A Rate-Based Borrowing Scheme for QoS Provisioning in Multimedia Wireless Networks *

Mona El-Kadi, Stephan Olariu, and Hussein Abdel-Wahab

Department of Computer Science, Old Dominion University, Norfolk, VA 23529-0162, U.S.A.

Abstract

Now that cellular networks are being called upon to support real-time interactive multimedia traffic, such as video tele-conferencing, these networks must be able to provide their users with quality-of-service (QoS) guarantees. Although the QoS provisioning problem arises in wireline networks as well, mobility of hosts and scarcity of bandwidth makes QoS provisioning a challenging task in wireless networks.

It has been noticed that multimedia applications can tolerate and gracefully adapt to transient fluctuations in the QoS that they receive from the network. The management of such adaptive multimedia applications is becoming a new research area in wireless networks. As it turns out, the additional flexibility afforded by the ability of multimedia applications to tolerate and adapt to transient changes in the QoS parameters can be exploited by protocol designers to significantly improve the overall performance of wireless systems.

The main contribution of this paper is to propose a novel, rate-based, borrowing scheme for QoS provisioning in high-speed cellular networks carrying multimedia traffic. Our scheme attempts to allocate the desired bandwidth to every multimedia connection originating in a cell or being handed off to the cell. The novelty of our scheme is that in case of insufficient bandwidth, in order not to deny service to requesting connections (new or handoff), bandwidth will be borrowed, on a temporary basis, from existing connections. Our borrowing scheme guarantees that no connection gives up more than its "fair share" of bandwidth, in the sense that the amount of bandwidth borrowed from a connection is proportional to its tolerance to bandwidth loss. Importantly, our scheme ensures that the borrowed bandwidth is promptly returned to the connections.

Extensive simulation results show that our rate-based QoS provisioning scheme outperforms the best previously-known schemes in terms of call dropping probability, call blocking probability, and bandwidth utilization.

Keywords: Bandwidth allocation, cellular networks, QoS provisioning, multimedia traffic, reservation schemes, rate-based schemes.

*Work supported by ONR grant N00014-97-1-0562.

1 Introduction

We are witnessing an unprecedented demand for wireless networks to support both data and real-time multimedia traffic. While best-effort service suffices for datagram traffic, the usability of real-time multimedia applications is vastly improved if the underlying network can provide adequate quality-of-service (QoS) guarantees. Admission control and bandwidth allocation schemes can offer wireline networks the ability to provide their users with such guarantees. Due to host mobility and notorious bandwidth limitations, the QoS provisioning problem is much more difficult in wireless networks. For example, a mobile host may be admitted into the network in a cell where its needs can easily be met, but the mobile host may eventually move to a cell that has little or no resources to offer. Since the user's itinerary and the availability of resources in various cells is usually not known in advance, global QoS guarantees are very hard to provide.

Admission control refers to the task of deciding if a connection should be admitted into, and supported by, the network. Admission control is necessary for real-time, continuous media connections since the amount of resources requested by these connections may not match the level of resources available at the time of connection setup. Admitting a connection into the network is tantamount to a contract between the network and the connection: on the one hand the network guarantees that a certain level of resources will be maintained for the duration of the connection. On the other hand, the connection is expected not to request additional resources over and above those negotiated at connection setup. The agreed-upon amount of resources that the network guarantees to a connection is commonly referred to as QoS. Traditional QoS parameters include bandwidth, end-to-end delay, and jitter. However, there are some QOS parameters that are specific to wireless networks.

It is typical in most admission schemes to deny service to a connection whose requests for resources cannot be met by the network. In such a case, the connection[1] is said to be *blocked*. In cellular networks, an important QoS parameter is the *call blocking probability* (CBP), denoting the chance that a new connection request will be denied admission into the network. A similar situation arises when an established connection in one cell attempts to migrate into a neighboring cell (i.e. a handoff is attempted). If the new cell cannot support the level of resources required by the connection, the handoff is denied and the connection is dropped. The *call dropping probability* (CDP) expresses the chance that an existing connection will be forcibly terminated during a handoff between cells due to a lack of resources in the target cell. The CBP and CDP together offer a good indication of a network's quality of service in the face of mobility.

