OPTIMAL BANDWIDTH ALLOCATION WITH BANDWIDTH RESERVATION AND ADAPTATION IN WIRELESS COMMUNICATION NETWORKS

Ali Amiri

Department of MSIS,College of Business,Oklahoma State University,Stillwater, OK, 74078, USA

ABSTRACT

Efficient management of bandwidth in wireless networks is a critical factor for a successful communication system. Special features of wireless networks such user mobility and growth of wireless applications and their high bandwidth intensity create a major challenge to utilize bandwidth resources optimally. In this research, we propose a model for an adaptable network bandwidth management method that combines bandwidth reservation and bandwidth adaptation to reduce call blocking and dropping probabilities. The model is an integer program that determines whether or not to accept new calls and decides how to allocate bandwidth optimally in a way to maximize user satisfaction. The results of a simulation study show that the proposed method outperforms an existing method with respect to key performance measures such as call blocking and dropping probabilities and call time survivability. This survivability indicator is a new measure that is introduced for the first time in this paper. We also present a second tradeoff model to allow the network manager to control call dropping probability. The results of a second simulation study show that network users are better off if a zero call dropping policy is adopted as proposed in the first model.

KEYWORDS

Wireless networks, Bandwidth allocation, User satisfaction, Integer programming

1. INTRODUCTION

Efficient management of bandwidth in wireless networks is a critical factor for a successful communication system to support the ever-increasing user demand involving a wide range of applications entailing video, voice and data. The number of wireless/mobile users is growing at a high rate and the applications are becoming more bandwidth intensive. These applications have varying quality of service (QoS) requirements. If bandwidth is efficiently allocated to network calls, then user satisfaction can be improved. One technical innovation in wireless networks involves the employment of smaller cells, called microcells and picocells, to allow radio channel reuse in cells sufficiently apart from each other to increase bandwidth utilization [1,2]. Lee [2] provides a detailed mathematical analysis of the improved performance of cellular networks using microcells as related to increased bandwidth capacity and improved quality of service to mobile users. The use of smaller cells implies, however, higher rate of handoff of mobile calls and creates a major challenge to provide continuous support of QoS guarantees for these calls. Such a support necessitates the deployment of an adaptable network bandwidth management system [3]. In this paper, we propose a model for an integrated admission control scheme for wireless networks as a central part of that system. This scheme combines bandwidth reservation and bandwidth adaptation to reduce rates of call blocking and handoff call dropping.

The bandwidth reservation technique allows cells to reserve bandwidth for admitted/accepted calls for the duration of the calls as users move from one cell to another. As a result, it is imperative to consider not only the availability of bandwidth at the cell where the call is initially connected but also the future availability of bandwidth at the cells the call moves into throughout the lifetime of the call. The bandwidth reservation technique uses both local and remote information about traffic conditions in the network to decide whether or not to admit a call by allocating bandwidth in the cell where the call originates and reserving bandwidth in the cells the call moves into. When the call moves into a new cell necessitating a call handoff, the bandwidth reserved at the new cell is used to handle the handoff connection.

The bandwidth adaptation technique is also an admission control scheme that considers the range of acceptable bandwidths to allocate to a call rather than just the normal bandwidth requirement of the call. Several admission control schemes [4] [5] [6] base their decisions solely on the normal bandwidth requirement; if this requirement can be satisfied by the available bandwidth at the time of the request, then the call is admitted; otherwise it is rejected. The bandwidth adaptation technique [3] [7] [8] [9] would, however, admit the call if the available bandwidth is simply greater than or equal to the bare minimum (i.e., lower limit of the range of acceptable bandwidths) requirement of the call. The technique attempts also to degrade, if necessary, the QoS of some existing calls to release enough bandwidth to admit the new call. The degradation is possible only when the degraded calls are allocated new bandwidths within their acceptable ranges. Thus, by allowing to reduce the bandwidth of an ongoing call and to reallocate the freed bandwidth to a new call or other calls, call blocking and handoff dropping can be reduced significantly, resulting in more satisfied users overall. In the other hand, when a call terminates at a particular cell or is handed off to another call, the released bandwidth is reallocated among ongoing calls (a process referred to as bandwidth compensation) or allocated to new calls.

