresholds, a reward-penalty model is designed, which provides different priorities between different service classes and call types through different reward/penalty policies according to network load and average call arrival rate. To speed up the running time of CAC, an Optimized Genetic Algorithm (OGA) is presented whose components such as encoding, population initialization, fitness function and mutation etc., are all optimized in terms of the traits of the CAC problem. The simulation demonstrates that the proposed CAC scheme outperforms the similar schemes, which means the optimization is realized. Finally, the simulation shows the efficiency of OGA.

Bo Rong et al [95] proposed a mobile agent (MA)-based handoff architecture for the WMN, where each mesh client has an MA residing on its registered mesh router to handle the handoff signaling process. To guarantee quality of service (QoS) and achieve differentiated priorities during the handoff, they develop a proportional threshold structured optimal effective bandwidth (PTOEB) policy for call admission control (CAC) on the mesh router, as well as a genetic algorithm (GA)-based approximation approach for the heuristic solution. The simulation study shows that the proposed CAC scheme can obtain a satisfying tradeoff between differentiated priorities and the statistical effective bandwidth in a WMN handoff environment.

1.7.3 Call Admission Control Using Neural Network (NN)

Conventional methods for developing CAC algorithms are based on mathematical and/or simulation modeling. These methods require making assumptions about the traffic processes. QoS prediction is then done using queuing models to reflect the buffering and transmission behavior. Since this approach can quickly become analytically involved, simplifying assumptions need to be made. For example, it is common to assume that traffic sources are Markovian, or stationary, or that cell arrival patterns depend upon some parametric models like Markov Modulated Poisson Processes (MMPP).

It has recently become known that high-speed network traffic is more complex, and that none of these assumptions is safe. Exact solutions based on analytical methods exist only for restricted traffic and system models. QoS estimation through analysis can also be inaccurate because declared and actual traffic parameters frequently differ.

Clearly results derived from analyses based on such assumptions have limited applicability and will generate inaccurate QoS estimates. To compensate, such CAC schemes force themselves to err on the side of being conservative and thus typically over allocate resources. This leads to inefficiency.

The easiest method for CAC is to accept a call if there is enough available bandwidth to allocate to the call its peak rate. This is the most inefficient method since it entirely ignores statistical multiplexing. The most well known analytical result for CAC is the equivalent bandwidth method [100,101]. This method provides a simple formula to compute the amount of bandwidth needed to meet a call's loss requirement, given its peak rate, mean rate, and average burst duration. The equivalent bandwidth yields a bandwidth that lies between the call's peak and mean rates. This method is exact only asymptotically, as the buffer size approaches infinity and the cell loss probability approaches zero. Although far superior to a CAC scheme that allocates peak bandwidth, the equivalent bandwidth method still over allocates bandwidth in most cases.

Neural networks are attractive for solving CAC problems because they are a class of approximators that are well suited for learning nonlinear functions. A neural network represents a multiple-input multiple-output nonlinear mapping. A NN can learn this mapping from a set of sample data. Feed-forward neural networks can approximate any piecewise-continuous function with arbitrary accuracy, given enough hidden neurons [102].

Some advantages of using neural networks for CAC are:

1. Neural networks do not require an accurate mathematical model of either the traffic or the system. No assumptions need to be made since the neural network is trained on observed data. Not assuming a specific traffic behavior a priori is a preferable approach because multimedia traffic is not well understood and continuously changing. NNs are also not affected by mistakes in declared traffic descriptors. These features allow a NN to yield more accurate QoS estimation which leads to greater efficiency and robustness.
2. When NNs are properly trained, they can generalize and extrapolate additional details of the function mapping the inputs to outputs. If the training set if sufficiently large, a NN will generalize accurately and will produce accurate outputs for inputs not in the training set. This also contributes to robustness of the CAC scheme.
3. NNs are adaptable since they can be retrained in real-time using the latest measurements

1.7.3.1 Implement CAC with Neural Networks

Neural networks can be used to solve the CAC problem and all its variants. First we describe a model using a NN for the single-buffer-per-link case, with a FIFO (first-in first-out) buffer. The NN contains n inputs, denoted $\{u_0, u_1, ...u_n\}$ and a single output Y = f(u), where f(u) denotes the transfer function from inputs to outputs. The output is the QoS estimate. This can be the estimate for the amount of bandwidth needed on the link to carry all the calls whose traffic enters the FIFO buffer, or the buffer delay, or the buffer's loss rate. This version of this model, with a single output representing a particular QoS estimate, only supports a service that makes guarantees on one QoS parameter. For services that provide guarantees on m QoS parameters, a version of this model with m outputs, one per QoS parameter, could be used.

In another version of this model, a single output is used that takes on binary values that represent accept or reject decision. In this version, particular QoS estimates are internal to the NN. Such a NN is called a classifier. With this choice for the output, the NN can be used to represent any definition of a feasible stream, including definitions that involve multiple QoS constraints (e.g., delay and CLR).

In the case of a single output representing a QoS estimate, a call is admitted to the network if the QoS estimate for the new candidate aggregate stream is below the most stringent QoS requirement (i.e., the decision threshold) for all calls in the stream. In the case of multiple outputs, each QoS estimate needs to be compared the relevant decision threshold, before a call is accepted. In the case of a single binary output (e.g., trained to learn 0 for accept and 1 for reject), the output is compared to a threshold such as 1/2. The choice of output influences by which CAC problem is being modeled. Consider the case in which the output is a queue related parameter, such as delay or loss. When FIFO queuing is used, there is only a single loss rate or distribution of delay associated with that queue.

The NN predicts the QoS of the aggregate stream of superposed traffic sources. When many sessions that have different QoS requirements are multiplexed together, the switch ensures the most stringent of all the delay and/or loss requirements. (Hence, some sessions will experience better QoS that they requested.) Thus this model can support multiple traffic classes, but they will all receive the same QoS. If a scheduling mechanism (e.g., earliest deadline first) is used to prioritize among traffic classes, then multiple loss rates (one for each class) could be computed for a single buffer. A NN used for this scenario would require multiple outputs, one for each class.

We now consider a second model for the multiple-buffers-per-link problem. In this case, there is at least one QoS estimate per buffer. If there are b buffers, and x QoS parameters per buffer, then this NN model could have b_x outputs. Alternatively, a single binary output (representing an accept or reject decision for a given call) could still be used in this case, since all call types and services share a single transmission link.

A third way of modeling CAC problems is to use a modular design in which multiple NN units are organized in a hierarchical fashion. Modularity is a means to solve a complex computational task by decomposing it into simpler subtasks and then combining the individual solutions. This approach is attractive if the functional relationship between the neural network inputs and outputs is very complex, and if parts of this function can naturally be separated. Consequently, the modules of the network tend to specialize by learning different regions of the input space. The decomposition should be structured so as to facilitate this.

