ELSEVIER

Computer Communications 26 (2003) 522–541

computer communications

www.elsevier.com/locate/comcom

Call admission control for wireless personal communications

Chi Wa Leong, Weihua Zhuang*

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont., Canada N2L 3G1

Received 26 November 2001; revised 10 May 2002; accepted 21 May 2002

Abstract

Call admission control (CAC) in future wireless personal communication systems is to simultaneously achieve quality-of-service satisfaction and high resource utilization. It poses significant technical challenges due to scarce radio spectrum, user mobility, hostile wireless propagation environment, end-to-end connectivity, and dynamic nature of multimedia traffic. This paper aims at providing a survey on the existing literature related to the works on CAC for future wireless systems, especially in the wireless and combined wireless/wireline domains. As the concepts of the virtual connection tree (VCT) and cell cluster have been proposed to handle user mobility, both centralized CAC policies for systems using static VCT static cell cluster and distributed CAC policies for systems using dynamic VCT dynamic cell cluster are discussed. Comparisons among the various CAC solutions are made, problems that have been dealt with and problems that need to be tackled are identified for perspective researchers in this area.

Keywords: Call admission control; Multimedia traffic; Quality-of-service; Virtual connection tree; Cell cluster; Internet

1. Introduction

The next generation wireless personal communication systems (PCS) are expected to revolutionize the way in which people communicate [4,7,9,14,22,29,45,54,73,74]. The revolution is brought about by allowing people to communicate with anyone, in any place, at any time, in a multimedia environment and with a pre-specified level of quality-of-service (QoS). Because of scarce radio spectrum and limited coverage of base stations, the wireless multimedia networks are connected to wireline networks to maximize the coverage area of the PCS [1,79]. That is, the PCS will be made up of interconnected regional wireless and wireline systems, as shown in Fig. 1, where WNRCS stands for Wireline Regional Communication System and WSRCS stands for Wireless Regional Communication System. A region can be a city, a province or state, or a country. The coverage areas of some of the regional systems can be overlapped, and far apart ones can be interconnected by high capacity and reliable dedicated links. These dedicated connections can be through satellites or under-ocean cables, or simply overland. The aggregate of all interconnected wireline regional systems can be treated as a giant backbone. Because wireline systems are more stable, much easier to upgrade and maintain, and have more capacity, the backbone will be a very important part of the PCS, acting as (i) a reliable passageway for separated wireless systems, and (ii) a connection point to large information database in the wireline domain. The backbone can be either ATM or IP based.[1]

QoS provisioning is a major feature of the future PCS. In most systems, call admission control (CAC) is the first control function imposed on a user for QoS provisioning. When a user requests a new connection, what the CAC usually does is to calculate the amount of resources required by (i) the users already in the system, and (ii) the pending user. If the sum of the two is not more than the total capacity, then the user's request will be acknowledged; otherwise the request will be rejected. This is equivalent to first reserving resources for the admitted users and then checking if the remaining resources are sufficient to support the new connection. How to make the two calculations depends on the particular system under consideration. CAC is important because its result is irreversible. A CAC mechanism is usually defined as the detailed work involved in the CAC function. This includes the decision process, signaling, routing table establishment, etc. The decision

---

* Corresponding author. Tel.: +1-519-888-4567x5354; fax: +1-519-747-3077.

E-mail address: wzhuang@bbcr.uwaterloo.ca (W. Zhuang).

[1] ATM and IP stand for asynchronous transfer mode and Internet protocol, respectively.

0140-3664/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0140-3664(02)00143-3

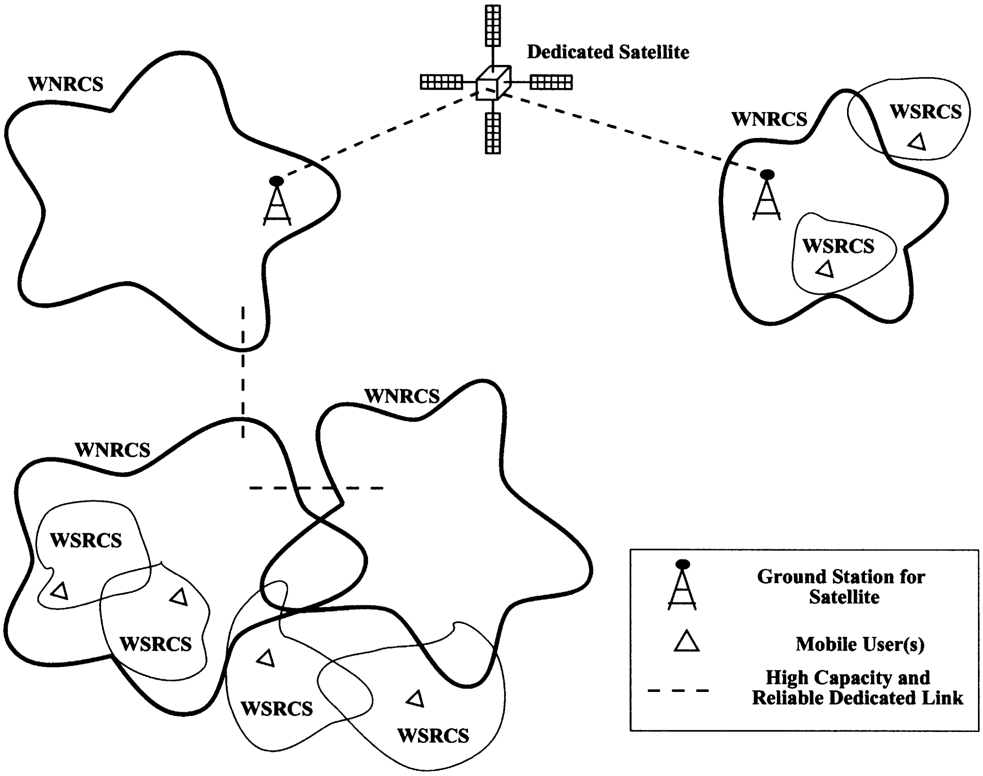


Fig. 1. The coverage area of the future PCS.

process of CAC can often be formulated in a high level representation called the CAC policy.[2] Whenever a user requests a new connection, the CAC policy takes the call request as input, and based on the current traffic conditions of the system, decides whether or not to accept the user, as illustrated in Fig. 2.

Established in the PCS environment, the entire end-to-end connection of each call, with the possibility of spanning both the wireline and the wireless domains, may consist of wireline hop(s) and wireless hop(s). During the call admission phase, the decision process of CAC is repeated over all the hops to ensure that sufficient resources are available to support the entire new connection. A significant amount of work on CAC has been done for the wireline domain alone [13,16,27,28,34,36,39,40,47,64,78,80], for the wireless domain alone [17,53,67], and for both [2,3, 19,21,49,65,66,85]. The works on CAC for connection oriented wireline networks (such as ATM networks) are more matured and there are usually less challenges involved. However, CAC in wireless, IP-based wireline, and interconnected wireless/wireline networks poses significant technical challenges due to user mobility, limited radio spectrum, dynamic nature of multimedia traffic, hostile wireless propagation environments, IP connectionless nature, etc. As a result, a robust CAC policy that can facilitate the provision of QoS in the PCS is hard to find. Most recent efforts on CAC are dedicated to the wireless portion of the end-to-end connection. This paper is aimed at providing a survey of the current existing proposals for CAC in the PCS, with emphasis placed on CAC in the wireless domain and in the combined wireless/wireline IP domain. In particular, the paper will provide a review of the previous works on CAC reported in Refs. [2,3,17,19–21,49,53, 65–67,85]. The survey will help to identify what problems have been and have not been studied. Various efforts aimed at a specific problem will be discussed and comments will be made in order to deliver a comprehensive insight to the previous studies. This is to provide an overview on CAC to general audience and, at the same time, to assist perspective researchers to focus their efforts on untackled problems.

User mobility is the most critical aspect that must be addressed in any literature related to wireless communications. Virtual connection tree (VCT) and cell cluster [1,3] are common strategies used to handle the mobility in PCS. The VCT and the cell cluster structures in a wireless network can be classified into two categories, namely the static VCT static cluster scheme and the dynamic VCT dynamic cluster scheme. The CAC policies for systems using each of the schemes will be discussed. The remainder of this paper is organized as follows. Section 2 presents the

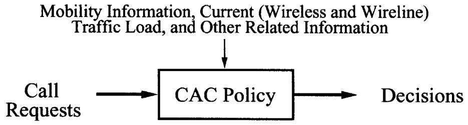


Fig. 2. The CAC decision process.

---

[2] A CAC policy is synonymous to a CAC algorithm.

major challenges in CAC for personal wireless communications. Section 3 describes the system infrastructures proposed to handle user mobility. It gives some important assumptions made in the literature and the definitions of the VCT and the cell cluster. The works on CAC in a system using the static scheme are reviewed in Section 4, while the works on CAC in a system using the dynamic scheme are reviewed in Section 5. Section 6 briefly reviews some preliminary works on CAC in wireless/IP interworking. Finally, Section 7 summarizes this survey and identifies the issues that should be investigated in the future.

2. Major challenges in CAC

There are many challenges in the wireless and the combined wireless/wireline environments. The limited radio spectrum, user mobility and fluctuation of usable bandwidth due to time-varying channel conditions are typical stumbling blocks found in such environments. Each of these problems is separately addressed in Refs. [26,37,44,56,77], [33,60,61] and [5,68], respectively. In addition, the complex interaction between the wireless and wireline networks [19,68] and the multiplexing of multimedia traffic in the same connection [18,79] render the QoS provisioning in the PCS more complex.

2.1. User mobility

In the PCS environment, services are expected to be enjoyed by users anywhere in the system. These users can be fixed or mobile. Fixed users are easier to handle because they can be served by wireline systems, and wireline transmission links are more reliable and have more bandwidth. Services to mobile users are wireless and difficult to maintain due to two main reasons: (i) wireless channels are time varying and (ii) radio bandwidth is limited. The former leads to various types of fading, which can be mitigated by well-known methods, such as diversity and error-correction coding [70]. The latter leads to congestion in the wireless spectrum, which can be relieved by introducing the cellular structure for frequency reuse. The smaller the cell size, the more often the frequency spectrum can be reused in the same coverage area. Because of this, cells of small size are expected to be deployed in most areas in the future PCS in order to handle high service demands from a large number of multimedia users. Due to user mobility, traffic load in each cell is dynamic and is difficult to accurately predict in any case. As a result, system resources allocated to each cell may not match to the traffic load in the cell. Once a certain frequency band is assigned to a particular cell, the total capacity of that cell is usually fixed. For example, in a code-division multiple access (CDMA) system with a certain signal-to-interference ratio, the allocated frequency spectrum is reused in every cell and, therefore, spare capacity available in a non-congested cell cannot be readily borrowed to relieve the conditions in a congested cell within the same regional system.

With the use of the cellular structure, there comes the problem of handoff. A handoff occurs when an active mobile user, who is being served by a radio cell, crosses the cell boundary and arrives at an adjacent cell. Handoff causes problems in resource allocation in both wireless and wireline domains. For the wireless domain, because the capacity of a radio cell is usually fixed and is separated from that of any other cells, the handoff user is required to release the resources obtained from the original cell and acquire resources from the new one in order to maintain the connection. This problem is called wireless handoff. For the wireline domain, the change in access points to the wireline backbone also requires the reallocation of wireline resources in order for the connection to continue. This problem is called wireline handoff.

Whenever a cellular system is used, there are two major types of calls that can arrive at any particular cell: new calls originated from within and handoff calls coming from adjacent cells. From the users' point of view, it is better to be blocked in the beginning rather than dropped in the middle of a connection. As a result, handoff calls should be given a higher priority than new calls. To do so, the usual approach is to reserve some capacity exclusively for them. To calculate the right amount of reserved resources requires the knowledge of the mobility information, such as the users' traveling directions and speeds, and the current traffic load, in the surroundings. Reserved resources for handoff calls are sometimes called guard bandwidth or guard channels. Because the rate of handoff occurrences is inversely proportional to the cell size, and small cells are expected to be used for the PCS, how handoffs are handled will become critical. If QoS is to be maintained, the consideration of handoff calls is mandatory in the CAC process.

