Dynamic QoS Provisioning in Wireless Data Networks

Samrat Ganguly
ganguly@cs.rutgers.edu

Dragos Niculescu
dnicules@cs.rutgers.edu

Brett Vickers
bvickers@cs.rutgers.edu

April 12, 2001

Abstract

Providing quality of service(QoS) guarantees in mobile data networks is an inherently challenging task. Mobility of users imposes a spatial demand on resources resulting in overloaded regions that are entirely dependent on mobility pattern of users, that is often unpredictable. Prior and ongoing work in this area of QoS relies on call admission control (CAC), or careful resource allocation based on mobility prediction. The latter approach is not scalable, as it requires constant monitoring of the mobility of individual users and per-user state. Furthermore, many of the CAC schemes assume random or uniform mobility patterns, and in most cased are based on local decisions. The challenge is to design a scalable scheme that can provide QoS under distinct mobility patterns.

In contrast to the standard notion of QoS which is based on the handoff dropping probability, another important notion of QoS is based on disallowing handoff drops but minimizing the cell congestion probability that may occur in a given cell. In this paper, we explore this notion of QoS by proposing a dynamic CAC scheme that uses an dynamically estimated mobility pattern and distribution of users in different cells to reach an admission decision with the objective of minimizing cell congestion probability. This scheme does not maintain per user state, and can be implemented in a distributed fashion. From simulation results, we show that across different mobility patterns, the proposed scheme performs better than existing schemes in terms of achieved QoS while providing the minimum level of overall target utilization.

1 Introduction

With the growth of data service in (GPRS, 3G, 4G) wireless networks, the need for better Quality of Service(QoS) provisioning to support data applications has received a lot of importance. Future wireless networks are evolving towards supporting a broad spectrum of data applications. One side of the spectrum are real-time applications needing strong bandwidth guarantees. The other side of the spectrum are applications which are adaptive in nature and can operate over a wide range of bandwidth. In a cellular network, when a user hand-offs to a new cell, there may not be sufficient available channels to support his bandwidth requirements. Under such a crisis, there are two possibilities based on the nature of the application. In the case of a non-adaptive application needing strict bandwidth guarantees, the call will be dropped, whereas in the case of adaptive applications, the call will not be dropped but will suffer bandwidth degradation. Most of the earlier work in the area of QoS provisioning is focused on supporting voice calls which falls under the case of non-adaptive applications. Primary goal of such work was to provide better QoS by reducing the probability of calls being dropped at hand-offs. In contrast, goal of our work is to provide better QoS to the adaptive applications by reducing the level of bandwidth degradation. It is important to note that one can achieve any desired level of QoS by sacrificing the total utilization. In that respect, the objective of our work is not to find the best trade-off between resource utilization and

QoS attained, but to find out what is the best level of QoS can be provided under a given level of utilization. It is entirely due to the mobility of users that certain cells get congested thus affecting the level QoS perceived by the user. Therefore the direction of this work is to find out how to avoid cells getting congested. This requires understanding the effect of mobility on perceived QoS and subsequently use this knowledge efficiently to admit new calls. Our understanding is that existing mobility pattern and spatial population distribution in cells are the main cause of congestion.

In this paper, we describe our call admission scheme which uses estimation of mobility patterns and spatial population distributions to admit new calls. Since in real life, both mobility pattern and the population distribution may change with time, our proposed admission scheme is designed to adapt to such changes dynamically. Another important property of the proposed scheme is that it attains efficiency in term of utilization by selectively admitting calls which do not lead to congestion in remote cells. One of the significant departure of this scheme from most of the earlier schemes is that they were based on making an admission decision relying on information from neighboring cells [6, 5]. The proposed dynamic scheme, being based on the current estimate of mobility pattern, queries dynamically defined disjoint regions when making the admission decision.

