QoS provisioning in micro-cellular networks supporting multiple classes of traffic

Mahmoud Naghshineh [a] and Anthony S. Acampora [b]

[a] *IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA*
[b] *Department of Electrical and Computer Engineering & Center for Wireless Communications, University of California, San Diego, LaJolla, CA 92093-0409, USA*

Abstract. We introduce an adaptive call admission control mechanism for wireless/mobile networks supporting multiple classes of traffic, and discuss a number of resource sharing schemes which can be used to allocate wireless bandwidth to different classes of traffic. The adaptive call admission control reacts to changing new call arrival rates, and the resource sharing mechanism reacts to rapidly changing traffic conditions in every radio cell due to mobility of mobile users. In addition, we have provided an analytical methodology which shows that the combination of the call admission control and the resource sharing schemes guarantees a predefined quality-of-service to each class of traffic. One major advantage of our approach is that it can be performed in a distributed fashion removing any bottlenecks that might arise due to frequent invocation of network call control functions.

1. Introduction

Multimedia applications necessitate support of different classes of traffic with diverse quality-of-service requirements which need to be guaranteed by the transmission network. Provisioning of QoS requires congestion and admission control which is a problem that has been studied extensively in the past (see, for example, [1–3]). Networks supporting multimedia wireless applications require dynamic resource allocation and admission control which is a much more complex problem in wireless networks as compared to wired networks. This is due to the fact that support of multimedia applications requires micro/picocellular architectures in order to provide a higher capacity [4]. However, micro/picocellular architectures give rise to a high rate of hand-offs due to the small coverage area of each base station. Frequent hand-off will result in rapidly changing traffic load in the network and frequent invocation of network control functions to allocate capacity to mobile connections dynamically. One major factor to be considered in the wireless networks is the complexity of the network control functions performing dynamic resource allocation and call admission. These control functions need to be simple and performed in a decentralized manner in order to avoid potential bottlenecks [5–8].

Hierarchical decomposition techniques have been proposed in the past for dynamic control of integrated networks with multiple classes of traffic [10]. Similarly, we break up the dynamic control and resource allocation in integrated wireless networks into two independent sub-problems: (1) we perform class-based wireless call admission control; and (2) within the area covered by each base station, we allocate wireless bandwidth to all admitted connections according to some scheduling (or resource sharing) algorithm which is directly related to the class of traffic and its quality-of-service requirements. This approach extends the scheduling and admission control separation principle introduced for broadband switching in [11] to wireless networks supporting multiple classes of traffic with QoS guarantees.

The ultimate goal of our scheme is to provide a predefined packet-level quality-of-service (defined e.g. based on packet delay, jitter, and loss statistics) to each class of traffic. In order to achieve this, we pre-determine the amount of bandwidth BW_i required by each connection in a class of traffic (i) in order to maintain its packet-level quality-of-service [9] [1]. Now, the wireless call admission control takes the following parameters into account to make an admission decision: (1) the class of traffic (i); (2) the amount of bandwidth (BW_i) required by each connection in that class; (3) the number of calls of each class in the radio cell where the call seeks admission and its surrounding cells; and (4) the amount of wireless resource available in radio cell where the call seeks admission and its surrounding cells.

As far as the wireless call admission control is concerned, the admission threshold for different classes of traffic (or the admission region) can be calculated based on the traffic and hand-off characteristics, call holding time statistics, desired quality-of-service of each class of

[1] Using this approach, we might over-allocate radio bandwidth to a connection since the multiplexing gain at the packet-level is not taken into account. This means that the packet-level QoS provided by the system might be higher (or better) than the anticipated or guaranteed level. Furthermore, we assume a reliable link which provides a very low bit-error-rate.

traffic, and the resource sharing algorithm used by each base station. In order to react to changing traffic patterns and to allocate the scarce wireless resource efficiently, the admission threshold of different classes (or the admission region) is recalculated periodically. We refer to the time between two consecutive admission threshold calculations as the fixed admission period which is relatively long compared to a typical call holding time. During the fixed admission period, admission thresholds of different classes remain fixed and each base station estimates the Erlang load of each class of traffic. These estimates are reported to the network call controller at the end of the fixed admission period and are used by the network admission controller to calculate admission thresholds for different classes of traffic in the next fixed admission period. Since the fixed admission period is relatively long, the invocation of network call control functions would not be of any concern. The resource sharing algorithm is used by each base station to allocate wireless bandwidth to every connection based on its quality-of-service requirements and the quality-of-service requirements of other connections sharing the same base station. The scheduling algorithm is performed in a decentralized fashion and ensures that base station capacity is efficiently shared among different classes of connections. It reacts to the frequent change of the number of mobile terminals within each radio cell.

It is important to note that provisioning of QoS to different classes of traffic necessitates a highly reliable radio link between the wireless terminals and its access point [5]. First of all, due to delay sensitivity of realtime traffic such as voice or video, temporary link outages need to be mitigated. Secondly, multipath and shadow fading which give rise to varying bit error rate of the radio link need to be combatted in order to maximize the radio channel throughput and minimize retransmission. This can be done for example by using array antennas and optimal combining [12,13]. In addition, the system must provide means for maintaining a highly available radio link, rapid hand-off, and signaling for resource management purposes [5]. In this paper, we focus on the mobility aspect of wireless/mobile networks supporting multiple classes of traffic and defer the discussion of the effect of time varying radio channel on multimedia traffic to a later paper.

This paper is organized as follows. In section 2, we discuss the class-based call admission control. In section 3, we consider three intra-cell resource sharing mechanisms and define how they are used to allocate wireless bandwidth to different classes of traffic. The model discussed in section 4 provides an analytical methodology to calculate the QoS for different classes of traffic subject to the call admission control and the scheduling algorithm. Finally, in section 5, we provide some numerical results and discuss different trade-offs associated with bandwidth requirements and scheduling mechanisms.