The following example illustrates the benefit of integrating bandwidth reservation and adaptation in managing bandwidth in wireless networks. Consider a small network with two cells, each has a bandwidth capacity of 30 units of bandwidths (BU's). Five calls arrive to the network at time 1 in order of 1, 2, 3, 4, and 5. The normal bandwidth requirements of the five calls are 10, 8, 10, 15, and 10 BU's, respectively. Each call has a duration of two time units (TU's). Calls 1, 2, and 3 are served by cell 1 and calls 4 and 5 are served by cell 2. All these calls are admitted at time 1 because their total bandwidth requirements are smaller than the cell capacities. Now, suppose that call 1 moves out of cell 1 and enters cell 2 at time 2. Then, it should be decided whether or not to handoff call 1 or drop it because of the network traffic conditions. If the admission control policy does not employ bandwidth reservation, then call 1 should be dropped, causing high dissatisfaction for the user. However, if the admission control policy employs bandwidth reservation (as it is the case in our model), then call 5 should have not been admitted in the first place at time 1 even though there is sufficient bandwidth to serve it and call 1 would have been successfully handed off to cell 2 at time 2. Better yet, if the bandwidth adaptation scheme is used (as it is the case in our model), then all calls would be admitted and served at both times 1 and 2 if call 4, for example, can be allocated a smaller bandwidth (i.e., 10 BU's) at time 2. Therefore, the integration of bandwidth reservation and adaptation in call admission control can lower call blocking and dropping rates. In addition, whenever, one of the ongoing calls is completed, its released bandwidth (or a portion of it) can be used to serve new/incoming class or it can be additionally allocated to ongoing calls to improve user satisfaction.

Several studies have dealt with bandwidth allocation in wireless networks. Very recently, Ahn and Kim [3] presented an optimization model to allocate bandwidth released by a completed call to ongoing calls within one cell to maximize user satisfaction. The model is essentially a multiple-choice knapsack problem and it is solved using Lagrangean relaxation. The model is used to manage bandwidth at an individual cell separately from the others. The model does not

incorporate bandwidth reservation and can result therefore in higher call dropping rates. Similarly, the bandwidth allocation mechanisms described in [9] [10] apply to an individual cell rather than all the cells in the network. The mechanism in [9] focus on bandwidth allocation equity by minimizing the number of calls in the cell which have current bandwidths less than their average requirements. The mechanism in [10] considers two levels of priority for the calls: high and low. High priority calls are allocated maximum bandwidths and low priority calls are allocated minimum bandwidths. The bandwidth management scheme proposed in [6] maintains/reserves a certain percentage of bandwidth capacity in a cell for future incoming calls or handoff calls based on the current bandwidth utilizations. The bandwidth does not, however, guarantee continuous connection for an admitted call throughout its lifetime; i.e., an admitted call can be dropped as it moves to a new cell because of insufficient available bandwidth. The model that we propose in this paper ensures continuous connection once a call is admitted; i.e., a call is never dropped once it is admitted.

Other research efforts have focused on the design and configuration aspects of wireless networks. Two studies [11] [12] addressed the problem of designing a wireless access network under capacity and reliability constraints over a multi-period planning horizon. More specifically, given the locations of the cells and hubs, the interconnections cost between cells and hubs, and the user demands at the cells, the goal is to find the cheapest interconnection between cells and hubs while the bandwidth capacity and reliability constraints are met. Heuristic solution methods based on integer programming formulations of the problem are proposed and tested. Kalvenes et al. [1] studied a more general version of the wireless network design problem that involves determining the locations and sizes of the cells and the channels to allocate to these cells in order to maximize revenue from user generated demand. A cutting-plane based method is used to generate solutions to the problem and verify their quality. Computational tests with 72 problem instances are reported to show the effectiveness of the proposed method. Tayi et al. [13] studied the problem of assigning users to cells in order to minimize data access costs under cell load constraints. They focused on the static/offline version of the problem where each user specifies a time interval during which data access is needed. They recognize that this version of the problem is interesting from a theoretical rather than practical perspective. Other studies have focused on bandwidth allocation to improve the energy efficiency for a wireless communication network [14,15]. For example, Huang et al. [14] proposed an improved energy efficient bandwidth expansion (IEEBE) scheme to effectively allocate the network bandwidth and improve energy consumption.
We adopt in this paper a scheme/method to efficiently manage bandwidth in a cellular network that has the following features.

1. It combines bandwidth adaptation and reservation to provide QoS guarantees and maximize user satisfaction.
2. It uses both local and remote information to decide whether to admit or reject new/incoming calls.
3. It ensures zero call handoff dropping tolerance when demand is known accurately.
4. It is implemented using an optimization model to achieve a high level of bandwidth utilization.
5. The proposed method outperforms the existing method in [3] with respect to key performance measures such as call blocking and dropping probabilities and call time survivability. The latter is a new measure that is introduced for the first time in this paper (see section 3).

As explained above, the proposed scheme is novel and more comprehensive than existing schemes as these possess none or only a small subset of the key features of the proposed scheme.

The optimization model for managing bandwidth is an integer programming model that determines whether or not to accept new calls and decides how to allocate bandwidth optimally in a way to maximize user satisfaction.

The remainder of the paper is organized as follows. In the next section, we present the optimization model for managing bandwidth in cellular networks. In section 3, we report the results of a simulation study to evaluate the performance of our proposed method for bandwidth allocation and compare it to that of a previous method proposed in [3]. In the next section, we develop a tradeoff model that can be used by the network manager to control call dropping probability. Finally, we provide some conclusions and future research directions.

