Dynamic QoS Provisioning in Wireless Data Networks

Samrat Ganguly, Dragos Niculescu, Brett Vickers
Rutgers University, USA
{ganguly,dnicules,bvickers}@cs.rutgers.edu

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 the users, that is often unpredictable. Prior and ongoing work in this area of QoS relies on call admission control (CAC) based on the assumption of random or uniform mobility patterns, and also in most cases are based on local information. 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 a 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 encompassing applications needing strong bandwidth guarantees and adapative applications. 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, in case of adaptive applications, the call may not be dropped but may suffer bandwidth degradation. Most of the earlier work in the area of QoS provisioning is focused on supporting voice calls which may be dropped at hand-off. 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 based on the understanding of the effect of mobility on perceived QoS.

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 section 2 we discuss related work in the area of QoS provisioning which led to both the motivation and focus of our work. In section 3, we describe our proposed dynamic QoS provisioning scheme. In section 4, we present simulation results and evaluate the performance of our scheme. Finally, in section 5, we provide conclusions about this work along with comments about possible future work.

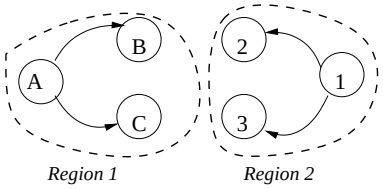

Figure 1. Region Formation

2. Related Work

The majority of the existing call admission schemes are based on using *guard channels* , where a certain amount of bandwidth is reserved for hand-off calls in each base station corresponding to a cell. [6] showed for a simplified queuing model of a single cell that "guard channel" type policies are optimal based on uniform mobility model. 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. 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 above CAC algorithms are mentioned in the literature [2] 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. The above schemes and their variations do not capture any existing global mobility pattern which leads to congestion in a cell and do not provide a way to adapt to changing mobility patterns.

There has also been an array of schemes proposed in [7, 3, 4], based on per user monitoring or bandwidth states. These schemes incurr high state space and message complexity [2] and are more suitable for supporting applications requiring strong guarantees.

3. 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 user 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.

3.1. Dynamic Region Formation

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 population of a remote cell due to mobility. The regions of *affect* can then be defined as groups 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 better explained by an example (fig. 1). 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 are only affecting cells 2 and 3 but not cells A,B and C. Therefore we need to group the cells into two regions as shown in fig. 1. 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 region 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 do 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. Instead, 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 of these conditions are defined between two adjacent cells i and j and are given as follows.

proactive condition:

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

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 prevents a possible overload. The threshold B_{thresh} can be set to a value in the interval [0.3 - 1] which is a function of how much a global traffic bias depends upon a local cell to cell 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 $Proximity_{i \to j}(d, h, \mu) = exp(-d^2 \frac{\mu}{h})$. 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 [8]. The reactive condition implies that, if a cell x gets overloaded, it is due to the users arriving from cells in the proximity. Therefore to what extent a user from these neighbouring cell can affect the overload in cell x is based on distance of these neighbouring cells from cell x and on the users degree of mobility given by $\frac{h}{\mu}$.

Finally the value of *affect* is set **True** if either of the above two conditions is met. Till now we have only defined a relationship *affect* between adjacent cells, 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** leads to an undirected link between nodes, then the connected components 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.

3.2. 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 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

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

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

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.

4. 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 (DCA) scheme. Simulation is done assuming rectangular, wrapped around maps, with square cells having four neighbors each and each cell with capacity of 40 bandwidth units(BU). The offered load with arrival rate λ is poisson modeled and exponential call holding time and hand-offs is assumed in the simulations. h and μ refers to mean call holding and hand-off rate respectively. Two types of scenarios were used in this study: transient state scenarios, where the the traffic pattern changes during the simulation, and steady state scenarios, in which patterns remain unchanged. The regions are being dynamically reshaped periodically, however, in the steady case, there are only changes due to Poisson randomness from one iteration to the next, the traffic patterns being stationary.

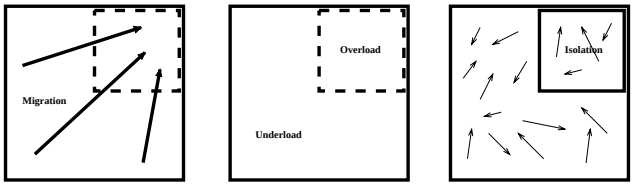

Figure 2. Transient state scenario - stadium

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. During the first time period (1000 time units in the figures 3 and 4), mobiles from all over the map are rushing towards one corner. Such a migration leads to the corner cells becoming heavily overloaded, while the rest of the map getting 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 the figures below we can see how the scenario is handled by the four CAC schemes in the dense region, the stadium (fig. 3), and averaged over the entire map (fig. 4). The QoS measure is in fact the ratio of bandwidth obtained by a mobile to the amount of bandwidth required with the assumption that bandwidth is fairly shared among users. Therefore, 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. Since cell-based decision does not stop accepting new calls in the underloaded region and therefore leads to maximum degradation during the migration period as seen in fig. 3. 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 follows 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 4 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. The dynamic call admission scheme manage to achieve a better QoS in the dense area still having an utilizationg of approx. 75%. This is achieved by selectively admitting calls which do not lead to overload due to convolution based admission decision. Another advantage of the dynamic scheme is that it reacts faster to changing conditions by creating a region around the stadium at isolation which