The traditional admission control process outlined above is, in many cases, too conservative and pessimistic. Indeed, multimedia applications are known to be able to tolerate and adapt to transient

[1] We will follow common practice and refer to connections as "calls".

fluctuations in QoS [3, 4, 8]. This adaptation is typically achieved by the use of an adjustable-rate codec or by employing hierarchical encoding of voice and/or video streams [3, 4, 11, 12]. The codec, along with appropriate buffering before play-out, can allow applications to gracefully adapt to temporary bandwidth fluctuations with little or no perceived degradation in overall quality. The graceful adaptation of applications to transient fluctuations in QoS is fundamental in wireless networks, where QoS provisioning is a very challenging task. As we shall demonstrate in this paper, the additional flexibility afforded by this ability to adapt can be exploited by protocol designers to significantly improve the overall performance of wireless systems.

As we briefly mentioned, once a connection is admitted into the network, resources must be *allocated*, at the negotiated level, for the duration of the connection. It is important to realize that in a cellular network where the user may move through the network traversing a sequence of cells, this commitment cannot be only local to the cell in which the connection originated. If the connection is to be maintained after the user crosses the boundary between neighboring cells (i.e. after a handoff), the network must guarantee an appropriate level of resources in each new cell that the user traverses [1, 5, 7, 9]. Without detailed knowledge about the intended destination of each connection, honoring this commitment is an extremely difficult task [5, 7, 9]. To address this problem, many QoS provisioning schemes *reserve* resources in cells on behalf of mobile hosts in anticipation of their arrival. Not surprisingly, the resource reservation problem has recently received well-deserved attention. We refer the reader to [7] and [9] for surveys of recent literature.

There are, essentially, two approaches to resource reservation:

- fixed reservation – where a certain percentage of the available resources in a cell are permanently reserved for handoff connections, and
- statistical reservation – where resources are reserved using a heuristic approach. These approaches range from allocating the maximum of the resource requirements of all connections in neighboring cells, to reserving only a fraction of this amount [5, 7].

In this paper, we propose a novel, rate-based, borrowing scheme for QoS provisioning in high-speed cellular networks carrying multimedia traffic. Our scheme includes a fixed reservation pool for handoffs. At call setup time, the connections are expected to specify (1) their desired amount of bandwidth, and (2) the minimum amount of bandwidth needed to ensure an adequate level of quality. Our scheme attempts to allocate the desired bandwidth to every multimedia connection originating in a cell or being handed off to that cell. The novelty of our scheme is that in case of insufficient bandwidth, in order not to deny service to a requesting connection (new or handoff), bandwidth will be borrowed, on a temporary basis, from existing connections. Our borrowing scheme guarantees that no connection will give up more than its "fair share" of bandwidth, in the sense that the amount of bandwidth borrowed from a connection is proportional to its tolerance to

bandwidth loss.

There are four important points to note about our scheme: first, our scheme guarantees that the bandwidth allocated to a real-time connection never drops below the minimum bandwidth requirement specified by the connection at call setup time. This is very critical to ensuring that the corresponding application can still function at an acceptable level. Second, our scheme guarantees that if bandwidth is borrowed from a connection, it is borrowed in small increments, allowing time for application-level adaptation. Third, our borrowing scheme is *fair* in the sense that if bandwidth is borrowed from one connection, it is also borrowed from the existing connections. Specifically, if borrowing is necessary in order to accommodate a requesting connection (new or handoff), every existing connection will give up bandwidth in proportion to its tolerance to bandwidth loss. This motivated us to refer to our scheme as *rate-based*. Finally, the borrowed bandwidth is returned to the connections as soon as possible.

Extensive simulation results show that our rate-based QoS provisioning scheme outperforms the best previously-known schemes in terms of call dropping probability and call blocking probability. In addition, our scheme ensures a high bandwidth utilization in the cellular system.

The remainder of this work is organized as follows: Section 2 reviews relevant results from the literature; Section 3 discusses the details of our rate-based QoS provisioning scheme: Subsection 3.1 describes the assumed system parameters, Subsection 3.2 discusses the new call admission protocol, while Subsection 3.3 gives the details of the handoff management protocol. Section 4 gives a detailed description of our simulation model. The experimental results obtained from extensive simulations are presented in Section 5. Finally, Section 6 offers concluding remarks and points out directions for further work.

2 State of the art

In order to set the stage for our rate-based QoS provisioning scheme, we now briefly review the bandwidth allocation and reservation schemes proposed in [7]. We chose these schemes as a benchmark since they are arguably better than other comparable bandwidth allocation and reservation schemes found in the literature [7].