An example, based on the work in [115], is given in Figure 1.23. There is a bank of NN units in level 1 of this model. Each NN unit is associated with a single traffic class, and its inputs correspond to descriptors for that class. Let's assume that one virtual path is assigned for each traffic class. The output gives the bandwidth estimate for all the calls in that class, i.e., for the VP assigned to that class. The NN unit at level 2 takes as inputs the bandwidth per class for each VP and outputs the link bandwidth needed to support the VPs.

The attraction of this model is that it naturally separates out the two levels of statistical multiplexing that occur in a switch that supports VPs. The NNs in level 1 determine the amount of multiplexing gain for mixing calls from the same class onto a single VP, while the NN in level 2 determines the amount of multiplexing gain for mixing VPs onto a link. Another example of a modular NN design can be found in [114].

Modular networks offer several advantages over a single neural network [102]. First, the training is faster, which allows the NN units to be more adaptable. Second, the representation of input data developed by a modular network tends to be easier to understand than in the case of an ordinary multilayer NN. Third, this type of design should lead to more accurate estimates since the input-output mapping that each NN unit has to learn is simpler. It has been proven [96] that the number of input-output patterns that a NN can deterministically learn is equal to twice the number of its weights. Fourth, a useful feature of a modular approach is that it also provides a better t to a discontinuous input-output mapping [102].

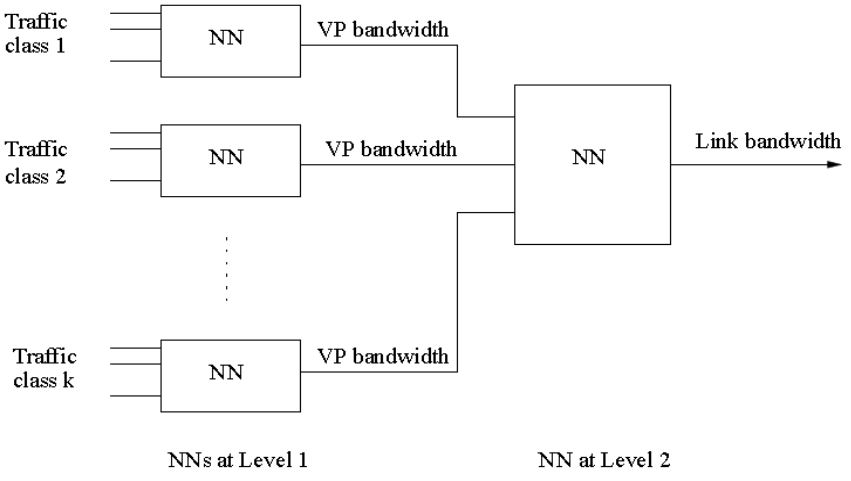


Fig. 1.23 Sample Modular NN Design for CAC

Architecture: A common architecture for all the models described above is a 3-layer feed forward neural network, with a layer of inputs, a single hidden layer of neurons, and an output layer. We establish the following notation for the neural network elements.

For a given set of inputs $u_1,.....,u_n$, the k-th output of the NN is given by

$$Y_k = g^{out}(\sum_{j=1}^{J} W_{jk} V_j + b_k^{out}) \tag{1.29}$$

where W_{jk} is the connection weight from the j-th hidden neuron to the k-th output, and b_k^{out} is the bias for the k-th output. $V_j, j = 1,...J$, V j , is the output of the jth hidden neuron, and is given by

$$V_j = g(\sum_{i=1}^{I} w_{i,j} u_i + b_j) \tag{1.30}$$

The w_{ij} are the weights from the inputs to the hidden neurons, and b j are the biases for the hidden neurons. The functions gout and g(x) are the activation functions for the output layer and the hidden neurons, respectively. These functions are typically either a linear function or a sigmoid function, such as the logistic function $_1 = (1+\exp(av))$ or the hyperbolic tangent function $\tanh(v = 2) = (1+\exp(v))$. In the case of a single output, k = 1, the subscript k can be dropped.

The key to a good NN design lies in the particular choice of inputs, outputs and training technique. Selecting a pair of inputs and outputs for which the desired output is not determinable from the input will clearly not yield an effective CAC algorithm.

1.7.3.2 NN Inputs

The NN inputs can be any input that helps the NN to predict the QoS of an aggregate traffic stream. The NN inputs typically include either traffic descriptors or system state parameters or some combination of both. In this section, we now expand the discussion of these types of inputs. The advantage of having the users supply traffic descriptors is that the network need not spend any resources or time to measure the traffic.

However the disadvantage comes from the fact that there is usually a difference between the declared traffic parameters and the actual traffic parameters since most applications today do not understand well the traffic they generate. When the number of connections in a network becomes large, the difference between the declared and actual traffic parameters can be quite large. The advantages and disadvantages of using measurements as NN inputs are exactly the reverse of the advantages and disadvantages of the user supplied approach.

Examples of system state parameters that can be used as inputs include buffer level, the number of existing calls for each traffic class, and buffer loss rates. It is desirable for the neural network inputs to have the following properties:

Capture key elements of traffic behavior that influence queuing. Many researchers believe that traffic descriptors that capture the correlation and burstiness properties of a traffic stream will be successful for CAC. In order to avoid over allocating or under-allocating resources, it is necessary estimating QoS well, which in turn requires proper traffic and system characterization.

When the NN input vector is additive, the current input vector can be updated efficiently by simply adding the traffic descriptors of the new call to that of the aggregate call (the current input vector). This additive property of traffic descriptors greatly speeds up the decision process of accepting or rejecting a call.

Support a large number of traffic classes. This will make the algorithm robust. Keep the number of inputs reasonably small. This will make the algorithm practical since the forward calculation speed is proportional to the number of weights. We now give examples of NN inputs that have been proposed in recent research efforts.

1.7.3.3 Number of Calls per Traffic Class

In this case, the NN input is a vector s = (s k) whose k^{th} component gives the number of calls in the stream that belong to the k^{th} traffic class. This input has been used by [105, 106, 114]. In [110], they use this input coupled with the link load level. We refer to this particular input as the call vector in subsequent sections of this chapter. The advantage of this approach is that the user need not supply any traffic descriptors at all, and the decision boundary is determinable from this input. The disadvantage of this approach is that it does not scale well in the number of traffic classes. There could be a very large number of traffic classes in any network. It has been suggested [106, 114] that a practical number of classes is less than 100.

1.7.3.4 Counts of Arrivals

In [115] they used on-line traffic measurements for the NN inputs. In each interval q, the number of cell arrivals, N S(q), is counted for each stream S. If one keeps track of the arrival process over consecutive intervals, and also provides this data to the NN, then the NN can be trained to capture the correlations that exist among cell arrivals. This approach requires a careful choice of the measurement interval. The advantages of this input are that this information entirely characterizes the input stream and that the user need not supply any traffic descriptors. The disadvantage of this choice for NN inputs is that is does not scale in the number of calls supported simultaneously. This choice of inputs is not considered very practical since the measurement intervals typically need to be very large.