The remainder of the paper is organized as follows. In Sections 2 and 3 we discuss the related work in the area of QoS provisioning which led to both the motivation and focus of our work. In Section 4, we describe our proposed dynamic QoS provisioning scheme. In section 5 we present simulation results to examine and compare the performance of our scheme. Finally, in section 5, we provide conclusions about this work along with comments about future work in this direction.

2 Related Work

The general problem of call admission aims to reconcile two opposite goals: maintaining a high utilization for the network as a cost maximization objective and providing a high QoS. The problem of provisioning bandwidth in a wireless network has been studied mostly for voice networks, and only recently for data networks.

Majority of the existing call admission schemes are based on using *guard channels*. Such schemes involve reserving a certain amount of bandwidth for hand-off calls in each base station corresponding to a cell. New calls and hand-off calls are therefore served from two separated, independent pools of bandwidth resources. [6] showed for a simplified queuing model of a single cell that "guard channel" type policies are optimal. The proof however, only holds for networks where all cells are behaving in the same manner and are sustaining the same incoming load, that is, for uniform models. Admission decision based on manipulating the guard channels are denoted as *cell-based* in this paper. [1] develops the *region-based* call admission scheme, which is an extension of the *cell-based* scheme in that it limits the number of new calls to a fixed fraction of the capacity of an entire region. This achieves a guaranteed utilization for the entire region, but fails to account for possible traffic patterns inside the region. Distributed call admission [5] admits a new call in a cell based on a probabilistic prediction of the future state of the cell and that of its neighbors. All these algorithms are mentioned in the literature [3] as cell-occupancy allocation algorithms and are characterized by models that monitor the arrival and departure rates at each cell, without regard to past of future location of a mobile user. They require a low complexity in the base station admission and signaling, and can be applied to both data and voice networks. They can also be adapted to support degraded calls instead of hand-off dropping.

A second approach, called "spatial-mobility allocation" [3], and is characterized temporal and spatial correlation of a user location. It involves per user, or per group states in each base station. [8] uses mobility specifications to represent collections of cells a user might visit during the lifetime

of a call. Users receive guaranteed QoS only as long as they move within the mobility specification. [4] proposes the "shadow-cluster" concept, in which a call is admitted if it can be supported by all the cells it might visit - the shadow. Base stations keep track of all overlapping shadows and their intensities when accepting new calls. The "most likely cluster" [2] approach extends the "shadow-cluster" idea by shaping the cluster using mobility prediction information and by reserving bandwidth ahead of the mobile only for some predicted time slots.

3 Motivation

With this large array of call admission schemes, why devise a new one? There are two trends that motivate the search of new, flexible, scalable call admission schemes. As bandwidth in wireless world does not seem to increase at the same pace as in the wired realm, it is foreseeable that in the future increases in wireless bandwidth will be achieved by reduction of cell size. Another trend is that as the wireless technology will become more pervasive, the wireless traffic will increase as well. This two trends pose two scalability problems: first, as the cell reduces in size, it means that for a fixed geographic area the number of cells increases, making less desirable the schemes that rely heavily on inter cell communication and synchronization. Second, as the user base increases, and this has been verified by the current explosion of the Internet, it becomes infeasible to manage resources at the granularity of each call/flow.

The main drawback of "spatial-mobility allocation" schemes lays in their increased signaling complexity, which renders them less usable for very large populations spread over many small cells. These schemes do not scale well with the number of users and with the size of their clusters - a base station might need to keep track of each user, for each time slot. In the shadow cluster scheme for example, admitting a call involves the entire shadow. In other schemes performance depends on the accuracy of prediction algorithms.

The "cell-occupancy" algorithms require a lower complexity, by not managing resources at granularity of each user. However, this is done at the expense of efficiency of the tradeoff between the utilization of the network, and the QoS perceived by the users. There is no adaptation to changing patterns of mobility, and most of these approaches are only suited for particular scenarios, biasing their tradeoffs one one side or another.