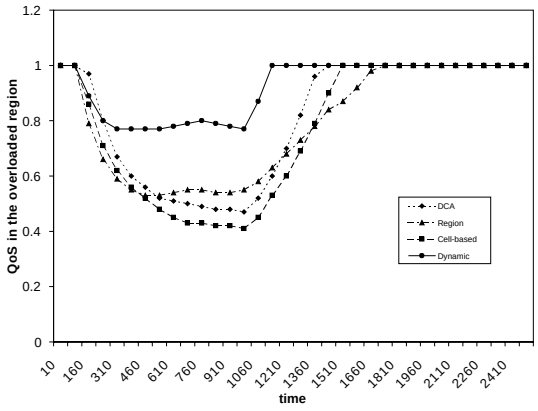

Figure 3. Transient state: QoS in the dense area

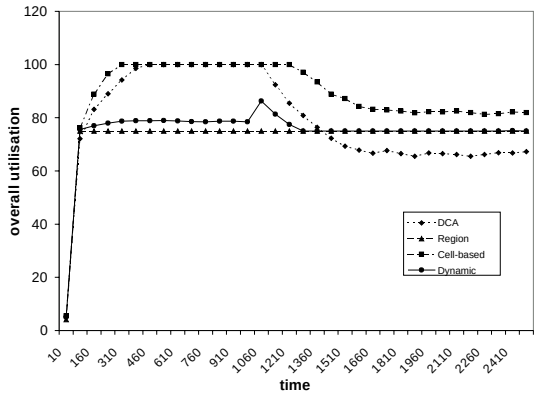

Figure 4. Transient state: overall utilization

allows faster de-population of the overloaded stadium.

For the steady state case we condsidered four scenarios, some of which are also used elsewhere in literature [2] : 1) *Highway:* a two lane highway with high hand-off rate, while the rest of the map has uniform $\lambda, h,$ 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 direction 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 5 shows average bandwidth received for users in the dense cells of the map under low mobility($h/\mu = 4$). Across all the scenarios, the dynamic scheme achieves better performance than other schemes. Similar performance is also observed from fig. 6 for the high mo-

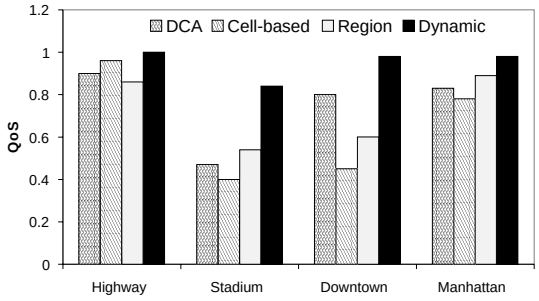

Figure 5. Steady state: low mobility($h/\mu = 4$)

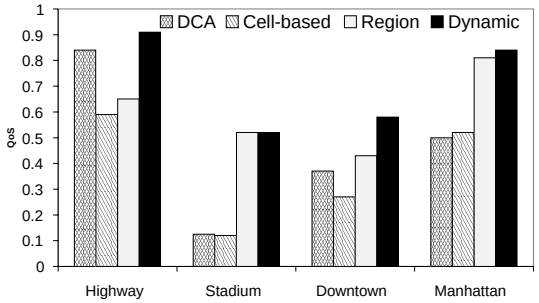

Figure 6. Steady state: high mobility($h/\mu = 40$)

bility($h/\mu = 40$) case. 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 in all cases. 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.

5. Conclusion and Future Work

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.

Future work includes supporting multiple classes of traffic with different levels of degradations, and a better tuning of the convolution to track the mobility patterns in a more accurate manner.

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] 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.

[3] 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.

[4] A. Aljadhai and T. Znati, "A framework for call admission control and QoS support in wireless environments, " in *Proc. of IEEE INFOCOM '99*,New York, NY, March 1999.

[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] S. Ganguly, D. Niculescu and B. Vickers, "Dynamic QOS provisioning in cellular data networks," Technical Report, Rutgers University.