2.2. Multimedia traffic

The multimedia services will be of any nature, including video, voice and data. In the PCS environment, the CAC function is required to take into account the behavior of multimedia traffic—the teletraffic generated by multimedia applications. In general, there are three major types of multimedia traffic: the constant-rate traffic for uncompressed voice and video, variable-rate traffic for compressed voice and video, and available-rate traffic for data[3] [80]. Each type of services can be further subdivided according to their bandwidth requirement. As an example, one constant-rate service may require 30 units of bandwidth[4] while

---

[3] In ATM networks, the traffic types are referred to as constant bit rate (CBR), variable bit rate (VBR), and available bit rate (ABR), respectively.

[4] For the simplicity of measurement, the bandwidth is discretized into many small but equal amounts called units of bandwidth, or bandwidth units.

another constant-rate service may require 40 units. Constant-rate traffic usually carries high priority information that always requires a constant amount of resources. Examples of such information are uncompressed voice or video. Variable-rate traffic carries high priority information that requires time-varying amount of resources. Examples of sources generating variable-rate traffic are voice or video compressors or encoders. Finally, available-rate traffic carries low priority information and occupies only leftover resources not used by high priority traffic, for delay insensitive data application, such as e-mails. There are also non-real-time delay-sensitive data services with a small delay bound (in the range of seconds), such as remote log-in, file transfer protocol (FTP), and similar applications associated with transport control protocols (TCP). For applications, such as web browsing [23] and wide-area TCP connections [69], it has been found via measurements that the data traffic flows have the long range dependence (i.e. the self-similar or the heavy tail) behavior and the length of a typical document is characterized by log-normal or Pareto distribution, not the exponential distribution.

The major challenge in dealing with multimedia applications is that they often do not have a constant transmission rate. If too many resources are reserved, they will be wasted when the sources are transmitting at low rates; otherwise, if insufficient resources are set aside, QoS to the users cannot be maintained when the sources are transmitting at high rates. There are two major approaches to dealing with multimedia traffic: (i) to reserve resources according to the average needs, or (ii) to adapt the CAC to the current needs of the sources. The first approach is simple but under or over utilization of resources will still occur from time to time. Complexity is the drawback of the second approach, but it will ensure better resource utilization. Besides the time-varying transmission rate, multimedia traffic usually comes with different QoS requirements. Because of limited resources, it is sometimes difficult, if not impossible, for the CAC to meet all these requirements of different users. As a result, the management of multimedia traffic is another difficulty lurking in CAC.

2.3. QoS provisioning and resource utilization

It is expected that the multimedia services will be provided with satisfactory QoS. QoS refers to the degree of satisfaction of the users with the communication services provided by the systems. When it is measured or calculated, it is more convenient to express it in terms of the amount of dissatisfaction of the users. In such context, QoS provisioning involves two major steps: (i) specifying the maximum allowed amount of various types of dissatisfaction, and (ii) designing the system so that the maximum allowed values are not exceeded. The dissatisfaction may result from call blocking and dropping, packet loss, transmission delay and delay jitter, and transmission error. The corresponding QoS measures commonly used are call blocking and dropping probability, packet loss probability, transmission delay and delay jitter, and error rate.

From the point of view of a service provider, resource utilization is, besides QoS provisioning, another important factor that needs to be considered. A highly utilized system that can provide a satisfactory amount of QoS to the users is always a desired solution. However, high resource utilization and QoS provisioning are always conflicting goals. As in the case of CAC, resources are set-aside for active users, so that their QoS can be maintained, but unused resources mean low utilization. In order to have a balance in the two conflicting goals, the amount of reserved resources should be calculated carefully.

In summary, the design of an efficient and practical CAC function or policy requires the consideration of complicated and sometimes conflicting factors, namely, QoS provisioning to mobile and fixed users, behavior and multiple QoS requirements of multimedia traffic, as well as resource utilization.

3. System infrastructure

CAC is a function closely related to the infrastructure of the system. Because of this, an interconnection structure for the wireless and wireline domains is required before the CAC functions can be implemented. Also, because of the convenience brought about by tetherless connections, the number of active mobile users in the future PCS is expected to be large. As a result, there will be a lot of information flowing between the wireless and the wireline domains. A carefully designed interconnection structure is therefore essential to the maintenance of high efficiency in the system. A promising interconnection structure for the PCS is proposed in Refs. [1–3,19,53,65,66]. The core of the interconnection structure is the concepts of VCT and cell cluster, which have been proposed to allow for fast handoffs through automatic detection and switching [2,3,17,21,49,53,65–67,85]. The static VCT static cluster scheme and the dynamic VCT dynamic cluster scheme are two popular arrangement schemes for the VCT and the cell cluster in a wireless network. Before the works on CAC under the two schemes are discussed, it is necessary to address some of the important assumptions made and definitions used in the literature.

Assumption 1. The transport mechanism in the wireline networks is connection oriented.

This can be either an ATM or an enhanced version of TCP/IP that can support connection oriented transmission. This assumption makes it not only easier to analyze the network but also possible to introduce control on each individual traffic flow. Indeed, QoS support in the IP-based networks has been an intensive research area in recent years [83]. Various mechanisms have been proposed to provision

QoS in such networks, e.g. the integrated services (IntServ) approach and the differentiated services (DiffServ) approach [11]. The IntServ approach uses the resource reservation protocol (RSVP) to explicitly signal and dynamically allocate resources at each intermediate node along the path for each traffic flow. In the DiffServ approach, packets are classified into a small number of service classes at the network edge. The packets of each class are marked and traffic conditioned by the edge router, according to the resource commitment negotiated in the service level agreement (SLA). The multiprotocol label switching (MPLS) technique can be used to establish a path-oriented environment in a DiffServ domain [6,20,81]. As a result, the assumption is valid for an ATM backbone at the call level and for an IP backbone at least at the flow level.

Assumption 2. The QoS of interest is at the connection level.

This includes the new call blocking probability, handoff call blocking probability, and radio cell overload probability. The other type of QoS is the cell (or packet) level QoS, which includes cell loss probability, cell delay, delay jitter, etc. Although there are direct relationships between both types of QoS, the connection level QoS is of the immediate interest to network providers and service subscribers.

Assumption 3. The base station considered is assumed to embrace the capability of switching traffic for the wireline network and managing resources for the radio cell.

This allows a base station to play the role of both a wireline network node and a wireless network resource manager. Both capabilities are essential to the proper functioning of the entire network because base stations always situate at the intersection of a wireline and wireless network. This definition also allows a base station to appear in the context of both a wireline and a wireless network without any conflicts.

In addition to the assumptions, three important entities used throughout the literature, namely the virtual connection, the VCT and the cell cluster, are defined in the following.

Definition 1

Virtual connection (VC). A virtual connection is an end-to-end connection set up from one network node to another in a wireline network which supports bandwidth on demand.

Note that the VC is similar to the traditional connection in a circuit switched network in the sense that end-to-end connectivity is always maintained, but is different in the sense that network resource is allocated to the VC only when it is needed.

Definition 2

Virtual connection tree (VCT). A VCT is a systematic organization of network links and nodes in a wireline network, where bandwidth on demand is supported.

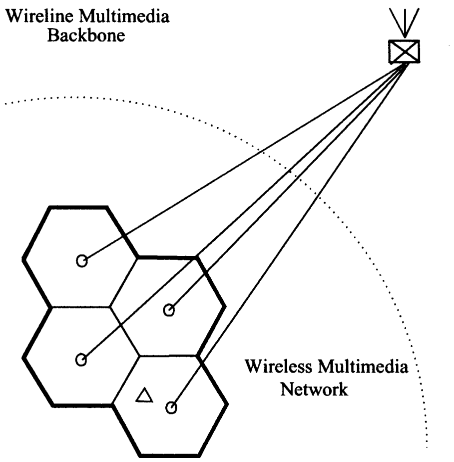
Definition 3

Cell cluster. A cell cluster, or simply a cluster, is a systematic organization of radio cells and their base stations which are adjacent to each other.

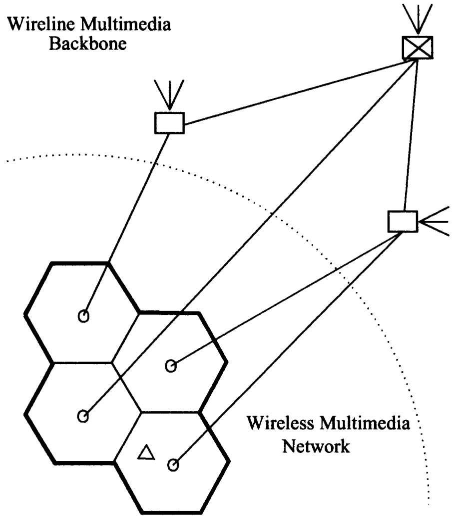
Note that according to the third assumption made earlier, a base station can be included as part of either a VCT or a cell cluster. Since in tradition a base station is related to its radio cell, in the following the base stations will be considered as part of a cell cluster.

In addition to allowing for fast routing and switching in a mobile environment, the VCT and the cell cluster can help to efficiently organize network resources in order to manage user mobility. First of all, the setting up of the VCT facilitates the pre-establishment of transmission paths in the wireline backbone, allowing bandwidth to be reserved in advance for the mobile terminal (mobile user) and packets to be rerouted from the old base station to the new one in the case that the mobile user hands off to an adjacent cell. On the other hand, the setting up of the cell cluster facilitates the reservation of wireless resources in the vicinity of the current base station of the mobile user. The existence of pre-established paths together with the reservation of wireline and wireless resources in advance reduces the chance of dropping handoff calls, ensuring the QoS provisioning of services for the mobile users. This, of course, comes with a cost of lowering the number of concurrent users that can be supported by the system, and reducing the network resource utilization. The dilemma between high QoS and low utilization is a problem that has been addressed in Refs. [49,65,85]. However, it has been shown in previous studies that a certain amount of reserved bandwidth is enough to maintain most QoS requirements, which in turn translates into higher user satisfaction and a steadier resource utilization in the long run. Therefore, the idea of bandwidth reservation for handoff calls will prevail in the rest of the paper.

Fig. 3 illustrates two examples of the VCT and the cell cluster structure. In both examples, the cell cluster consists of four cells and has the same structure. On the other hand, the structure of the VCT is different. In Fig. 3(a), the root node of the VCT is directly connected to the base stations in the cell cluster. The direct connections, however, may not be possible when the number of radio cells in a cell cluster is large. When this is the case, some intermediate nodes are needed in between to join the root node and the base stations together. The resultant VCT is shown in Fig. 3(b). Because the VCT in Fig. 3(a) can be considered as a special case of the one in Fig. 3(b), in the rest of the paper the more general structure of the VCT in Fig. 3(b) is assumed.



(a) With direct connections



(b) Without direct connections

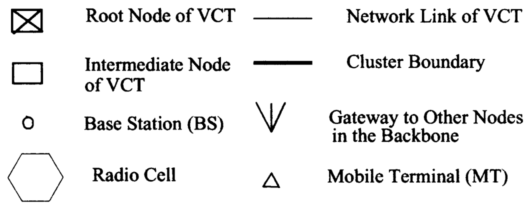


Fig. 3. Two examples of the VCT and the cell cluster.

To ensure the provision of QoS in the PCS, the end-to-end connection can be established in the following ways. In the situation when a mobile user is communicating with a fixed user, the end-to-end connection is made up of a cell cluster, a VCT and a point-to-point VC from the root node of the VCT to the fixed user. The situation is shown in Fig. 4(a). In the situation when two mobile users, each situated in a different cluster, are communicating with each other, as is shown in Fig. 4(b), the end-to-end connection is composed of two cell clusters and two VCTs, each of which corresponds to a mobile user, and a point-to-point VC connecting the two root nodes of the VCTs together. Note that the two types of connections, the mobile user to fixed user one and the mobile user to mobile user one, are different from their respective counterparts in a traditional cellular network, where an end-to-end connection involves only a single radio cell or base station in either end and a single circuit switched path in between. In comparison to the traditional network, the stronger capability of a network utilizing VCTs and cell clusters to guarantee the provision of QoS becomes evident.