2. PROBLEM FORMULATION

In this paper, we employ the notion of bandwidth adaptation that allows the bandwidth of an ongoing call to vary during its lifetime. Following [3] [7], the bandwidth of a call can be varied and takes any value in a set of acceptable values. The lowest value is the bare minimum required for the call to be admitted. When a new call is attempted at a particular cell, it is admitted if the available bandwidth at the cell is greater than or equal to that minimum and enough bandwidth can be reserved for the call when it moves into new cells (bandwidth reservation). If these two conditions are not met, the call is blocked/rejected. It is important to note that bandwidth is reserved only when it is needed. For example, if the call lasts two time units (TU's): 1 TU in cell 1 where it originates and 1 TU in cell 2 where it moves next, then the bandwidth is reserved for the call at cell 2 only at time 2 and no bandwidth is reserved for the call at cell 2 at time unit 1.

When a call is terminated/completed or handed off from a current cell to another, the released bandwidth can be reallocated among ongoing calls to improve user satisfaction or used to handle new calls. As in [3] [6], we consider the multi-class case where calls in different classes have different sets of bandwidth levels and calls within the same class have the same set of bandwidth levels. Classes can correspond to types of applications of the calls, i.e., video, voice, and data.
We also assume that the movement patterns and durations of the calls are known with great accuracy. Efforts to approximate even roughly these parameters would definitely be beneficial to the cellular network operator. Indeed, the issue of mobility and hence the issue of the estimation of the parameter a_{ij}^t (defined later in this section) are important. As long real system measurements of the parameter are not available, one can choose to estimate (and refine the estimation of) this parameter based on the application type, the average speed of the mobile call, the historical movement patterns, and the geographical features of the region where the call originates, among others [4]. For example, if the caller's position coincides with that of a highway, both the general speed and direction of the call can be predicted to a significant extent based on signal strengths received by the base stations [4]. It was observed that mobile terminal users have well defined routes and behavior when driving [4]. The introduction of vehicle automatic navigation systems and future deployment of intelligent highway systems will increase the accuracy of the estimation of the parameter a_{ij}^t. Traffic measurements regarding movements and durations of mobile calls are instantaneously collected in a wireless network. Hence, the estimation of the parameter a_{ij}^t can always be refined based on statistical analysis of such measurements. Indeed, Levine at al. [4] developed a stochastic method, called shadow cluster, to estimate these parameters with great precision based on such factors as geographical features of the current position, velocity, type of application, and historical behavior of the callers.

In addition and as suggested in [3] [6], a certain portion of the network bandwidth capacity, called guard bandwidth, can be left aside to handle demand of admitted calls that deviate from their approximated patterns used in the model. The benefit of the network being able to handle handoffs of admitted calls can exceed the loss of revenue of not admitting some calls despite the current availability of sufficient bandwidth to admit them. Our model can easily handle this situation by having the parameter B_j (defined later in this section) represent the net bandwidth capacity of cell *j*, that is the bandwidth capacity minus the guard bandwidth that is reserved to handle possible deviations of real movement patterns of the calls from the ones used in the model. Historical data and simulation studies can be used to estimate the guard bandwidth at each cell and the estimation can be constantly refined as more accurate, updated data becomes available. It is also important to bear in mind that the amount of guard bandwidth changes with time of the day and geographical area, among other factors. For example, this amount may be relatively higher during peak hours and in cells where call mobility is higher.

Moreover, the tradeoff model presented in section 4 allows calls to be dropped during handoffs in case of insufficient bandwidth due to, for example, deviations from the patterns used in the model. If a call moves to a cell different from the one used in the model, then the call handoff can succeed (i.e., the call is not dropped) if sufficient bandwidth is available at the new cell (and of course bandwidth reserved for the call in the cell used in the model is released) or the call is dropped if bandwidth at the new cell is insufficient to handle the call (and of course, in this case, the bandwidth reserved for the call in the remaining duration of the call assumed in the model is released).

We consider a wireless network where the scheduling window/horizon/period *T* is divided in time units/intervals $T=\{1,2, \ldots, |T|\}$ [4]. The network includes a set of cells *M*, each cell *j* has a bandwidth capacity B_j. Note that there is a one-to-one correspondence between the set of cells and the set of base stations in the network and therefore we use cells and base stations interchangeably. Let *N* denote the set of calls that are attempted during scheduling window *T*. Each call is characterized by its duration d_i, start time S_i, and finish time F_i (note: $F_i = S_i + d_i - 1$). A set of parameters a_{ij}^t is used to keep track of the locations (in terms of cells) of each call during its lifetime. Specifically, a_{ij}^t equals 1 if call *i* at time *t* is located at and served by cell *j*; and zero otherwise. The set of acceptable bandwidths for call $i \in N$ is $R_i = \{b_i^1, b_i^2, ..., b_i^{h_i}\}$, where $h_i = |R_i|$. Following [3], when a call i is allocated bandwidth $b_i^k \in R_i$, it produces a degree of satisfaction U_i^k. Note that $b_i^k < b_i^{k+1}$ and $U_i^k < U_i^{k+1}$.