The traffic offered to the cellular system is assumed to belong to two classes:

1. Class I traffic – real-time multimedia traffic, such as interactive voice and video applications,
2. Class II traffic – non real-time data traffic, such as email or ftp.

When a mobile host requests a new connection in a given cell, it provides the following parameters:

- the desired class of traffic (either I or II),

- the desired amount of bandwidth for the connection, and
- the minimum acceptable amount of bandwidth, that is the smallest amount of bandwidth that the source requires in order to maintain acceptable quality, e.g. the smallest encoding rate of its codec.

One of the significant features of the admission control and bandwidth reservation schemes in [7] is that in order to admit the connection, bandwidth must be allocated in the originating cell and, at the same time, bandwidth must be reserved for the connection in all the neighboring cells. Specifically, for a new connection to be admitted in a cell, the cell must be able to allocate the connection its desired bandwidth. For Class I connections, the call will be blocked unless the desired bandwidth can be allocated to it in the original cell, and some bandwidth can be reserved for it in each of its six neighboring cells.

During a handoff, an established Class I connection is dropped if its minimum bandwidth requirement cannot be met in the new cell or if appropriate reservations cannot be made on its behalf in the new set of neighboring cells. However, Class II traffic has no minimum bandwidth requirement in the case of a handoff, and a call will be continued if there is any free bandwidth available in the new cell.

Numerous approaches for reserving bandwidth have been reported in the literature [1, 2, 4, 5, 6, 7, 9]. The schemes presented in [7] use statistical reservation techniques based on the number of connections in neighboring cells, the size of the connections in neighboring cells, the predicted movement of mobile hosts, and combinations of these factors. It is worth noting that the reservation schemes in [7] keep the dropping probability for Class I connections very low, since the mobile host should find bandwidth reserved for it, regardless of the cell to which it moves. But bandwidth may be wasted in the neighboring cells (the host can only move into one neighbor), and the blocking probability in those cells may increase because unused bandwidth is being kept in reserve. In general, the schemes described in [7] favor minimizing the CDP at the expense of the CBP and give Class I traffic precedence over Class II traffic.

3 The rate-based borrowing scheme

It is clear that keeping a small pool of bandwidth always reserved for handoffs, as in [7], yields low CDP. However, in our scheme, the size of the reserved pool is not determined by requests from neighboring cells, but is *fixed* at a certain percentage of the total amount of bandwidth available in the cell. We found that this produced results similar to the best results reported in [7], without the overhead of communication between neighboring base stations to request and release reservations. To further reduce the CDP in our scheme, we treat the reserved pool very carefully. We do not allow bandwidth from the reserved pool to be allocated to incoming handoffs unless the bandwidth is

needed to meet the minimum bandwidth requirements of the connection. Like [7], our scheme gives precedence to Class I connections; Class II traffic does not make use of the reserved bandwidth. In order to lower the call blocking probability as well as the dropping probability, our scheme allows for borrowing resources (i.e. bandwidth) from existing connections. Our borrowing strategy has the following interesting features:

1. No Class I connection will ever have to give up bandwidth beyond the minimum level negotiated at call setup time.
2. If the cell does not have enough residual bandwidth to accommodate an incoming call, the existing connections will temporarily have to give up a certain amount of bandwidth (see Subsection 3.2 for details).
3. If bandwidth must be borrowed, it is borrowed gradually, in small increments, to allow time for application-level adaptation.
4. As soon as bandwidth becomes available due to a terminating call or to a mobile host leaving the cell, the borrowed bandwidth will be returned to the connections.
5. Our scheme is fair, in the sense that if bandwidth is borrowed, all connections will give up an amount of bandwidth proportional to their tolerance to bandwidth loss.
6. Our scheme only requires minimal computational overhead and no communication overhead.

3.1 Cell and connection parameters

Each cell maintains a pool of bandwidth reserved for Class I handoffs which, initially, represents r percent of the total bandwidth. At setup time, each connection specifies to the cell in which it originates a *maximum bandwidth* M (termed the *desired bandwidth*) and a *minimum bandwidth* m. The difference between these two values is the *bandwidth_loss tolerance* (BLT) of the connection. Thus, $BLT = M - m$. We note that for constant bit rate (CBR) connections $M = m$, indicating no bandwidth_loss tolerance and, thus, $BLT = 0$.

