Wireless Networks 8, 85–95, 2002

QoS Performance Bounds and Efficient Connection Admission Control for Heterogeneous Services in Wireless Cellular Networks

DONGMEI ZHAO, XUEMIN SHEN and JON W. MARK
Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

Abstract. Quality-of-Service (QoS) performance and connection admission control (CAC) for heterogeneous services in wireless multiple access networks are investigated. The heterogeneous services include constant bit rate (CBR), variable bit rate (VBR) and available bit rate (ABR) services. Multiple access control is handled by a polling-based scheme with non-preemptive priority. Tight delay variation (jitter) bounds for CBR connections and delay bounds for VBR connections are derived. A CAC scheme based on the derived bounds is developed. The CAC makes use of user mobility information to reserve an appropriate amount of system resources for potential handoff connections to achieve low handoff connection dropping rate (HCDR). Simulation results show that the proposed CAC scheme can achieve both low HCDR and high resource utilization.

Keywords: cellular networks, Quality-of-Service, multiple access control, performance bound, connection admission control

1. Introduction

The internetworking of broadband wireline networks and wireless cellular networks is expected to provide adequate multimedia service support for mobile users anywhere at anytime. Because of user mobility, limited radio frequency spectrum, radio channel impairment, etc., how to efficiently utilize the precious radio resources and provide more users with guaranteed Quality-of-Service (QoS) levels becomes a challenging issue in the wireless segment of the integrated networks. Connection admission control (CAC) is to make a decision about whether a connection should be admitted or not, and has effects on both the system resource utilization and the QoS provided to the users. QoS performance bounds, such as the maximum delay or delay variation (jitter) experienced by a connection if it is admitted in the system, may be used as the basis for making the admission decision. A CAC scheme based on QoS provisioning by using tight performance bounds can achieve more accurate resource allocation and improve the system resource utilization. A polling-based multiple access control scheme with non-preemptive priority is proposed in [1] and a slightly different version is in [2]. With these schemes, sufficient conditions are derived for all the constant bit rate (CBR) connections to satisfy their jitter constraints and all the variable bit rate (VBR) connections to satisfy their delay constraints in both [1] and [2]. However, these bounds are too conservative to be used in practical systems.

User mobility is a special issue in cellular networks. To efficiently utilize the limited radio spectrum and maximize the system capacity, the cell size of future cellular networks tends to be smaller. As a result, frequent handoffs may occur during a connection's lifetime. When a connection requests to handoff to a new cell, sufficient resource must be available in the cell in order to accept the handoff connection and maintain a continuous connection. Because interrupting an on-going connection is much more undesirable than refusing to admit a new connection from the user's point of view, admission decisions should give a higher priority to a handoff connection than a new request. The higher priority can be achieved by reserving a certain amount of system resources for handling potential handoff connections. By doing so, lower handoff connection dropping rate (HCDR) than the new connection blocking rate (NCBR) can be achieved. Different approaches have been proposed in the literature to do the resource reservation. The *Guard channel* approach is proposed in [3], where a fixed amount of resource is reserved for handoff connections. The *Virtual connection tree* (VCT) approach is proposed in [4] to support high rate handoffs in wireless ATM networks. The VCT is a group of pre-established connections between a fixed switch and a set of base stations (BSs) with which the mobiles could potentially associate. Each of the BSs in the VCT reserves 100% resources for each of the connections in the VCT. A mobile can freely handoff to any cell within the VCT without being subject to a further admission control. However, the resource utilization is relatively low for both guard channel and VCT approaches due to the potential waste of the reserved resource. The *shadow cluster* approach is proposed in [5]. A shadow cluster is a set of BSs that a mobile may influence in the near future and is updated based on the user mobility information. The influence area moves along with the mobile, like a shadow. As the mobile moves, new BSs which are within the mobile's new influence area are included in its shadow cluster, while the old ones which are out of the influence area are removed from the shadow cluster. The shadow cluster concept is used in the CAC scheme in [5] to predict the resource demand for homogeneous services in the near future and to reserve resources accordingly.

In general, admission decisions in wireless cellular networks are made to ensure guaranteed QoS for heterogeneous traffic, while maintaining high resource utilization and low

HCDR. Reserving more resource achieves lower HCDR, but reduces the system resource utilization. Therefore, how to balance the tradeoff between the efficiency of resource utilization and the satisfaction of QoS to mobile users is a very important issue. Handoffs are mainly caused by user movement. The more likely users will handoff to a particular cell, the more resources should be reserved in that cell. In other words, effective and efficient resource reservation should be based on user mobility information, which is the probability that a mobile user may reside in a particular cell at future moments and determined by the users' movement, including initial locations, speeds and directions [7,8]. So far, little work has been done to reserve resources for potential handoff connections based on user mobility information in the wireless systems.

In this paper, tighter QoS performance bounds, compared to those in [1] and [2], are derived for both CBR and VBR traffic. A CAC scheme based on the derived bounds is developed. The CAC scheme gives a higher priority to handoff connections by reserving an appropriate amount of system resources for the handoff connections in order to achieve lower HCDR. The resource reservation in the CAC scheme makes use of the user mobility information to ensure efficient system resource utilization. The remainder of the paper is organized as follows. Section 2 describes the system model. Tight jitter bounds for CBR traffic and delay bounds for VBR traffic are derived in section 3. Section 4 presents the proposed CAC scheme. Simulation results are shown in section 5, followed by the conclusions in section 6.

2. System model

We consider a TDMA cellular network connected to a wireline backbone network through a mobile switching center (MSC). The service area of an MSC consists of several radio cells, each of which is the coverage area of a base station (BS). A mobile station (MS) selects its associated BS according to the received signal strength. This research is confined to the coverage area of the BSs under one MSC and focused on the uplink, since the uplink usually has worse propagation and interference conditions than the downlink, and the system capacity is restricted by the uplink. All the data packets are assumed to have the same length. In general, packet transmission accuracy can be improved by techniques such as diversity reception, forward error correction, automatic retransmission request at the physical and the link layers. For the CAC problem under consideration, we assume that the channel impairment mitigation strategies are in place and the residual error rate at the output of the BS receiver is negligible.