Two types of binary decision variables are used in the model, namely Y_i that takes the value 1 if call i is admitted and 0 otherwise and X_i^{kt} that takes the value 1 if call i is assigned bandwidth b_i^k at time $t \in [S_i, F_i]$. For convenience the parameters and decision variables used in the model are summarized below.

T	scheduling window
N	set of calls (indexed by i)
M	set of cells (indexed by j)
B_j	bandwidth capacity of cell j
$R_i = \{b_i^1, b_i^2, ..., b_i^{h_i}\}$	set of acceptable bandwidths for call i
d_i	duration of call i

S_i start time of call i

F_i finish time of call i (note: $F_i = S_i + d_i - 1$)

U_i^k satisfaction of user making call i when the call is assigned bandwidth b_i^k

$$a_i^{jt} = \begin{cases} 1 & of\ call\ i\ is\ located\ at\ cell\ j\ at\ time\ t \\ 0 & otherwise \end{cases}$$

Decision variables

$$Y_i = \begin{cases} 1 & if\ call\ i\ is\ admitted \\ 0 & otherwise \end{cases}$$

$$X_i^{kt} = \begin{cases} 1 & if\ call\ i\ is\ allocated\ bandwidth\ b_i^k \in R_i\ at\ time\ t \in [S_i, F_i] \\ 0 & otherwise \end{cases}$$

The optimal bandwidth allocation problem with bandwidth adaptation and reservation can be formulated as the following integer linear programming problem.

$$\text{Max} \sum_{i \in N} \sum_{t \in [S_i, F_i]} \sum_{k=1}^{h_i} U_i^k X_i^{kt} \quad (1)$$

Subject to

$$\sum_{k=1}^{h_i} X_i^{kt} = Y_i \qquad \forall\ i \in N\ and\ t \in [S_i, F_i] \quad (2)$$

$$\sum_{i \in N} a_{ij}^t \sum_{k=1}^{h_i} b_i^{kt} X_i^{kt} \le B_j \qquad \forall\ j \in M\ and\ t \in T \quad (3)$$

$$Y_i \in \{0,1\} \qquad \forall\ i \in N \quad (4)$$

$$X_i^{kt} \in \{0,1\} \qquad \forall\ i \in N, t \in [S_i, F_i]\ and\ 1 \le k \le h_i \quad (5)$$

The objective function maximizes the total satisfaction of admitted calls. Constraints in set (2) ensure that an acceptable bandwidth is allocated to each admitted call in each time unit of the duration of the call. They also guarantee a zero tolerance call dropping policy; that is, once a call is admitted it is never dropped because of lack of bandwidth. Constraint set (3) represents the cell bandwidth capacity constraints; it ensures that the total bandwidth allocated to admitted calls in a cell does not exceed the bandwidth capacity of the cell. Constraint sets (4) and (5) are the binary integer constraints.

The problem is NP-complete as the special version of the problem with one cell and one scheduling period reduces to the problem studied in [3] which is equivalent to a multiple-choice knapsack problem which is known to be NP-complete.

Computational tests using CPLEX [16] show that the formulation can be strengthened by adding the following constraints.

$$\sum_{k=1}^{h_i} X_i^{kt} \leq 1 \qquad \forall\ i \in N \text{ and } t \in [S_i, F_i] \qquad (6)$$

$$\sum_{t \in [S^i, F^i]} \sum_{k=1}^{h_i} X_i^{kt} = d_i Y_i \qquad \forall\ i \in N \qquad (7)$$

Constraint set (6) ensures that at most one acceptable bandwidth level is selected for each call in each time unit of the duration of the call. Constraint set (7) can be seen as the aggregation of the constraints (2); they set the number of acceptable bandwidth levels to be selected for an admitted call to the length (in terms of time units) of the duration of the call (d_i).

The model is used to help decide which new calls to admit and their bandwidth allocations. The model takes into consideration the status of the network as related to the number of existing calls, their bandwidth allocations, the approximated durations and movements of the calls in order to make call admission decisions. The model can be solved at the network call processor after collecting the necessary information. Based on the simulation study (discussed in the next section), CPLEX [16] seems to be very efficient in solving the model optimally. However, if the solution time becomes an issue, the model can be solved sub-optimally, say within 5% or 10% of optimality. I believe that the model solution would be preferred to a manual solution or ad hoc solution that is based solely on "common practice". This is true even if only rough approximations of the values of the model parameters are used. The model can be solved in (near) real time at the network call processor because of the efficiency in solving the model. However, it may be sufficient to solve the model mainly during the peak or busy hours when bandwidth utilization becomes a crucial factor of network performance. The utility of the solution model may decrease during the off-peak hours as bandwidth utilization is usually low during these hours. But, I believe that the use of the model during these hours can be, in the other hand, beneficial with respect to improving ways of estimating the parameters of the model and learning how to incorporate additional, new aspects of operations in the model itself. In addition, the model can be implemented in a "semi-distributed" manner by solving it at the call controller of each cell cluster, which is a group of neighboring cells [5].