Besides scalability, network management is another important factor in designing a wireless network. Many of the aforementioned schemes require either operator assistance to set the appropriate thresholds depending on traffic conditions, or reliable input from prediction models. Our solution, while requiring a low complexity, like the other algorithms in the "cell-occupancy" class, has the capacity to adapt to changing mobility patterns and can make use of prediction models, although in our simulation we did not assume such external input.

4 Proposed Scheme

The dynamic scheme operates in two stages: *dynamic region formation* and *convolution based call admission.* The first stage divides the geographic area into regions based on estimating and characterizing users mobility patterns. The second stage involves a call admission decision in a given cell based on the spatial population distribution inside the region to which the cell belongs. In the following sections we describe each stage in detail.

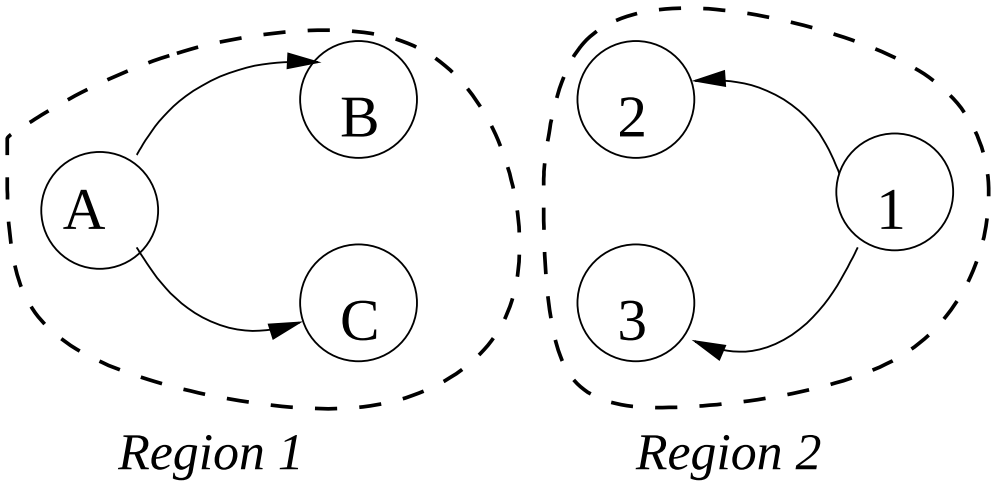


Figure 1: Region Formation

4.1 Dynamic Region Formation

The aim of this stage is to dynamically group together into regions the cells that affect each other. For example, users from certain cells producing a heavily directed traffic may cause overload in some remote cells. The objective of the region formation is to group these cells which are getting affected along with the cells who are causing the affect. The *affect* is therefore defined as users from a given cell affecting the user population of a remote cell due to mobility. The regions of *affect* can then be defined as group of cells, where users from any cell can *affect* any other cell in the same region. An important point to note here is that such mobility pattern may change with time and therefore region formation should be dynamic in defining the regions of *affect.*

The need for such a region formation can be explained by an example (figure 4). Due to directed traffic as shown by arrows, users from cell A can lead to overload in cell C and also in cell B. On the other hand, users from cell 1 is only affecting cells 2 and 3 but not cells A, B and C. Therefore we need to group the cells into two region as shown in figure 4. The regions are basically the extent of information about cells required to make the efficient call admission decision to prevent any overload or congestion in any cell. In other words, once the region is formed, each cell will use information from the entire region to decide for the acceptance of a newly arrived call. A salient feature of region formation is that one can selectively control the admission in each region. For example, one can have strict admission control in an overloaded regions to provide better QoS and looser admission control in underloaded region to increase utilization. Such a selective control is possible due to the underlying basis of region formation which ensures that two cells belonging to different region does not affect each other.

In a cellular network, mobility pattern is a global behaviour and therefore region formation based on estimating or characterizing such pattern may require a centralized scheme. Nevertheless, we make our region formation scheme distributed by only monitoring traffic in a given cell and its neighbouring cells with the aim of capturing the global mobility pattern.