The cell maintains a parameter, f, $(0 \leq f \leq 1)$, which represents the fraction of the BLT that a connection may have to give up, in the worst case. This fraction is the *actual borrowable bandwidth* (ABB) of the connection. Thus,

$$ABB = f \times BLT = f(M - m).$$

By accepting a new call, the cell agrees that the supplied bandwidth will not fall below a certain level that we call the *minimum expected* (MEX) bandwidth. By definition, $MEX = M - ABB$. It is worth noting that $MEX \geq m$. Simple computation shows that

$$MEX = (1 - f)M + fm. \tag{1}$$

To prevent borrowing from producing noticeable changes in a connection's QoS, we introduce another network parameter, λ. The ABB is divided into λ *shares*, each share being equal to $\frac{M-m}{\lambda}$. This provides the basis for a method of borrowing bandwidth gradually from a set of connections whose allocated resources may be quite different.

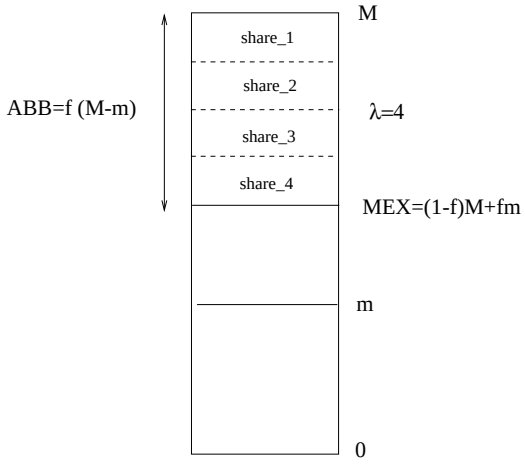

Figure 1: *Illustrating the main connection parameters*

A cell is said to be operating at level L, $(0 \leq L \leq \lambda)$, when all its ongoing connections have had L (or more) shares borrowed from them. Observe that for any connection the ratio between the amount of bandwidth given up and its bandwidth_loss tolerance is a *constant*. Specifically, we have

$$\frac{\text{amount given up}}{\text{bandwidth\_loss tolerance}} = \frac{\frac{Lf}{\lambda} \times (M-m)}{M-m} = \frac{Lf}{\lambda}, \tag{2}$$

which is a constant independent of the connection parameters. This is the sense in which we consider our borrowing scheme to be fair.

We note, however, that it is possible for a connection to be missing more than L shares after a handoff, due to the sacrifices made to prevent call dropping. However, our scheme attempts to restore bandwidth to handoff connections as soon as it becomes available. Figure 1 illustrates the concepts discussed above: this particular connection is in a cell with $f = \frac{2}{3}$ and $\lambda = 4$. We refer the reader to Appendix A for the pseudocode of the initialization protocol needed to begin QoS negotiations.

3.2 New call admission protocol

When a new call requests admission into the network in a cell operating at level L, the cell first attempts to provide the connection with an amount of bandwidth equal to its desired bandwidth minus L shares of its ABB, that is

$$M - \frac{ABB \times L}{\lambda} = \left(1 - \frac{Lf}{\lambda}\right) M + \frac{Lf}{\lambda} m. \tag{3}$$

If the amount of bandwidth specified in (3) exceeds the amount of bandwidth available, the cell tests to see if the call could be admitted if the cell progressed to level $L + 1$. If transition to level $L + 1$ will provide enough bandwidth to admit the call, the bandwidth is borrowed, the level is incremented, and the call is admitted; otherwise, the call is blocked. When the cell is operating at level $L = \lambda$, no more borrowing is allowed. It is important to note that our scheme never borrows from CBR connections or from connections that have already lost more than L shares.

Every time bandwidth becomes available in a cell due to a connection releasing its bandwidth allocation, the cell will attempt to make a transition to the next lower level. As a result, the available bandwidth is returned to the connections that have lost bandwidth due to borrowing. All fluctuations in a connection's allocated bandwidth are gradual as only one share can be borrowed or returned at a time. We refer the reader to Appendix B for the pseudo-code for new call admission.

3.3 Handoff management

The handoff admission policies differentiate between Class I and Class II connections. The reserved bandwidth is used only for Class I connections, which are admitted only if their minimum bandwidth needs can be met. When a Class I connection requests admission into a cell as a handoff, the cell checks to see if the minimum bandwidth requirement can be met with the sum of the available free and reserved bandwidth in the cell. If such is the case, the call is admitted into the cell and given bandwidth from the free bandwidth up to its desired level minus L shares. The connection is given bandwidth from the reserved bandwidth pool only if it is needed to reach its minimum requirement. If the minimum cannot be met using the free and reserved bandwidth, the cell tests to see if scaling to level $L + 1$ would free up enough bandwidth to admit the call. If so, the cell scales the other calls in the cell and provides the handoff call with bandwidth according to the guidelines described above.