The multiple access control strategy under consideration is polling-based with non-preemptive priority as shown in figure 1 [2]. Each connection has one ready-to-transmit (RTT) buffer located at the MS, referred to as MS RTT Buffer, to temporarily store its generated packets before transmission. The time difference between the instant a packet arrives to and the instant it departs from the MS RTT Buffer is its experienced delay. The difference between the experienced delay of any two successive packets is called delay variation, or jitter. Three types of traffic are considered: CBR, VBR, and available bit rate (ABR). All the connections with the same traffic parameters are grouped together as a class. An ith class CBR connection is characterized by the 2-tuple (γ_i, δ_i), $i = 1, \ldots, N_c$, where γ_i is the packet generation rate, δ_i is the maximum jitter tolerance, and N_c is the total number of CBR classes. The packet generation rate of a CBR connection is constant, so there is no packet burstiness. The ith class CBR connection is provided with guaranteed QoS if its maximum experienced jitter is less than δ_i. Without loss of generality, we assume that $\delta_{i_1} \leqslant \delta_{i_2}$ and $\gamma_{i_1} \leqslant \gamma_{i_2}$ for all $1 \leqslant i_1 < i_2 \leqslant N_c$. An ith class VBR connection is characterized by the triple (ρ_i, σ_i, d_i), $i = N_c + 1, \ldots, N_c + N_v$, where ρ_i is the average packet generation rate, σ_i is the maximum burst tolerance, d_i is the maximum tolerable transit delay, and N_v is the total number of VBR classes. The ith class VBR connection is provided with guaranteed QoS if the experienced delay for any of its packets is less than d_i. Each ith class VBR connection is also assumed to be regulated by a leaky bucket (LB) with parameters (σ_i, ρ_i) at the BS, where σ_i corresponds to the BS polling token (PT) buffer size, and ρ_i corresponds to the BS PT generation rate. Without loss of generality, we assume that $d_{i_1} \leqslant d_{i_2}$, $\delta_{i_1} \leqslant \delta_{i_2}$, and $\rho_{i_1} \leqslant \rho_{i_2}$ for all $N_c + 1 \leqslant i_1 < i_2 \leqslant N_c + N_v$. An ABR connection has no delay or delay variation specification and its minimum packet rate (MPR) is set to zero. All the ABR connections are grouped together as a class, indexed by $i = N_c + N_v + 1$. There are n_i connections in the ith class, $i = 1, \ldots, N_c + N_v + 1$. Class i_1 is given a higher priority than class i_2 for all $1 \leqslant i_1 < i_2 \leqslant N_c + N_v + 1$. Packet transmission is directed by the BS according to the preset priority. We assume that there is a separate channel for the transmission of control signals. The execution of the multiple access algorithm adheres to the following rules.

1. For each CBR or VBR connection in the system, there is a PT buffer at the BS, called BS PT Buffer.

2. For a connection in the ith CBR class, its PT is generated every $1/\gamma_i$ seconds at the BS and stored in the BS PT Buffer.

3. For a connection in the ith VBR class, its PT is generated every $1/\rho_i$ seconds at the BS and stored in the BS PT Buffer of size σ_i. For each VBR connection, there is also one RTT counter at the BS, called BS RTT Counter, to count the number of the packets waiting for transmission. The BS RTT Counter is initially set to zero. When a data packet is available at a VBR connection, it is stored in the MS RTT Buffer. At the same time, a packet-ready command is transmitted from the MS to the BS via the signaling channel. Upon receiving the command, the BS PT Buffer drops one token if it is not empty, and the BS RTT Counter is increased by one.

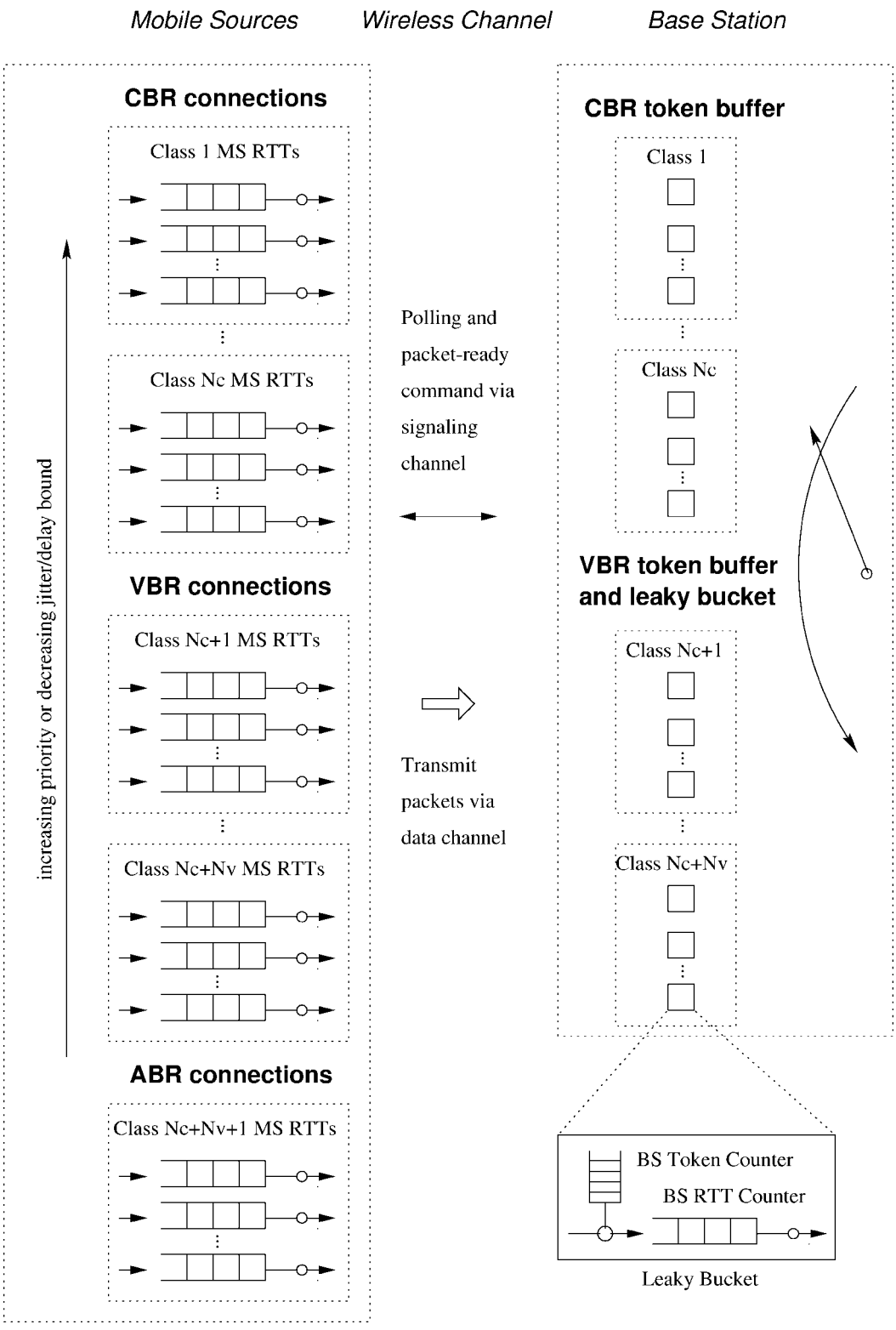

Figure 1. System model of a polling scheme with non-preemptive priority.

4. Whenever the channel is cleared, the BS scans the BS PT Buffers of the connections from the first CBR class to the last VBR class.

5. If a PT is found for a CBR connection, the BS removes the PT and polls the CBR connection.

6. If a PT is found for a VBR connection, the BS polls the VBR connection without removing the PT. At the same time, the BS RTT Counter is decreased by one. As long as the BS RTT Counter is larger than zero, the MS RTT Buffer is not empty, since the MS RTT Buffer contains at least as many packets as the BS RTT Counter value.

7. Every time when a connection is polled, it transmits at most one packet from its MS RTT Buffer.
8. Within each CBR or VBR class, the connections are served according to a round-robin scheme.
9. When there is no PT found for CBR and VBR connections, a connection from the ABR class is allowed to transmit at most one packet. How to schedule the packet transmission of the ABR connections to effectively and efficiently utilize the remaining available resources after serving the CBR and VBR traffic can be found in [1,9–11].

With the access control scheme, it is observed that all computations of the scheduling steps are performed at the BS, and the MSs only react to the commands from the BS. Therefore, MSs can be made with low power and light weight. A similar polling scheme is considered in [1], where the LB controllers of the VBR connections reside in each MS, and each CBR and VBR class has only one connection.