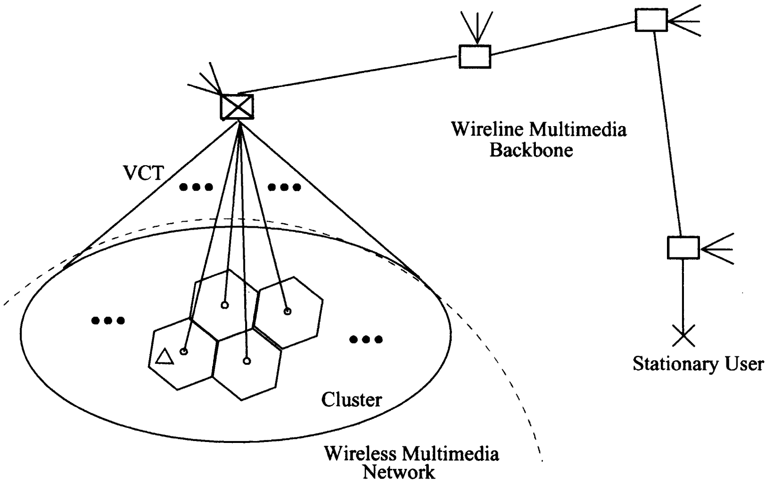
With the help of the VCT and the cell cluster, the goal of CAC to achieve a guaranteed level of QoS in the PCS becomes less formidable. Also, there can be different arrangements of the VCT and the cell cluster. For the CAC work to be reviewed in the following, two major arrangement schemes have been used. The first one is called the static VCT static cluster scheme and the second one the dynamic VCT dynamic cluster scheme. The works on CAC will be discussed according to the arrangement scheme used.

4. Centralized CAC with static VCT static cluster

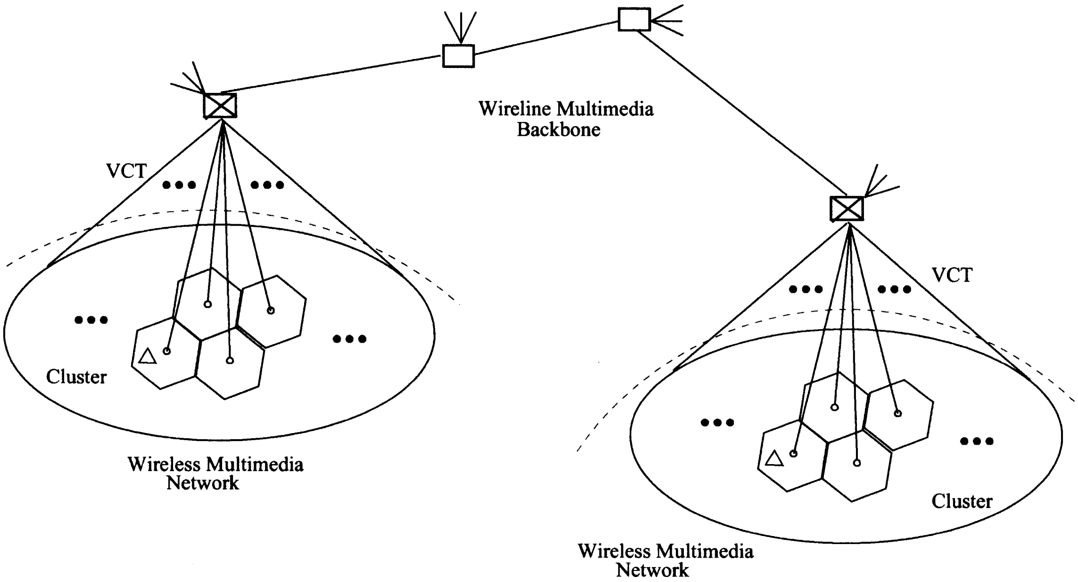
4.1. The static VCT static cluster structure

The arrangements of the VCT and the cell cluster in Refs. [2,3,17,21,65,66] are similar to each other. For illustration, consider a single wireless regional system and the backbone. The coverage area of the wireless regional system is first divided into fixed subregions. Each of these subregions can be looked at as a cell cluster. For each subregion, dedicated wireline links are then used to connect together all the base stations within, via a central wireline switch. The bunch of dedicated links in a subregion can be looked at as a VCT, with the central switch as the root node. The VCT has the responsibility to coordinate the transmission of packets and messages to and from *all* the mobile users within the cell cluster. The root node is the major access point to the backbone. The structure is illustrated by the example in Fig. 5, which shows how the geographical coverage area of the wireless portion of a PCS is divided into four static clusters. The procedure is performed in the construction phase of the PCS; both the boundary of the cell clusters and the layout of the VCTs remain unchanged during the actual operation of the PCS [2,3,21]. Reconfiguration is necessary only when there is a major catastrophe in the network. Because of the static nature, the arrangement scheme is called the static VCT static cluster scheme.

There are two kinds of handoff that a mobile user may initiate: the intra-cluster handoff and the inter-cluster handoff. Both handoffs are shown in Fig. 6. The former is the handoff of a mobile user from a radio cell to another within the same cluster, whereas the latter is the handoff of a mobile user from a radio cell in one cluster to another in an adjacent cluster. The intra-cluster handoff is easily handled with the coordination of the VCT. On the other hand, the



(a) Connection between an MT and a stationary user



(b) Connection between two MTs

Fig. 4. Two cases of an end-to-end network connection.

inter-cluster handoff is difficult to manage because the VCT of one cluster is not related to another VCT at all. One way to avoid inter-cluster handoff is to include as many radio cells in a cluster as possible in the hope of a mobile user finishing its connection before handing off to an adjacent cluster. Even if the size of a cluster is large, however, an inter-cluster handoff may eventually occur. To overcome this problem, two solutions have been proposed [21]:

(i) the root nodes of the VCTs in adjacent clusters are connected together by dedicated links, as shown in Fig. 7;
(ii) the base stations of the boundary cells are cross-connected to the root node of the VCT in the adjacent cluster, as shown in Fig. 8.

By using either solution, the VCTs of adjacent clusters can cooperate with each other during the handoff of a mobile user to minimize the chance of dropping the call. The effect of mobility between two adjacent clusters on handoff call dropping probability is investigated in Ref. [17], which will be discussed in more detail later.

Because the amount of resources constituting a cell cluster and the corresponding VCT is large, the initiation process performed by the network controller, including user registration and bandwidth allocation, etc. for a newly admitted mobile user is usually burdensome [2,3]. The burdensome initiation or admission process has to be invoked again during an inter-cluster handoff. Fortunately, the inter-cluster handoff can be avoided most of the time provided that the cell cluster is large enough and the

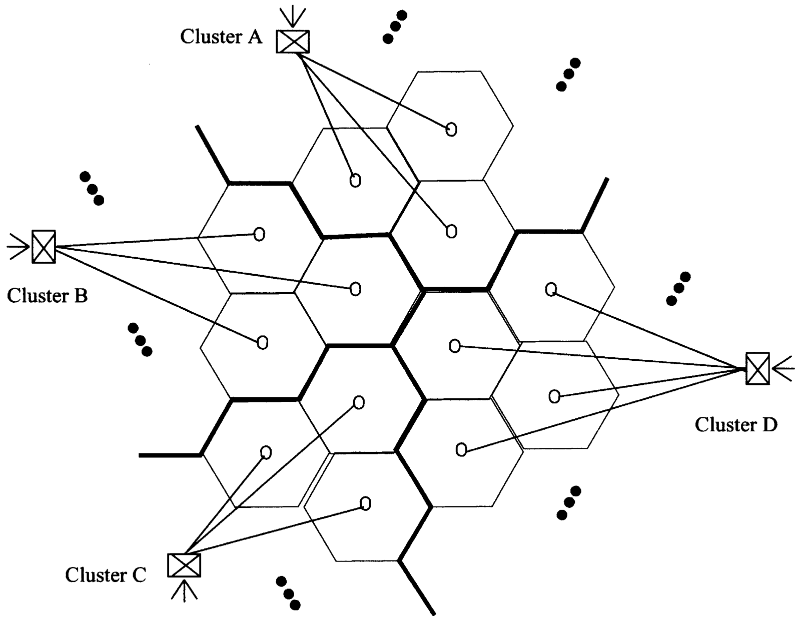


Fig. 5. Division of the coverage area into static clusters.

boundary is well positioned. In a network with carefully designed cell clusters, therefore, the call admission process will very likely be invoked only once for every network connection during its life-time [3].

The advantage of this type of interconnection is that the boundary of each subregion can be specified in such a way that the amount of traffic moving from one subregion to another is minimized. In other words, most mobile users within a subregion will stay in the same subregion for the entire life-time of the connections. Consequently, the major access point to the backbone, the central switch of the subregion, does not change for most users and the effect of wireline handoff on the system is minimized. Inter-subregion handoffs may still occur occasionally, and if that is the case, the handoff calls can be treated as new calls in CAC. On the other hand, the establishment of VCT facilitates: (i) the transmission of information between the base stations and the backbone via the central switch, and (ii) the exchange of information among the base stations within a subregion for coordinating handoffs of the mobile users. Communications among adjacent base stations are not mandatory, but are essential when the following tasks

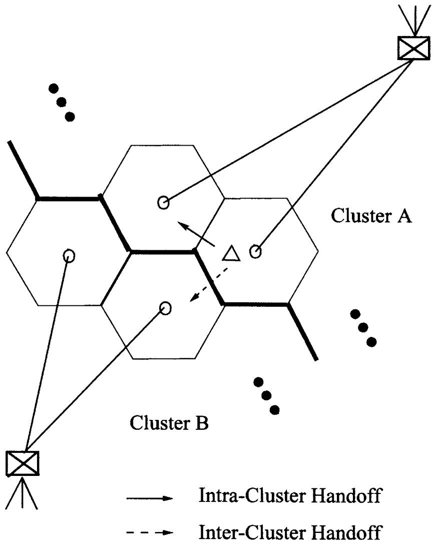


Fig. 6. The two possible kinds of handoff in the static scheme.

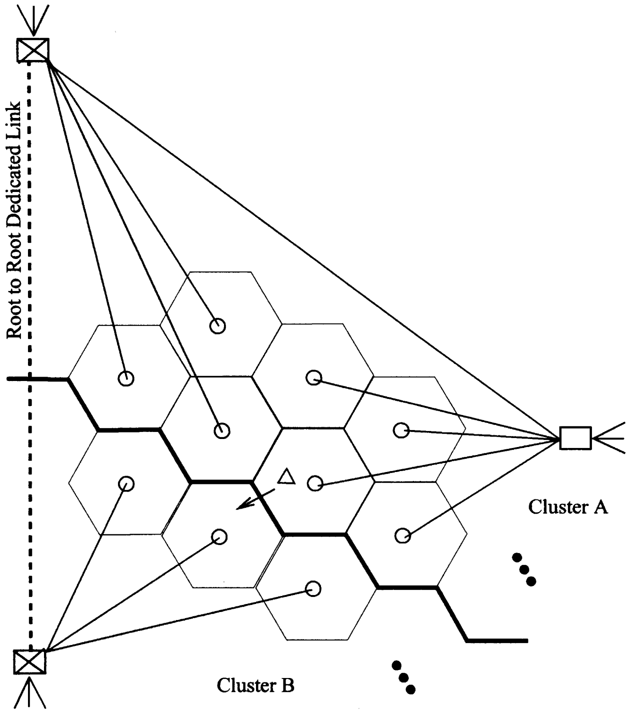


Fig. 7. The first solution for the inter-cluster handoff.

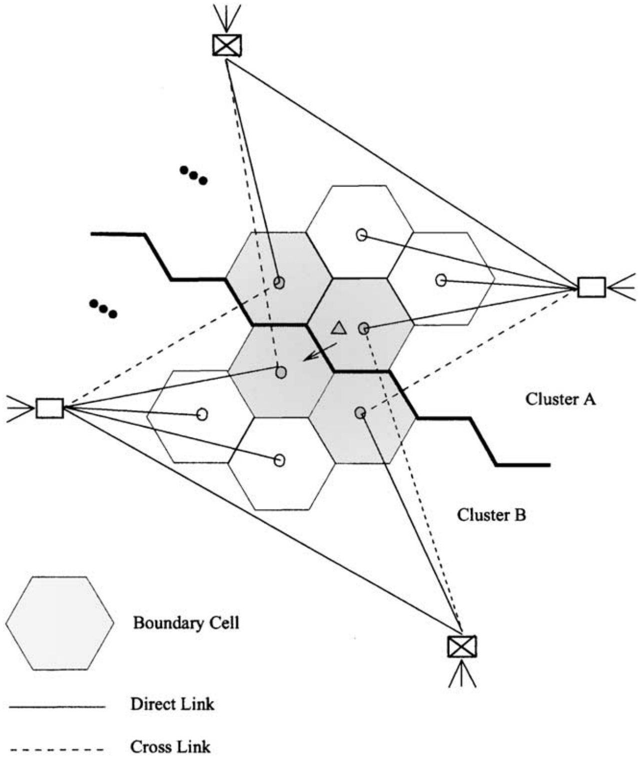


Fig. 8. The second solution for the inter-cluster handoff.

are required for the system: (i) soft handoff; (ii) information rerouting from old to new base stations, and (iii) calculation of required reserved resources for handoff calls at the radio cell level.