On the other hand, Class II traffic will only be dropped if there is no free bandwidth left in the cell at all. The reserved pool is not available to these connections, because, as in [7], we assume that Class II traffic is able and willing to incur a possibly substantial fluctuation in service rather than be disconnected. Calls that have suffered a lowering of bandwidth due to a handoff will eventually be brought back to a reasonable level as their new cell has free bandwidth to give them. This is in contrast to the schemes presented in [7], which have no facility to improve connections which have been degraded due to a handoff. We refer the reader to Appendix C for the pseudocode for handoff management.

4 Simulation model

In order to evaluate the performance of our rate-based borrowing scheme, we implemented and simulated two other schemes for comparison. First, we implemented a request-based statistical reservation scheme from [7], termed the uniform and bandwidth-based model. According to this scheme, when reservations are made on behalf of a connection in neighboring cells, an equal amount of bandwidth is reserved in each neighboring cell, with no consideration of the most likely cell to which the host might travel. A cell does not reserve the sum of all the bandwidth it is asked to reserve, but just the largest of all the current requests.

We also simulated a simple scheme that reserves 5% of the total bandwidth in each cell for handoffs. New calls are admitted into the network if their desired bandwidth can be met, otherwise they are blocked. Class I handoffs are admitted if at least their minimum bandwidth requirements can be met. They are only given enough bandwidth from the reserved pool to meet their minimum, if there is too little free bandwidth available. Class II handoffs are admitted if there is any free bandwidth in the cell.

To simulate our rate-based borrowing scheme, we used a fixed reservation pool representing 5% of the total bandwidth. We set f to 0.5, thus permitting borrowing up to half of the bandwidth_loss tolerance. And we set λ to 10, so that each call had 10 shares to give.

To fairly contrast our scheme to the one in [7], we used the traffic types and characteristics given in [7], and modeled traffic behavior just as described there, with the exception of the handoffs. In [7], a handoff would occur during a connection with some given probability, and that probability would decrease exponentially with each successive handoff during the connection. We chose a different approach that seemed more realistic. We gave each mobile host a speed characteristic specifying the amount of time that will be spent in each cell during a call. Thus, longer calls are likely to experience more handoffs than shorter ones. Even with this minor change our results for the scheme from [7] correspond very closely to the results given there.

Table 1 shows the exact characteristics of the traffic used in our model. Each of the six types occurs with equal probability. For the results discussed in the following section, the speed was set to a host spending from 1 to 15 minutes in a cell, with an average of 5 minutes per cell. Each cell has 30Mbps of bandwidth. The network is a hexagonal grid of size 6×6 consisting of 36 cells. Traffic is provided to each cell at the level being measured. If a host moves out of the 6×6 grid, it is as though the connection ended normally – hosts do not "bounce" back into the network.

5 Experimental Results

Figure 2 compares the values of bandwidth utilization for the request-based reservation scheme from [7], for a fixed reservation scheme with $r = 5\%$, and for our rate-based borrowing scheme

CLASS	AVG BPS	MIN BPS	MAX BPS	AVG CALL	MIN CALL	MAX CALL
Class I	30Kbps	30Kbps	30Kbps	180s	60s	600s
Class I	256Kbps	256Kbps	256Kbps	300s	60s	1800s
Class I	3000Kbps	1000Kbps	6000Kbps	600s	300s	18000s
Class II	10Kbps	5Kbps	20Kbps	30s	10s	120s
Class II	256Kbps	64Kbps	512Kbps	180s	30s	36000s
Class II	5000Kbps	1000Kbps	10000Kbps	120s	30s	1200s

Table 1: *Traffic characteristics for our simulation model*

with $r = 5\%$, $\lambda = 10$ and $f = 0.5$, so that at most half of a call's bandwidth_loss tolerance can be borrowed. For the fixed reservation scheme and the rate-based borrowing scheme, at the maximum connection rate, the bandwidth utilization comes close to equaling the bandwidth outside of the reserved pool. The results for the request-based reservation scheme are worse than for the other two, because we did not implement a cap on the size of the reserved pool.