In order to form a region, we define a boolean measure *affect* between a given cell i and its neighbouring cell j. The value of *affect* depends upon two conditions: *Proactive* and *Reactive.* Both these conditions are defined between two adjacent cell i and j and given as follows.

Proactive Condition:

$$Bias_{i \rightarrow j} > B_{thresh}$$

Reactive Condition:

$$Proximity_{i \to j}(d, h, \mu) > P_{thresh}$$

where d refers to the distance of cell i to the closest overloaded cell through cell j, h and μ are mean handoff and departure rate respectively. In the above proactive condition, the $Bias_{i \to j}$ refers to the ratio of outgoing traffic from cell i to cell j to the total outgoing traffic from cell i. The proactive condition tries to capture the existence of directed traffic which leads to overload in a remote cell and therefore tries to prevent a possible overload. The threshold B_{thresh} can be set to a value in the interval [0.3 - 1] which depends on how much a global traffic bias depends upon a local cell to cell bias. In our simulation, cells are modelled to be square, having four neighbours, therefore a fraction larger than 0.25 with one of the neighbours would indicate a traffic bias.

In many cases the proactive condition just by itself cannot ensure the prevention of overloading a cell. The reactive condition comes into play when a cell gets overloaded and therefore acts to ameliorate the cell from overloading. The proximity function in this condition is given as $Proximity_{i \to j}(d, h, \mu) = exp(-d^2 \frac{h}{\mu})$. The distance d is evaluated at each cell based on a distance propagation scheme initiated by the overloaded cell which is presented in Apendix I. The reactive condition implies that, if a cell x gets overloaded, it is due to users arriving from neighbouring cells. Therefore how much a user from these neighbouring cells can affect the overload in cell x is based on distance of these neighbouring cells from cell x and on the users's degree of mobility given by $\frac{h}{\mu}$.

Finally, the value of *affect* is set to**True** if either of the above two conditions is met. Till now we have only defined a relationship *affect* between adjacent cell, next we describe how regions are formed using the *affect* relationship. If we represent the cells in a geographic area as nodes of a graph, and the above described *affect* = **true** relations as undirected link between nodes, then the connected components of this graph will describe our resulting regions. The central property of these regions is that any two cells in a region are either affecting each other (possibly in an indirect manner), or are both being affected by a common source. The regions are dependent on both traffic patterns and distribution of mobiles in the infrastructure, therefore it is necessary to reevaluate the regions periodically. This must be done often enough to reflect changing patterns, and seldom enough not to pose an overhead problem. Once the regions are formed, it is the call admission process, described in the next subsection, which evaluates the admission, based on the spatial population distribution in the region.

4.2 Convolution Based Call Admission

The objective of the call admission decision is to control the load in all the cells in a given region R. Just using the population of a single cell as in [1] or the region as in [2] in order to decide in admitting a call is not efficient. It may happen that a given cell x remain underloaded but leads to overload in some other cells due to users mobility. Therefore, admission decision needs to take into account overload and population states of other cells in the region. A counter scenario can also be valid where an overloaded cell is not affected by the underloaded cell x by being far away from cell x, or by users having a low degree of mobility. In such cases, blocking a call in cell x will lead to poor utilization. The idea is to use the spatially distribution of population with respect to a given cell x and the current degree of mobility to admit a call in that cell. It should also be noted that users from different cells have different degree of affect on given cell x. Therefore the degree of $affect$ of users from any cell j on cell x is quantified using a weight function $W(x, j)$ which depends on the distance d between cell j and x and the degree of mobility (h/μ). The purpose of having $W(x, j)$ is to define a convoluted sum of influence the population of cell x exerts on cell j.