4.2. Centralized CAC policies

CAC for systems with the static VCT static cell cluster structure is performed in a centralized manner at the cluster level. The most prominent proposals for such centralized CAC can be found in Refs. [3,65,66], where a single subregion threshold check is performed at the central switch. For any new connection request, the check simply examines whether the number of active users within the subregion is below a certain threshold. The request will be acknowledged if the answer is positive; otherwise it will be turned down. This threshold is usually much larger than the capacity of a single radio cell to achieve a statistical multiplexing gain. Also, the threshold is periodically calculated to allow the network to adapt to the current traffic load. There is, however, no proposal in the papers on how to calculate the threshold. On the other hand, the function of resource allocation is decoupled from admission control and is performed by each individual cell within the cluster. Once a mobile user is admitted to a cluster, the base station of the cell, (where the mobile user is situated) is responsible to allocate an adequate amount of resources, if available, to the mobile user [65]. If there is no sufficient resource to meet the requirement, depending on how the situation is handled, the call may be dropped by the base station. Therefore, both the admission controller at the cluster level and the base stations within the cluster have the capability to drop the calls when necessary. On the other hand, if the request for a connection is acknowledged by both the admission controller and the base station, the mobile user will be allowed to roam from cell to cell within the cluster without the meddling of the admission controller. As mentioned earlier, the call, depending on how a particular cell handles congestion, will still be dropped if the mobile user arrives at an overloaded cell [65]. The probability of a cell being overloaded is a function of both the capacity of a radio cell and the admission threshold imposed at the cluster level. Besides, the mobile user will also subject to another instance of CAC if it hands off to an adjacent cluster. The discussion of how the effect of inter-cluster handoff on call dropping probability is investigated, together with other important performance issues found in Refs. [3,65,66], are provided in the following.

CAC for constant-rate services. Consider a system supporting only a single type of constant-rate services. The constant-rate service is for real-time applications and cannot tolerate any bandwidth fluctuation. Therefore, a constant-rate call must be dropped whenever the mobile user moves to a congested radio cell. By assuming a user is equally likely to be found in any radio cell within a cell cluster, a model can be set up to obtain the overload probability and the resource utilization efficiency, both as functions of the cell capacity and the admission threshold [3]. By approximating the tail of a binomial distribution using a Gaussian distribution with the appropriate mean and variance, the analysis can be simplified. For a cell cluster of any size, the results show that the overload probability increases rapidly as a function of the admission threshold. For any given cell capacity and overload probability, the results also show that a cluster consisting of more radio cells is more efficient in terms of resource utilization than a cluster with less cells. This illustrates the phenomenon of trunking efficiency. CAC for constant-rate services is relatively straightforward; however, it provides some basic performance results, many of which have been used as a foundation for further research on CAC with the static cluster scheme [65,66].

CAC for available-rate services. Next, consider a system supporting only a single type of available-rate services. The available-rate service is for non-real-time applications and can withstand any change in bandwidth availability. Thus, by making all of the available-rate calls in a radio cell share the available bandwidth, there is no need to drop any calls, even in a congested cell. This prevents the retransmission of the lost data if the calls were dropped and were reinitiated again. With the assumption that the call arrival is Poisson and the service time is exponentially distributed, analytical results can be obtained by mathematical modeling [66]. Especially, a model which consists of B truncated $M/M/\infty$ queues can be used to represent a cluster with B radio cells and with the admission control performed at the cluster level. The steady state probability of there being i mobile connections in any radio cell is then found, from which the probability of a bandwidth r being available to any mobile connections is calculated. This probabilistic information leads to the derivation of QoS measures, such as the overload probability of a radio cell, P_o, the probability of a mobile user receiving a bandwidth less than a minimum threshold ($\tau_{\min}$), $P[r < r_{\min}]$, and the new call blocking probability. Results are obtained for the controlled system and are compared to the results for a system with no admission control. The comparison shows that both P_o and $P[r < r_{\min}]$ for the controlled system level off under heavy traffic load condition, while the same QoS measures for the uncontrolled system quickly approaches one when the traffic load increases. On the other hand, the new call blocking probability for the controlled system is acceptably low under most traffic conditions. The overall results show that the system is capable of maintaining a certain level of QoS for available-rate users.

CAC for constant-rate and available-rate services. To make the system model more realistic, both constant-rate and available-rate services should be supported [65]. In particular, consider a system accommodating a single type of constant-rate service together with a single type of available-rate service. The constant-rate service is for real-time connections and is called Class I service, while the available-rate service is for non-real-time connections and is called Class II service. Class I connections are allowed to use up to a total of C_I units of bandwidth in any radio cell within a cell cluster. Depending on the resource allocation scheme used, the total amount of bandwidth allocated to Class II connections, C_{II}, is different. How the value of C_{II} is determined will be clarified later. Both Class I and Class II new connections are subject to CAC at the cluster level. This is done by limiting the total numbers of Class I calls and Class II calls, respectively, to N_I and N_{II}. In addition, new and handoff Class I connections will be dropped if the mobile user is in a congested cell. By making all Class II connections share the resources available to them in a radio cell, Class II new and handoff connections do not have to be dropped, even in congested areas [66]. For the resource allocation schemes, three have been considered. The schemes are shown in Fig. 9 and are described below:

Complete partitioning (CP)—the radio cell capacity C is completely partitioned so that Class I connections can occupy up to C_I bandwidth units while Class II connections can occupy up to $C_{II} = C - C_I$ bandwidth units.

Class I complete access (CA)—Class I connections can use up to the total cell capacity, which is C bandwidth units, with pre-emptive priority over Class II connections. At any given time, the Class II connections in the cell will use the remaining capacity not consumed by Class I connections.

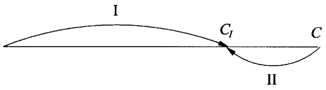
Class I restricted access (RA)—Class I connections can use at most $C_I < C$ units of bandwidth with pre-emptive priority over Class II connections. Similar to the CA scheme, at any given time unused bandwidth is allocated to the Class II connections. On the other hand, unlike the CA scheme, Class I connections do not have access to all the resources in a radio cell.

For the CP scheme, the QoS measures for both Class I and Class II connections can be calculated independently of each other. For the CA and RA scheme, only the QoS measures for Class I connections can be calculated independently of Class II connections because of the pre-emptive property. To calculate the QoS measures for Class II connections, however, for both the CA and RA schemes, the amount of Class I traffic currently in the system has to be taken into account. In particular, the statistics for the leftover capacity by Class I calls are found out, from which the QoS measures for Class II calls can be derived. Numerical results show that although Class I connections have priority over Class II connections, the QoS level for Class II connections can be guaranteed under heavy traffic conditions. The results also show that, in terms of ensuring the QoS level for Class II calls, the RA scheme always outperforms the CP scheme and sometimes the CA scheme. This is because in the case of the CP scheme unused bandwidth by Class I calls is always wasted, and in the case of the CA scheme when the Class I traffic load is heavy, Class I calls occupy majority of the bandwidth most of the time, leaving only an imperceptible amount for Class II calls. By showing all the analytical results, some insight into how the multiplexing of both constant-rate and available-rate traffic can be done is provided in Ref. [65].

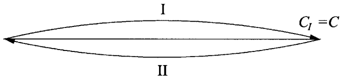
CAC with inter-cluster handoffs. Besides multimedia traffic, the problem of inter-cluster handoff in the static cluster scheme also needs to be addressed. The investigation of inter-cluster handoff is done in Ref. [17]. In particular, an analytical model, which consists of two adjacent clusters as shown in Fig. 10, is used. Multiple constant-rate sources are considered and user mobility within each cluster is ignored. Poisson arrivals and exponential service times are assumed and the resultant Markov chain for the two clusters has, unfortunately, too many states. An approximate analysis is then used, in which the original Markov chain is decoupled into two dependent but smaller chains, one for each of the two clusters. An iterative procedure is then used to find the steady state probabilities of the chains, from which the handoff dropping probability can be calculated. The results show that user movements between adjacent clusters have a significant impact on the handoff dropping probability, reassuring the fact that mobility is an important factor to be considered in wireless communications.

4.3. Summary

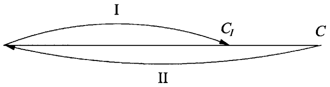
By using the static VCT static cluster scheme, CAC can be done in a centralized manner at the cluster level rather



(a) Complete Partitioning (CP)



(b) Class I Complete Access (CA)



(c) Class I Restricted Access (RA)

Fig. 9. Three resource allocation schemes used in Ref. [65].

than for each individual radio cell. The advantage of the centralized CAC approach is simplicity. Once a call is admitted to a subregion, the function of CAC does not have to be invoked again for the mobile user. This can significantly reduce the burden of the system if the number of wireless connections is large. The disadvantage of this approach is that no specific resources are reserved at the base station level for handoff calls. If the threshold is too large, QoS of the mobile users will be severely degraded due to excessive handoff dropping. Consequently, the problem of wireless handoff is not addressed.

5. Distributed CAC with dynamic VCT dynamic cluster

5.1. The dynamic VCT dynamic cluster structure

The work on CAC in Refs. [49,53,67,85] will be examined together in this section, on the basis that the arrangements of the VCT and the cell cluster in these papers are similar to each other. In particular, the arrangement scheme used in these papers is different from the static scheme described in Section 4 in three aspects. First of all, in the dynamic scheme a VCT and a cell cluster is set up for *every single* mobile user admitted to the wireless network [53,85]. Secondly, the sizes of the VCT and the cell cluster are significantly smaller in the dynamic scheme than in the static scheme. Finally, the VCT and the cell cluster are always changing according to the movement of the mobile user [53,85]. Because of its dynamic nature, the VCT and the cell cluster used in this scheme are, respectively, called dynamic VCT and dynamic cluster.

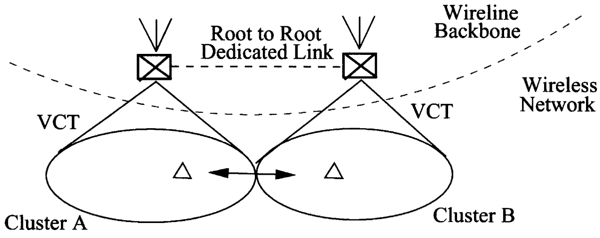


Fig. 10. The model used to analyze the inter-cluster handoff.

The fundamental idea of a dynamic cluster comes from the fact that every mobile user with an active wireless connection exerts an influence on the cells in the vicinity of its present location and along its traveling direction [53,67,85]. As an active mobile user moves to other cells, the region of influence also moves. Because the region of influence dynamically follows the movements of the mobile user like a shadow, the cells currently under the influence are said to form a dynamic cluster or shadow cluster [53]. An example of a dynamic cluster for an arbitrary mobile user is shown in Fig. 11, where the level of influence is illustrated using different level of shade: the darker the cell the more influenced it is by the mobile user. The level of influence is strongest near the mobile user and fades away as a function of the distance from the mobile user and the mobile user's traveling direction. The darkness or the level of shade of a cell indicates how many resources should be reserved for that particular mobile user in case it hands off to that cell in the future [53,85]. The chunk of resources reserved for handoff mobile users is called guard bandwidth in Ref. [85].