3. QoS performance bounds of CBR and VBR connections

In this section, jitter bounds for all CBR connections and delay bounds for all VBR connections are derived. Let τ_{p} be the time for the BS to poll a connection and τ_{d} be the time for a connection to transmit a data packet, respectively. The following theorems provide sufficient conditions for CBR and VBR connections to satisfy their performance constraints, where $a = \tau_{\mathrm{p}} + \tau_{\mathrm{d}}$.

Theorem 1 (Jitter bounds for CBR connections). Let

$$\delta_i' = \frac{\sum_{k=1}^{i} n_k}{1/a - \sum_{k=1}^{i-1} n_k \gamma_k}. \tag{1}$$

If

$$\delta_i' + a < 1/\gamma_i \tag{2}$$

holds for all $i = 1, \ldots, N_{\mathrm{c}}$, then the jitter of an ith class CBR connection is upper bounded by δ_i'. If, furthermore, $\delta_i' \leqslant \delta_i$ for all $i = 1, \ldots, N_{\mathrm{c}}$, then all the packets generated by these $\sum_{k=1}^{N_{\mathrm{c}}} n_k$ CBR connections meet their jitter constraints.

Theorem 2 (Delay bounds for VBR connections). Define recursively

$$d_i' = \frac{n_i + \sigma_i + 1 + \sum_{k=N_{\mathrm{c}}+1}^{i-1} n_k \widehat{\sigma}_k + \sum_{k=1}^{N_{\mathrm{c}}} n_k}{1/a - (\sum_{k=1}^{N_{\mathrm{c}}} n_k \gamma_k + \sum_{k=N_{\mathrm{c}}+1}^{i-1} n_k \rho_k)} \tag{3}$$

for $i = N_{\mathrm{c}} + 1, \ldots, N_{\mathrm{c}} + N_{\mathrm{v}}$, where

$$\widehat{\sigma}_i = \rho_i d_i' + \sigma_i + 1 \tag{4}$$

for $i = N_{\mathrm{c}} + 1, \ldots, N_{\mathrm{c}} + N_{\mathrm{v}}$. If $1/a - (\sum_{k=1}^{N_{\mathrm{c}}} n_k \gamma_k + \sum_{k=N_{\mathrm{c}}+1}^{i-1} n_k \rho_k) > 0$ holds, then the delay of the packets from an ith class VBR connection is upper bounded by d_i'. If, furthermore, $d_i' \leqslant d_i$ for all $i = N_{\mathrm{c}} + 1, \ldots, N_{\mathrm{c}} + N_{\mathrm{v}}$, then all the packets generated by these $\sum_{k=N_{\mathrm{c}}+1}^{N_{\mathrm{c}}+N_{\mathrm{v}}} n_k$ VBR connections meet their delay constraints.

Proof of theorem 1. Mark the time when a PT for an ith class CBR connection is generated as 0. Let $\widetilde{\delta}_i$ be the amount of time that the PT needs to wait before it is polled. It follows that during the interval $[0, \widetilde{\delta}_i)$, the channel must be busy serving all the connections in the $(i-1)$ higher priority classes. The total number of CBR connections from the first class to the $(i-1)$th class is equal to $\sum_{k=1}^{i-1} n_k$. Moreover, because of the round-robin scheduling within the same class, a ready-to-transmit connection has to wait for a maximum of $(n_i - 1)$ same-class connections to finish transmission. Thus, the total number of CBR connections that can be served within $[0, \widetilde{\delta}_i)$ is at most $\sum_{k=1}^{i-1} n_k + n_i - 1$. Then, the total number of PTs, or packets, that can be served in the interval is at most $\sum_{k=1}^{i-1} n_k \lceil \gamma_k \widetilde{\delta}_i \rceil + n_i - 1$. With channel clearing time for non-preemptive priority, the total amount of time to serve these packets is upper bounded by

$$\widetilde{\delta}_i \leqslant \left(\sum_{k=1}^{i-1} n_k \lceil \gamma_k \widetilde{\delta}_i \rceil + n_i - 1 \right) a + a. \tag{5}$$

Since $\lceil \gamma_k \widetilde{\delta}_i \rceil \leqslant \gamma_k \widetilde{\delta}_i + 1$, (5) becomes

$$\widetilde{\delta}_i \leqslant \left[\sum_{k=1}^{i-1} n_k \left(\gamma_k \widetilde{\delta}_i + 1 \right) + n_i - 1 \right] a + a, \tag{6}$$

then $\widetilde{\delta}_i$ is upper bounded from above by

$$\widetilde{\delta}_i \leqslant \frac{\sum_{k=1}^{i} n_k}{1/a - \sum_{k=1}^{i-1} n_k \gamma_k}. \tag{7}$$

Note that the right-hand side of (7) is δ_i'. This shows that $\widetilde{\delta}_i \leqslant \delta_i'$. Since condition $\delta_i' + a < 1/\gamma_i$ holds for all $i = 1, \ldots, N_{\mathrm{c}}$, each polling token of an ith class CBR connection must be used before its next arrival, i.e., there is at most one PT for each ith class CBR connection in the BS at any time. Then, the marked PT will be removed before the generation of the next PT. Since the maximum jitter cannot be larger than the maximum delay, the jitter for the CBR connection is upper bounded by δ_i'. □

Proof of theorem 2. The connections in each VBR class can only use the remaining bandwidth after serving all the CBR classes and the VBR classes with higher priorities. For a connection in the first VBR class (indexed by $i = N_{\mathrm{c}} + 1$), its minimum available bandwidth is the remaining bandwidth after serving all the CBR connections minus that consumed by the $(n_{N_{\mathrm{c}}+1} - 1)$ same-class connections. The maximum number of packets from the N_{c} CBR classes that can be served in an interval $(t_1, t_2]$ is at most $\sum_{k=1}^{N_{\mathrm{c}}} n_k \lceil \gamma_k (t_2 - t_1) \rceil$, which is upper bounded by $\sum_{k=1}^{N_{\mathrm{c}}} n_k [\gamma_k (t_2 - t_1) + 1]$. Adding a for non-preemptive strategy, the available bandwidth for

a connection in the first VBR class, $C_{N_c+1}(t_1, t_2)$, is lower bounded by

$$(t_2 - t_1) - a - a\left\{\sum_{k=1}^{N_c} n_k\big[\gamma_k(t_2 - t_1) + 1\big] + n_{N_c+1} - 1\right\} = (1 - a\widehat{\rho}_0)(t_2 - t_1) - a(\widehat{\sigma}_0 + n_{N_c+1}), \quad (8)$$

where

$$\widehat{\rho}_0 = \sum_{k=1}^{N_c} n_k \gamma_k \quad (9)$$

and

$$\widehat{\sigma}_0 = \sum_{k=1}^{N_c} n_k. \quad (10)$$

Therefore, $C_{N_c+1}(t_1, t_2)$ is $[a(\widehat{\sigma}_0 + n_{N_c+1}), (1 - a\widehat{\rho}_0)]$ lower constrained.