2. Class-based wireless call admission

Our approach to class-based wireless call admission is based on the cell-cluster concept [5] which is a group of adjacent cells. Whenever a new mobile connection seeks admission to a cell-cluster, the cell-cluster call controller will admit (or reject) the call based on (1) the class of the mobile connection, (2) the number of wireless connections of each class already admitted to that cell-cluster. It is important to note that once a wireless connection is admitted to a cell-cluster, it can freely hand-off from one base station to another without the involvement of the call controller. As discussed in section 1, the admission threshold for different classes is updated periodically to adapt to changing traffic conditions. However, once a wireless connection is admitted, the probability of encountering a congestion state which can arise due to momentary high load conditions is limited to a pre-defined level.[2]

Based on the total mobile connection load in a given geographical area (such as the area covered by a number of neighboring cell-clusters), the number of base stations and the coverage area of each need to be designed in such a way that under normal conditions, the blocking probability of new calls due to admission control is suitably low [7].

2.1. Class-based quality-of-service metrics

Generally, our methodology can be applied to wireless networks supporting a large number of traffic classes. However, we consider two major classes of traffic, i.e. realtime and non-realtime, in this paper. These two classes of wireless connections are differentiated on the basis of the action initiated when a radio congestion state is encountered. For real time or class I connections such as voice or video, the connection must be dropped if the mobile moves into a congested area where no wireless channel is available. Hence, we define the QoS metric of a class I connection to be the hand-off dropping and forced call termination probability [14].

Class II calls are data connections which support applications requiring reliable transport and are characterized by their delay tolerance and loss sensitivity. Transmission Control Protocol (TCP) might typify such connections. Data connections typically do not require realtime transport of the information through the network and are therefore more tolerant to delay as compared to voice and video connections. Delay tolerance is

[2] Whenever a mobile connection leaves its associated cell-cluster in the middle of the call, it seeks admission to a new cell-cluster. We refer to this as the cell-cluster hand-off. By giving cell-cluster hand-off calls priority over new calls, the probability of encountering a congested radio state during a cell-cluster hand-off is reduced. The priority mechanism is designed based on the arrival rate of hand-off as well as new calls so that enough capacity remains available for the cell-cluster hand-off calls.

equivalent to acceptance of a variable service rate from the transport network, and data transport protocols are designed in such a way that when the network is congested, the arrival rate of new packets to the network is reduced (by reducing the window size, or other flow control mechanisms) [15]. Once the congestion passes, the normal flow of packets into the network is resumed. Thus, once a wireless data connection encounters a congested area wherein the total available wireless capacity is smaller than the total instantaneous capacity required by wireless terminals in that area, the end-to-end control entity (e.g. transport protocol) will observe the congestion and will reduce the flow of new packets into the network. In other words, instead of queuing or dropping a wireless data connection which hands-off to a congested radio cell, the wireless bandwidth of that cell can be "shared" by all wireless connections within that cell. The average available rate and the probability that the average available rate is smaller than a certain threshold are two quality-of-service metrics which we consider for a mobile/wireless data connection [7].

Regarding the wireless resources, the purpose of call admission control is to limit the number of in-progress wireless calls such that, once a wireless call is admitted, the probability of its encountering a congested radio state is acceptably low as to provide the required QoS. This is done by blocking new wireless call setup requests when the number of existing calls has reached this limit. Thus, the new-call blocking probability is another QoS metric which needs to be considered for all classes of wireless traffic.

3. Scheduling

In this section, we discuss three different scheduling mechanisms which define how the wireless spectrum of any radio cell is shared among realtime and non-realtime connections within its domain. In our model, we define bandwidth in terms of Units of Bandwidth (UB), where each realtime connection uses a fixed amount of bandwidth equal to BW_I UB, and a total bandwidth of C_I UB is assigned to realtime connections in each base station. A realtime connection is dropped if it hands-off to a base station in which the bandwidth used by all realtime connections is equal to C_I UB. At any given time, *the bandwidth available to non-realtime connections is shared equally between all of them.*

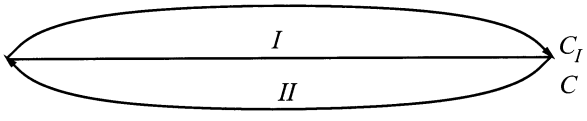
Sharing schemes for different types of connections have been studied in the past in the context of integrated networks [16–18]. In this chapter, we consider three simple intra-cell scheduling mechanisms which all have two attributes in common: (1) realtime connections can use up to C_I UB of a base station with pre-emptive priority over non-realtime connections to meet their strict delay requirements; (2) the value of C_I is updated at the beginning of each fixed admission period and remains unchanged during that period. In the following, we describe these sharing mechanisms which differ in the way non-realtime connections use the wireless bandwidth not utilized by realtime connections. We denote the total base station wireless capacity by C.

- *Complete Partitioning (CP)* – In this scheme, the base station capacity is completely partitioned where realtime connections use up to C_I UB, and $C - C_I$ UB is assigned to non-realtime connections. This scheme is illustrated in Fig. 1(a). As we discuss later, this scheme is the least efficient scheme.

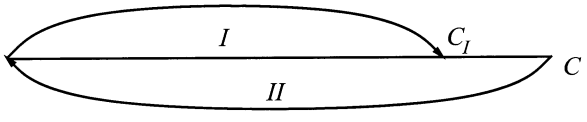
- *Class I Complete Access (CA)* – In this scheme, realtime connections can use up to the total base station capacity ($C_I = C$) with pre-emptive priority over non-realtime connections. At any given time, the capacity not utilized by realtime connections is equally shared between all non-realtime connections. This scheduling mechanism is shown in Fig. 1(b).