The weight function here follows a gaussian kernel approximation and is given by

$$W(x,j) = exp^{-(d^2)/(h/\mu)} \quad \forall j \in R, j \neq x$$

$$W(x,x) = 1$$

For a given cell x, in region R a convoluted population is then obtained as

$$P_c(x) = \sum_{j \in R} W(x,j)P(j)$$

where $P(j)$ denotes the population of cell j. Normalized convoluted population P_{nc} is obtained from dividing $P_c(x)$ by $\sum_{j \in R} W(x,j)\dot{C}(j)$ where $C(j)$ is capacity of cell j. A call is admitted in a cell x if $P_{nc} \leq l$ where l is the required utilization level. Visually, the weight function applied on the neighbors's population is bell shaped, thus giving more weight to nearby, loaded cells, and less to farther, underloaded cells.

5 Simulation Results

In order to evaluate the performance of the proposed scheme(Dynamic), we simulated it against three other major CAC schemes: the "guard channel" cell-based scheme(Cell), the region based scheme(Region), and the distributed call admission scheme(DCA). Our event-based simulator supports rectangular, wrapped around maps, with square cells having four neighbors each. The threshold value to determine "heavy" traffic is set to 0.4, as 0.25 means equal amount of traffic for all four directions. Values $\mu, h \, and \, \lambda$ and direction probabilities are associated with each cell thus allowing the description of various cell occupancy scenarios. New calls, hand off, and call termination are Poisson modeled processes with their respective rates. The call admission process uses smoothed averages of traffic measurements to estimate local traffic patterns and values of $\mu \, and \, h$. There is no per-mobile state in each cell, the global state of the network being described by the number of mobiles in each cell and in each region, and each cell permanently knows to which region it belongs. Per cell parameters used are :

- λ - new call rate (Poisson modeled process)
- $1/\mu$ - mean call holding time (exponential)
- h - hand off rate (exponential)
- B - number of channels available in each cell
- α - fraction of channels reserved for hand off
- h/μ - the average number of hand offs a mobile makes
- direction probabilities describe traffic patterns $\{d_N, d_S, d_E, d_W\}$
- region it belongs to and region population

Two types of scenarios were used in this study: transient state scenarios, where the the traffic patterns are changing throughout the simulation, and steady state scenarios, in which patterns remain unchanged. The regions are being dynamically reshaped periodically, however, in the steady

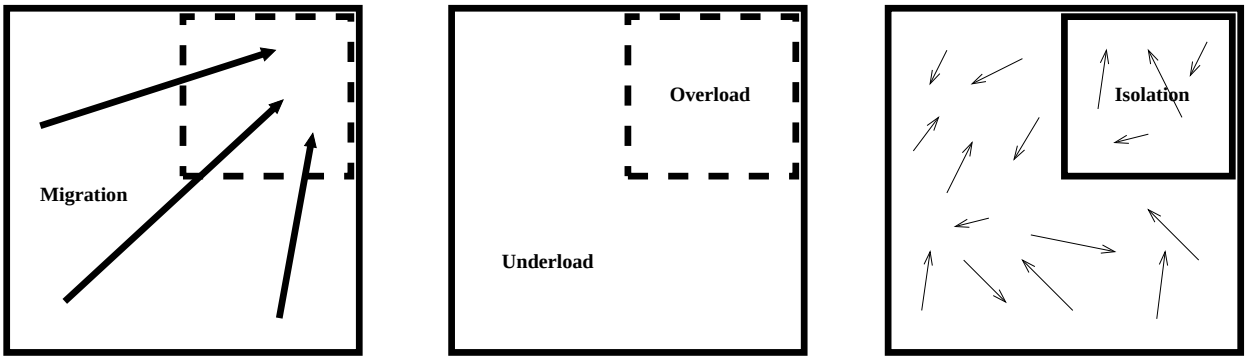
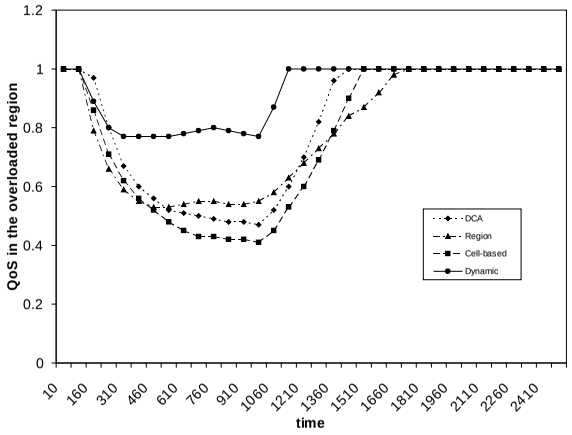