1.7.3.5 Variance of Counts

In [111], we used the variance of counts (VOCs) as NN inputs. To calculate or measure VOCs, time is divided into intervals of equal length. As above, N S(q) denotes the number of cells or packets arriving in interval q for stream S. Let Ns (q; h) denote the number of cells arriving in the interval consisting of intervals q through h. Let S denote the mean of Ns(q). The variance of counts for an interval of length m is defined by

$$VOC_s(m) = \frac{\text{var}\{N_s(q+1, q+m)\}}{m} \tag{1.31}$$

where q is an arbitrary time slot. For a given stream S, the NN inputs are scaled versions of S and $VOC_S(m)$ for m = 1; 2; 4; :::; 2 M , where M + 1 is the number of VOCs used. To limit the number of NN inputs while considering a representative set of VOCs, we used VOCs over intervals of exponentially increasing length.

The VOC traffic descriptor is an not normalized IDC as described in [103], that is, VOC S(m) = IDC S(m) S. Not normalized IDCs are preferable because then the VOCs are additive, i.e., if S is the sum of statistically independent streams S_i , then VOC S(m) is the sum of VOC Si(m) over the Si .

The advantages of these NN inputs include: (i) VOCs characterize all second-order statistics of the stream, (ii) they are additive, (iii) moments of interval counts have been shown to accurately predict queuing delay for some models [103], and (iv) this method is independent of the number of traffic classes.

It is known [103] that IDC S(m) converges to var(X) = E^2(X), the squared coefficient of variation of the inter-arrival time X, and so VOC S(m) also converges to a constant if the inter-arrival time has finite variance. One can thus choose the number M + 1 of VOCs to be large enough so that $VOC_S(2M)$ is close to the limit for most streams. This limits the number of NN inputs to M + 2, which is typically much smaller than the number of traffic classes as used in

These NN inputs can be measured by either the user or the network. They could be calculated by the user when the user's traffic is a bursty on-off process. VOCs can be prior computed for a set of traffic classes and stored in a table, in which case a user only needs to indicate a traffic class in the call request.

1.7.3.6 Power Spectral Density Parameters

Another type of input that can capture correlation and burstiness properties of traffic streams is the power-spectral-density (PSD) function in the frequency domain. Using the PSD as inputs was first proposed in [87]. All of the above inputs discussed focus on the time-domain. The PSD is the Fourier transform of the autocorrelation function of the input process. A traffic source can be characterized by a PSD can be described by three parameters (u;v;w): the DC component (u), the half-power bandwidth (v), and the average power (w). As u increases, the traffic loads increases; as v decreases, the input power in the low frequency band increases; and as w increases, the variance of the input rate increases.

There are two advantages to this approach. First, it has been shown in [107] that the low frequency band of the input PSD has a dominant impact on the queuing performance while the high frequency band can usually be neglected. This is because the low frequency component of the PSD contains the correlation component. Larger the low frequency component becomes, the burstier the traffic source. Second, since the PSD has the additive property, so do these three parameters.

In [97], the CAC controller is designed so that the user can input three simple parameters: its peak rate, mean rate and peak cell rate duration. The controller applies the Fast Fourier Transform (FFT) to these inputs and outputs the three PSD parameters (u; v; w) which are in turn fed as inputs to a NN.

1.7.3.7 Entropy

Entropy has been proposed as a tra c descriptor in [99, 112]. The entropy of traffic streams is attractive as a descriptor because it can capture the behaviour of correlations over many time scales. The entropy has been used as an input for CAC in [99, 112] but it has not yet been tried in a CAC algorithm based on neural networks.

1.7.3.8 NN Outputs

A NN output variable can represent any of the following:

1. **Accept/Reject Decision.** With this output, the NN has to learn the boundary between the feasible and infeasible performance regions for a given input space.
2. **Loss Rate.** In this case, the NN predicts the average buffer overflow rate. Since the loss rate can have an exponentially wide range, from 10-1 to 10-9 it is common to use log(loss) instead.
3. **Delay.** In this case, the NN predict the average buffer delay, or average buffer occupancy.
4. **Jitter.** In this case, the NN predicts the variation of the buffer delay. 5. Bandwidth. In this case, the NN predicts the amount of bandwidth needed to achieve a specific QoS level for the given input stream.

5. **P^{th} percentile delay.** In this case, the NN predicts the value D such that the probability that a cell or packet experiences a delay less than or equal to D is p%. The percentile is typically chosen to be around 90%. In this approach, all the calls need to have the same percentile requirement.
6. **Probability distribution of delay.** In this case, the NN output represents the probability that the delay experienced will be less than the requested delay (which can be one of the NN inputs), conditioned on the buffer status (represented by the other NN inputs). This method, proposed in [113], works well when the probability is conditioned with respect to the number of active connections. This method allows calls to have different delay requirements.

1.7.3.9 Compression of Neural Network Inputs

In Section 1.7.3.5, we discussed the mapping of the call vector to a vector of parameters related to second order statistics of the aggregate traffic process (based on VOCs or the PSD). Such a mapping can be considered a compression of the call vector (whose dimension is the number of traffic classes, which can be in the hundreds) to a smaller vector whose dimension is independent of the number of traffic classes. One benefit of reducing the number of NN inputs is that the NN output can be computed in less time, thus allowing more calls to be processed per second. The best compression is one that maps the call vector to the fewest parameters without reducing performance significantly.

In this section, we present a method for finding a linear compression of the call vector to I parameters (where I is any positive integer) that is optimal in the sense of minimizing the output error function. Linear compression is chosen so that the compressed parameters are additive. A similar method was presented in [111] to compress vectors of VOCs, where it was shown that compression to three parameters did not result in a significant reduction of performance. An advantage of compressing call vectors instead of VOCs is that the statistics of calls need not be known or measured.

Although we focus on the call vector in this section, the method can be applied to any choice of NN inputs that are additive. The compression method involves adding a new hidden layer to the NN, between the input layer and the current hidden layer, as shown in Fig. 1.24. The neurons in the new hidden layer have linear activation functions, and there is one such neuron for each compressed parameter.

For this design, the previous equation described in architecture is replaced by the following two equations

$$V_j = g\left(\sum_{i=1}^{I} w_{ij} + b_j\right) \tag{1.32}$$

$$v_i = \sum_{l=1}^{L} \alpha_{li} s_l \tag{1.33}$$

The outputs v_i of the hidden linear neurons form the compressed vector corresponding to the call vector s, and the weights l_i from the l^{th} input to the first hidden layer form the compressed vector corresponding to the l^{th} traffic class. Above equation can be expressed in vector form as

$$v = \sum_{l=1}^{L} \alpha_l s_l \tag{1.34}$$

where v is the compressed vector for call vector s and l is the compressed vector for a single call of traffic class 1.