For a particular mobile user, the level of shade or the amount of bandwidth to be reserved in any cell within the dynamic cluster can be computed using the mobility information associated with that mobile user [53,67,85]. The mobility information can be estimated[5] by the current base station serving the mobile user and then forwarded to the corresponding cell, which then calculates the appropriate amount of bandwidth to be reserved in that cell. The setup of the dynamic VCT on top of the cells in a dynamic cluster allows efficient transmission of this mobility information through the wireline backbone. The same VCT also handles the proper reroutings of packets to and from the backbone once the mobile user hands off to a new cell [49,85]. Because the coverage area of the dynamic cluster follows the movement of the mobile user, the layout of the corresponding VCT needs to be constantly updated to ensure the proper connections between the cells within the new cluster [49,85]. The example in Fig. 12 shows how the corresponding VCT and the cluster evolve during the handoff of a mobile user. As illustrated in the figure, radio cells that are not in the new dynamic cluster must be deleted from the VCT to free any resources previously reserved for the mobile user, whereas those that become part of the new dynamic cluster must be appended to the VCT and given the appropriate mobility information to reserve bandwidth.

Depending on the size of the dynamic cluster and the amount of mobility information required to calculate the reserved bandwidth, the amount of information transferred from one base station to another may not be large. However, when all the mobile users in the system are taken into consideration, the situation can be different, and the amount of

---

[5] An efficiency way to estimate the mobility information for the dynamic scheme can be found in Ref. [55], where a hierarchical location prediction scheme is proposed.

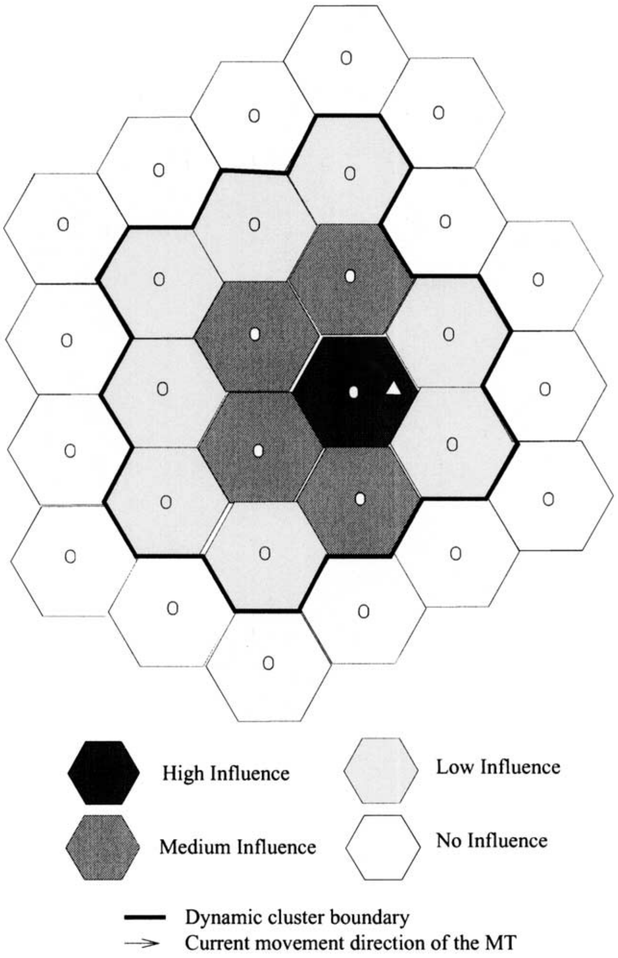
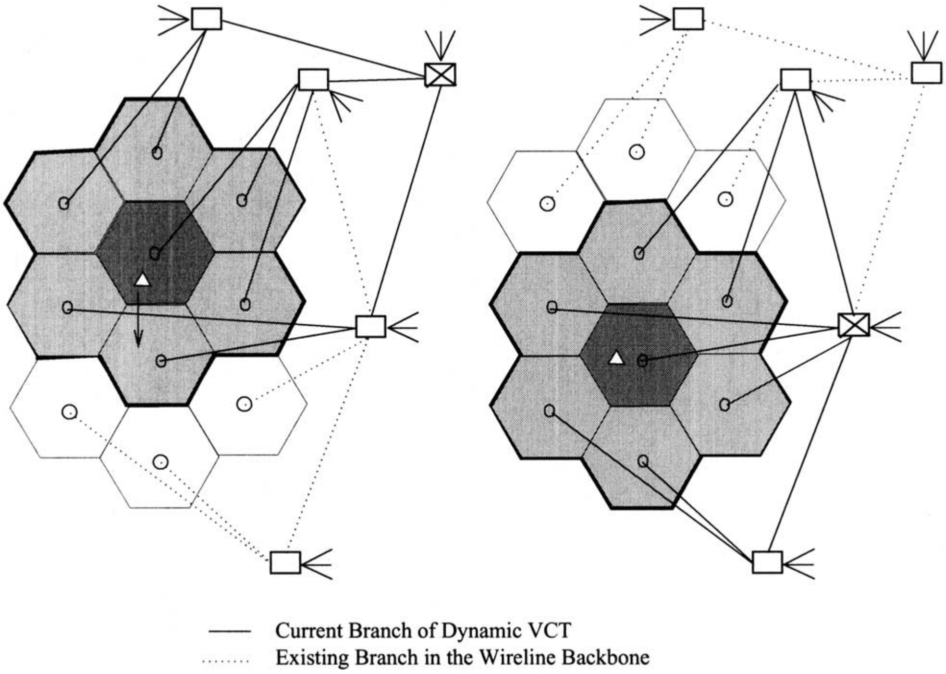


Fig. 11. An example of the dynamic or shadow cluster [53].

mobility information transferred can be very large [53,85]. This cannot be ignored, especially in a congested network, where the transfer of an unnecessary amount of mobility information can further slow down other connections in the wireline backbone. Therefore, the size of the dynamic cluster and the way in which the reserved bandwidth is calculated must be chosen carefully to minimize the flow of mobility information between base stations.

5.2. Distributed CAC policies

With the help of the dynamic VCT and the dynamic cluster, CAC can be done in a distributed fashion by distributing the task to each base station within a subregion. In this case, each base station handles new and handoff call requests within the cell boundary. Because of the importance of handoff calls, exclusive resources are usually reserved for them. To calculate this amount accurately, each base station is allowed to exchange users' mobility information with others in the vicinity, possibly through the backbone, in order to predict how many handoff calls will occur. Consequently, the problem of wireless handoff is addressed, with a price on the complexity due to: (i) invocation of the CAC function for each handoff call, and (ii) communication overhead among the base stations. Unlike the centralized CAC, there have been many previous studies on the distributed CAC. This is because the modeling of a single test cell is, in most cases, simpler than that of a group of radio cells together as in the case of centralized CAC. The major advantage of distributed CAC



(a) Before Handoff (b) After Handoff

Fig. 12. The dynamic VCT and the dynamic cluster before and after handoff.

is that each radio cell is capable of reserving resources for local handoff calls.

CAC for single type of constant-rate services. Consider a system supporting only a single type of constant-rate services with one and two-dimensional mobility models, respectively [67]. The one-dimensional model as shown in Fig. 13 corresponds to the highway scenario, whereas the two-dimensional one as shown in Fig. 14 corresponds to the normal street scene. The traffic pattern is assumed to be uniform and a mobile user is assumed to be equally likely to initiate a handoff to any of the adjacent cells. To determine whether to admit a mobile user or not, two criteria are used: (i) by accepting the new call the desired QoS of existing calls must be maintained; (ii) the system must provide the new call its desired QoS before and after it hands off to an adjacent cell. To evaluate the criteria, any base station has to be in constant communication with its neighboring base stations in order to exchange mobility information of the mobile users. Based on these criteria, the admission threshold for new calls in each radio cell under stationary traffic conditions can be found for any specified QoS level and traffic load. Simulation results show that the admission threshold obtained based on the criteria is better than other arbitrary values in the sense that the system has lower new call blocking probability under light traffic conditions and lower handoff call dropping probability under heavy traffic conditions.

CAC for multiple types of constant-rate services. In a system providing multiple types of constant-rate services, different criteria can be used to determine whether or not a mobile user should be admitted [53]. Two entities, the availability and survivability estimates, are defined, respectively, for each radio cell and mobile user in the system. The two estimates are functions of the mobility information of the mobile users, their desired QoS levels, and the capacity of the radio cells. Whether a mobile user is admitted to a radio cell or not depends on its survivability and the resource availability of the destined radio cell. The higher the survivability of a mobile user the higher the chance its QoS level can be maintained without affecting the QoS of others. On the other hand, the higher the availability estimate of a radio cell the more probable its capacity is available for future calls. By evaluating these two estimates each base station in the system performs its own admission control. Simulation results show that, when compared to a random admission strategy, where a mobile user is admitted to the system randomly, the proposed admission control can guarantee the desired level of QoS without indefinitely sacrificing the resource utilization.

The limited fractional guard channel policy (LFGCP). With both handoff calls and new calls, CAC should give a higher priority to handoff calls than that to new calls. The LFGCP is such a CAC policy [71]. This policy is originally proposed to work with a single type of traffic, and can be abbreviated as $g^{\beta}_{T,M}$, where M is the number of channels available in a radio cell, $T (< M)$ is the number of occupied channels over which no new calls are accepted, and β is a constant denoting the probability of accepting a new call when the channel occupancy in the cell is T. Let i denote the current number of occupied channels. The policy can be summarized as: (1) a new call is always accepted if $i < T$, is accepted with probability β if $i = T$, and is always rejected if $i > T$, and (2) a handoff call is always accepted if $i < M$. The advantage of the LFGCP is that, given traffic conditions and constraints on the QoS requirements, the optimal parameters of the policy can be found easily. After QoS requirements are specified, the resource utilization of the system can be maximized by using the minimum amount of resources to satisfy those requirements. An algorithm called `Min M` is proposed to find the minimum value of M and the corresponding values of T and β, under a certain traffic conditions and constraints on QoS requirements.

CAC for constant-rate and available-rate services. The LFGCP can be extended to a system supporting both constant-rate and available-rate services [50]. The CAC policy consists of two LFGCPs, one for each type of traffic. The mobility model consists of a single test cell, characterized by a new call arrival rate, a lumped handoff arrival rate and a service rate for each type of traffic. The RA scheme shown in Fig. 9(c) is used for resource allocation. The real-time constant-rate traffic has pre-emptive priority over the non-real-time available-rate traffic, and can occupy up to C_{I} out of the total capacity. The leftover resources in the test cell are equally shared among all the admitted available-rate users. The amount of resources received by each available-rate user is therefore random. Because the amount of resources received by an available-rate user varies, there is a chance that this amount drops to a critical level below which a connection will suffer severe quality degradation. This phenomenon is called overload. Thus, in addition to the new call blocking and handoff call dropping probabilities for both types of calls, the overload probability is a QoS measure for the admitted available-rate users. Given certain QoS requirements for both types of calls and the total capacity C of the radio cell, the minimum value of C_{I} required can be calculated. A minimum value is optimal in the sense that no resource is wasted. In particular, an algorithm called `Min M_2` is proposed that can find the optimal parameters for the two LFGCPs. Numerical results show that the new call blocking and handoff call dropping probabilities for both types of calls can be maintained below the required upper bounds under all traffic conditions considered.

5.3. Soft QoS and adaptive CAC policies

CAC policies are usually designed to achieve the optimal performance for a target traffic condition. If the actual traffic condition deviates from the target one, the policies may result in dissatisfactory service quality or under-utilized resources. One solution to the problem is to make the CAC adaptive to changes in the environment. Furthermore, soft

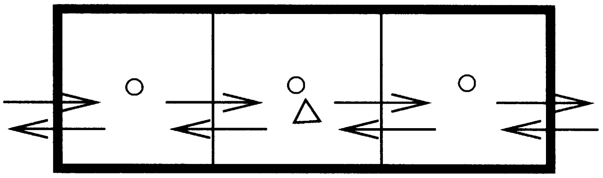


Fig. 13. One-dimensional mobility model (VCT and shadows not shown).

QoS can be incorporated with adaptive CAC to make a compromise among various QoS requirements of multimedia traffic in a wireless environment. The idea of soft QoS was first introduced in Ref. [85], where the relaxed QoS for handoff users was suggested to allow more users into the system. An idea similar to soft QoS was pursued in Ref. [49], where an adaptive CAC policy was devised, allowing users to sacrifice their bandwidth requirements in order to enjoy a lower handoff call dropping probability. Effectiveness of soft QoS and adaptive CAC for a system with both voice and data traffic is demonstrated numerically in Refs. [51,52].