- *Class I Restricted Access (RA)* – In the restricted access mechanism illustrated in Fig. 1(c), realtime connections can use up to (and not more than) $C_I < C$ UB with pre-emptive priority over non-realtime connections in any radio cell, and $C - C_I$ UB is dedicated to non-realtime connections. Similar to the CA mechanism, the capacity not utilized by realtime connections is available to non-realtime connections at any given time.

(a) Complete Partitioning (CP)



(b) Class I Complete Access (CA)



(c) Class I Restricted Access (RA)

Fig. 1. Resource sharing schemes used by each based station.

4. Analysis

In this section, we provide a model to calculate the QoS of class I (realtime) and class II (non-realtime) traffic subject to call admission control described in section 2. The analysis is carried out for each of the three resource sharing mechanisms discussed in section 3. We consider cellular wireless systems where the full radio spectrum is reused in every cell [13], and each base station covers a fixed geographical area. This model resembles a Fixed Channel Allocation (FCA) scheme in systems with spectrum partitioning. We perform call admission control for class I and class II traffic. A new call admission request is rejected if the number of existing calls of the same class (as the class of the new call) in the cell-cluster is equal to the admission threshold corresponding to that class of traffic. We denote N_I as the admission threshold of class I (or realtime) traffic, and N_{II} as the admission threshold of class II (or non-realtime) traffic. In addition, we denote B as the number of base stations in a cell-cluster.

4.1. Realtime traffic

For realtime connections, we calculate the steady state probability of the number of connections per base station, the new call blocking probability, hand-off dropping probability, and forced call termination probability in a wireless/mobile network subject to wireless call admission control using the analysis presented in [14]. We consider a homogeneous system in which the new realtime call arrivals are Poisson with rate λ_I per radio cell, the call duration is exponentially distributed with mean $1/\mu_I$, and the time a call spends with any base station before handing-off to another base station is exponentially distributed with mean $1/h_I$ where the hand-off rate is denoted by h_I. We assume that the hand-off rate from any cell to any other cell is such that all cells experience the same rate of arrival of hand-off calls. This would arise, for example, when a call hands-off to any adjacent cell with equal probability, or when the hand-off probability to any cell at distance r from the original cell is P_r such that $\sum_{all\,r} P_r = 1$, or for any other hand-off pattern that would result in a homogenous load distribution among all cells. (Our model can be extended to incorporate different rates in a non-homogeneous system.) Furthermore, we assume that all wireless real-time connections are of the same type (e.g. 10 kbps voice connections), and that any base station can support up to m_I calls. The hand-off dropping can be taken into account by considering the fact that the effective or actual call departure rate μ_{I_e} is higher than the "natural" call departure rate μ_I. Thus, we use the following approximation to derive the call blocking probability of the controlled system. A new call can be blocked by the network admission controller if the total number of calls in the cluster exceeds a predetermined threshold, or if the base station covering its geographical area is full and cannot support any additional connections. We denote the probability of being blocked by the admission controller by P_{AB}. Since λ_I is the new call arrival rate per radio cell in a cell-cluster, the admission blocking probability of new calls is given by $P_{AB} = E((B\lambda_I/\mu_{I_e}), N_I)$, where $E(\rho, m_I)$ represents the Erlang loss formula defined as $(\rho^{m_I}/m_I!)/(\sum_{i=0}^{m_I} \rho^i/i!)$. In addition, the effective departure rate of wireless calls from the system is given by $\mu_{I_e} = \mu_I + h_I P_{HD}$, where P_{HD} denotes the hand-off dropping probability. Thus, the effective Erlang load in any radio cell is given by $\rho_e = \lambda_{I_e}/\mu_{I_e}$, where $\lambda_{I_e} = \lambda_I(1 - P_{AB})$. Assuming that mobile hand-offs (or movement patterns) are independent and identical, and that the arrival of newly admitted calls to any cell is Poisson, we model the aggregate arrival of calls to any cell in a cell-cluster subject to call admission control as Poisson. As a result, the hand-off dropping probability can be calculated as $P_{HD} = E(\rho_e, m_I)$. Hence, P_{AB} and P_{HD} can be calculated by solving the following set of non-linear equations:

$$\mu_{I_e} = \mu_I + h_I P_{HD}\,, \tag{1}$$

$$\lambda_{I_e} = \lambda_I(1 - P_{AB})\,, \tag{2}$$

$$P_{AB} = E((B\lambda_I/\mu_{I_e}), N_I)\,, \tag{3}$$

$$P_{HD} = E((\lambda_{I_e}/\mu_{I_e}), m_I)\,. \tag{4}$$

The above set of equations can be solved by repeated substitution [14].[3] Based on the above and since a new call can be blocked by the admission controller, or after being admitted, by a base station which is fully loaded, the total blocking probability of a new call is given by $P_B \simeq 1 - (1 - P_{AB})(1 - P_{HD})$. Under heavy load conditions ($\lambda \to \infty$), eq. (4) approaches an asymptote and the hand-off dropping probability can be calculated as follows:

$$\begin{aligned} P_{HD,\infty} &= \lim_{\lambda_I \to \infty} P_{HD} = E((\lambda_{I_{e,\infty}}/\mu_{I_e}), m_I) \\ &= E(N_I/B,\ m_I)\,. \end{aligned} \tag{5}$$