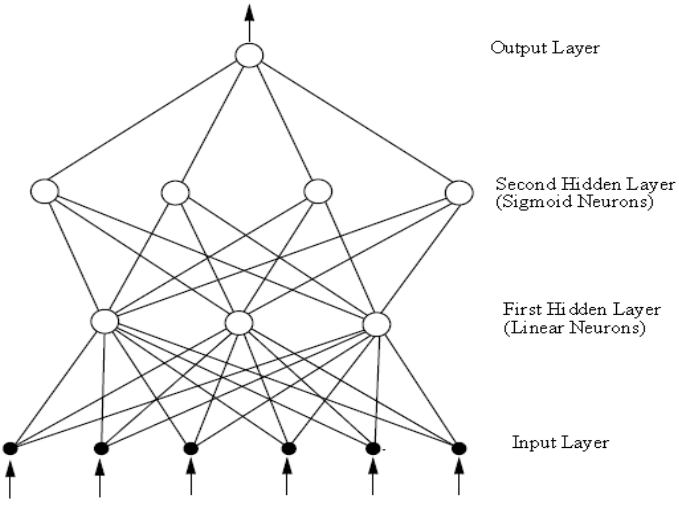


Fig. 1.24 NN with Additional Hidden Layer for Compressing Input

This method was motivated by the well-known technique of using a hidden layer of linear neurons for image or data compression [102, 104, 108]. Assuming that NN weights (including the l_i) are found that minimize the output error function, the compressed parameters v i are optimal by definition. Once this NN is trained and thus the matrix $\{\alpha_{li}\}$ is obtained, the original input layer is no longer required. The first hidden layer becomes the new input layer; i.e., the compressed parameters v_i are now used as the NN inputs.

Since the compressed parameters are additive, when a new call of class l arrives, the compressed vector for the new call vector is updated simply by adding the compressed vector l for the new call. To perform this operation quickly, the compressed vector l corresponding to each traffic class l can be stored in a look-up table.

Consider the special case in which only a single compressed parameter is used. Thus, a single parameter α_l is computed for each traffic class l, and the compressed parameter for a given call vector is equal to the sum of α_l over all calls

represented by the call vector. Moreover, it is possible to scale the α_l so that a given call vector falls into the call acceptance region (learned by the neural network) if and only if this sum is less than the link bandwidth. Therefore, compression to a single parameter corresponds to learning the equivalent bandwidth for each traffic class.

1.7.3.10 Design of CAC Controller

Fig. 1.25 depicts a method of incorporating a NN into the design of a CAC algorithm.

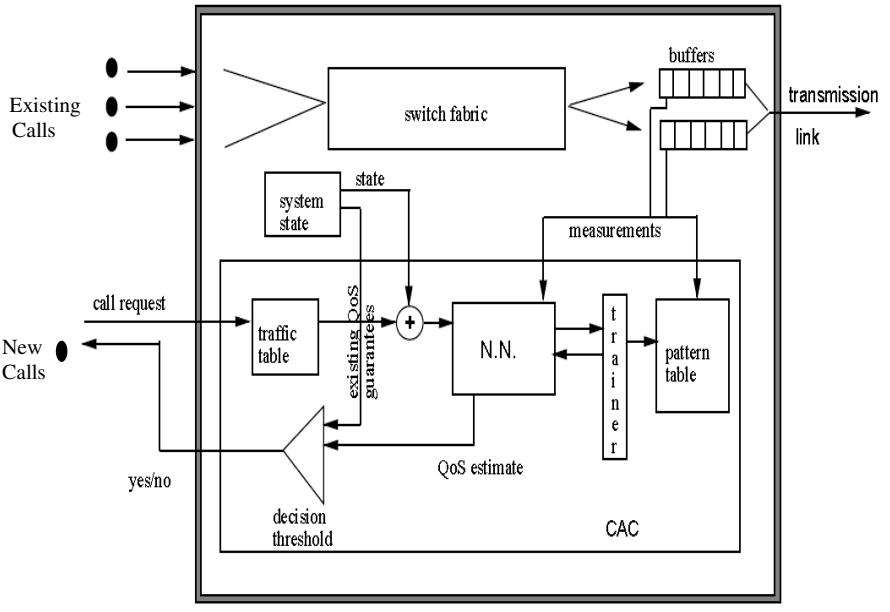


Fig. 1.25 Design of CAC Controller using a Neural Network

This figure is general in that it illustrates potential CAC inputs, and not required inputs. The CAC inputs can come from the user, the system state or measurements, or some combination thereof. A traffic table can be used to convert user supplied traffic descriptors into VOCs or PSDs, if needed. If both the system state and the user's traffic descriptors (or the mapped version of the user's descriptors) are used, we assume they have the additive property (indicated by the adder in Fig. 1.25). The system state is updated each time new connections are established or torn down. The pattern table and trainer are only present when the system is designed for on-line training.

1.7.3.11 Off-Line Training

The feed forward NNs described above can be trained using standard back propagation algorithms and their variations. In this section we brie y discuss methods for generating the training set (the set of input-output patterns) used for back propagation and present a non-standard error function that helps to achieve the asymmetric goal of the CAC problem. Methods for the more difficult problem of online training are presented in the next section. When a NN is trained offline for CAC, the training set consists of a large number (typically 1000 or more) of input-output patterns (u^k, y_k) that are usually generated by simulating the traffic and queuing processes. The traffic can be modeled as an on-o Markov chain, a Markov-modulated Poisson process, or a self-similar traffic process. More complex traffic models or traces of traffic from actual network can also be used. A fixed number of traffic classes can be defined by specifying the parameters for each class, or an infinite number of classes can be obtained by allowing any choice of parameters within some range.

Aggregate traffic streams must be generated that cover all regions of the space of input patterns u^k. If the number of traffic classes and the maximum number of calls per class are small, it may be possible to use a training set that includes every possible call state. Otherwise, one way to generate an aggregate stream is to rst randomly select the number of calls in the stream (between 1 and the maximum possible number of calls), and then randomly select the traffic class for each call. If the input pattern u k is the call state, then it is known immediately. If some other choice is used for u^k (e.g., VOCs), then u^k must be obtained using analysis or simulation. For each aggregate stream, a simulation can be run to determine the resulting performance measure y_k. Each simulation should be run long enough to obtain an accurate estimate of the performance measure. For example, 10^8 packets would need to be observed to accurately estimate a loss rate of 10^{-6}. If such long simulations are not feasible, the virtual output buffer method discussed in the next section can be used to improve the estimate through extrapolation.

Once the training set is generated, the NN can be trained using a version of back propagation. A commonly used error function for training is

$$J = \sum_{k=1}^{K} \left| f(u^k) - y_k \right|^N \tag{1.35}$$

where, usually N = 2 (equivalent to the mean squared error).