3. SIMULATION STUDY

We use the commercial optimization software CPLEX [16] to solve the problem. Data in this study is generated in a similar way as in [3] [6]. We consider an hexagonal micro cellular network (Figure 1) that is composed of a set of cells (M) and each cell keeps in contact with its six neighboring cells. One base station is located in the center of each cell which has a bandwidth capacity of 30 Mbps. Six call types are assumed based on the call duration, bandwidth requirement, and application type (table 1). These types correspond to six multimedia application groups that are widely used in previous simulation studies [3] [6]. It is assumed that new calls from all six applications are generated with equal probabilities. Each call can originate from any cell in the network with equal probability. The handoff probability (p_h) is used to capture user mobility. Three values 0.30, 0.50, and 0.70 are used to represent low, medium, and high mobility in the network [6]. This handoff probability (p_h) for a call means that the call has (1- p_h) probability to remain in the current cell in the next time unit and p_h probability to move to one of its neighboring cells (with equal probabilities) in the next time unit

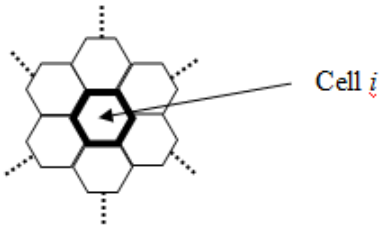

Figure 1. Cellular network configuration

Table 1. Call traffic information

Application group	Bandwidth requirement	Average bandwidth requirement	Call duration	Average call duration	Example
1	30 Kbps		1-10 mn	3 mn	Voice service and audio-phone
2	256 Kbps		1-30 mn	5 mn	Video-phone and Video-conference
3	1-6 Mbps (average) 2.5-9 Mbps (peak)	3 Mbps	5 mn – 5 hr	10 mn	Interactive multimedia and video on demand
4	5-20 Kbps	10Kbps	10-120 sec	30 sec	Email, paging, and fax
5	64-512 kbps	256 Kbps	30 sec – 10 hr	3 mn	Remote long & Data on demand
6	1-10 Mbps	5 Mbps	30 sec – 20 mn	2 mn	File Transfer and retrieval service

Our proposed method for bandwidth allocation in wireless networks is compared with the previous method that is recently developed by Ahn and Kim [3]. The performance measures obtained through the simulation are blocking probability of new calls, dropping probability of ongoing calls, bandwidth utilization, and the call time survivability which is a new measure that is introduced for the first time in this paper. The time survivability for a call is defined as the ratio of the effective time the call lasts by the duration of the call. In our proposed method, the call time survivability is 100% for all calls because a call is never dropped once it is admitted. Whereas in the previous method, the time survivability for an admitted call can be lower than 100% because the call can be dropped prior to its normal completion. Simulation results for a medium mobility environment (p_h =0.5) and a high mobility environment (p_h =0.7) showed similar behaviors as seen in a low-mobility environment. Due to space limitations, those results (for p_h =0.5 and p_h =0.7) are not reported here.

Figure 2 shows the new call blocking probability of the two methods as a function of the call arrival rate. As this rate increases the blocking probability increases but at a decreasing rate for both methods. Our proposed method blocks more calls than the previous method. This is expected as our proposed method admits a new call only if it guarantees that the call won't be dropped during its lifetime, whereas the previous method can admit a new call despite the possibility of dropping the call prior to its normal completion. It is commonly recognized that a dropped call produces more dissatisfaction with network services than a blocked call.

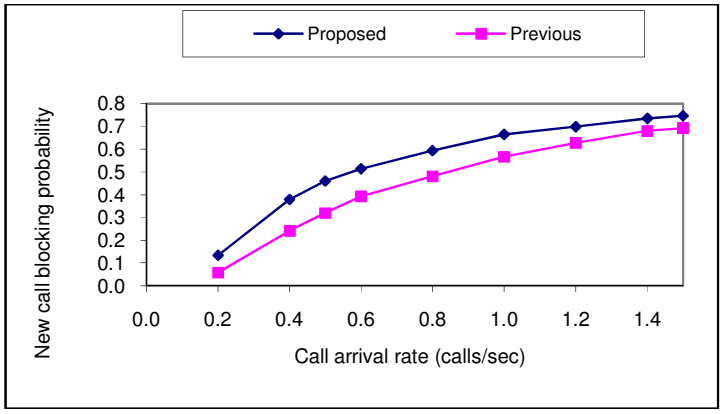

Figure 2. Call blocking probability

Figure 3 shows the dropping probability of ongoing calls for both methods as a function of the call arrival rate. As just mentioned, our method restricts the dropping probability to zero, that is, once a call is admitted, it is never dropped before its normal completion. The dropping probability for the previous method increases dramatically with the call arrival rate. Indeed, when the call arrival rate is 0.2, 10% of admitted calls are dropped prior to their normal completion and when the rate is 1.5, 43% of the admitted calls are dropped. This obviously leads to a high degree of dissatisfaction among network users.