Figure 2: Transient state scenario - Stadium game

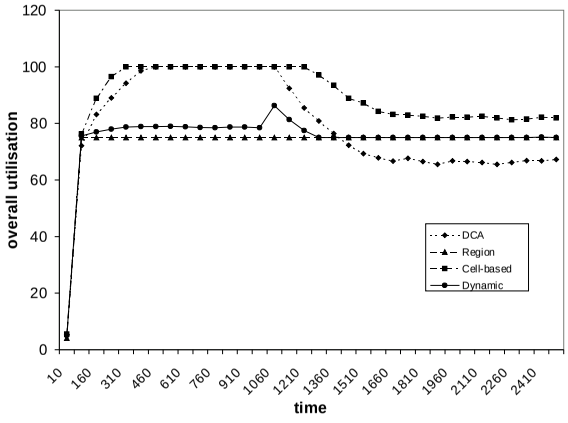
case, there are only changes due to Poisson randomness from one iteration to the next, the traffic patterns being stationary.

The transient state scenario we focused on is called the "stadium scenario" because of its resemblance with the real life event of a sports game(figure 5). During the first time period (1000 time units in the graphs below), mobiles from all over the map are rushing towards one corner. During this migration the corner becomes heavily overloaded, while the rest of the map is mostly underutilized. As the game starts, at time 1000, there is not much traffic between the corner and the rest of the map, and the stadium becomes isolated. At this point, an uniform movement pattern is restored both outside and inside the stadium, while the isolation is maintained. In figure 3 we can see how the scenario is handled by the four CAC schemes in the dense region, the stadium(a), and averaged over the entire map(b). The QoS measure is in fact the ratio of bandwidth obtained by a mobile to the amount of bandwidth required - the closer this ratio is to 1, the better quality the mobile will perceive. Our goal is to maximize the QoS under a given utilization, which was chosen to be 75% in this simulation. The worst performer is the cell-based CAC, mostly due to the locality of its decision base and its inability to counter the effects of large scale mobility. In other words, there is nothing to stop acceptance of new calls at unloaded locations of the map, even if their migration will eventually lead to heavy degradation on the stadium. The region based scheme behaves somewhat better in that it does not continue to degrade the global average after the overall capacity of 75% is reached. However due to its global averaging, it will continue to accept calls in the dense region even when overloaded. The distributed call admission tracks the evolution of the cell-based admission, because it takes a decision based on the neighbors of a cell only, therefore on a larger scale still taking a local decision.

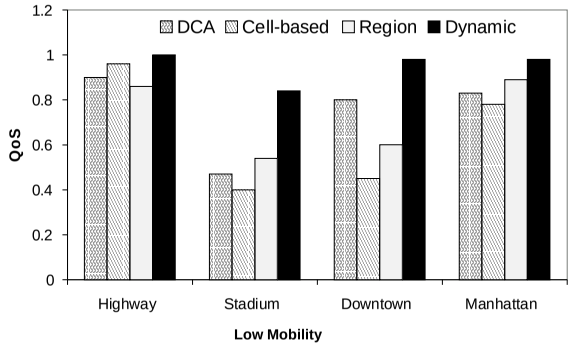
Figure 3b shows the utilization averaged over the entire map during the transient state "stadium scenario". The cell-based and distributed call admission schemes are manually tuned to achieve an utilization of approximately 75%, thus making them hard to manage in large networks with a lot of traffic patterns variety. How does the dynamic call admission scheme manage to achieve a better QoS in the dense area under the same conditions, and still utilize 75% of the network? During the migration period, the dynamic region will in fact cover the entire map, just like in region based case. The population accepted throughout the map however, is distributed in a different fashion by the convolution part of the admission. Another advantage of the dynamic scheme is that it reacts faster to changing conditions. Once the isolation phase begins, at time 1000, the dynamic regions created are separate for the loaded corner and for the rest of the map, mostly under loaded. This allows for faster re-populating of the sparse area, while not accepting new calls in the dense area till the existing calls die off. For the steady state case we condsidered four scenarios, some of which