Every connection in the first VBR class is regulated by a $(\sigma_{N_c+1}, \rho_{N_c+1})$-LB, and the number of packets at the output of the LB is $(\sigma_{N_c+1} + 1, \rho_{N_c+1})$ upper constrained. Therefore, the backlog of a connection in the first VBR class is $B_{N_c+1}(t_1, t_2) \leqslant \rho_{N_c+1}(t_2 - t_1) + \sigma_{N_c+1} + 1$. The time required to clear the backlog is $T_{N_c+1}(t_1, t_2) = aB_{N_c+1}(t_1, t_2) \leqslant a\rho_{N_c+1}(t_2 - t_1) + a(\sigma_{N_c+1} + 1)$. This indicates that $T_{N_c+1}(t_1, t_2)$ is $[a(\sigma_{N_c+1} + 1), a\rho_{N_c+1}]$ upper constrained. If $1 - a\widehat{\rho}_0 > a\rho_{N_c+1}$, that is, the remaining bandwidth after serving all the CBR connections and $(n_{N_c+1} - 1)$ same-class VBR connections is larger than the current VBR transmission bandwidth, then the delay experienced by a connection in the first VBR class is upper bounded by [12]

$$d'_{N_c+1} = \frac{a(\widehat{\sigma}_0 + n_{N_c+1}) + a(\sigma_{N_c+1} + 1)}{1 - a\widehat{\rho}_0} = \frac{n_{N_c+1} + \sigma_{N_c+1} + 1 + \sum_{k=1}^{N_c} n_k}{1/a - \sum_{k=1}^{N_c} n_k \gamma_k}. \quad (11)$$

The argument for the delay of an ith $(N_c + 1 < i \leqslant N_c + N_v)$ class VBR connection is similar, except that the remaining bandwidth is the total bandwidth minus the bandwidth to serve all the CBR connections, all the connections in the $(i - 1 - N_c)$ higher priority VBR classes and the $(n_i - 1)$ same-class connections. Therefore, the remaining bandwidth to serve an ith class VBR connection, $C_i(t_1, t_2)$, is lower bounded by

$$(t_2 - t_1) - \left\{a\sum_{k=1}^{N_c} n_k\big[\gamma_k(t_2 - t_1) + 1\big] + a\sum_{k=N_c+1}^{i-1} V_k + a(n_i - 1) + a\right\}, \quad (12)$$

where V_k is the maximum number of packets from the kth VBR class that can be served in $(t_1, t_2]$. Let the maximum delay be upper bounded by d'_k for a kth class VBR connection, then we have

$$V_k = n_k\big[\rho_k(t_2 - t_1) + (\sigma_k + 1) + \rho_k d'_k\big]. \quad (13)$$

Table 1
Jitter bounds of CBR classes when $n_i = 5$, where $i = 1, \ldots, 5$.

(CBR) class index	Parameters (γ_i, δ_i)	Bound in [2] (δ_i^*)	Bound (δ'_i)	Simulation results
1	(0.05, 12)	5.25	5.25	4.20
2	(0.01, 60)	30.25	14.24	9.45
3	(0.0075, 80)	50.25	22.99	10.50
4	(0.0064, 100)	65.25	32.53	15.75
5	(0.0032, 200)	130.25	42.89	21.00

Substituting V_k in (12) by (13) and rearranging yields

$$C_i(t_1, t_2) \leqslant \left(1 - a\widehat{\rho}_0 - a\sum_{k=N_c+1}^{i-1} n_k\rho_k\right)(t_2 - t_1) - a\left[\sum_{k=N_c+1}^{i-1} n_k\big(\rho_k d'_k + \sigma_k + 1\big) + \widehat{\sigma}_0 + n_i\right]. \quad (14)$$

By using (4) we have

$$C_i(t_1, t_2) \leqslant \left(1 - a\widehat{\rho}_0 - a\sum_{k=N_c+1}^{i-1} n_k\rho_k\right)(t_2 - t_1) - a\left(\sum_{k=N_c+1}^{i-1} n_k\widehat{\sigma}_k + \widehat{\sigma}_0 + n_i\right). \quad (15)$$

Therefore, $C_i(t_1, t_2)$ is $[a(\sum_{k=N_c+1}^{i-1} n_k\widehat{\sigma}_k + \widehat{\sigma}_0 + n_i), 1 - a\widehat{\rho}_0 - a\sum_{k=N_c+1}^{i-1} n_k\rho_k]$ lower constrained. Analogous to the argument for the first VBR class, the backlog of a connection in the ith VBR class is $B_i(t_1, t_2) \leqslant \rho_i(t_2 - t_1) + \sigma_i + 1$. The time required to clear the backlog is $T_i(t_1, t_2) = aB_i(t_1, t_2) \leqslant a\rho_i(t_2 - t_1) + a(\sigma_i + 1)$. This indicates that $T_i(t_1, t_2)$ is $[a(\sigma_i + 1), a\rho_i]$ upper constrained. If $1 - a\widehat{\rho}_0 - a\sum_{k=N_c+1}^{i-1} n_k\rho_k > a\rho_i$, that is, the remaining bandwidth is larger than the current VBR transmission bandwidth, then the delay experienced by the ith class VBR connection is upper bounded by

$$d'_i = \frac{a[\sum_{k=N_c+1}^{i-1} n_k\widehat{\sigma}_k + \widehat{\sigma}_0 + n_i] + a(\sigma_i + 1)}{1 - a\widehat{\rho}_0 - a\sum_{k=N_c+1}^{i-1} n_k\rho_k} = \frac{n_i + \sigma_i + 1 + \sum_{k=N_c+1}^{i-1} n_k\widehat{\sigma}_k + \sum_{k=1}^{N_c} n_k}{1/a - (\sum_{k=1}^{N_c} n_k\gamma_k + \sum_{k=N_c+1}^{i-1} n_k\rho_k)}. \quad (16)$$

□

Similar jitter bounds for CBR traffic and delay bounds for VBR traffic are derived in [1,2]. Table 1 shows the jitter bounds for CBR traffic, where δ'_i is calculated from (1), and δ_i^* is the jitter bounds in [2]. Table 2 shows the delay bounds for VBR traffic, where d'_i is calculated from (3), and d_i^* is the delay bounds in [1]. The simulation results in tables 1 and 2 are based on the system model in section 2. The parameters used to obtain the jitter and delay bounds are listed as follows: the link speed of the wireless channel is 10 Mb/s, the data packets are 1 kbits in length, and the length of a polling message is

Table 2
Delay bounds of VBR classes when $n_i = 1$, where $i = 6, \ldots, 10$.

(VBR) class index	ON prob.	Parameters (ρ_i, σ_i, d_i)	Bound in [1] (d_i^*)	Bound (d_i')	Simulation results
6	0.62	(0.00196, 7, 1200)	646.52	59.98	31.50
7	0.57	(0.00183, 6, 1200)	665.36	72.79	32.55
8	0.55	(0.00177, 6, 1200)	684.23	85.69	33.60
9	0.55	(0.00177, 6, 1200)	703.25	98.77	34.65
10	0.53	(0.00168, 5, 1200)	720.25	109.71	35.70

