Mobile Networks and Applications 7, 143–151, 2002

A Prioritized Real-Time Wireless Call Degradation Framework for Optimal Call Mix Selection

GERGELY V. ZÁRUBA and IMRICH CHLAMTAC
Center for Advanced Telecommunications Systems and Services (CATSS), University of Texas at Dallas, Richardson, TX 75080, USA

SAJAL K. DAS
Center for Research in Wireless Mobility and Networking (CReWMaN), University of Texas at Arlington, Arlington, TX 76019, USA

Abstract. This paper describes a framework for selecting the optimal call mix to be admitted while employing a bandwidth degradation policy in a wireless cellular network. The optimal property is achieved by maximizing the revenue generated by different calls in a cell for the service provider. By *degradation*, we mean that: (1) some channels can be taken away from ongoing calls that are assigned multiple channels, and/or (2) newly admitted calls that require multiple channels get fewer than what they requested. To avoid removing more channels from calls than they could tolerate, we incorporate a new call attribute: the *degradation tolerance*, i.e., the number of channels a call can be degraded without sacrificing the acceptable level of quality. We also consider *priorities* over calls to influence the admission and/or degradation decision. Our analytical framework includes both static and dynamic scenarios. The dynamic case is enhanced with the ability to *select the optimal call mix* using incoming and departing handoffs, new calls, and call terminations in a recursive way, thus, resulting in a call admission policy. We also discuss how to accommodate non-real-time calls into our system. To evaluate the performance of the proposed scheme, a discrete event simulation tool has been developed that models our dynamic framework built on a customized simulated annealing optimization function. Simulation results demonstrate that not only does the proposed degradation framework maximize the total revenue generated by the admitted calls in the cells, but also reduce the handoff and new call blocking probabilities.

Keywords: call degradation, wireless cellular networks, admission control, simulated annealing

1. Introduction

Rapid growth in wireless network access technology and people's desire to access various kinds of information while on the move, have necessitated the voice transmission based cellular wireless systems to accommodate data traffic as well. Due to the fact that most cellular networks already have gateways to, or are built on the Internet, data transmission issues on the wireless interface are becoming increasingly important. The rapidly decreasing cost of portable computing devices, such as laptops and personal digital assistants (PDAs), has resulted in data services such as e-mail and WWW information access to mobile users. Recent studies done by wireline service providers document that Internet traffic overtook the traffic generated by plain old telephone services. Furthermore, Internet traffic is increasing exponentially by a rate that even seems to beat Moore's law. Observing these trends, it is predicted that the traffic in the next generation of high-speed wireless networks will be mostly generated by high-bandwidth multimedia applications, including such services as video-conferencing, tele-medicine, health-monitoring systems, and tele-education [22]. For a network to be able to satisfy the requirements to transfer real-time multimedia data, quality of service (QoS) guarantees between the end systems must also be provided.

By QoS guarantee of a data stream we mean that, by pre-negotiation, this stream will receive a predictable transmission service from the communication system. The QoS of a system can be generally specified by *bandwidth*, *delay* and *reliability*. There has been a significant amount of research on QoS issues in wireline networks (see [9] for a good overview). However, QoS provisioning in wireless networks is more challenging and requires more attention. This is partly due to

(1) the limited resources (e.g., bandwidth) available in wireless systems,

(2) the poor link characteristics of wireless transmissions compared to the wireline counterparts, and

(3) mobility issues that are not present in wireline networks.

Wireless links have transmission rates of Kbps to Mbps compared to Gbps and higher for wireline links (such as optical), while the bit error rate (BER) difference between wireline and wireless transmission is about 7–10 orders of magnitude. Additionally, in wireless networks the errors are more likely to come in bursts, thus, requiring a more complex error detection or forward error correction. User mobility offers another and even more important challenge for dealing with a wireless network. The user–network interface (UNI) is not fixed in most wireless situations implying that the UNI can change its network attachment point resulting in the need for mobility tracking. Therefore, the conclusion is that most QoS results achieved for wireline networks do not directly apply to wireless networks. Hence, there is a need to revise, if not redesign, QoS provisioning techniques. This need motivates our work.