Recall that the CAC objective is asymmetric: to accept as many feasible input patterns as possible while rejecting all infeasible input patterns. After the NN is trained, the decision threshold can be adjusted so that all infeasible input patterns in the training set are rejected. That is, assuming that a NN output greater than the decision threshold corresponds to a reject decision, the threshold is chosen so that it is slightly less than the smallest NN output for an infeasible input pattern in the training set.

In order to accept as many feasible streams as possible, the maximum error over all infeasible training patterns u should be minimized, so that the decision

threshold can be chosen as large as possible without accepting any infeasible patterns. One way to achieve this objective is to use the following asymmetric error function

$$J = \sum_{feasible\,k} \left| f(u^k) - y_k \right|^2 + \sum_{infeasible\,k} \left| f(u^k) - y_k \right|^N \qquad 1.36$$

where N > 2. For large N, minimizing the second sum will tend to minimize the maximum error $\left| f(u^k) - y_k \right|$ over all infeasible patterns.

1.7.3.12 Online Training

In online training, the NN is trained continuously or frequently, based on actual measurements obtained while the NN is being used for CAC. Online training is useful if the CAC needs to adapt to changing network characteristics or new traffic classes or to fine tune an offline-trained NN using more accurate measurements.

Online training is more difficult than offline training in part because the call state changes frequently. Thus, the performance measurement (NN output) obtained for a given training pattern is based on fewer packets (or cells) and is therefore subject to more statistical variability (i.e., noise). In online training, unlike offline training, one cannot x the call state and observe the performance of a large number of packets in order to obtain a good estimate of the resulting loss rate or average delay. This difficulty can be reduced by exploiting the ability of NNs to learn the average of several different measurements associated with the same call state, as explained below.

However, online training can be slow because a large number of packets must be observed before a small loss rate can be estimated with any accuracy. For example, more than a billion packets must be observed to accurately estimate a loss rate on the order of 107. In this case, a packet loss can be considered a rare event. The online training can be made faster by using the ability of a neural network to extrapolate from estimates of measures that are based on more common events. For example, in the virtual output buffer method, discussed below, the NN learns the loss rate that would occur if the packets were fed into imaginary queues with smaller service rates than the actual queue, and extrapolates that knowledge to improve the estimate of the actual loss rate.

For a NN CAC to be adaptive, old measurements for a given input pattern must eventually be “forgotten" and replaced by new measurements for similar input patterns. There is a tradeoff between adaptability and accuracy: If the NN is trained using past measurements made over a large time window, it can achieve good accuracy but will not adapt quickly to network changes. If a small time window is used (so that the NN quickly forgets past measurements), faster adaptation is achieved at the cost of less accuracy.

Another reason online training is more difficult than offline training is that input patterns that are marginally unacceptable occur rarely or never, assuming the CAC is performing well. If the NN remembers that these patterns are

unacceptable, it may not be able to adapt to changing network conditions that cause these patterns to become acceptable. If the NN eventually “forgets" these patterns, it will start to accept infeasible call vectors and thus perform poorly for some period of time while it is relearning the decision boundary. Such behavior should be avoided, since it is more important to reject a given infeasible call vector than to accept a given feasible call vector, i.e., the NN should achieve “safe-side" control. One way to help achieve this goal to start with a NN CAC that has been trained offline to perform a conservative version of CAC, such as one based on peak rate or equivalent bandwidth, and then use online training with a slow learning rate so that the CAC gradually learns to accept more calls. The virtual buffer method also helps to achieve safe-side control by learning more quickly that the call vector is approaching the decision boundary.

Other problems that will be addressed in this section include how to decide which patterns to store, given a bounded storage capacity, and summarizing different measurements for the same input pattern, in order to reduce the training set and thus reduce the time required to train the NN. In the following subsections, we assume that the NN input is the call vector and that the NN has a single output. However, the methods are applicable to other choices for the NN input and can easily be extended to multiple NN outputs.

1.7.3.13 Procedures for Online Training

In online training, we cannot assume that the queuing system reaches equilibrium between changes in the call vector. Therefore, the neural network can only learn, for each possible state of the call vector, the average performance of packets that are admitted when the call vector is in that state. A training pattern will therefore consist of a pair (s^k, y_k) and a size m_k where y_k is a measurement of the average performance of m_k packets that were admitted when the call vector was s^k. Since the call vector can change at arbitrary times, the intervals over which these measurements are performed need not have equal length. In an extreme case, each measurement can correspond to a single packet.

The simplest method for online training is to perform one step of gradient descent (back-propagation) per measurement, in the same order that the measurements are observed. However, this method does not provide the benefit of storing a large window of past measurements, including measurements for call vectors that rarely occur, and using each pattern repeatedly for training, so that better convergence is achieved and past measurements are not quickly forgotten.

Therefore, training patterns (each consisting of a call state and a corresponding measurement) should be stored in a pattern table. (See Fig 1.25) The time window over which past measurements are stored can be selected depending on the desired degree of adaptability. If the network characteristics are not expected to change and good estimation accuracy is desired, then a very large time window can be used. In addition, the total number of patterns that are stored should be limited, either because of storage limitations or to limit the time required to train the NN. To achieve this, a circular buffer can be used, so that the newest pattern replaces the oldest pattern.

1.8 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, adaptively 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.

References

1. Viterbi, A.: Principles of Spread Spectrum Communication. Addison–Wesley (1995)
2. TIA/EIA/IS–95, Mobile Station–Base Station Compatibility Standard for Dual–Mode Wideband Spread Spectrum Cellular System. Telecommunication Industry Association (May 1995)
3. Cooper, G.R., Nettleton, R.W.: A Spread–Spectrum Technique for High–Capacity Mobile Communications. IEEE Transactions on Vehicular Technology VT–27, 264–275 (1978)
4. Rappaport, T.S.: Wireless communications. Prentice Hall (1996)
5. Gilhousen, K.S., Jacobs, I.M., Padovani, R., Viterbi, A.J., Weaver Jr., L.A., Wheatly III, C.E.: On the capacity of a cellular CDMA system. IEEE Transactions on Vehicular Technology 40, 303–312 (1991)
6. Viterbi, A.J.: Wireless Digital Communication: A View Based on Three Lessons Learned. IEEE Communications Magazine 29, 33–36 (1991)
7. Jung, P., Baier, P.W., Steil, A.: Advantages of CDMA and Spread Spectrum Techniques over FDMA and TDMA in Cellular Mobile Applications. IEEE Transactions on Vehicular Technology 42, 357–364 (1993)
8. Corazza, G.E., De Maio, G., Vatalaro, F.: CDMA Cellular Systems Performance with Fading, Shadowing, and Imperfect Power Control. IEEE Transactions on Vehicular Technology 47, 450–459 (1998)
9. Wong, D., Lim, T.J.: Soft handoffs in CDMA mobile systems. IEEE Personal Communications Magazine 4(6), 6–17 (1997)
10. Prakash, R., Veeravalli, V.V.: Locally optimal soft handoff algorithms. IEEE Transactions on Vehicular Technology 52(2), 231–260 (2003)
11. Lin, Y.-B., Pang, A.-C.: Comparing soft and hard handoffs. IEEE Transactions on Vehicular Technology 49(3), 792–798 (2000)
12. Ramjee, R., Towsley, D., Nagarajan, R.: On optimal call admission control in cellular networks. Wireless Networks 3(1), 29–41 (1997)