CAC adapted to traffic load. Consider a system with one type of constant-rate traffic, where mobility information of the mobile users residing in a radio cell is passed to all its surrounding cells. At any particular cell, the instantaneous handoff arrival rate is estimated from the mobility information, which the cell receives from its immediate neighbors. The instantaneous handoff arrival rate can then be used to calculate the amount of guard bandwidth reserved for handoff calls [85]. This mobility information is constantly updated, so that the amount of guard bandwidth required adapts to the current needs of the system. Simulation results show that the adaptive policy works better than the traditional non-adaptive one under both stationary and non-stationary traffic load conditions. Soft QoS can be introduced by slightly raising the upper bound requirement for the handoff blocking probability. To maintain a certain level of handoff blocking probability requires guard bandwidth to be set aside for handoff calls, reducing the resource utilization. By slightly relaxing this requirement, resource utilization will not be sacrificed indefinitely. Let η_n and η_h, both not less than unity, be the factors of increase over the original target QoS requirements for new call blocking and handoff call dropping probabilities, respectively. In general, the utilization efficiency increases as η_n and/or η_h increases. However, there should be upper bound values for η_h and η_n; otherwise the purpose of having soft QoS becomes fallacious. Numerical results demonstrate that resource utilization efficiency can be easily increased by 10% with soft QoS. For example, for $\eta_n = 5$ and $\eta_h = 4$, a gain of 11.19% in resource utilization can be achieved [51].

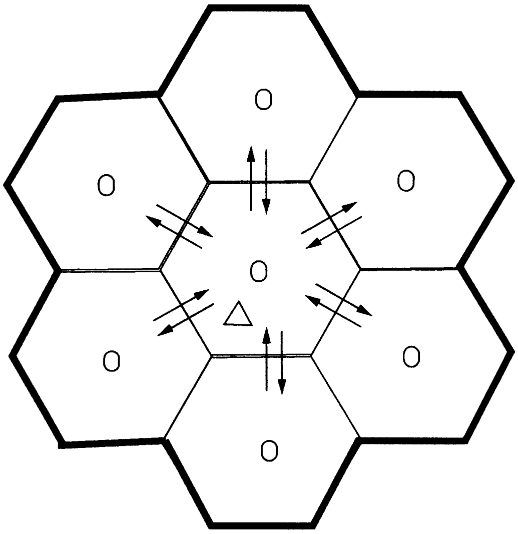


Fig. 14. Two-dimensional mobility model (VCT and shadows not shown).

CAC adapted to channel resource fluctuation. In addition to the traffic load adaptation, a CAC policy can be adaptive to the fluctuation of channel resources [49]. Consider a new type of service called the adaptive reserved service. Connections requesting this type of service are required to have a flexible QoS requirement, which can be specified in terms of two vectors: the throughput window $(b_{\min}, b_{\max})$ and the quality variation window $(q_{\min}, q_{\max})$. The quality of an adaptive reserved connection lies between $q_{\max}$ and $q_{\min}$ and is determined by the amount of bandwidth it receives from the system. The connection will be dropped only when the bandwidth it receives is below $b_{\min}$. The maximum amount of bandwidth that a connection ever needs is $b_{\max}$. The admission policy, depending on the current traffic load and channel conditions, may assign a bandwidth anywhere from $b_{\min}$ to $b_{\max}$ to an admitted mobile user. Once admitted, the mobile user may either renegotiate with or be required by the system to upgrade or downgrade its bandwidth. The QoS requirement of an adaptive reserved connection is flexible in the sense that the connection can tolerate different QoS levels (Fig. 15). As a result, a mobile user asking for the adaptive reserved service will have a smaller chance of being dropped than a mobile user asking for other types of services. The lower dropping probability will result in a larger number of users that can be supported by the system, which in turn translates into higher utilization of system resources. With the new service type, the adaptive policy is based on a bandwidth segregation scheme, which strategically divides the total capacity of a radio cell into different portions. The scheme is shown in Fig. 16, where C represents the total capacity of a radio cell. The bandwidth currently being used is denoted by $C_u(t)$. The unused capacity is divided into the primary portion and the secondary portion. The former can be used for all kinds of calls while the latter is reserved for handoff calls and call upgrades only. At each base station, the amount of secondary bandwidth, $C_2(t)$, is estimated periodically and the estimation is based on two factors: (i) the expected demand on bandwidth by possible handoff calls from adjacent radio cells; and (ii) the expected demand on bandwidth by possible upgrade requests from its own mobile users. Once $C_2(t)$ is determined, the adaptive policy

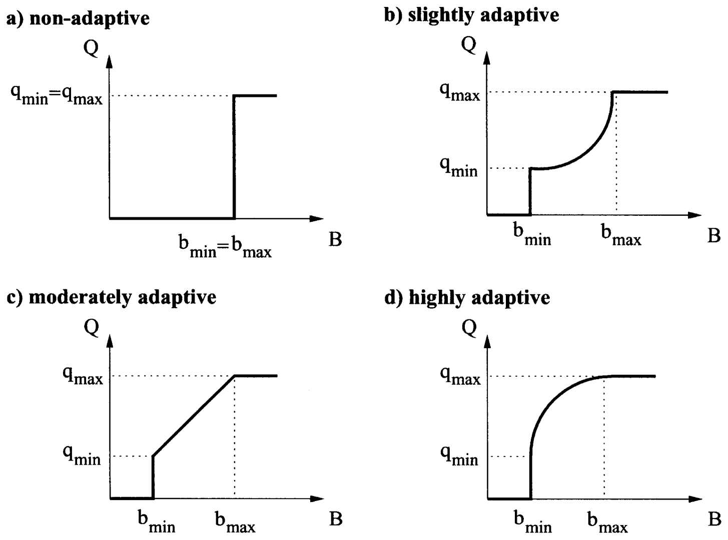


Fig. 15. QB plots for four different adaptive reserved connections [49].

accepts a new call only if $C_1(t)$, which is equal to $C - C_u(t) - C_2(t)$, is larger than the minimum bandwidth requirement, $q_{\min}$, of the new call request. Simulation results show that, when compared to a non-adaptive guard bandwidth policy, the adaptive one has a lower new call blocking probability for light traffic conditions and a lower handoff call dropping probability for heavy traffic conditions. The former translates into a higher utilization of network resources, whereas the latter construes into better provision of QoS to the admitted users.

CAC with both load adaptation and bandwidth allocation adjustment. Consider a system supporting both constant-rate and available-rate traffic. A CAC policy consisting of two correlated LFGCPs for voice and data users, respectively, can be used [51,52]. The traffic model for voice and data users under the control of the CAC policy is described by a two-dimensional Markov chain. The parameters of the policy can be determined using two adaptive approaches: (i) load adaptation and (ii) bandwidth allocation adjustment. In load adaptation, factors related to the traffic load are periodically monitored. If a significant change is detected, the procedure to find the new parameters of the CAC policy is re-executed. Due to scarce wireless spectrum and expected high number of subscribers in the future PCS, load adaptation alone may not be able to simultaneously deliver satisfactory services to both voice and data users. For example, when the traffic load of voice users becomes very large, the resources left for data users starts to decrease and eventually either (i) delay suffered by data users becomes completely intolerable if the same number of data users are supported, or (ii) the new call blocking probability has to be increased if the number of data users allowed into the system is reduced. The second scenario also leads to lower throughput and reduced resource utilization for the system. This problem can be mitigated by bandwidth allocation adjustment, which evolves from the fact that voice calls can tolerate a certain amount of reduction in transmission rate or capacity before their quality drops to an unacceptable level. When traffic load increases, it may be possible to take a small amount of resources from the users already in the system and use the aggregate to compensate for the extra demand. This can be done so long as the degradation in service quality caused by the reduced resource allocated to each admitted user is tolerable. Numerical results demonstrate that the bandwidth adjustment can be used together with load adaptation to achieve simultaneous QoS provisioning to multimedia traffic and to improve the efficiency of utilizing resources.

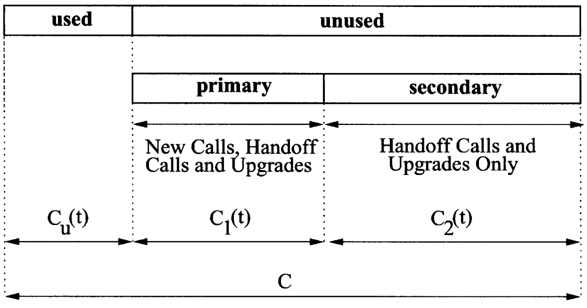


Fig. 16. The bandwidth segregation scheme.

Another use of soft QoS can be introduced in the multimedia environment, where each type of traffic has its own QoS requirements. Introducing soft QoS to some of the traffic types allows the requirements for other traffic easier to be met. For example, soft QoS can be applied to constant-rate traffic in the CAC policy in order to reduce the overload probability of available-rate users. When the frequency of overload is decreased, more resources are allocated to available-rate users and therefore the overall traffic throughput in the system is increased. Numerical results presented in Ref. [51] demonstrate that introducing soft QoS to constant-rate calls in the CAC policy can improve the overload probability and thus the throughput for available-rate calls.

5.4. Summary

The dynamic VCT dynamic cluster scheme allows the CAC to be done in a decentralized fashion. The CAC process requires the efficient exchanges of mobility information between radio cells and the accurate estimation of future resources required by a mobile user. Once the resource estimation is done, individual cell can predict future demands, reserve bandwidth accordingly, and admit only those mobile users that can be supported adequately. As the radio spectrum is very limited, the objective of CAC is to achieve high radio resource utilization and, at the same time, to ensure QoS satisfaction to multimedia mobile users. Due to the dynamic nature of multimedia traffic and the characteristics of a wireless mobile environment, the available resources in a base station can change dramatically as a mobile user enters or leaves the cell coverage area. As a result, simultaneously providing consistent QoS satisfaction to mobile connections with heterogeneous requirements and achieving maximal utilization of system resources are not realistic. Certain compromise between the two objectives has to be made. Adaptive CAC with soft QoS is one approach to make a good compromise.

6. Wireless/IP interworking

The pervasiveness of the Internet and the flexibility of the wireless communication network to support user roaming make ‘interworking’ of these two information transport platforms imperative for the support of multimedia services between remotely located mobile users. By design, the Internet only offers best effort service for fixed users, while the wireless environment supports user mobility and is prone to noise and interference. The interworking of these two communications subnets for end-to-end QoS provisioning presents a new set of challenges.

6.1. CAC in DiffServ networks

The classic best-effort Internet is evolving into a versatile network that can provide various multimedia real-time services in addition to the traditional data services, and can provision certain QoS guarantee to different Internet applications. Architectural frameworks for supporting DiffServ-based end-to-end QoS in the Internet, assuming underlying multiprotocol label switching (MPLS)-based explicit routed paths, are proposed in Refs. [20,82].

A simple and efficient approach to differentiate services is to use a set of buffers served with priorities. There are two levels of priority. One is the *inter-buffer priority*, where each class of traffic with a certain QoS requirement enters a separate buffer granted a certain priority, and the traffic in a buffer of higher priority is served before that of lower priority. Typically in a DiffServ core router, the buffer with the highest priority is used to serve the traffic for the *premium service* [42] to guarantee its low loss, low delay and low jitter; another buffer with the secondly high priority is used for the *assured service* [38]; the third buffer with the lowest priority can be used to serve the best-effort traffic. The other level of priority, *intra-buffer priority*, is to serve traffic with a *partitioned* buffer [48], which provides different loss priorities while keeping the order of packets from the same microflow. In the DiffServ model, the buffer for the assured service is usually a partitioned buffer, which allows the *in* profile traffic to use the whole buffer and the *out* profile traffic to use only part of the buffer. Thus, the out traffic always suffers a higher loss probability than the in traffic. In the performance analysis and capacity planning of such a multiclass multipriority DiffServ network, the priority structure should be considered.