50 bits. Time is normalized with respect to the time to transmit a packet which is 0.1 ms. In the simulation, each VBR connection is generated in the same way as that in [1]. In specific, it is assumed that there is an output from an ON–OFF coder with a code rate of 32 kb/s. There is a packet coming out from the coder every $1/32$ s when the coder is ON, and both the ON and OFF periods are $1/32$ s. The probability that a period is an ON period is listed in table 2. From tables 1 and 2 it can be seen that the QoS performance bounds given in (1) and (3) are much tighter than those in [2] and [1], respectively. In the proof of jitter bounds for CBR traffic in [2], condition $\lceil \gamma_k \delta_i^* \rceil < \lceil \gamma_k / \gamma_i \rceil$ is used, which makes the derived bounds very loose. The reason is that, in order to provide guaranteed QoS for all the connections, during the time interval after serving a packet from the ith class connection and before the arrival of its next packet, all the higher priority CBR connections must be served if their MS RTT Buffers are not empty. For a lower priority connection, there are more connections which have higher priorities than it. Therefore, $1/\gamma_i$ is much larger than δ_i^*, i.e., $\lceil \gamma_k \delta_i^* \rceil \ll \lceil \gamma_k / \gamma_i \rceil$ for $i > 1$ (and $i \leqslant N_c$). In contrast, our derived bounds are based on $\lceil \gamma_k \widetilde{\delta}_i \rceil$ which is limited by $\gamma_k \widetilde{\delta}_i + 1$ in equation (6), and thus are much tighter. Because the polling model in [1] places the LB for each VBR connection in each mobile, the BS has no information about whether there are packets or not from a VBR connection. Because of this, a ready-to-transmit packet from a VBR connection may have to wait for a longer time to be transmitted. The polling model used in this paper is from [2], where the LBs are placed at the BS, so the BS always knows whether there are packets to transmit or not from each VBR connection. The significance of the improvement in the derived bounds is that a CAC scheme by using the improved bounds for QoS provisioning may result in lower NCBR and HCDR and allow more connections in a system. The tables also show that the difference between the derived bounds and the simulated performance becomes large as the class priority decreases. The reason is due to error accumulation, because the jitter or delay bound for a particular class is based on the jitter or delay bounds of all the higher priority classes.

4. CAC with user mobility information

Three aspects are included in this CAC scheme: resource reservation for potential handoff connections based on user mobility information, admission control for a new connection, and admission control for a handoff connection. To admit a connection (new or handoff), two conditions must be satisfied: guaranteed QoS for the particular connection should be provided, and QoS performance of all the existing connections should still be satisfied after the admission. The QoS provisioning in the CAC is based on the derived jitter bounds for CBR traffic and delay bounds for VBR traffic. From the theorems in section 3 it can be seen that the QoS performance bound (jitter or delay) of each traffic class is affected by all the traffic classes with higher priorities, but not the classes with lower priorities. This indicates that admitting any connection only has impact on the QoS performance bounds of all the lower priority classes. The proposed CAC scheme is for CBR and VBR traffic. For ABR traffic, all new and handoff requests are treated the same, and no resource is reserved for potential ABR handoff connections.

4.1. Resource reservation for handoff CBR and VBR connections

Since user mobility information changes with time, the resource reservation process should be updated periodically based on the current user mobility information. To achieve this, system time is divided into equal intervals in length τ beginning at $t = 0, \tau, 2\tau, \ldots$. Smaller value of τ leads to frequent updating, while larger value of τ affects the accuracy of reservation. The value of τ is chosen so that the probability of having more than one handoff event for any mobile in any τ interval is negligible. It is assumed that, at the beginning of every τ interval, reasonably accurate handoff probabilities of each mobile from its currently serving cell to its neighboring cells are known to the mobile's current BS [7,8], and all the neighboring cells can exchange the probabilistic information with each other. All the reserved resources are shared by all the handoff connections to efficiently utilize the reserved resource. Without loss of generality, let cell o be the reference cell. Define p_{box} as the probability that a mobile x will handoff from its current cell b to cell o in the next τ interval. Then

$$h_i = \sum_{b \in B_o} \sum_{x \in \chi_{b,i}} p_{box}, \quad i = 1, \ldots, N_c + N_v, \tag{17}$$

is the accumulated handoff probability of all the mobiles carrying an ith class connection from the neighboring cells of cell o to cell o in the next τ interval, where B_o is the set of all the neighboring cells of cell o, and $\chi_{b,i}$ is the set of all the mobiles carrying an ith class connection in cell b. The amount of the reserved resources for potential ith class handoff connections in cell o is equal to the resources required by h_i ith class connections, and h_i is also defined as the equivalent number of potential handoff connections. At the beginning of every τ interval, resource reservation is updated at each BS as follows.

R1: Resource reservation starts from the first CBR class: `resvd_cls` $= 1$.

R2: Treat the potential handoff connections as existing connections: $n_{\texttt{resvd\_cls}} = n_{\texttt{resvd\_cls}} + h_{\texttt{resvd\_cls}}$.

R3: Let START_CLS = resvd_cls. Check if the system can reserve the required resource for class resvd_cls by performing pseudocode 1.

R4: If QOS_grt = 1, then $R_{\texttt{resvd\_cls}} = 1$, and resource reservation for class resvd_cls is successful. Otherwise, $R_{\texttt{resvd\_cls}} = 0$, $n_{\texttt{resvd\_cls}} = n_{\texttt{resvd\_cls}} - h_{\texttt{resvd\_cls}}$, and the reservation process is failed.

R5: If resvd_cls $= N_c + N_v$, then $n_i = n_i - R_i h_i$ for $i = 1, \ldots, N_c + N_v$, and the process ends. Otherwise, let resvd_cls = resvd_cls + 1 and go to step R2.

4.2. Admission decision for a new CBR or VBR connection

Suppose the new connection is an αth class connection. The following procedure is used to make the admission decision for the connection.

N1: Treat the new connection as an existing connection: $n_\alpha = n_\alpha + 1$.

N2: Treat the reserved resources as allocated resources: $n_i = n_i + R_i h_i$ for $i = 1, \ldots, N_c + N_v$.

N3: Let START_CLS $= \alpha$. Check if the connection and all the other connections in the system can receive guaranteed QoS by performing pseudocode 1.

N4: If QOS_grt $= 1$, then $n_i = n_i - R_i h_i$ for $i = 1, \ldots, N_c + N_v$, and the connection is accepted. Otherwise, $n_\alpha = n_\alpha - 1$, $n_i = n_i - R_i h_i$ for $i = 1, \ldots, N_c + N_v$, and the connection is rejected.

4.3. Admission decision for a handoff CBR or VBR connection

Suppose the handoff connection is an αth class connection, the following procedure is used to make the admission decision for the connection.

H1: Treat the handoff connection as an existing connection: $n_\alpha = n_\alpha + 1$.

H2: Let START_CLS $= \alpha$. Check if the connection and all the other connections in the system can receive guaranteed QoS by performing pseudocode 1.

H3: If QOS_grt $= 1$, then the connection is accepted. Otherwise, $n_\alpha = n_\alpha - 1$, and the connection is rejected.

Pseudocode 1.

```
LAST_flag = 1;
QOS_grt = 1;
/* Bound calculations start from the
   START_CLS/
current_cls = START_CLS;
while(QoS_grt & LAST_flag)
{
  if (current_cls <= Nc)
  /* current_cls is a CBR class */
  {
    Calculate the jitter bound DELTA of
       current_cls;
      if(DELTA > delta(current_cls))
        /* Cannot accept */
        QoS_grt = 0;
      end
  }
  else
  {
    Calculate the delay bound D of
     current_cls;
    if(D > d(current_cls))
      /* Cannot accept */
      QoS_grt = 0;
    end
  }
  end
  if (current_cls == Nc+Nv)
    LAST_flag = 0;
  else
    current_cls = current_cls + 1;
  end
}
```

5. Simulation results and discussions

Without loss of generality, we consider a one dimensional (1-D) cellular array as shown in figure 2, where 5 radio cells are arranged on a circle to avoid the boundary effect. The linear dimension of each cell is $D = 1500$ m. Mobile x can be in any one of the cells with equal probability. Let the initial location where mobile x originates its connection request be uniformly distributed in its current cell. The mobile can move in either direction of the 1-D region with equal probability, and its velocity is uniformly distributed between 5 m/s and 20 m/s. The time interval used to update the resource reservation is 60 s. The handoff probabilities of x from cell o to its two neighboring cells, cell r and cell l, p_{orx} and p_{olx}, respectively, are given by

$$p_{orx} = \frac{D'}{D}, \tag{18}$$

$$p_{olx} = 1 - \frac{D'}{D}, \tag{19}$$

where D' is the distance between the mobile's current location and the left edge of cell o.