13. Valko, A.G., Campbell, A.T.: An efficiency limit of cellular mobile systems. Computer Communications Journal 23(5-6), 441–451 (2000)
14. Talukdar, A.K., Badrinath, B., Acharya, A.: Integrated services packet networks with mobile hosts: Architecture and performance. Wireless Networks 5(2), 111–124 (1999)
15. Lu, S., Bharghavan, V.: Adaptive resource management algorithms for indoor mobile computing environments. In: Proc. ACM SIGCOMM 1996, Palo Alto, USA, pp. 231–242 (August 1996)
16. Hong, D., Rappaport, S.S.: Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. IEEE Transactions on Vehicular Technology 35(3), 77–92 (1999)
17. Katzela, I., Naghshineh, M.: Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey. IEEE Personal Communications Magazine 3(3), 10–31 (1996)
18. Chang, C.-J., Huang, P.-C., Su, T.-T.: A channel borrowing scheme in a cellular radio system with guard channels and finite queues. In: Proc. IEEE ICC 1996, Dallas, USA, vol. 2, pp. 1168–1172 (June 1996)
19. Wu, X., Yeung, K.L.: Efficient channel borrowing strategy for multimedia wireless networks. In: Proc. IEEE GLOBECOM 1998, Sydney, Australia, vol. 1, pp. 126–131 (November 1998)
20. Chu, T.-P., Rappaport, S.S.: Generalized fixed channel assignment in microcellular communication systems. IEEE Transactions on Vehicular Technology 43(3), 713–721 (1994)
21. Chang, C.-J., Su, T.-T., Chiang, Y.-Y.: Analysis of a cutoff priority cellular radio system with finite queueing and reneging/dropping. IEEE/ACM Transactions on Networking 2(2), 166–175 (1994)
22. Li, W., Chao, X.: Modeling and performance evaluation of a cellular mobile network. IEEE/ACM Transactions on Networking 12(1), 131–145 (2004)
23. Li, B., Chanson, S., Lin, C.: Analysis of a hybrid cutoff priority scheme for multiple classes of traffic in multimedia wireless networks. Wireless Networks 4(4), 279–290 (1998)
24. Lin, P., Lin, Y.-B.: Channel allocation for GPRS. IEEE Transactions on Vehicular Technology 50(2), 375–384 (2001)
25. Fang, Y., Zhang, Y.: Call admission control schemes and performance analysis in wireless mobile networks. IEEE Transactions on Vehicular Technology 51(2), 371–382 (2002)
26. Moorman, J.R., Lockwood, J.W.: Wireless call admission control using threshold access sharing. In: Proc. IEEE GLOBECOM 2001, San Antonio, USA, vol. 6, pp. 3698–3703 (November 2001)
27. Levine, D., Akyildiz, I., Naghshineh, M.: A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept. IEEE/ACM Transactions on Networking 5(1), 1–12 (1997)
28. Choi, S., Shin, K.G.: Predictive and adaptive bandwidth reservation for handoffs in QoS-sensitive cellular networks. In: Proc. ACM SIGCOMM 1998, Vancouver, Canada, vol. 27, pp. 155–166 (October 1998)
29. Epstein, B.M., Schwartz, M.: Predictive QoS-based admission control for multiclass traffic in cellular wirelessnetworks. IEEE Journal on Selected Areas in Communications 18(3), 523–534 (2000)

30. Wu, S., Wong, K.Y.M., Li, B.: A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks. IEEE/ACM Transactions on Networking 10(2), 257–271 (2002)
31. Peha, J.M., Sutivong, A.: Admission control algorithms for cellular systems. Wireless Networks 7(2), 117–125 (2001)
32. Epstein, B., Schwartz, M.: Reservation strategies for multi-media traffic in a wireless environment. In: Proc. IEEE VTC 1995, Chicago, USA, vol. 1, pp. 165–169 (July 1995)
33. Box, G.E., Jenkins, G.M.: Time Series Analysis: Forecasting and Control, 2nd edn. Holden-Day (1976)
34. Zhang, T., Berg, E., Chennikara, J., Agrawal, P., Chen, J.C., Kodama, T.: Local predictive resource reservation for handoff in multimedia wireless IP networks. IEEE Journal on Selected Areas in Communications 19(10), 1931–1941 (2001)
35. Hosking, J.R.M.: Fractional differencing. Biometrika 83(1), 165–176 (1981)
36. Brockwell, P.J., Davis, R.A.: Time Series: Theory and Methods, 2nd edn. Springer (1991)
37. Gripenberg, G., Norros, I.: On the prediction of fractional brownian motion. Journal of Applied Probability 33, 400–410 (1996)
38. Norros, I.: On the use of fractional brownian motion in the theory of connectionless networks. IEEE Journal on Selected Areas in Communications 13(6), 953–962 (1995)
39. Leland, W.E., Taque, M., Willinger, W., Wilson, D.: On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Transactions on Networking 2(1), 1–15 (1994)
40. Crovella, M.E., Bestavros, A.: Self-similarity in world wide web traffic: Evidence and possible causes. IEEE/ACM Transactions on Networking 5(6), 835–846 (1997)
41. Beran, J., et al.: Long-range dependence in variable-bit-rate video traffic. IEEE Transactions on Communications 43(2), 1566–1579 (1995)
42. Shu, Y., Jin, Z., Wang, J., Yang, O.W.: Prediction-based admission control using FARIMA models. In: Proc. IEEE ICC 2000, New Orleans, USA, vol. 3, pp. 1325–1329 (June 2000)
43. Shu, Y., et al.: Traffic prediction using FARIMA models. In: Proc. IEEE ICC 1999, Vancouver, Canada, vol. 2, pp. 891–895 (June 1999)
44. Oliveira, C., Kim, J.B., Suda, T.: An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks. IEEE Journal on Selected Areas in Communications 16(6), 858–874 (1998)
45. Aljadhai, A., Znati, T.F.: Predictive mobility support for QoS provisioning in mobile wireless networks. IEEE Journal on Selected Areas in Communications 19(10), 1915–1930 (2001)
46. Acampora, A., Naghshineh, M.: An architecture and methodology for mobile-executed handoff in cellular ATM networks. IEEE Journal on Selected Areas in Communications 12(8), 1365–1375 (1994)
47. Ross, S.M.: Stochastic Process, 2nd edn. American Mathematical Society (1997)
48. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press (1992)
49. Iraqi, Y., Boutaba, R.: When is it worth involving several cells in the call admission control process for multimedia cellular networks? In: Proc. IEEE ICC 2001, Helsinki, Finland, vol. 2, pp. 336–340 (June 2001)