The concept of effective bandwidth has provided a useful practical framework for CAC and capacity planning [10,30,46]. Given the traffic characteristics of a source, the effective bandwidth is the minimum link capacity required for the source under QoS constraint. For example, for real-time services, the peak rate of a source can be used as the effective bandwidth to achieve the premium service; for assured services of non-real-time traffic with packet loss requirements, the minimal channel capacity obtained from the buffer partitioning optimization can be used as the effective bandwidth. In Ref. [18], admission control at the flow level for QoS support is investigated based on the effective bandwidth concept. Buffer partitioning techniques are proposed to dynamically adjust the buffer partition thresholds according to the input traffic, in order to minimize the link capacity required. With the dynamic buffer partitioning, the admitted sources can be served with QoS guarantee. Numerical studies demonstrate that the effective bandwidth can be used for admission control in an additive way for heterogeneous multiclass Markovian sources.

6.2. CAC in mobile Internet

Provision of various real-time multimedia services to mobile users is the main objective of the next-generation wireless networks, which will be IP based and are expected to interwork with the Internet backbone seamlessly [12]. The establishment of such wireless mobile Internet is technically very challenging. Two major tasks are the support of fast handoff and the provision of QoS guarantee over IP-based wireless access networks.

The next-generation wireless networks will adopt micro/picocellular architectures for various advantages including higher data throughput, greater frequency reuse, location information with finer granularity. In this environment, the handoff rate grows rapidly and fast handoff support is essential. Especially for real-time traffic, the handoff call processing should be fast enough to avoid high loss of delay sensitive packets. To achieve fast handoff requires both a fast location/mobility update scheme and a fast resource allocation scheme. The popular scheme for fast location update is a *registration-domain-based* architecture, which basically is a cell-cluster based architecture. The radio cells (or the related base stations) in a geographic area are organized into a registration domain (e.g. a cellular IP network in the cellular IP scheme [15], a foreign domain in the HAWAII approach [72], and a foreign network in the TeleMIP architecture [24]), and the domain connects to the Internet through a gateway [15] (a foreign root router [72], or a mobility agent [24]). When a mobile host (MH)[6] moves into a registration domain for the first time, it will register the new care-of-address (the address of the gateway) to its home agent. While it migrates within the domain, the mobility updates messages will only be sent to the gateway, without registration with the home agent, which often locates far away.

Consider a registration-domain-based mobility management architecture as illustrated in Fig. 17 [19], where DiffServ is used to provision QoS. All the registration domains are DiffServ administrative domains in which all the routers are DiffServ IP routers. The gateway and base stations are edge routers, and they are connected through core routers. The gateway is the interface connecting to the DiffServ Internet backbone, where a service level agreements (SLA) is negotiated to specify the resources allocated by the Internet service provider to serve the aggregate traffic flowing from/into the gateway. Consider wireless links as bottleneck links in the domain and the SLA is negotiated mainly based on the wireless resource availability. The gateway conditions the aggregate traffic for each class according to the SLA resource commitments. The base stations provide MHs the access points to the Internet, and perform per-flow traffic conditioning and marking when data flow in the uplink direction. All the base stations in the same registration domain are connected to the same gateway router. All DiffServ routers use three separate queues to provide the premium service, the assured service and the best-effort service, respectively. The three buffers are served under priority scheduling or weighted fair queue (WFQ) scheduling. The traffic classes provided by the next-generation wireless networks can be mapped to these three DiffServ classes. For example, in a Universal Mobile Telecommunications System (UMTS) wireless network [25], the conversational class and the streaming class can be mapped to the premium service and the assured service, respectively, while the interactive class or the background traffic can be mapped to the best effort class. A bandwidth broker in the gateway router is responsible for the resource allocation and CAC over the DiffServ registration domain.

Using the effective bandwidth to characterize both the traffic characteristics and the QoS requirements, the resource commitments specified in the SLA can then be represented in terms of how many calls for each class are allowed in the registration domain. As a result, the admission control procedure is straightforward: whenever a new MH requests admission to a registration domain, the bandwidth broker determines whether to admit or reject the new call, based on the number of the calls currently in service and the SLA allocation for the service class to which the new call subscribes. The new call has to be blocked if all the SLA allocation has been occupied. This procedure requires very simple communications between the edge router (the base station) and the bandwidth broker (in the gateway router) and can be executed very fast. Furthermore, once an MH is admitted to a registration domain, it can hand off to other cells within the domain without the involvement of further CAC in the bandwidth broker. Such a simple resource allocation scheme in fact implies a very complicated design problem. The number of base stations in a domain, the resources allocated to each service class in each base station, and the resource commitment in the SLA should be determined carefully so that the new-call blocking and handoff dropping probabilities are reasonably low, while considering the traffic load in the registration domain, the mobility information and the call duration statistics. This design problem can be solved using the approaches discussed in Section 4. For example, for the situations that the interval between call arrivals, cell residence time and call duration are independently and exponentially distributed, the method presented in Ref. [66] can be used. To further decrease the handoff dropping probability, the guard channel scheme [85] can be used to reserve a fixed percentage of each base station's resources for handoff calls [20].

6.3. Summary

As the third generation (and beyond) wireless systems will employ packet switching in an all-IP infrastructure, CAC in wireless and wireline IP interworking is a new but important research area. In addition to the challenges introduced by the wireless domain, the connectionless nature of IP networks makes it difficult to guarantee QoS and to implement CAC. With the infrastructure of IntServ/RSVP or DiffServ based on MPLS for a path-oriented environment, CAC at the flow level can be implemented at edge routers to avoid network congestion.

7. Conclusions

Interconnection of the wireless and wireline domains using VCTs and cell clusters is geared to the solution of wireline handoff. This particular advantage comes from the fact that the major access point to the wireline backbone is fixed, making an end-to-end connection easier to maintain.

---

[6] In mobile IP, the term MH is often used instead of mobile terminal or mobile user.

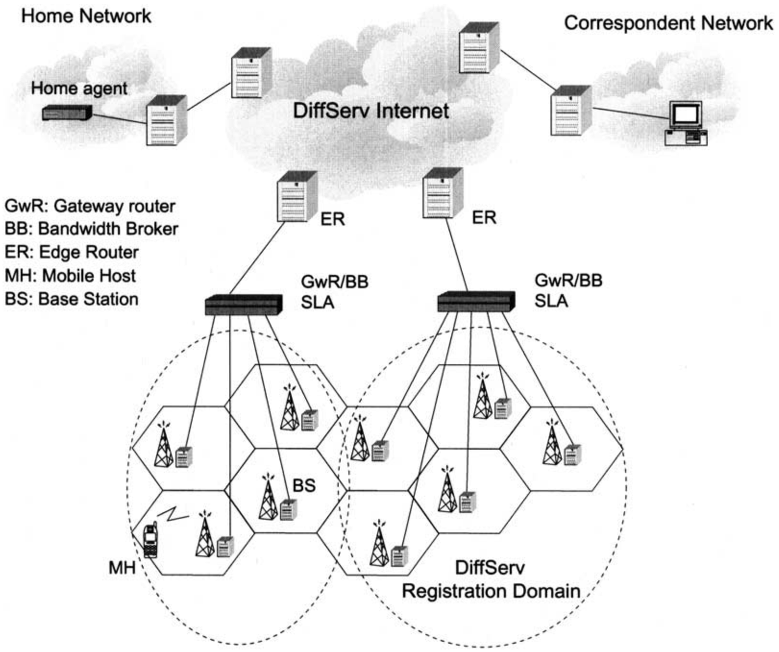


Fig. 17. A conceptual model of DiffServ registration-domain-based wireless network architecture.

According to the resource arrangement scheme used, the interconnection structure can be divided into two categories: the static VCT static cluster scheme and the dynamic VCT dynamic cluster scheme.

The works on CAC for the wireless and combined wireless/wireline domain are reviewed for each of the interconnection structure category. The static scheme allows the CAC to be done in a centralized fashion. Especially, radio cells are put into groups called static clusters and admission decision is done at the cluster level, rather than at the individual cell level. In contrast, the dynamic scheme allows CAC to be done in a decentralized manner. In particular, each radio cell is completely in control of its own resources and is allowed to make decision on whether or not to admit a particular mobile user. Compared to the dynamic scheme, the static scheme allows mobility information within a cluster to be partially ignored, which results in a simpler mobility model for the system. Moreover, the statistical multiplexing spawned within the static cluster increases network efficiency. On the other hand, the dynamic scheme allows the use of a more general mobility model to be applied to the system. The mobility model usually includes a small number of radio cells surrounding a center core, which consists of a single radio cell. By using this model, the dependence between radio cells that are far away can be neglected. This allows the same model to be applied to almost anywhere within the system, except in the boundary of the coverage area. Also, the more general model expedites future analyses when reconfiguration of the network is required. Consequently, both centralized CAC and distributed CAC approaches have their pros and cons.

In addition to what have been done on CAC, there are other aspects that must be looked at before the picture of CAC in the PCS is complete. For the static scheme, adaptive CAC, CAC that can handles variable-rate sources, and CAC based on both wireless and wireline resource availability should be investigated. For the dynamic scheme, CAC that can handle truly multimedia traffic should be studied. Also, most of the CAC policies proposed are not optimal. Optimization of a particular policy in terms of resource utilization with constraints on QoS requirements enhances the performance of the network. Optimization of network resources have been investigated for the wireline networks [31,32,35,62,63,75,76] and for the wireless networks [8,59,71,84]. Most of the problems are formulated in standard forms and can be solved by using software packages. The application of these optimization approaches to enhance the CAC policies for the PCS, which includes both wireline and wireless networks, should be explored in the near future. Finally, research efforts should be devoted to the application of the CAC strategies to future multimedia wideband CDMA wireless systems [41,57,58], interworking with the Internet [19], taking into account practical traffic models for packetized data applications [23,69] and asymmetric traffic loads between the uplink and downlink [43].

Acknowledgements

This work was partially supported by Communications and Information Technology Ontario (CITO) and by Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] A. Acampora, Wireless ATM: a perspective on issues and prospects, IEEE Pers. Commun. 3 (1996) 8–17.

[2] A. Acampora, M. Naghshineh, Control and quality-of-service provisioning in high-speed microcellular networks, IEEE Pers. Commun. 1 (1994) 36–43. 2nd Quarter.

[3] A. Acampora, M. Naghshineh, An architecture and methodology for mobile-executed handoff in cellular ATM networks, IEEE J. Select. Areas Commun. 12 (1994) 1365–1375.

[4] P. Agrawal, E. Hyden, P. Krzyzanowski, P. Mishra, M. Srivastava, J. Trotter, SWAN: a mobile multimedia wireless network, IEEE Pers. Commun. 3 (1996) 18–33.

[5] A. Alwan, R. Bagrodia, N. Bambos, M. Gerla, L. Kleinrock, J. Short, J. Villasenor, Adaptive mobile multimedia networks, IEEE Pers. Commun. 3 (1996) 34–51.

[6] G. Armitage, MPLS: the magic behind the myths, IEEE Commun. Mag. 38 (1) (2000) 124–131.

[7] K. Asatani, S. Nogami, Trends in standardization on multimedia communications, IEEE Commun. Mag. 35 (1997) 112–116.

[8] M. Asawa, W. Stark, Optimal scheduling of handoffs in cellular networks, IEEE/ACM Trans. Networking 4 (1996) 428–441.

[9] E. Ayanoglu, K. Eng, M. Karol, Wireless ATM: limits, challenges and proposals, IEEE Pers. Commun. 3 (1996) 18–34.

[10] A.W. Berger, W. Whitt, Effective bandwidths with priorities, IEEE/ACM Trans. Networking 6 (4) (1998) 447–460.

[11] Y. Bernet, The complementary roles of RSVP and differentiated services in the full-service QoS network, IEEE Commun. Mag. 38 (2) (2000) 154–162.

[12] L. Bos, S. Leroy, Toward an all-IP-based UMTS system architecture, IEEE Networks 15 (1) (2001) 36–45.

[13] J. Burgin, Dynamic capacity management in the BISDN, Int. J. Digit. Anal. Commun. Syst. 3 (1990) 161–165.

[14] S. Bush, S. Jagannath, R. Sanchez, J. Evans, V. Frost, G. Minden, K. Shanmugan, A control and management network for wireless ATM systems, Wireless Networks 3 (1997) 267–283.