In the simulation, it is assumed that the connection arrival process is a Poisson process and the connection duration follows an exponential distribution. The average connection arrival interval $1/\lambda$ varies from zero to 10 s, and the average connection duration $1/\mu$ is fixed at 50 s. Three CBR classes and three VBR classes are considered, and the parameters of each of the classes are given in table 3. For any particular connection, the traffic class to which it belongs is randomly chosen from the six classes with equal probability. For each VBR connection, its packets are generated in the same way as that in section 3. Two resource reservation approaches are used in the simulation: *full reservation* (FR) in which 100% resource is reserved for each connection in each of its neighboring cells, and *partial reservation* (PR) in which resource is

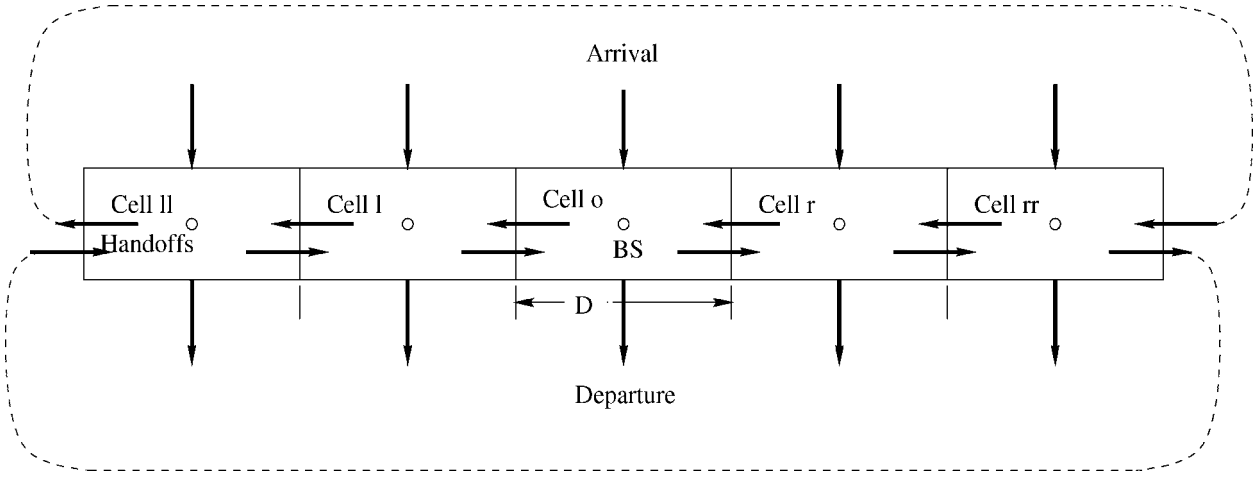

Figure 2. 1-D cell model.

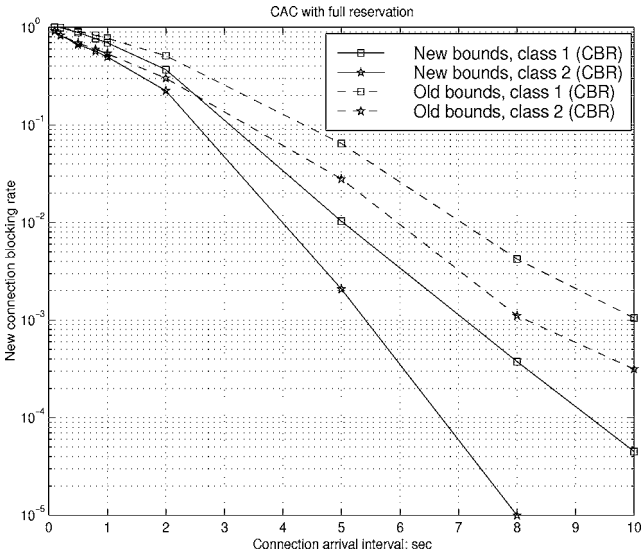

Figure 3. Comparison of different bounds: full reservation case.

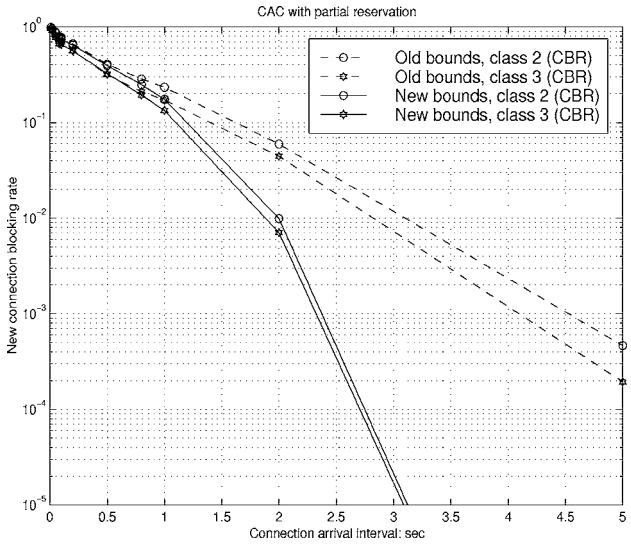

Figure 4. Comparison of different bounds in partial reservation case: new connection blocking rate.

Table 3
Traffic parameters for simulation (time unit: s).

(CBR) class index	Parameters (γ_i, δ_i)	(VBR) class index	ON prob.	Parameters (ρ_i, σ_i, d_i)
1	(500, 0.0012)	4	0.62	(19.6, 7, 0.12)
2	(100, 0.006)	5	0.57	(18.3, 6, 0.12)
3	(75, 0.008)	6	0.55	(17.7, 6, 0.12)

reserved by using mobility information for potential handoff connections. The performance of the CAC schemes using the tighter bounds and the bounds given in [1,2] is compared for both the reservation approaches.

Figures 3 and 4 show the NCBR for CBR classes in FR and PR cases, respectively. It can be seen that the CAC schemes by using the improved new jitter bounds to do the QoS provisioning can admit more new CBR connections in both FR and PR cases. Figure 5 indicates that, for the PR case, the CAC scheme using the improved jitter bounds can reduce the HCDR of both CBR and VBR connections, compared with that using the bounds in [2]. The HCDR in the FR case is

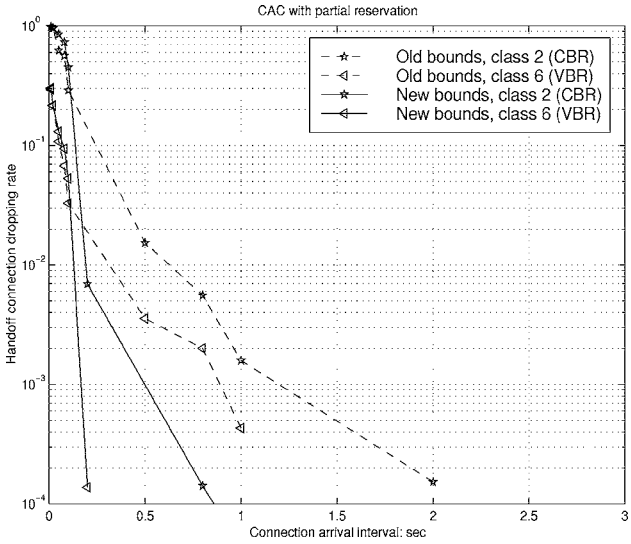

Figure 5. Comparison of different bounds in partial reservation case: handoff connection dropping rate.