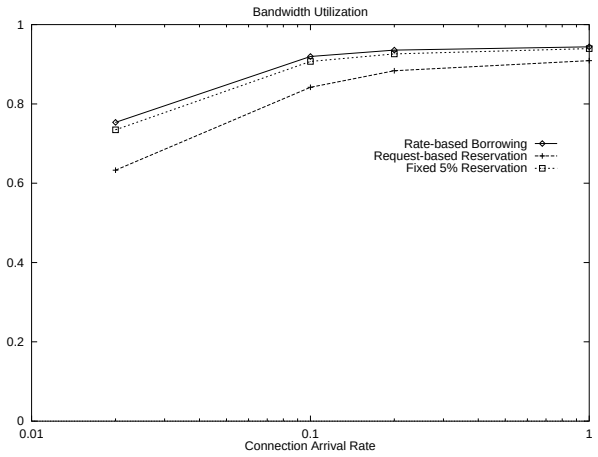

Figure 2: *A comparison of bandwidth utilization by the three schemes*

Figures 3 and 4 show, respectively, the CDP for Class I traffic alone and for Class I and II traffic combined. The borrowing scheme outperforms the other two schemes in both cases. In fact, the dropping probability for Class I connections is very close to zero. The motivation, of course, for favoring Class I connections by giving them exclusive use of the handoff reserves, is that real-time connections would suffer an actual loss by being dropped. We assume that a Class II application, although inconvenienced by being dropped, would be able to resume its transmission at a later time, without any significant loss. Despite this, Class II traffic fares significantly better under our rate-based borrowing scheme than under the others; it is especially important that our scheme

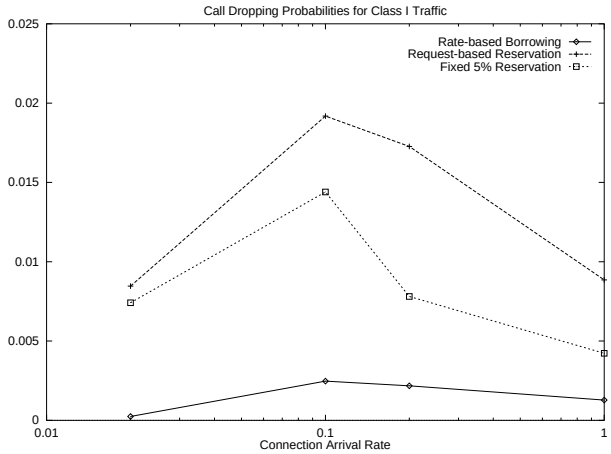

Figure 3: *Illustrating call dropping probabilities for Class I traffic*

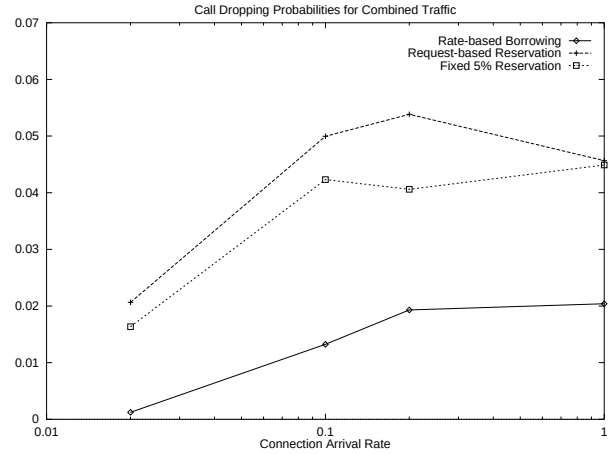

Figure 4: *Illustrating call dropping probabilities for Class I and Class II traffic combined*

returns bandwidth to connections who have suffered cuts during a handoff. The values chosen for r, f and λ do have a marked impact on the results. Some of our future research will involve finding optimal values for these parameters, understanding how they relate to each other and to the QoS parameters, and determining whether they can be adjusted dynamically to further increase network performance.