The above result can be explained as follows. Under heavy load conditions, the cell cluster will always have N_I calls. Hence, the Erlang load on any cell is given by N_I/B and the hand-off dropping probability is given by $E(N_I/B, m_I)$. Assuming an exponential call duration time and independent hand-off dropping probability in radio cells, the forced termination probability (the prob-

[3] Strictly speaking, our analysis does not incorporate hands-offs between cell-clusters explicitly. However, since the aggregate hand-off traffic leaving a cell-cluster at every radio cell is statistically equivalent to the hand-off traffic arriving to that radio cell from other cell-clusters in a homogeneous system, our model incorporates the hand-off traffic between cell-clusters implicitly and captures the effect of reduced Erlang load on every radio cell in a system subject to call admission control.

ability that a call is forcefully terminated prior to its completion due to a hand-off dropping) is given by [19]

$$P_T = \frac{P_{HD}}{\frac{\mu_I}{h_I} + P_{HD}} . \tag{6}$$

The steady-state distribution of number of calls in any cell of the cell-cluster subject to call admission control can also be approximated by

$$P_I(i) = ((\rho_e)^i/i!)/(\sum_{i=0}^{m_I}(\rho_e)^i/i!) , \tag{7}$$

where $\rho_e = \lambda_{I_e}/\mu_{I_e}$ is calculated based on the fixed point approximation as described in the above and $P_I(i)$ is the steady state probability that i calls are within any test cell in the cell cluster. Generally speaking, as the number of cells in a cell cluster and the number of calls that each base station can support (m_I) increases, the Poisson arrival assumption becomes stronger and the calculation of the steady-state distribution as described in the above would provide very accurate results. Under heavy load conditions, this assumption is not valid for small cell clusters where each base station can support only a very small number of connections (such as high bandwidth video traffic). In [14], we have provided a model where the call arrival to any cell is represented by a Markov-Modulated Poisson Process (MMPP) which can be used for a more accurate calculation of the steady-state distribution under such conditions.

4.2. Non-realtime traffic

Regarding the data traffic, the model in [6] is used to calculate the steady state probability of number of non-realtime connections. Let us consider a homogeneous system in which the new non-realtime call arrivals are Poisson with rate λ_{II} per radio cell, the call duration is exponentially distributed with mean $1/\mu_{II}$, and the time a call spends with any base station before handing-off to another is exponentially distributed. Similar to the analysis of realtime connections, we assume that the hand-off rate from any cell to any other cell is such that all cells experience the same rate of arrival of hand-off calls. Furthermore, we assume that all wireless non-realtime connections are of the same type and that each base station has a wireless bandwidth of C. In addition, we assume that the total base station capacity is *shared equally* among all active mobile users within its domain at any given time.

A wireless network without call admission control can be modeled as an open queuing network of $M/M/\infty$ queues, and the steady-state distribution of the number of mobiles per radio cell is given by a Poisson distribution $\pi(n) = \rho^n e^{(-\rho)}/n!$, where $\rho = \lambda_{II}/\mu_{II}$ is the Erlang load per radio cell. Now, if we perform call admission control and limit the number of calls admitted to a cell-cluster with B base stations to a value not greater than N_{II}, then the state-space of the system is a truncation of the state-space of the above open queuing network and has a product form equilibrium distribution where the probability of there being i mobiles in any cell, $P_{II}(i)$, is given by [6]

$$P_{II}(i) = [(\rho)^i/i!]\frac{\sum_{k=0}^{N_{II}-i}[(B-1)\rho]^k/k!}{\sum_{k=0}^{N_{II}}(B\rho)^k/k!} . \tag{8}$$

Also, we note that, under heavy traffic conditions, the steady-state distribution of the number of mobiles within any cell approaches a Binomial distribution:

$$\begin{aligned}\lim_{(\rho)\to\infty} P_{II}(i) &= \lim_{(\rho)\to\infty} (\rho)^i/i!\frac{\sum_{k=0}^{N_{II}-i}[(B-1)\rho]^k/k!}{\sum_{k=0}^{N_{II}}(B\rho)^k/k!} \\ &= \frac{(B-1)^{N_{II}-i}/[i!(N_{II}-i)!](\rho)^{N_{II}}}{(B\rho)^{N_{II}}/N_{II}!} \\ &= \binom{N_{II}}{i}\left(\frac{1}{B-1}\right)^i\left(\frac{B-1}{B}\right)^{N_{II}} . \end{aligned} \tag{9}$$

The wireless bandwidth available to any mobile terminal is a discrete random variable P_r where r is the available rate and belongs to the discrete space $\{C, C/2, C/3, \ldots, C/N_{II}\}$. The steady state probability of a rate $r = C/i$ being available to a mobile terminal is given by

$$P_{r=C/i} = \frac{i\,P_{II}(i)}{G} , \tag{10}$$

where $G = \sum_{i=0}^{N} iP_{II}(i)$ is the normalization constant and $P_{II}(i)$ is given by eq. (8). In eq. (10), the reason why the state distribution probability $P_{II}(i)$ is multiplied by the state i is simply that the probability that any given test mobile is in a cell with $k-1$ other mobiles is $kP_{II}(k)/(lP_{II}(l))$ times the probability that the test mobile is in a cell with $l-1$ other mobiles. Hence, the probability that the available rate to any mobile is less than a threshold r_{min} is given by

$$P[r < r_{min}] = \frac{1}{G}\sum_{i=n}^{N_{II}} i\,P_{II}(i) , \tag{11}$$

where n is the smallest integer greater than C/r_{min} and $P_{II}(i)$ is given by eq. (8).