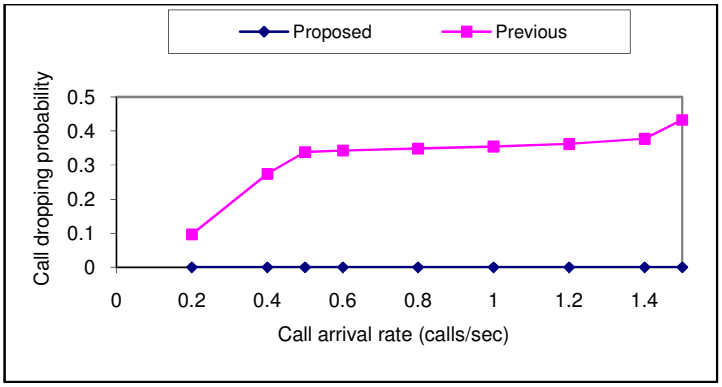

Figure 3. Call dropping probability

Figure 4 shows the cell bandwidth utilization for both methods as a function of the call arrival rate. Both methods utilize bandwidth more efficiently when more calls arrive to the networks. For the same call arrival rate, the previous method utilizes bandwidth slightly more efficiently than our proposed method. For example, when the call arrival rate is 1.5, bandwidth utilization is 93.52% for our method and 96.82% for the previous method. This small advantage of the previous method is achieved, however, at the significant expense of user satisfaction as the previous method ends up dropping as many as 43% of the admitted calls, whereas our method does not drop any admitted calls.

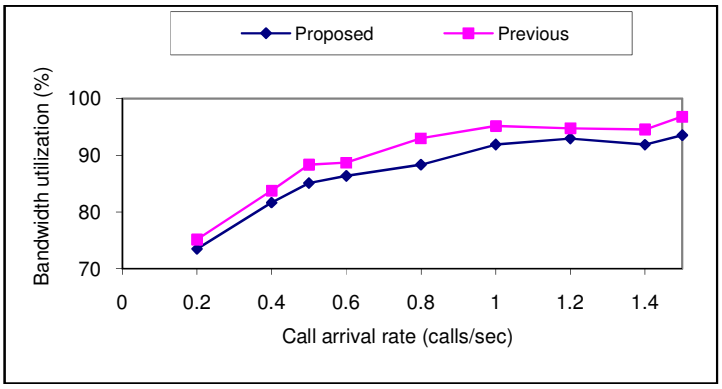

Figure 4. Bandwidth utilization

Since competition in the cellular network industry is based mainly on price and customer service, among others. A 100% bandwidth utilization can hurt quality of service to customers and consequently revenue. A bandwidth utilization very close to 100% causes network congestion and can ultimately drive away customers to other competitors. As noted in [5], "handoff dropping and cell overload are consequences of congestion in wireless networks". If a cellular network starts to experience a very high bandwidth utilization, then the expansion of the network capacity becomes a priority and can be achieved by either increasing bandwidth of existing facilities or installing new facilities such as new base stations. From a computational point of view, an utilization very close to 100% can be achieved if a larger number calls with varying traffic requirements, movements, and durations arrive to the network, so that the admission controller can selectively "pack" as many of them as possible to achieve a high level of bandwidth utilization.

Finally, Figure 5 depicts the call time survivability for admitted calls for both methods as a function of the call arrival rate. The figure shows clearly the superiority of our proposed method over the previous one. Indeed, our method guarantees that an admitted call would last for its entire duration because it never gets dropped; whereas the call time survivability drops significantly for the previous method as the call arrival rate increases. Indeed, using the previous method, when the call arrival rate is 0.2, a call survives on average for only 95.28% of its duration and when the call arrival rate is 1.5, a call survives on average for only 76.78% of its duration. This obviously creates a higher degree of dissatisfaction among network users.

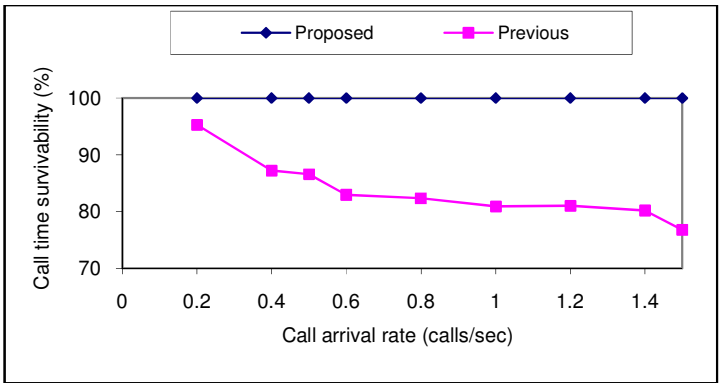

Figure 5. Call time survivability

4. A TRADEOFF MODEL: CONTROLLING CALL DROPPING PROBABILITY

Our proposed model prohibits call dropping and the model in [3] allows a call to be dropped freely due to insufficient bandwidth. In this section, we present a model that is a compromise between the two models. Instead of totally forbidding call dropping or freely dropping calls, the tradeoff model lets the network manager set a dropping threshold (δ) for accepted calls; i.e., an upper limit on the percentage of calls that can be dropped once they are admitted. (1-δ) is referred to as the survival rate of admitted calls. The following is the tradeoff model that controls call dropping.