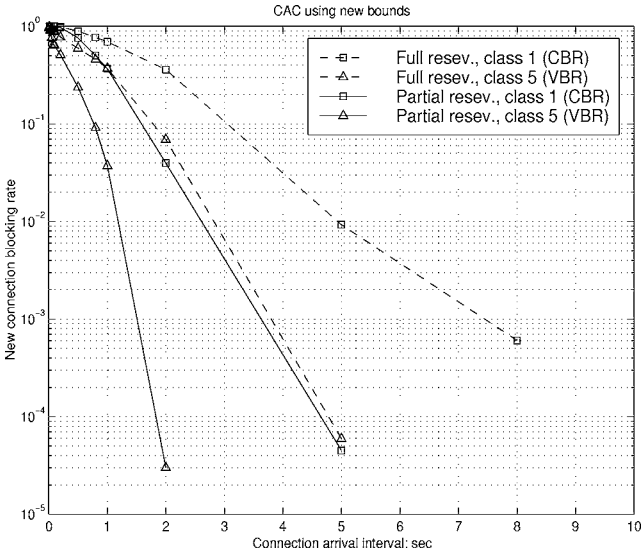

Figure 6. Comparison between different resource reservation schemes based on new bounds.

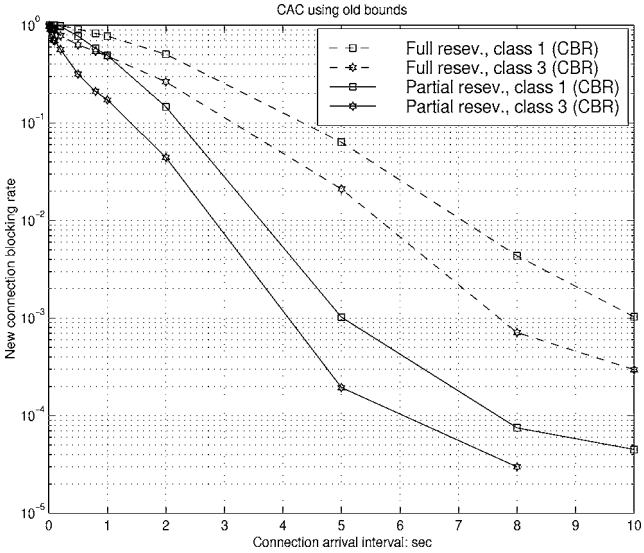

Figure 7. Comparison between different resource reservation schemes based on old bounds.

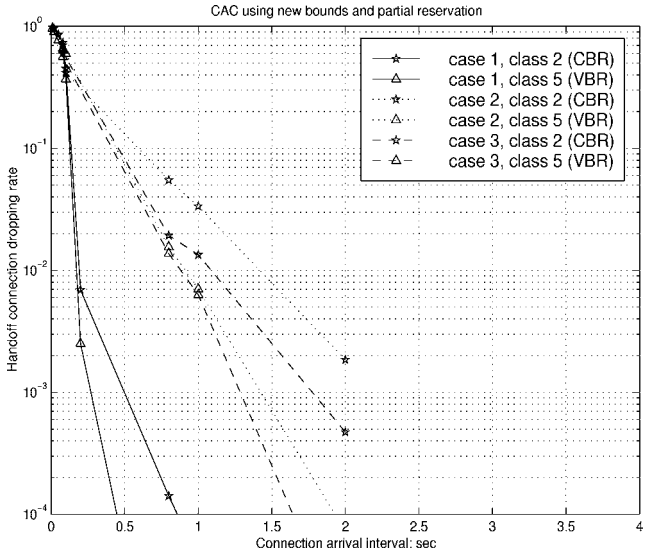

Figure 8. Comparison among different reserved resource sharing schemes: handoff connection dropping rate.

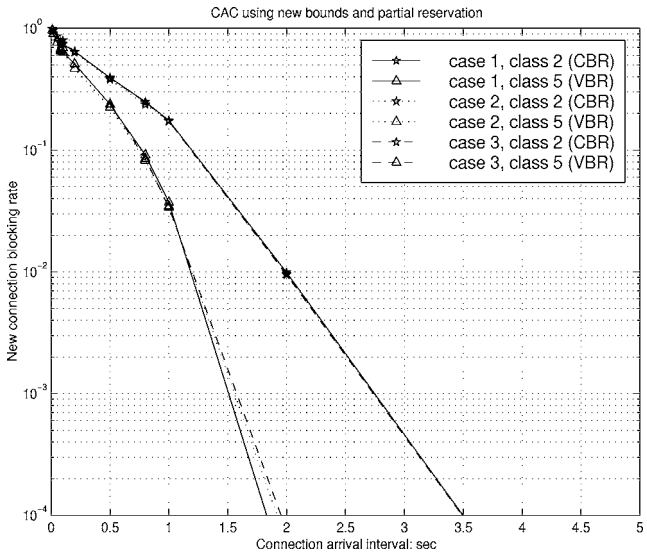

Figure 9. Comparison among different reserved resource sharing schemes: new connection blocking rate.

zero. Figures 6 and 7 show that the CAC with PR can achieve much lower NCBR and higher resource utilization (figure 6 is for the case using the improved bounds and figure 7 is for the case using the bounds in [2]) while keeping the HCDR at a very low level.

In the proposed CAC scheme, the reserved resources for all classes are completely shared by all handoff connections. To study the effect of the completely sharing on the performance of the CAC scheme, the following cases are studied. Case 1: when a handoff occurs, a handoff connection can use all the resources that the BS has reserved for potential handoff connections in all classes. Case 2: the handoff connection can only use the resources that the BS has reserved for the class to which it belongs. Case 3: the handoff connection can use the resources that the BS has reserved for the class to which it belongs and all the lower priority classes. Figure 8 shows that case 1 can achieve much less HCDR than the other two cases because all reserved resources are completely shared. Figure 9 shows that the NCBR values are almost the same for all three cases because the amount of reserved resources is almost the same.