Finally, the new-call blocking probability caused by limiting the number of mobiles to a threshold N_{II} per cell-cluster is readily calculated to be

$$P_B = \frac{(B\rho)^{N_{II}}/N_{II}!}{\sum_{j=0}^{N_{II}}(B\rho)^j/j!} . \tag{12}$$

In a system that supports both classes of traffic using the intra-cell resource sharing schemes described in section 3, the QoS and the steady-state distribution of realtime connections can be calculated independently of non-realtime connections by using the analysis of the Realtime traffic where each base station can support up to $m_I = C_I/BW_I$ (we assume that C_I is a multiple of

BW_I) realtime connections, and C_I is the bandwidth assigned to class I in any scheduling scheme discussed in section 3. This is due to the fact that class I connections have pre-emptive priority to class II connections. In addition, the hand-off dropping and forced call termination probability of class I connections is guaranteed to be bounded based on eq. (5) independent of the offered load. We calculate the QoS of class II connections by using eqs. (8)–(11), and incorporating the steady-state distribution of realtime connections in any radio cell. In addition, for all three scheduling mechanism, we prove that certain QoS can be guaranteed to class II connections independent of the load of either class of traffic.[4]

- *Complete Partitioning (CP)* – In this scheduling mechanism, the QoS of class II traffic can be calculated independently of class I traffic using eqs. (10) and (11), where the base station capacity is $C - C_I$ UB. Since under heavy traffic conditions, the steady state distribution of the number of class II connection approaches the binomial distribution given by eq. (9), the average available rate and the probability that the available rate is above a threshold can be guaranteed to class II connections[5].
- *Class I Complete Access (CA)* – In this scheme, since the whole base station capacity can be used by class I connections and class II connections share the available base station bandwidth not used by class I connections equally, the probability of a rate r being available to a class II connection can be calculated as follows:

$$P(r) = \frac{1}{\sum_{k=0}^{N} k P_{II}(k)} \sum_{i,j \in A_s} P_I(i)\, j\, P_{II}(j)\,, \qquad (13)$$

where A_s defines set of all combinations of i $(0 \le i \le C/BW_I)$ and j $(0 \le j \le N_{II})$ such that $(C - i\,BW_I)/j = r$, $P_I(i)$ is the steady state probability of having i class I connections in any radio cell given by eq. (7), $P_{II}(j)$ is the steady state probability of having j class II connections in any radio cell given by eq. (8). The finite set of all possible rates A_R can be calculated by considering all combinations i and j. Hence, the probability that the available rate to a class II connection is smaller than a threshold r_{min} can be calculated:

$$P[r < r_{min}] = \sum_{r \in A_R \text{ and } r < r_{min}} P(r)\,, \qquad (14)$$

where $P(r)$ is calculated in eq. (13). The average available rate to a class II call can be readily calculated as well. The new call blocking of class II calls can be calculated as specified in eq. (12). Since under heavy load conditions the steady state probability of the number of class I connection approaches the steady state probability given by eqs. (5) and (7), and the steady state distribution of class II connections approaches the binomial distribution given by (9), $P[r < r_{min}]$ has a limiting distribution and as a result a certain QoS can be guaranteed to class II connections independent of the load of class I or class II connections.

- *Class I Restricted Access (RA)* – The calculation of the QoS for this scheduling mechanism is very similar to the complete access (CA) case. The only difference is that the amount of capacity that can be used by class I calls is limited to $C_I < C$ and all possible values of i are given by $0 \le i \le C_I/BW_I$. Hence, corresponding to any C_I, A_s and A_R can be identified and the rate probability ($P(r)$) as well as $P[r < r_{min}]$ can be calculated using eqs. (13) and (14). Using the same argument as in the complete sharing mechanism, a predefined quality of service can be guaranteed to class II calls independent of load conditions.

5. Numerical results and discussion

In this section, we provide some numerical results for the QoS provided in a cellular system with cell-cluster-based call admission control supporting realtime and non-realtime traffic using the results of section 4. In addition, we compare the three scheduling mechanisms discussed in section 3, and discuss the results.

First of all, as we discussed in section 4, the QoS of class I is independent of the class II traffic and depending on realtime traffic characteristics, the admission threshold, and the amount of bandwidth C_I assigned to class I traffic in each cell, a predefined QoS can be guaranteed. We showed in section 4 that for any scheduling mechanism a certain QoS can be guaranteed to class II calls. This is shown in Fig. 2 where we have plotted the probability that the available rate per class II connection is less than 0.8 units of bandwidth ($P[r < 0.8$ UB]) as a function of class I and II traffic loads. In this figure, we consider a cell-cluster consisting of seven cells ($B = 7$) each providing $C = 20$ UB. Restricted access (RA) is used in every base station where $C_I = 15$ UB can be used by class I connections, and each class I connection utilizes $BW_I = 1$ UB. The admission threshold for both class I and class II calls is 60 calls per cell-cluster($N_I = N_{II} = 60$). As one can see, at high load, $P[r < 0.8$ UB] approaches an asymptote very rapidly protecting the already admitted connections from congestion.

[4] We note that the steady state distribution of the number of connections of either class can be used in our analysis since during the relatively long admission period the admission threshold of different classes remains unchanged.

[5] We assume that the call holding time is large compared to average time between two hand-off events in the system.

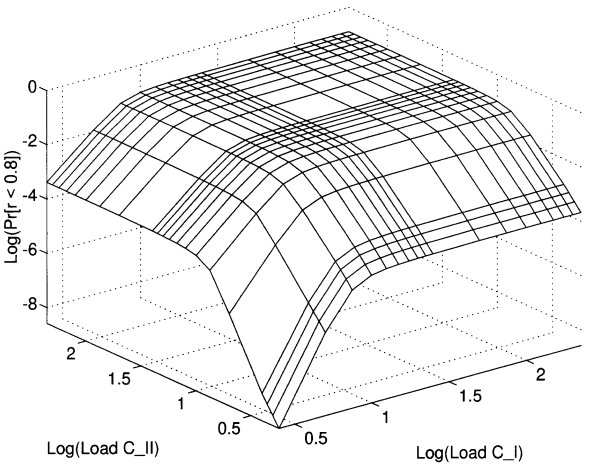


Fig. 2. Class II QoS as a function of class I and II Erlang loads.