$$\text{Max} \sum_{i\in N} \sum_{t\in[S_i,F_i]} \sum_{k=1}^{h_i} U_i^k X_i^{kt} \tag{8}$$

Subject to

$$\sum_{k=1}^{h_i} X_i^{kt} \le 1 \qquad \forall\ i\in N \text{ and } t\in[S_i,F_i] \tag{9}$$

$$\sum_{i\in N} a_{ij}^t \sum_{k=1}^{h_i} b_i^{kt} X_i^{kt} \le B_j \qquad \forall\ j\in M \text{ and } t\in T \tag{10}$$

$$\sum_{k=1}^{h_i} X_i^{k(t+1)} \le \sum_{k=1}^{h_i} X_i^{kt} \qquad \forall\ i\in N \text{ and } t\in[S_i,F_i-1] \tag{11}$$

$$(1-\delta)\sum_{i\in N} \sum_{k=1}^{h_i} X_i^{kS_i} \le \sum_{i\in N} \sum_{k=1}^{h_i} X_i^{kF_i} \tag{12}$$

$$X_i^{kt} \in \{0,1\} \qquad \forall\ i\in N, t\in[S_i,F_i] \text{ and } k\in R_i \tag{13}$$

The objective function (8) maximizes total satisfaction of accepted calls. Constraint set (9) ensures that at most one acceptable bandwidth level is selected for each admitted call in each time unit of the duration of the call. Constraint set (10) represents the usual bandwidth capacity constraints. Constraints in set (11) ensure that if a call is dropped at a particular time, it is automatically dropped for the remainder of its normal duration. Constraint (12) is the policy constraint that lets the network manager control call dropping. Indeed, $\sum_{i\in N} \sum_{k=1}^{h_i} X_i^{kS_i}$ represents the number of admitted calls at their start times and $\sum_{i\in N} \sum_{k=1}^{h_i} X_i^{kF_i}$ represents the number of accepted calls that survive until their last finish times. Constraint (12) ensures that the ratio $\frac{\sum_{i\in N} \sum_{k=1}^{h_i} X_i^{kF_i}}{\sum_{i\in N} \sum_{k=1}^{h_i} X_i^{kS_i}}$ for admitted calls that survive for their entire durations is greater than or equal to the survival rate (1-δ).

We conducted a simulation to study the effect of call dropping policy on key network performance measures. Figure 6 shows call blocking and dropping probabilities, bandwidth utilization, and call time survivability (as percentages) as a function of call dropping threshold using a fixed call arrival rate (0.5 calls/second). As expected, call blocking probability decreases, call dropping probability increases, and call time survivability drops when the call dropping threshold (δ) increases. For instance, when δ =0, the call blocking probability is 46% and the call dropping probability is zero and when δ =40%, the call blocking probability is 33% and the call dropping probability is 32.84%. The call dropping threshold seems, however, to have insignificant effect on bandwidth utilization. Indeed, when δ =0, the average bandwidth utilization is 85% and when δ =40%, the average bandwidth utilization is around 89%. This gain in bandwidth utilization is to small to justify the surge in call dropping probability and the decline of the call time survivability. Therefore, we recommend that the call dropping threshold be set to zero and hence guarantee that an admitted call is never dropped during its lifetime as proposed by our original model (section 2), i.e., zero call dropping tolerance.

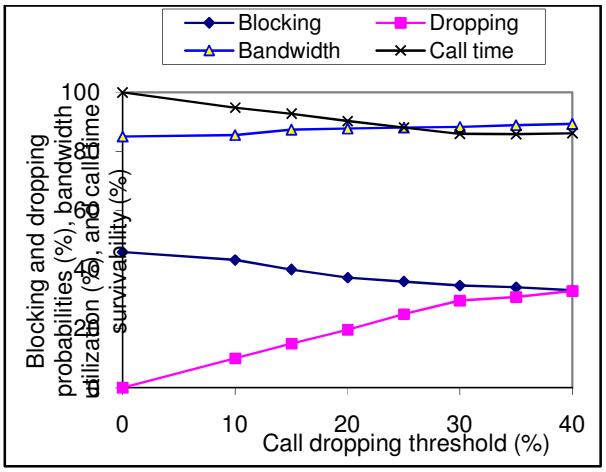

Figure 6. Performance measures for the tradeoff model

5. CONCLUSION AND FUTURE WORK

The provision of QoS guarantees in wireless communication networks is vital to the satisfaction of network users. In this paper, a bandwidth allocation method based on bandwidth reservation and bandwidth adaptation has been proposed to provide those guarantees. The bandwidth reservation technique allows cells to reserve bandwidth for admitted/accepted calls for the duration of the calls as users move from one cell to another. When the user moves from a current cell into a new cell necessitating a call handoff, the bandwidth reserved at the new cell is used to handle the handoff connection and the bandwidth at the current cell is released. The bandwidth adaptation technique is also admission control scheme that considers the range of acceptable bandwidths to allocate to a call rather than just the normal bandwidth requirement of the call. This technique would admit the call if the available bandwidth is simply greater than or equal to the bare minimum requirement of the call. It attempts also to degrade, if necessary, the QoS of some existing calls to release enough bandwidth to admit a new call and to reallocate bandwidth released by a terminated call among ongoing calls (a process referred to as bandwidth compensation) and/or new calls.