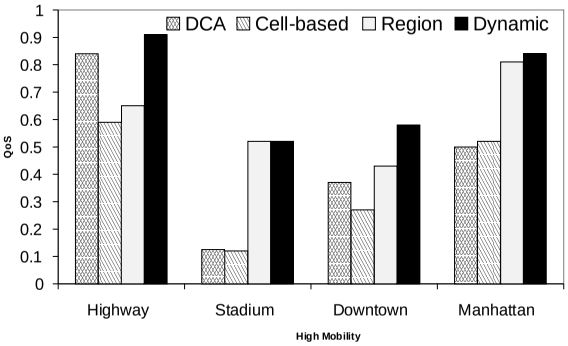
(a) transient state: qos in dense area



(b) transient state: overall utilization



(c) steady state: low mobility ($h/\mu = 4$)



(d) steady state: high mobility ($h/\mu = 40$)

Figure 3: Performance measurements

are also used elsewhere in literature [3] : 1) *Highway:* where a two lane highway has a high hand-off rate, while the rest of the map has uniform λ, h, and μ; 2) *Stadium:* this scenario is similar to the transient state scenario, but it consists only of the migration phase – the traffic pattern doesn't change; 3) *Downtown:* a "belt" around the middle of the map has the property that once there, the users move with equal probability towards the interior, or towards the exterior; 4) *Manhattan:* vertical and horizontal meshed lanes with opposite one way streets. Simulation experiments were run for sufficient long time, so there is no significant change in population per cell in order to gather steady state results. The offered Erlang load in each simulation experiment was $\lambda/\mu = 0.7$ with each bandwidth requirement to be one unit. Figure 3c shows average bandwidth received for users in the overloaded parts of the map under low mobility, when $h/\mu = 4$ – across all scenarios, the dynamic scheme has achieved better performance than other schemes. Similar behaviour is also observed from figure 3d for high mobility, $h/\mu = 40$. In the stadium scenario, high mobility, we notice that our performance is similar to that of region based admission. This is justified by the traffic pattern dictating a region encompassing the entire region, and the convolution giving equal weight to all cells in the map, due to the high value of h/μ. In the low mobility case however, as the h/μ decreases, the dynamic scheme takes a better decision admission by not weighting the entire map in the same manner. In all the other cases, the dynamic scheme choses a region partitioning that is different from the entire region, a single cell, or a cell and its neighbors, thus achieving better results than the three mentioned schemes. This class of simulations is in fact proving that, when a traffic pattern rarely changes, the dynamic scheme stabilizes on a region partitioning that provides a better balance between admission in sparse regions and degradation in overloaded ones.

6 Future Work

We forsee at least two possible directions to continue this work. The first would envision QoS for multiple classes under the same assumptions of degradable service. The issues to solve here are how to provision for each class when patterns of movement are similar or different between classes. One possible example would be a class of users requesting QoS for paging or messaging requests, but with a high mobility pattern, while the other class, while having much less and intermittent movement, has higher bandwith and jitter requirements. A second direction is the improvement of our scheme to completely eliminate the regions as a discrete partitioning of the map, and replace them instead with smoothly decreasing probabilities of overload. This would account more acurately for pattern of movement where mobiles follow nonlinear trajectories. In a naive solution, by simply distributing those probabilities, away from each cell, would incur a high amount of signalling.