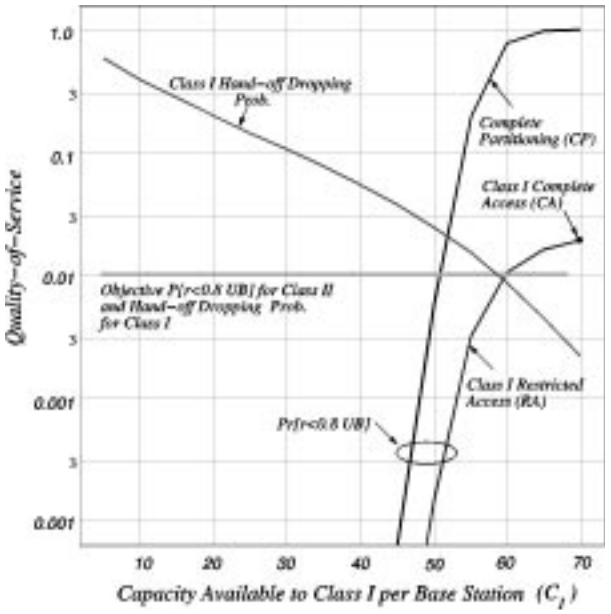


Fig. 3. A performance comparison of different resource sharing schemes.

An interesting and valuable study is the comparison of the three resource sharing schemes. As an example, we consider a system where each cell-cluster consists of seven base stations each having a capacity of $C = 70$ UB. In every radio cell, the arrival rate of class I calls is $\lambda_I = 6$ calls per unit of time and the arrival rate of class II calls is $\lambda_{II} = 14.5$ calls per unit of time. In addition, the average call holding time of both classes is 1 unit of time ($1/\mu_I = 1/\mu_{II} = 1$), and the average time a call spends with any base station prior to a handing-off to another is 0.2 units of time ($h_I = h_{II} = 5$). Furthermore, each class I connection requires $BW_I = 5$ UB. We would like to provide a QoS as defined by: (1) the new call blocking probability to be less than 1% for both classes of traffic; (2) the probability that the average available bandwidth to a class II call is less than 0.8 UB to be smaller than 0.01 ($P[r < 0.8 \text{ UB}] < 0.01$); (3) the average available bandwidth to any class II connection to be greater than 1 UB; (4) and, class I hand-off dropping probability to be less than 1%. We have limited the number of class I calls to $N_I = 70$ and the number of class II calls to $N_{II} = 119$ per cell-cluster. Based on the specified traffic and system parameters ($N_I, N_{II}, C, C_I, BW_I$) the new call blocking probability of either class of traffic is less than 1% and the average available rate to a class II connection is greater than 1 UB for all three scheduling schemes. In Fig. 3, we have plotted the QoS (in terms of hand-off dropping probability for class I and $P[r < 0.8 \text{ UB}]$ for class II) of CP, and RA sharing mechanisms as a function of the base station capacity assigned to class I traffic (C_I) and compared them to the QoS of the CA mechanism. As is shown in Fig. 3, only the Restricted Access (RA) scheme can provide both a $P[r < 0.8 \text{ UB}] < 0.01$ for class II connections , and a class I hand-off dropping probability smaller than 1%. In other words, only the restricted access scheme can satisfy all four QoS objectives. It can be easily shown that both CA and RA scheduling mechanisms provide a better QoS than the partitioning scheme. This is due to the fact that in the CP scheme the base station capacity not utilized by class I traffic is wasted. The reason why RA scheduling mechanism performs better than the CA scheme in this example is that often the desired QoS of class I calls can be achieved by allocating them a capacity smaller than the total base station capacity ($C_I < C$). This would result in more bandwidth being available to class II connections. In other words, by making the total base station capacity available to class I connections, a QoS better than their objective can be achieved at the expense of an unacceptable QoS to class II connections.

Another important factor that is usually considered in traffic integration problems is the ratio between bandwidth used by different classes and the effect of statistical multiplexing [17]. We discuss this problem in the following example. We consider a cell-cluster consisting of seven base stations ($B = 7$) each providing 50, 100, or 500 UB under heavy load conditions. The call admission controller sets the admission threshold for class I and class II calls such that the forced termination probability of class I calls is less than 1%, the probability that the available bandwidth to class II calls is less than 0.8 UB is less than 1%, and average available bandwidth to class II calls ($BW_{II,desired}$) is greater than 1 UB . We define the efficiency of each class to be the ratio $N_i\ BW_i/(B\ C)$, where N_i the number of class i connections, BW_i is equal to BW_I for class I, or equal to $BW_{II,desired}$ for class II traffic, B is the number of base stations, and C is the total base station capacity. In Fig. 4, We have plotted the efficiency of class II traffic as a function of the efficiency of class I traffic for Complete Access (CA) scheduling mechanism. Obviously, as we reduce the number of class I connections in the system, a larger number of class II connections can be admitted while satisfying the desired