50. Mitchell, K., Sohraby, K.: An analysis of the effects of mobility on bandwidth allocation strategies in multi-class cellular wireless networks. In: Proc. IEEE INFOCOM 2001, Anchorage, USA, vol. 2, pp. 1005–1011 (April 2001)
51. Puterman, M.L.: Markov decision processes: Discrete stochastic dynamic programming. John Wiley & Sons (1994)
52. Haas, Z., Halpern, J.Y., Li, L., Wicker, S.B.: A decision-theoretic approach to resource allocation in wireless multimedia networks. In: Proc. ACM 4th Workshop Discrete Alg. Mobile Comput. Commun., Boston, USA, pp. 86–95 (August 2000)
53. Tijms, H.C.: Stochastic Modeling and Analysis: A Computational Approach. John Wiley & Sons (1989)
54. Saquib, M., Yates, R.: Optimal call admission to a mobile cellular network. In: Proc. IEEE VTC 1995, Chicago, USA, vol. 1, pp. 190–194 (July 1995)
55. Chen, D., Hee, S.B., Trivedi, K.S.: Optimal call admission control policy for wireless communication networks. In: Proc. International Conference on Information, Communication and Signal Processing, ICICS 2001, Singapore (December 2001)
56. Gao, Q., Acampora, A.: Performance comparisons of admission control strategies for future wireless networks. In: Proc. IEEE WCNC 2002, Orlando, USA, vol. 1, pp. 317–321 (March 2002)
57. Kwon, T., Choi, Y., Naghshineh, M.: Optimal distributed call admission control for multimedia services in mobile cellular networks. In: Proc. Mobile Multimedia Ccommunication, MoMuC 1998, Berlin, Germany (October 1998)
58. Xiao, Y., Chen, C.L.P., Wang, Y.: An optimal distributed call admission control for adaptive multimedia in wireless/mobile networks. In: Proc. IEEE MASCOTS 2000, San Francisco, USA, pp. 477–482 (August 2000)
59. Kwon, T., Choi, Y., Naghshineh, M.: Call admission control for adaptive multimedia in wireless/mobile networks. In: Proc. ACM WOWMOM 1998, Dallas, USA, pp. 111–116 (October 1998)
60. Kwon, T., Park, I., Choi, Y., Das, S.: Bandwidth adaptation algorithms with multi-objectives for adaptive multimedia services in wireless/mobile networks. In: Proc. ACM WOWMOM 1999, Seattle, USA, pp. 51–59 (August 1999)
61. Choi, J., Kwon, T., Choi, Y., Naghshineh, M.: Call admission control for multimedia service in mobile cellular networks: A markov decision approach. In: Proc. IEEE ISCC 2000, Antibes, France, pp. 594–599 (July 2000)
62. Yoon, I.-S., Lee, B.G.: A distributed dynamic call admission control that supports mobility of wireless multimedia users. In: Proc. IEEE ICC 1999, Vancouver, Canada, pp. 1442–1446 (June 1999)
63. Kwon, T., Choi, J., Choi, Y., Das, S.: Near optimal bandwidth adaptation algorithm for adaptive multimedia services in wireless/mobile networks. In: Proc. IEEE VTC 1999, Amsterdam, Netherlands, vol. 2, pp. 874–878 (September 1999)
64. Yener, A., Rose, C.: Near optimal call admission policies for cellular networks using genetic algorithms. In: Proc. IEEE Wireless 1994, Calgary, Canada, pp. 398–410 (July 1994)
65. Xiao, Y., Chen, C.L.P., Wang, Y.: A near optimal call admission control with genetic algorithm for multimedia services in wireless/mobile networks. In: Proc. IEEE NAECON 2000, Dayton, USA, pp. 787–792 (October 2000)
66. El-Alfy, E.-S., Yao, Y.-D., Heffes, H.: A learning approach for call admission control with prioritized handoff in mobile multimedia networks. In: Proc. IEEE VTC 2001, Rhodes, Greece, vol. 2, pp. 972–976 (May 2001)

67. Li, B., Li, L., Li, B., Cao, X.-R.: On handoff performance for an integrated voice/data cellular system. Wireless Networks 9(4), 393–402 (2003)
68. Fang, Y.: Thinning schemes for call admission control in wireless networks. IEEE Transactions on Computers 52(5), 686–688 (2003)
69. Pavlidou, F.-N.: Two-dimensional traffic models for cellular mobile systems. IEEE Transactions on Communications 42(2/3/4), 1505–1511 (1994)
70. Wu, H., Li, L., Li, B., Yin, L., Chlamtac, I., Li, B.: On handoff performance for an integrated voice/data cellular system. In: Proc. IEEE PIMRC 2002, Lisboa, Portugal, vol. 5, pp. 2180–2184 (September 2002)
71. Wieselthier, J.E., Ephremides, A.: Fixed- and movable-boundary channel-access schemes for integrated voice/data wireless networks. IEEE Transactions on Communications 43(1), 64–74 (1995)
72. Young, M.C., Haung, Y.-R.: Bandwidth assignment paradigms for broadband integrated voice/data networks. Computer Communications Journal 21(3), 243–253 (1998)
73. Chih-Lin, I., Greenstein, L.J., Gitlin, R.D.: A microcell/macrocell architecture for low and high mobility wireless users. IEEE Journal on Selected Areas in Communications 11(6), 885–891 (1993)
74. Rappaport, S.S., Hu, L.-R.: Microcellular communication systems with hierarchical macrocell overlays: Traffic performance models and analysis. Proc. of the IEEE 82, 1383–1397 (1994)
75. Hu, L.-R., Rappaport, S.S.: Personal communication systems using multiple hierarchical cellular overlays. IEEE Journal on Selected Areas in Communications 13(2), 406–415 (1995)
76. Yeung, K.L., Nanda, S.: Channel management in microcell/macrocell cellular radio systems. IEEE Transactions on Vehicular Technology 45(4), 601–612 (1996)
77. Chang, C., Chang, C.J., Lo, K.-R.: Analysis of a hierarchical cellular system with reneging and dropping for waiting new and handoff calls. IEEE Transactions on Vehicular Technology 48(4), 1080–1091 (1999)
78. Marsan, M.A., Ginella, G., Maglione, R., Meo, M.: Performance analysis of hierarchical cellular networks with generally distributed times and dwell times. IEEE Transactions on Wireless Communications 3(1), 248–257 (2004)
79. Jain, R., Knightly, E.W.: A framework for design and evaluation of admission control algorithms in multi-service mobile networks. In: Proc. IEEE INFOCOM 1999, New York, USA, vol. 3, pp. 1027–1035 (March 1999)
80. Yu, F., Leung, V.C.: Mobility-based predictive call admission control and bandwidth reservation in wireless cellular networks. In: Proc. IEEE INFOCOM 2001, Anchorage, USA, vol. 1, pp. 518–526 (April 2001)
81. Lim, S., Cao, G., Das, C.: An admission control scheme for QoS-sensitive cellular networks. In: Proc. IEEE WCNC 2002, Orlando, USA, vol. 1, pp. 296–300 (March 2002)
82. Soh, W.-S., Kim, H.S.: Qos provisioning in cellular networks based on mobility prediction techniques. IEEE Communications Magazine 41(1), 86–92 (2003)
83. Shen, X., Mark, J.W., Ye, J.: User mobility profile prediction: An adaptive fuzzy inference approach. Wireless Networks 6(5), 363–374 (2000)
84. Evci, C., Fino, B.: Spectrum management, pricing, and efficiency control in broadband wireless communications. Proc. of the IEEE 89, 105–115 (2001)
85. Heikkinen, T.: Congestion based pricing in a dynamic wireless network. In: Proc. IEEE VTC 2001, Rhodes, Greece, vol. 2, pp. 1073–1076 (May 2001)