7 Conclusion

We presented a new call admission scheme to accommodate adaptive calls in a wireless network. Calls can be degraded, but not dropped, being therefore appropriate for adaptive multimedia. The dynamic scheme first divides the map into disjoint regions which contain both the cause and the effect of mobility driven overload. The actual admission is performed using a convolution based scheme that favors accepting of new calls in zones that are far from overloaded zones. The advantages of the proposed scheme are that it pro-actively reduces the probability of congestion and reactively ameliorates existing congestion. In transient scenarios it reacts faster to changing conditions, and in steady state scenarios it captures advantages of both cell based and region based schemes. Instead of setting individual thresholds for separate cells or regions based on the current mobility pattern, the dynamic scheme achieves uniform levels of utilization by resizing regions.

References

[1] A. Acampora and M. Naghshineh, "Design and control of micro-cellular networks with QOS provisioning for data traffic," *Wireless Network,* vol. 3, pp. 249-256, September 1997.

[2] A. Aljadhai and T. Znati, "A Framework for Call Admission Control and QoS Support in Wireless Environments", in *IEEE INFOCOM'99,* New York, March 1999.

[3] R. Jain and E.W. Knightly, "A Framework for Design and Evaluation of Admission Control Algorithms in Multi-Service Mobile Networks," in *Proc. IEEE INFOCOM '99,*New York, March 1999

[4] 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 Trans. on Networking,* vol. 5, pp. 1-12, February 1997.

[5] M. Naghshineh and M. Schwartz, "Distributed Call Admission in Mobile/Wireless Networks," *IEEE Journal for Selected Areas in Communications,"* 14(4), pp. 711-717, 1996.

[6] R. Ramjee, R. Nagarajan, and D. Towsley, " On optimal call admission in cellular networks," in *Proc. IEEE INFOCOM '96,* pp. 43-50, San Francisco, March 1996.

[7] S Choi and K. G. Shin, "Predictive and Adaptive Bandwidth Reservation for Hand-Offs in QoS-Sensitive Cellular Networks," in *Proc. ACM SIGCOMM'98,* pp. 155-166, Vancouver, September 1998.

[8] Anup Kumar Talukdar, B. R. Badrinath and A. Acharya, "On accommodating mobile hosts in an integrated services packet network," in *Proceedings of The IEEE INFOCOM 1997,*Kobe, Japan, April 1997.

8 Appendix 1

8.1 Distance propagation algorithm

Description of distance propagation algorithm assumes sqaure cells but can be easily extended to the case of hexagonal cells. The purpose of this algorithm is find the distance of a cell to the nearest overloaded cell along four different directions *north, south, east, west.* In order to do find the distance, we define for each cell say i, four distances: $d_{north}(i), d_{south}(i), d_{east}(i), d_{west}(i)$. Initially for every cell i, the above distances are initialized to a sufficiently large value. If a cell x gets overloaded, all the four distances are set to zero. For given cell i which is not overloaded, following steps are executed. Consider a given direction (say *north*), and the adjacent cell j located along that direction. Let d_{min} refer to the minimum of $d_{north}(j), d_{east}(j), d_{west}(j)$. If there exist nonzero traffic along *north* direction from cell i then $d_{north}(i)$ is set to d_{min}. The distance in the opposite direction ($d_{south}(j)$) is not considered for finding d_{min} in order to avoid looping in the distance propagation scheme. Similarly, distances along other directions for cell i is evaluated. Once a overloaded cell gets underloaded, all the four distances for this cell is set to a sufficiently large value. Such a distance propagation scheme can be applied in cellular network where population in a cell follows a stationary process. Therefore assuming the population is stationary for a period, each cell will find the distance to the nearest overloaded cell after a few steps of distance evaluation which is required for distance to propagate from the overloaded cell.