[15] A.T. Campbell, J. Gomez, S. Kim, A.G. Valko, C.-Y. Wan, Z.R. Turanyi, Design, implementation and evaluation of cellular IP, IEEE Pers. Commun. 7 (4) (2000) 42–49.

[16] J. Causey, H. Kim, Comparison of admission control schemes in ATM networks, Int. J. Commun. Syst. 8 (1995) 165–184.

[17] C. Chao, W. Chen, Connection admission control for mobile multiple-class personal communications networks, IEEE J. Select. Areas Commun. 15 (1997) 1618–1626.

[18] Y. Cheng, W. Zhuang, Optimal buffer partitioning for multiclass Markovian traffic sources, Proceedings of the IEEE Globecom 2001, San Antonio, TX, December, 2001, pp. 1852–1856.

[19] Y. Cheng, W. Zhuang, DiffServ resource allocation for fast handoff in wireless mobile Internet, IEEE Commun. Mag. 40 (5) (2002) 130–136.

[20] Y. Cheng, W. Zhuang, Efficient Resource Allocation over MPLS Differentiated Services Networks, Proceedings of 3Gwireless'02, San Francisco, CA, May, 2002 p. 749–754.

[21] B. Cheung, V. Leung, Network configurations for seamless support of CDMA soft handoffs between cell cluster, IEEE J. Select. Areas Commun. 15 (1997) 1276–1288.

[22] D. Cox, Wireless personal communications: what is it? IEEE Pers. Commun. 2 (1995) 20–35.

[23] M.E. Crovella, A. Bestavros, Self-similarity in world wide web traffic: evidence and possible causes, IEEE/ACM Trans. Networking 5 (6) (1997) 835–846.

[24] S. Das, A. Misra, P. Agrawal, S.K. Das, TeleMIP: telecommunications-enhanced mobile IP architecture for fast intradomain mobility, IEEE Pers. Commun. 7 (4) (2000) 50–58.

[25] S. Dixit, Y. Guo, Z. Antoniou, Resource management and quality of service in third-generation wireless networks, IEEE Commun. Mag. 39 (2) (2002) 125–133.

[26] J. Dunlop, J. Irvine, D. Robertson, P. Cosimini, Performance of a statistically multiplexed access mechanism for a TDMA radio interface, IEEE Pers. Commun. 2 (1995) 56–64.

[27] Z. Dziong, J. Choquette, K. Liao, L. Mason, Admission control and routing in ATM networks, Comput. Networks ISDN Syst. 20 (1990) 189–196.

[28] A. Elwalid, D. Mitra, Effective bandwidth of general Markovian traffic sources and admission control of high-speed networks, IEEE/ACM Trans. Networking 1 (1993) 329–343.

[29] E. Eng, M. Karol, M. Veeraraghavan, E. Ayanoglu, C. Woodworth, P. Pancha, R. Valenzuela, A wireless broadband ad-hoc ATM local-area network, Wireless Networks 1 (1995) 161–174.

[30] J.S. Evans, D. Everitt, Effective bandwidth-based admission control for multiservice CDMA cellular networks, IEEE Trans. Veh. Technol. 48 (1) (1999) 36–46.

[31] G. Foschini, B. Gopinath, Sharing memory optimally, IEEE Trans. Commun. 31 (1983) 352–360.

[32] G. Foschini, B. Gopinath, J. Hayes, Optimum allocation of servers to two types of competing customers, IEEE Trans. Commun. 29 (1981) 1051–1055.

[33] G. Foschini, B. Gopinath, Z. Miljanic, Channel cost of mobility, IEEE Trans. Veh. Technol. 42 (1993) 414–424.

[34] R. Gibbens, F. Kelly, P. Key, A decision-theoretic approach to call admission control in ATM Networks, IEEE J. Select. Areas Commun. 13 (1995) 1101–1114.

[35] I. Gopal, T. Stern, Optimal call blocking policies in an integrated services environment, in: R. Melville, H. Weinert (Eds.), Proceedings of the Conference Information Science System, Baltimore, MD, USA, 1983, pp. 383–388.

[36] R. Guerin, H. Ahmadi, M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, IEEE J. Select. Areas Commun. 9 (1991) 968–981.

[37] N. Guo, S. Morgera, P. Mermelstein, Common packet data channel (CPDC) for integrated wireless DS-CDMA networks, IEEE J. Select. Areas Commun. 14 (1996) 735–749.

[38] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, Assured Forwarding PHB Group, Internet RFC 2597, June 1999.

[39] J. Hui, Resource allocation for broadband networks, IEEE J. Select. Areas Commun. 6 (1988) 1598–1608.

[40] J. Hyman, A. Lazar, G. Pacifici, A separation principle between scheduling and admission control for broadband switching, IEEE J. Select. Areas Commun. 11 (1993) 605–616.

[41] Y. Ishikawa, N. Umeda, Capacity design and performance of call admission control in cellular CDMA systems, IEEE J. Select. Areas Commun. 15 (1997) 1627–1635.

[42] V. Jacobson, K. Nichols, K. Poduri, An Expedited Forwarding PHB, Internet RFC 2598, June 1999.

[43] W.S. Jeon, D.G. Jeong, Call admission control for mobile multimedia communications with traffic asymmetry between uplink and downlink, IEEE Trans. Veh. Technol. 50 (1) (2001) 59–666.

[44] D. Jeong, C. Choi, W. Jeon, Design and performance evaluation of a new medium access control protocol for local wireless data communications, IEEE/ACM Trans. Networking 3 (1995) 742–752.

[45] R. Katz, Adaptation and mobility in wireless information systems, IEEE Pers. Commun. 1 (1994) 6–17. 1st Quarter.

[46] F.P. Kelly, Effective bandwidths of multiclass queues, Queueing Syst. 9 (1991) 5–16.

[47] G. Kesidis, ATM Network Performance, Kluwer Academic Publishers, New York, 1996.

[48] V.G. Kulkarni, L. Gun, P.F. Chimento, Effective bandwidth vectors for multiclass traffic multiplexed in a partitioned buffer, IEEE J. Select. Areas Commun. 13 (6) (1995) 1039–1047.

[49] K. Lee, Supporting mobile multimedia in integrated services networks, Wireless Networks 2 (1996) 205–217.

[50] C.W. Leong, W. Zhuang, Optimal call admission policy for wireless multimedia networks, Comput. Commun. 25 (10) (2002) 972–979.

[51] C.W. Leong, W. Zhuang, Soft QoS in CAC for wireless personal communications, Wireless Pers. Commun. 20 (2) (2002) 127–144.

[52] C.W. Leong, W. Zhuang, Novel system modeling in call admission control for wireless personal communications, Proceedings of IEEE Globecom'00, pp. 177–181, San Francisco, Nov.–Dec. 2000.

[53] D. Levine, I. Akyildiz, M. Naghshineh, A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept, IEEE/ACM Trans. Networking 5 (1997) 1–12.

[54] V. Li, X. Qiu, Personal communication system (PCS), Proc. IEEE 83 (1995) 1210–1243.

[55] T. Liu, P. Bahl, I. Chlamtac, Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks, IEEE J. Select. Areas Commun. 16 (1998) 922–936.

[56] Z. Liu, M. Karol, M. El Zarki, K. Eng, Channel access and interference issues in multicode DS-CDMA wireless packet (ATM) networks, Wireless Networks 2 (1996) 173–193.

[57] Z. Liu, M. El Zarki, SIR-based call admission control for DS/CDMA cellular systems, IEEE J. Select. Areas Commun. 12 (1994) 638–644.

[58] Y. Ma, J. Han, K. Trivedi, Call Admission Control for Reducing Dropped Calls in Code Division Multiple Access (Cdma) Cellular Systems, Proceedings of the IEEE Infocom'00, Tel-Aviv, Israel, March 26–30, 2000.

[59] U. Madhow, M. Honig, K. Steiglitz, Optimization of wireless resources for personal communications mobility tracking, IEEE/ACM Trans. Networking 3 (1995) 698–707.

[60] M. Marsan, C. Chiasserini, R. Cigno, M. Munafó, Local and global handovers for mobility management in wireless ATM networks, IEEE Pers. Commun. 4 (1997) 16–24.

[61] D. McMillan, Traffic modeling and analysis for cellular mobile networks, in: A. Jensen, V. Iversen (Eds.), Proceedings of ITC-13, 1991, pp. 627–632.

[62] G. Meempat, M. Sundareshan, Optimal channel allocation policies for access control of circuit-switched traffic in ISDN environments, IEEE Trans. Commun. 41 (1993) 338–350.

[63] B. Miller, A queueing reward system with several customer classes, Man. Sci. 16 (1969) 234–245.

[64] T. Murase, H. Suzuki, S. Sato, T. Takeuchi, A call admission control scheme for ATM networks using a simple quality estimate, IEEE J. Select. Areas Commun. 9 (1991) 1461–1470.

[65] M. Naghshineh, A. Acampora, QoS provisioning in micro-cellular networks supporting multiple classes of traffic, Wireless Networks 2 (1996) 195–203.

[66] M. Naghshineh, A. Acampora, Design and control of micro-cellular networks with QOS provisioning for data traffic, Wireless Networks 3 (1997) 249–256.

[67] M. Naghshineh, M. Schwartz, Distributed call admission control in mobile/wireless networks, IEEE J. Select. Areas Commun. 14 (1996) 711–717.

[68] M. Naghshineh, M. Willebeek-LeMair, End-to-end QoS provisioning in multimedia wireless/mobile networks using an adaptive framework, IEEE Commun. Mag. 35 (1997) 72–81.

[69] V. Paxson, Empirically derived analytical models of wide-area TCP connections, IEEE/ACM Trans. Networking 2 (4) (1994) 316–336.

[70] J. Proakis, Digital Communications, third ed., McGraw-Hill, New York, 1995.

[71] R. Ramjee, D. Towsley, R. Nagarajan, On optimal call admission control in cellular networks, Wireless Networks 3 (1997) 29–41.

[72] R. Ramjee, T. La Porta, S. Thuel, K. Varadhan, S.Y. Wang, HAWAII: a domain-based approach for supporting mobility in wide-area wireless networks, Proc. IEEE Int'l Conf. Network Protocols (1999) 283–292.

[73] D. Raychaudhuri, Wireless ATM: an enabling technology for multimedia personal communication, Wireless Networks 2 (1996) 163–171.

[74] D. Raychaudhuri, N. Wilson, ATM-based transport architecture for multiservices wireless personal communication networks, IEEE J. Select. Areas Commun. 12 (1992) 1401–1414.

[75] K. Ross, D. Tsang, The stochastic knapsack problem, IEEE Trans. Commun. 37 (1989) 740–747.

[76] K. Ross, D. Tsang, Optimal circuit access policies in an ISDN environment: a Markov decision approach, IEEE Trans. Commun. 37 (1989) 934–939.

[77] I. Rubin, S. Shambayati, Performance evaluation of a reservation random access scheme for packetized wireless systems with call control and handoff loading, Wireless Networks 1 (1995) 147–160.

[78] H. Saito, K. Shiomoto, Dynamic call admission control in ATM networks, IEEE J. Select. Areas Commun. 9 (1991) 982–989.

[79] M. Schwartz, Network management and control issues in multimedia wireless networks, IEEE Pers. Commun. 2 (1995) 8–16.

[80] M. Schwartz, Broadband Integrated Networks, Prentice-Hall, Englewood Cliffs, NJ, 1996.

[81] G. Swallow, MPLS advantages for traffic engineering, IEEE Commun. Mag. 37 (12) (1999) 54–57.

[82] P. Trimintzios, I. Andrikopoulos, G. Pavlou, P. Flegkas, D. Griffin, P. Georgatsos, D. Goderis, Y. T'Joens, L. Georgiadis, C. Jacquenet, R. Egan, A management and control architecture for providing IP differential services in MPLS-based networks, IEEE Commun. Mag. 39 (5) (2001) 80–88.

[83] X. Xiao, L.M. Ni, Internet QoS: a big picture, IEEE Networking 13 (1999) 8–18.

[84] A. Yener, C. Rose, Genetic algorithms applied to cellular call admission: local policies, IEEE Trans. Veh. Technol. 46 (1997) 72–79.

[85] O. Yu, V. Leung, Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN, IEEE J. Select. Areas Commun. 15 (1997) 1208–1225.