The effect of cell size and mobile velocity on the performance of the proposed CAC scheme is also studied. Figure 10 shows the NCBR for different cell size and mobile velocity. It can be seen that the change in cell size and mobile velocity has no significant impact on the performance of the proposed CAC scheme. This can be explained as follows. In general, as the cell size decreases or the mobile velocity increases, the frequency of handoffs for the mobile users increases. Consequently, more resources are reserved for handoff connections. This results in fewer new connections to be admitted in the system. On the other hand, fewer admitted new connections

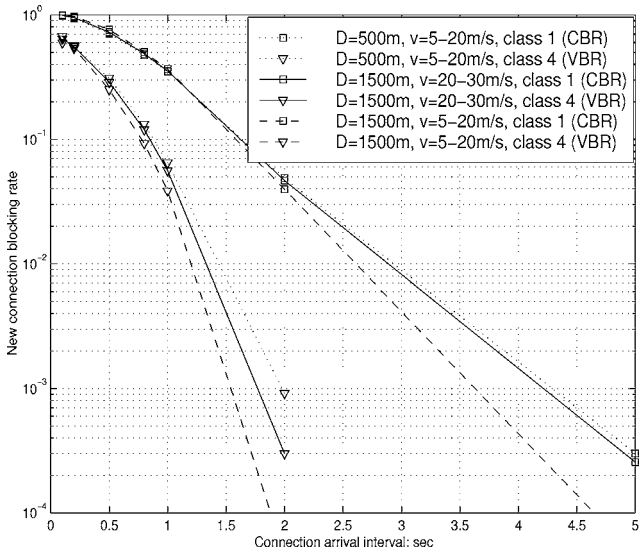

Figure 10. Effect of cell size and mobile velocity.

will result in fewer total number of ongoing connections in the system. Consequently, less resource needs to be reserved for handoff connections.

From figures 3–10 it can be seen that, when the other parameters are fixed, the HCDR/NCBR for a traffic class decreases as the traffic load decreases (connection arrival interval increases). It can also be seen that lower priority classes experience lower HCDR/NCBR compared to the higher priority classes, because the connections in lower priority classes require less system resource than that in higher priority classes.

6. Conclusions

We have developed an efficient CAC scheme with QoS provisioning based on user mobility information and tighter jitter and delay bounds. Since the proposed bounds are deterministic (the worst case) ones, it is anticipated that the performance of the CAC scheme could be further improved with the stochastic bounds. Research on finding the stochastic bounds is under way.

Acknowledgements

This work has been supported by a grant from the Communications and Information Technology of Ontario (CITO), Ontario, Canada.

References

[1] C.S. Chang, K.C. Chen, M.Y. You and J.F. Chang, Guaranteed quality-of-service wireless access to ATM networks, IEEE Journal on Selected Areas in Communications 15(1) (January 1997) 106–117.

[2] J.X. Qiu and J.W. Mark, Service scheduling and CAC for QoS guarantee in future PCS, in: *IEEE Global Telecommunications Conference*, Sydney, Australia (November 1998) pp. 2039–2044.

[3] Y.B. Lin, A. Noerpel and D. Harasty, A nonblocking channel assignment strategy for hand-offs, in: *IEEE ICUPC'94*, San Diego, CA (September 1994).

[4] A.S. Acampora and M. Naghshineh, An architecture and methodology for mobile-executed handoff in cellular ATM networks, IEEE Journal on Selected Areas in Communications 12(8) (October 1994) 1365–1374.

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

[6] D. Zhao, X. Shen and J.W. Mark, Improved QoS performance bounds for a wireless ATM network, in: *Fifth Asia-Pacific Conference on Communications*, Beijing, China (October 1999) CD-ROM file `(173)APOC_1.ps`.

[7] X. Shen and J.W. Mark, Mobility information for resource management in wireless ATM networks, Computer Networks 31 (1999) 1049–1062.

[8] T. Liu, P. Bahl and I. Chlamtac, Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks, IEEE Journal on Selected Areas in Communications 16(6) (August 1998) 922–936.

[9] Y. Shimizu and H. Sato, Proposal of flow and resource control schemes for ABR service in wireless ATM, in: *Proceedings of PIMRC'99*, Vol. 3 (1999) pp. 1237–1241.

[10] Y.H. Long, T.K. Ho and A.B. Rad, Explicit rate allocation algorithm of generalized max-min fairness for ATM ABR services, Electronics Letters 35(7) (1999) 530–531.

[11] G. Bianchi, L. Fratta and L. Musumeci, Congestion control algorithms for the ABR service in ATM networks, in: *Proceedings of GLOBECOM'96*, Vol. 2 (1996) pp. 1080–1084.

[12] R.L. Cruz, A calculus for network delay, Part I: Network elements in isolation, IEEE Transactions on Information Theory 37(1) (January 1991) 114–131.

Dongmei Zhao received a B.S. degree in electrical engineering from Northern Jiaotong University, Beijing, China, in 1992. Since 1999 she has been a Ph.D. student in the Department of Electrical and Computer Engineering at University of Waterloo, Waterloo, Ontario, Canada. Her research interests include QoS performance for heterogeneous services and radio resource management in wireless cellular networks.
E-mail: dzhao@bbcr.uwaterloo.ca

Xuemin Shen received the B.Sc. (1982) degree from Dalian Marine University (China) and the M.Sc. (1987) and Ph.D. degrees (1990) from Rutgers University, New Jersey (USA), all in electrical engineering. From September 1990 to September 1993, he was first with Howard University, Washington D.C., and then University of Alberta, Edmonton. Since October 1993, he has been with the Department of Electrical and Computer Engineering, University of Waterloo, where he is an Associate Professor. Dr. Shen's research focuses on control algorithm development for mobility and resource management in interconnected wireless/wireline networks. In specific, his interests are traffic flow control, connection admission and access control, handoff, user location estimation, end-to-end performance modeling and evaluation, voice over mobile IP, stochastic process and H_∞ filtering. He is the coauthor of the books *Singular Perturbed and Weakly Coupled Linear Systems – A Recursive Approach* (Springer-Verlag, 1990) and *Parallel Algorithms for Optimal Control of Large Scale Linear Systems* (Springer-Verlag, 1993).
E-mail: xshen@bbcr.uwaterloo.ca

Jon W. Mark received the B.S.Sc. degree from the University of Toronto, Toronto, Ontario, Canada, in 1962, and the M.Eng. and Ph.D. degrees from McMaster University, Hamilton, Ontario, Canada, in 1968 and 1970, respectively, all in electrical engineering. From 1962 to 1970, he was with Canadian Westinghouse Co. Ltd. in Hamilton, Ontario, Canada, where he was an engineer and then a senior engineer. He took leave of absence from Westinghouse in October 1968 to pursue Ph.D. studies at McMaster University under the auspices of an NRC PIER Fellowship. Since September 1970 he has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada, where he is currently a Professor. He was the Department Chairman from July 1984 to June 1990. He had been on sabbatical leaves at the IBM Thomas Watson Research Center, Yorktown Heights, NY (1976–1977), where he was a Visiting Research Scientist, AT&T Bell Labs., Murray Hill, NJ (1982–1983), where he was a Resident Consultant, Laboratoire MASI, Université Pierre et Marie Curie, Paris, France (1990–1991), where he was an Invited Professor, and the National University of Singapore (1994–1995 and in 1999), where he was a Visiting Professor. In 1996 he established the Centre for Wireless Communications at the University of Waterloo where he is currently serving as its founding Director. He had previously worked in the areas of sonar signal processing, adaptive equalization, image and video coding, spread spectrum communications, computer communication networks, ATM switch design and traffic management. His current research interests are in broadband communications, wireless communications and wireless/wireline interworking. Prof. Mark is an IEEE Fellow. He was a former editor of the IEEE Transactions on Communications. He is currently a member of the Inter-Society Steering Committee of the IEEE/ACM Transactions on Networking, an Editor of Wireless Networks, and an Associate Editor of Telecommunication Systems.

E-mail: jwmark@bbcr.uwaterloo.ca