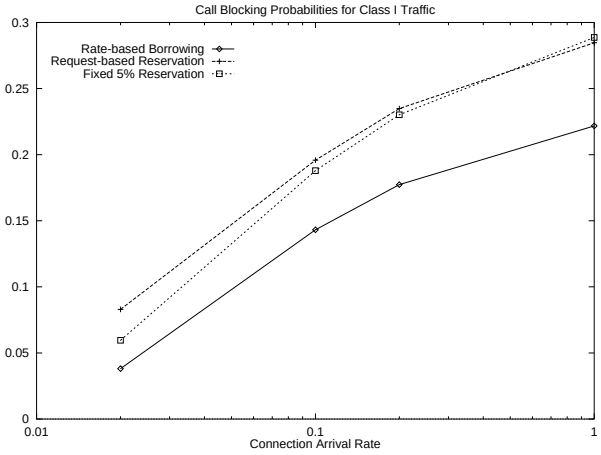

Figure 5: *Illustrating call blocking probabilities for Class I traffic*

Next, Figures 5 and 6 illustrate, respectively, the call blocking probabilities for Class I traffic alone and for Class I and II traffic combined. They demonstrate how borrowing allows a significant improvement in the CBP while also improving the dropping probability. As with CDP, the combined traffic also fares worse than Class I traffic alone in terms of CBP. However this is not due to any bias in the algorithms, but rather to the characteristics of the traffic being simulated. The Class II traffic requires more bandwidth on average.

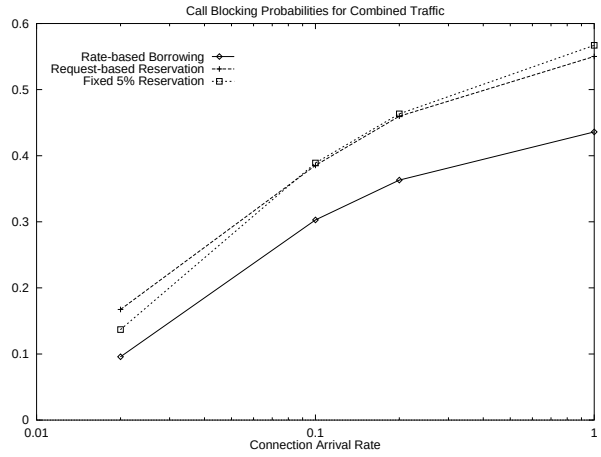

Figure 6: *Illustrating call blocking probabilities for Class I and Class II traffic combined*

6 Concluding remarks and directions for future work

Since multimedia traffic is intended mainly for human consumption [8], and since human senses are most forgiving, multimedia applications can tolerate and gracefully adapt to transient fluctuations in the QoS that they receive from the network. We have demonstrated that the additional flexibility afforded by the ability of multimedia applications to tolerate and adapt to transient changes in the QoS parameters can be exploited by protocol designers to significantly improve the overall performance of wireless systems.

Our main contribution is a novel, rate-based, borrowing scheme for QoS provisioning in high-speed cellular networks carrying multimedia traffic. To the largest extent possible, our scheme attempts to allocate the desired bandwidth to every multimedia connection originating in a cell or being handed off to that cell. The novelty of our scheme resides in the fact that, in case of insufficient bandwidth, in order not to deny service to requesting connections (new or handoff), bandwidth is borrowed, on a temporary basis, from existing connections.

One important characteristic of our rate-based borrowing scheme is that no connection gives up more than its "fair share" of bandwidth, in the sense that the amount of bandwidth borrowed is proportional to the connection's tolerance to bandwidth loss. Importantly, our scheme ensures that the borrowed bandwidth is returned promptly to the connections.

Extensive simulation results reveal that our scheme features very low call dropping probability, low call blocking probability, good bandwidth utilization, and reasonable success with keeping both classes of connections operating steadily near their desired bandwidth.

But our successes do not come without a price. Bandwidth borrowing subjects connections to possibly frequent fluctuations in the amount of bandwidth they are provided. In the two comparison schemes, the only fluctuation in the bandwidth provided would occur due to a handoff. In a simulation run at a rate of one connection per second, we found that the bandwidth supplied to a connection fluctuated an average of once every 10 seconds. It is clear that this issue requires more research.

Bandwidth borrowing decreases the probability that calls will always be provided their desired amount of bandwidth. In simulation we noticed that the calls that lost bandwidth during a handoff were always steadily replenished; however, the network wide average bandwidth provided to each call was about 85% of the desired amount. In the case of gracefully adaptive multimedia applications it makes sense to introduce a novel QoS parameter that specifies, for each connection, the fraction of time that it can tolerate operating below its expected bandwidth level. We are working on incorporating such a QoS parameter into cellular networks. This promises to be an exciting area for further research.

References

[1] A. Acampora and M. Naghshineh, Control and Quality-of-Service provisioning in high-speed microcellular networks, *IEEE Personal Communications*, Vol. 1, Second Quarter 1994.