The proposed bandwidth allocation method is based on an integer programming model that determines whether or not to accept new calls and decides how to allocate bandwidth optimally in a way to maximize user satisfaction. The method ensures zero call handoff dropping tolerance

when demand is known accurately. The results of a simulation study show that the proposed method outperforms an existing method with respect to key performance measures such as call blocking and dropping probabilities and call time survivability. This indicator is a new measure that is introduced for the first time in this paper. We also presented a second tradeoff model to allow the network manager to control call dropping probability. The results of a second simulation study show that network users are better off if a zero call dropping policy is adopted as proposed in the first model.

As shown by the simulation tests, our method utilizes bandwidth efficiently and some new calls are rejected because of insufficient bandwidth. As user demand increases and network applications become more bandwidth intensive over time, bandwidth capacity expansion of the cellular network becomes a necessity. Therefore, an important extension of this research in the future is to incorporate decisions about creating new cells and expanding the bandwidth capacities of existing ones.

REFERENCES

[1] J. Kalvenes, J. Kennington, E. Olinick, "Hierarchical cellular network design with channel allocation", European Journal of Operational Research, vol. 160, no. 1, pp. 3–18 (2005).

[2] W. C. Y. Lee, "Smaller Cells for Greater Performance", IEEE Communications Magazine, vol. 29, no. 11, pp. 19-23 (1991).

[3] K. Ahn, S. Kim, "Optimal bandwidth allocation for bandwidth adaptation in wireless multimedia networks", Computers & Operations Research, vol. 30, no. 13, pp. 1917-1929 (2003).

[4] D. Levine, I. Akyildiz, M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster", IEEE/ACM Transactions on Networking, vol. 5, no. 1, pp. 1–12 (1997).

[5] M. Naghshineh, M. Schwartz, "Distributed call admission control in mobile/wireless networks", IEEE Journal on Selected Areas in Communications, vol. 14, no. 4, pp. 711–717 (1996).

[6] C. Oliveira, J. B. Kim, T. Suda, "An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks", IEEE Journal on Selected Areas in Communications, vol. 16, no. 6, pp. 858–874 (1998).

[7] V. Bharghavan, K. Lee, S. Lu, S. Ha, J. Li, D. Dwyer, "The timely adaptive resource management architecture", IEEE Personal Communications, vol. 5, no. 4, pp. 20–31 (1998).

[8] T. Kwon, J. Choi, Y. Choi, "Near optimal bandwidth adaptation for adaptive multimedia services in wireless/mobile networks", IEEE 50th Conference of Vehicular Technology,, v. 2, pp. 874–878 (1999).

[9] J. Lee, T. Jung, S.Yoon, S. Youm, C. Kang, An adaptive resource allocation mechanism including fast and reliable handoff in IP-based 3G wireless networks", IEEE Personal Communications, vol. 7, no. 6, pp. 42–47 (2000).

[10] A. Aljadhar, T. Znati, "A bandwidth adaptation scheme to support QoS requirement of mobile users in wireless environment", Proceedings of the Ninth International Conference on Computer Communications and Networks, pp. 34–39 (2000).

[11] I. Bose, E. Eryarsoy, L. He, Multi-period design of survivable wireless access networks under capacity constraints, Decision Support Systems, vol. 38, no. 4, pp. 529-538 (2005).

[12] P. Kubat, J. MacGregor Smith, C. Yum, "Design of cellular networks with diversity and capacity constraints", IEEE Transactions on Reliability, vol. 49, no. 2, pp.165–175 (2000).

[13] G. Tayi, D. Rosenkrantz, S. Ravi, "Local base station assignment with time intervals in mobile computing environments", European Journal of Operational Research, vol. 157, no. 2, pp. 267–285 (2004).

[14] YF Huang, Yung-Fa, Che-Hao Li, Hua-Jui Yang, and Ho-Lung Hung, "Performance of an Energy Efficient Bandwidth Allocation for Wireless Communication Systems," Intelligent Information and Database Systems, pp. 312-322 (2014).

[15] Zhang, Xing, Kun Yang, Pingyang Wang, Xuefen Hong. "Energy Efficient Bandwidth Allocation in Heterogeneous Wireless Networks." Mobile Networks and Applications, vol. 20, no. 2, pp. 137-146 (2015).

[16] CPLEX 12.0 User's Manual, IBM., http://www.ibm.com, Mountain View, CA (2014).