86. Hou, J., Yang, J., Papavassiliou, S.: Integration of pricing with call admission control for wireless networks. In: Proc. IEEE VTC 2001, Atlantic City, USA, vol. 3, pp. 1344–1348 (October 2001)
87. Haung, Y.-R., Lin, Y.-B., Ho, J.-M.: Performance analysis for voice/data integration on a finite-buffer mobile system. IEEE Transactions on Vehicular Technology 49(2), 367–378 (2000)
88. Ye, J., Shen, X., Mark, J.W.: Call Admission Control in Wideband CDMA Cellular Networks by Using Fuzzy Logic. IEEE Transactions on Mobile Computing 4(2) (March/April 2005)
89. Chen, Y.H., Chang, C.J., Shen, S.: Outage-based fuzzy call admission controller with multi-user detection for WCDMA systems. IEE Proc. Commun. 152(5) (October 2005)
90. Chung-Ju, Chang, L.-C., Kuo, Y.-S., Chen, S.S.: Neural Fuzzy Call Admission and Rate Controller for WCDMA Cellular Systems Providing Multirate Services. In: IWCMC 2006, Vancouver, British Columbia, Canada, July 3-6 (2006)
91. Huang, C.-J., Chuang, Y.-T., Yang, D.-X.: Implementation of call admission control scheme in next generation mobile communication networks using particle swarm optimization and fuzzy logic systems. Expert Systems with Applications 35(3), 1246–1251 (2008)
92. Thilakawardana, S., Tafazolli, R.: Efficient Call Admission Control and SchedulingTechnique for GPRS Using Genetic Algorithms. Mobile Communications Research Group, Centre for Communications Systems Research, CCSR (2004)
93. Wang, S.-L., Hou, Y.-B., Huang, J.-H., Huang, Z.-Q.: Adaptive Call Admission Control Based on Enhanced Genetic Algorithm in Wireless/Mobile Network. In: Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence (2006)
94. Wang, S., Cui, Y., Koodli, R., Hou, Y., Huang, Z.: Dynamic Multiple-Threshold Call Admission Control Based on Optimized Genetic Algorithm in Wireless/Mobile Networks. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences E91-A(7), 1597–1608 (2008)
95. Rong, B., Qian, Y., Lu, K., Hu, R.Q., Kadoch, M.: Mobile-Agent-Based Handoff in Wireless Mesh Networks: Architecture and Call Admission Control. IEEE Transactions on Vehicular Technology 58(8), 4565–4575
96. Bernard, W.: 30 Years of Adaptive Neural Networks: Perception, Madalines and Back Propagation. Proceedings of the IEEE 78(9) (September 1990)
97. Chang, C.J., Lin, S.Y., Cheng, R.G., Shiue, Y.R.: PSD-based Neural-net Connection Admission Control. In: Proceedings of IEEE Infocom (April 1997)
98. Diaz-Estrella, A., Jurado, A., Sandoval, F.: New Training Pattern Selection Method for ATM Call Admission Neural Control. Electronic Letters 330 (March 1994)
99. Dueld, N.G., Lewis, J.T., O'Connell, N., Russell, R., Toomey, F.: Entropy of ATM Traffic Streams: A Tool for Estimating QoS Parameters. IEEE Journal on Selected Areas in Communications (August 1995)
100. Elwalid, A., Mitra, D.: Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks. In: Proceedings of IEEE Infocom (June 1994)
101. Guerin, R., Ahmadi, H., Naghshineh, M.: Equivalent Capacity and Its Application to Bandwith Allocation in High-Speed Networks. IEEE Journal on Selected Areas in Communications (September 1991)

102. Haykin, S.: Neural Networks, A Comprehensive Foundation. Macmillan Publishing Company (1994)
103. Hees, H., Lucantoni, D.: A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance. IEEE Journal on Selected Areas in Communications (September 1986)
104. Hertz, J., Krogh, A., Palmer, R.: Introduction to the Theory of Neural Computation. Addison-Wesley Publishing Company (1991)
105. Hiramatsu, A.: ATM Call Admission Control Using a Neural Network Trained with Virtual Output Buffer Method. In: IEEE International Conference on Neural Networks, vol. 6 (1994)
106. Hiramatsu, A.: Training Techniques for Neural Network Applications in ATM. IEEE Communications Magazine (October 1995)
107. Li, S.Q., Hwang, C.L.: Queue Response to Input Correlation Functions: Discrete Spectral Analysis. IEEE/ACT Transactions on Networking (October 1993)
108. Masters, T.: Practical Neural Network Recipes in C++. Academic Press (1993)
109. Morris, R., Samadi, B.: Neural Network Control of Communications Systems. IEEE Transactions on Neural Networks (1994)
110. Nordstrom, E., Carlstrom, J., Gallmo, O., Asplund, L.: Neural Networks for Adaptive Traffic Control in ATM Networks. IEEE Communications Magazine (October 1995)
111. Ogier, R., Plotkin, N.T., Khan, I.: Neural Network Methods with Traffic Descriptor Compression for Call Admission Control. In: Proceedings of IEEE Infocom (March 1996)
112. Plotkin, N.T., Roche, C.: The Entropy of Cells Streams as a Traffic Descriptor in ATM Networks. IFIP Performance of Communications Systems (October 1995)
113. Sarajedini, A., Chau, P.M.: Quality of Service Prediction Using Neural Networks. In: Proceedings of MILCOM, vol. 2 (1996)
114. Tham, C.-K., Soh, W.-S.: Multi-service Connection Admission Control using Modular Neural Networks. In: Proceedings of IEEE Infocom (March 1998)
115. Youssef, S., Habib, I., Saadawi, T.: A Neurocomputing Controller for Bandwdith Allocation in ATM Networks. IEEE Journal on Selected Areas in Communications (February 1997)