[2] P. Agrawal, D. K. Anvekar, and B. Narendran, Channel management policies for handovers in cellular networks, *Bell Labs Technical Journal*, 1 (1996), 96–109.

[3] S. Chen and K. Nahrstedt, Distributed Quality-of-Service routing in ad-hoc networks, *IEEE J. Select. Areas in Communications*, 17, (1999), 1488–1505.

[4] H. Kanakis, P. P. Mishra, and A. Reibman, An adaptive congestion control scheme for real-time video packet transport, *IEEE/ACM Transactions on Networking*, 3, (1996),

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

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

[7] C. Oliviera, J. Kim and T. Suda, An Adaptive Bandwidth Reservation Scheme for High Speed Multimedia Wireless Networks, *IEEE J. Select. Areas in Communications*, 16, (1998), 858–874.

[8] S. V. Raghavan and S. K. Tripathy, *Networked Multimedia Systems*, Prentice-Hall, 1998.

[9] H. G. Perros, K. M. Elsayyed, Call admission control schemes: A review, *IEEE Communications Magazine*, (1996), 82–91.

[10] L. Trajković and A. Neidhardt, Effect of traffic knowledge on the efficiency of admission-control policies, *ACM Computer Communication Review*, (1999), 5–34.

[11] N. Tran and K. Nahrstedt, Adaptive adaptation by program delegation in VOD, *Proc. Int. Conf. Multimedia Computing and Systems*, 1998, 96–107.

[12] B. J. Vickers, M. Lee, and T. Suda, Feedback control mechanism for real-time multipoint video services, *IEEE J. Select. Areas in Communications*, 15, (1997).

Appendix

A Initialization – pseudocode

```
Protocol Initialize
  call_ABB = (call_desired_bw - call_min_bw) * f;
  // Find the size of a share for this call
  call_share = call_ABB/λ;
  // Scale down the desired bandwidth to the operating level L of the cell
  call_scaled_bw = call_desired_bw - L * call_share;
```

B New call admission – pseudocode

```
Protocol Admit_new_connection
  // Total the values of one share of each existing connection in the cell
  one_level_bw = sum_existing_calls();
  free_bw = total_bw - bw_used - bw_reserved;
  if ( call_scaled_bw ≤ free_bw )
    call is ACCEPTED;
  else if (L < λ AND call_scaled_bw - call_share ≤ free_bw + one_level_bw)
    {
      //if the cell is not yet at its last level
      //AND the call will fit after a round of borrowing
      call is ACCEPTED;
      // scale down
      scale_down_existing_connections(); L = L+1;
      call_scaled_bw = call_scaled_bw - call_share;
    }
  else
    call is BLOCKED
  if ( call is accepted )
    call_granted_bw = call_scaled_bw;
```

C Handoff management – pseudocode

```
Protocol Admit_handoff_connection
  one_level_bw = sum_existing_calls();
  free_bw = total_bw - bw_used - bw_reserved;
  if ( call is CLASS I )
  {
    if ( call_min_bw ≤ free_bw + reserved_bw )
      //if cell can provide its minimum bandwidth using the free and reserved bandwidth
      call is ACCEPTED;
    else if (L < λ AND call_min_bw ≤ free_bw + reserved_bw + one_level_bw )
    {
      // if the cell is not yet at its last level
      // AND the call will fit after a round of borrowing
      call is ACCEPTED;
      scale_down_existing_connections(); L = L + 1;
      free_bw = total_bw - bw_used - bw_reserved; // recompute the free bandwidth
      call_scaled_bw = call_scaled_bw - call_share;
    }
    else call is DROPPED;
    if ( call was ACCEPTED )
    {
      call_granted_bw = MIN( call_scaled_bw, free_bw + reserved_bw );
      if ( call_granted_bw > free_bw );
        call_granted_bw = MAX( free_bw, call_min_bw );
    }
  }
  else // call is CLASS II
  {
    if ( free_bw > 0 ) accept_call
    else if (L < λ AND free_bw + one_level_bw > 0 )
    {
      scale_down_existing_connections(); L = L + 1;
      free_bw = total_bw - bw_used - bw_reserved;
      call_scaled_bw = call_scaled_bw - call_share;
    }
    else call is DROPPED
    if ( call was ACCEPTED )
      call_granted_bw = MIN( call_scaled_bw, free_bw );
  }
```