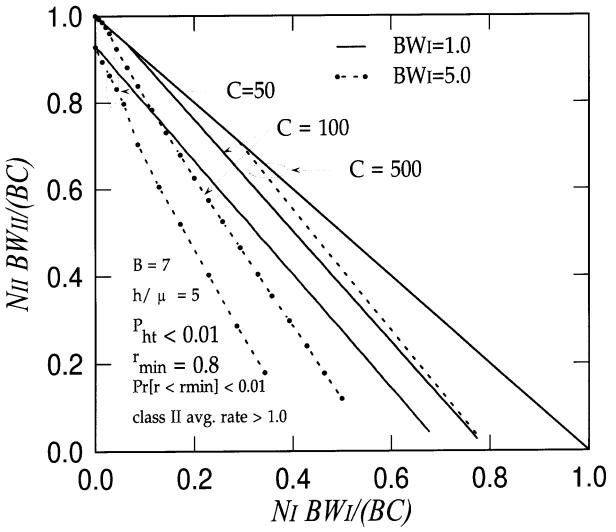


Fig. 4. Effect of base station capacity and required bandwidth for class I on the admission region using class I complete access scheduling mechanism.

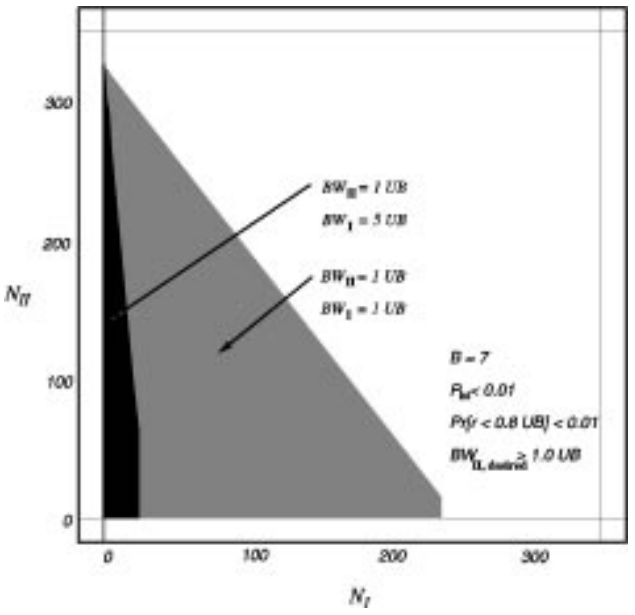


Fig. 5. Admission region of a cell-cluster with seven base station each with a capacity of $C = 50$ UB for class I complete access scheduling mechanism.

QoS. As one can see, by increasing the base station capacity, the wireless resources can be utilized more efficiently which is a result of trunking efficiency. Secondly, an increase in the bandwidth of class I connections, results in a reduction of the system utilization. The percentage of this reduction depends on the base station capacity, as well the number of class I calls.

In a sense, results shown in Fig. 4 identify an admission region for the wireless network supporting multiple classes of traffic with QoS guarantees. This is an extension of the admissible load region defined for broadband switching in [11]. The admission region of a cell-cluster in a mobile/wireless network defines all possible combinations of the number of wireless/mobile connections with different classes of traffic that can be admitted to a cell-cluster such that the QoS desired by each class of traffic is maintained. In Fig. 5, we have plotted two admission regions corresponding to two sets of class I and class II connections for a cell-cluster with $B = 7$ base station. In this example, each base station provides 50 UB, and the call admission controller sets the admission threshold for class I and class II calls such that the forced termination probability of class I calls is less than 1%, the probability that the available bandwidth to class II calls is less than 0.8 UB is less than 1%, and average available bandwidth to class II calls ($BW_{II,desired}$) is greater than 1 UB. Similar to Fig. 4, we consider the system under heavy load conditions using a CA scheduling mechanism. The inner region corresponds to the case where each class I connection requires 5 UB, and the outer region corresponds to the case where each class I connection requires 1 UB. Based on Figs. 4 and 5, and the results of section 4, we note that the size and the shape of a wireless admission region depends on bandwidth requirements of each class of traffic (BW_i), base station capacity (C), the size of the cell-cluster (B), and the scheduling mechanism (CP, CA, RA). The wireless admission region is defined based on the connection-level quality of service as defined in section 2, as compared to the admissible load region defined for example in [11] which is based on the packet-level quality of service. Both of these regions need to be taken into account in order to provide packet- and connection-level quality-of-service in wireless/mobile networks.

6. Conclusion

Quality-of-service provisioning in micro/picocellular networks supporting multiple classes of traffic necessitates distributed control of wireless network resources. These resources need to be allocated is such a way that when a wireless terminal hands-off from one access point to another, network resources are available with an acceptably high probability. In addition, these control functions need to be simple enough such that a high rate of hand-off events can be accommodated. The call admission region and parameters of the resource sharing scheme need to be updated periodically in order to adapt to changing traffic conditions and to make use of the scarce radio spectrum efficiently. In addition, the resource sharing in each base station needs to be performed in a decentralized fashion in order to adapt to rapidly changing number of connections in each radio cell. The combination of these two elements enable us to guarantee a predefined QoS to different traffic classes in the complex environment of microcellular networks supporting multimedia traffic. The adaptive control methodology presented in this paper can be extended to support a large number of connection classes and to

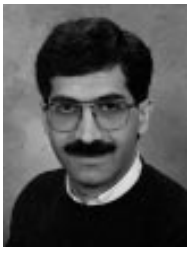
incorporate dynamic/hybrid channel allocation schemes.

Acknowledgements

Many thanks are due to Prof. Bruce Hajek for his helpful discussions, and to Prof. Mischa Schwartz for his careful reading of the manuscript and his helpful comments.

References

[1] IEEE J. Select. Areas Commun., Special Issue on Congestion Control in High-Speed Packet Switched Networks (September 1991).
[2] IEEE Commun. Mag., Special Issue on Congestion Control in High-Speed Packet Switched Networks (December 1991).
[3] A. Campbell, G. Coulson, F. Garcia, D. Hutchison, and H. Leopold, Integrated quality of service for multimedia communications, *Proceedings IEEE Infocom'93*.
[4] W.C.Y. Lee, Smaller cells for greater performance, IEEE Commun. Mag. (November 1991).
[5] A.S. Acampora, M. Naghshineh, Control and quality-of-service provisioning in high-speed microcellular networks ,IEEE Personal Commun. 1(2) (1994).
[6] A.S. Acampora and M. Naghshineh, An architecture and methodology for mobile-executed hand-off in cellular ATM networks, IEEE J. Select. Areas Commun. 12(8) (1994).
[7] M. Naghshineh, Distributed control of wireless/mobile networks, Doctoral Thesis, Columbia University (1994).
[8] D. Goodman, Personal communications, *Proc. 1994 Int. Zurich Seminar on Digital Communications* (1994).
[9] R. Guerin, H. Ahmadi and M. Naghshineh, Equivalent capacity and its applications to bandwidth allocation in high-speed networks, IEEE JSAC 9(7) (1991).
[10] R. Bolla, F. Danovaro, F. Davoli and M. Marchese, An integrated dynamic resource allocation scheme for ATM networks, *Proc. IEEE Infocom '94*.
[11] J.M. Hyman, A.A. Lazar and G. Pacifici, A separation principle between scheduling and admission control for broadband switching, IEEE J. Select. Areas Commun. 11(4) (1993).
[12] W.C. Jakes, *Microwave Mobile Communications* (Wiley, New York, 1974; Reprinted by IEEE Press, 1994).
[13] A.S. Acampora and J.H. Winters, A wireless network for wide-band indoor communications, IEEE JSAC 5 (June 1987).
[14] M. Naghshineh and A.S. Acampora, Design and control of micro-cellular networks with QoS provisioning for realtime traffic, to appear in J. High-Speed Networks, Special Issue on PCN (1996).
[15] M. Schwartz, *Telecommunication Networks: Protocols, Modeling and Analysis* (Addison-Wesley, 1987).
[16] Y.H. Kim and C.K. Un, Analysis of bandwidth allocation strategies with access restriction in broadband ISDN, IEEE Trans. Commun. 41(5) (1993).
[17] B. Kraimeche and M. Schwartz, Bandwidth allocation strategies in wide-band integrated networks, IEEE Select. Areas Commun. SAC-4(6) (1986).
[18] J.M. Aein, A multi-user-class, blocked-calls-cleared, demand access model, IEEE Trans. Commun. 26(3) (1978).
[19] S. Rappaport, The multi-call hand-off problem in high capacity cellular communications systems, IEEE Trans. Veh. Tech. 40(3) (1991).



Mahmoud Naghshineh is a Research Staff Member at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, where he currently works in the wireless and mobile networks group. He joined IBM in 1988. From 1988 to 1991, he worked on a variety of research and development projects dealing with design and analysis of local area networks, communication protocols, and fast packet-switched/broadband networks. Since 1991, he has been working in the area of wireless and mobile ATM, wireless access broadband networks, and mobile and wireless local area networks. He received his doctoral degree form Columbia University, New York, in 1994, M.S. in electrical engineering and B.S. in computer engineering from Polytechnic University, New York, in 1991 and 1988, respectively, and the Vordiplom degree in electrical engineering from RWTH Aachen, Germany, in 1985. He is a member of the IEEE communication society, and a member of the IEEE technical committee on computer communications as well as the technical committee on personal communications. He has served as a member of technical program committees, session organizer and chairperson for many IEEE conferences and workshops. He is also an editor of the IEEE Personal Communications Magazine. Currently, he is an adjunct faculty member of the Department of Electrical Engineering at Columbia University, where he teaches a course on wireless/mobile communications and networking. He has published numerous technical papers and holds a number of IBM awards and patents in the area of high-speed and wireless/mobile networks.
E-mail: mahmoud@watson.ibm.com

Anthony Acampora is a Professor of Electrical and Computer Engineering at the University of California, San Diego, and Director of UCSD's Center for Wireless Communications, where he is involved in wireless access research and education programs focused on antennas and progagation, circuits, communication theory, telecommunication networks, and multimedia applications. All of the work at the center involves strong participation and support from the wireless communications industry. Prior to joining the faculty at UCSD in 1995, he was Professor of Electrical Engineering and Director of the Center for Telecommunications Research at Columbia University. He joined the faculty at Columbia in 1988 following a 20-year career at AT& T Bell Laboratories, most of which was spent in basic research where his interests included radio and satellite communications, local and metropolitan area networks, packet switching, wireless access systems, and lightwave networks. His most recent position at Bell Labs was Director of the Transmission Technology Laboratory where he was responsible for a wide range of projects, including broadband networks, image communications, and digital signal processing. At Columbia, he was involved in research and education programs concerning broadband networks, wireless access networks, network management, optical networks and multimedia applications. Many of these projects enjoyed active industrial participation and involved cross-disciplinary research teams to develop new system approaches, analytical methodologies, VLSI circuitry, lightwave devices, and telecommunications software. He received his Ph.D in electrical engineering from the Polytechnic Institute of Brooklyn and is a Fellow of the IEEE and a former member of the IEEE Communication Society Board of Governors. Professor Acampora has published over 140 papers, holds 24 patents, and has authored a recently completed textbook entitled "An Introduction to Broadband Networks: MANs, ATM, B-ISDN, Self-Routing Switches, Optical Networks, and Network Control for Voice, Data, Image, and HDTV Telecommunications." He sits on numerous telecommunications advisory committees and frequently serves as a consultant to government and industry.