Existing public data networks such as Cellular Digital Packet Data (CDPD) [3] or General Packet Radio Service (GPRS) [6] utilize the unused voice capacity to support non-real-time data. Using the channels not allocated to calls, these services cannot satisfy delay constraints. Furthermore, since they work as secondary users of the bandwidth, the data transferred by them has lower priority, i.e., they can only provide "best effort" services, thus the "non-real-time" attribute. The way to transmit real-time data is to use the circuits provided for voice transmission, although in most cases this means: one channel with its restricted bandwidth (e.g., 4.8–14.4 Kbps in GSM). Recent systems and standards already deal with this issue allowing calls to use more than one channel for communication. For example, the High Speed Circuit Switched Data (HSCSD) [5] approach of GSM can assign several channels to one mobile user with bandwidths up to 64 Kbps. The HSCSD standard is accommodated in the "2.5th" generation of GSM called phase 2+ (released in 1996)[1]. Third generation (3G) systems will be able to provide several channels to mobile units in a more transparent way, thus increasing the bandwidth available to users equipped with only one mobile station (MS). According to the respective specifications, this multiple channel assignment required by our work will be possible in both future European (e.g., GSM, HSCSD) and American (e.g., IS95B, CDMA2000, EDGE, WCDMA) enhanced data networks.

1.1. Related work

In recent years, QoS provisioning in wireless networks has attracted significant attention. For related work on multimedia traffic support in wireless cellular networks, refer to [1,4,7,15,17,18,20]. In this section we will only address results directly relevant to our context. Rappaport and Purzynski [15] considered cellular systems with mixed platforms that support calls with differing resource requirements. They provided a framework for performance analyses of systems with priority access. Acampora and Nagshineh [1] introduced an adaptive call admission control for cellular networks carrying multimedia data. They combined admission control with resource sharing and hence were able to guarantee QoS requirements for different kinds of multimedia traffic. Oliveira et al. [14] proposed a new admission control scheme for real-time calls in wireless networks supporting QoS. This scheme considered both local information (available bandwidth in the cell the MS resides) and remote information (from neighboring cells). By reserving channels for calls at neighboring cells, they can guarantee better QoS requirements. They calculated the optimal amount of the reserved bandwidth of cells. Seal and Singh [17] were among the first addressing a problem close to the essence of this paper. They identified a new QoS parameter applied only to the wireless domain called *loss profiles*. Loss profiles help keep the communication of endpoints stable even when the bandwidth between them is significantly varying. This ensures graceful degradation of calls. They considered cellular networks in which each call can use more than one "slot" in a cell for its communication flow. If a cell becomes saturated, the network is allowed to remove bandwidth (i.e., slots) from users, thus degrading their calls. Sen et al. [18] introduced a novel framework for cellular systems to degrade calls on demand depending on their bandwidth requirement. They calculated revenue functions and showed that a saturated cell can generate more revenue for the system provider by degrading ongoing calls to be able to admit more calls. This scheme considered call degradation by one channel only, therefore, it might not be appropriate in most cases. They also provided a framework for non-real-time call distribution. Das et al. [4] addressed QoS provisioning techniques in cellular networks for both real-time and non-real-time data-flows at the link layer and proposed schemes for issues like bandwidth compaction, channel reservation and degradation to satisfy QoS requirements for multimedia traffic. Bharghavan et al. [2] considered service classes in the Internet and, thus, attempted to achieve optimal bandwidth allocation over the entire network. A more generalized approach of bandwidth adaptation was proposed by Talukdar et al. [19] who investigated the tradeoff between the network overhead caused by the adaptation and optimality of the bandwidth allocation. Kwon et al. [10] considered bandwidth adaptation from a different viewpoint. They described adaptive multimedia streams and provided a bandwidth adaptation algorithm for these streams.

1.2. Our contributions

In this paper we propose a framework for modeling call *degradation* for real-time traffic, by deriving formulas for calculating the revenue generated by calls for the service provider given the incoming and ongoing call mixes. Each admitted call in our framework generates a revenue for the service provider based upon the parameters of the call. The sum of the revenues generated by all admitted calls is considered as the total revenue in a cell.

The first parameter of a call is the preferred (and maximum) number of channels requested for the given call. Our model considers closely a more QoS-centric view of calls than the model presented in [18], by defining *priorities* over different channel demanding calls as a second parameter. This is very important, since most 3G services will have some type of priority classifications. In GSM, for example, a sophisticated way to handle different priority calls is already included in phase 2+, called the enhanced multi-level priority and preemption (EMLPP) [22]. This can be used to preempt calls or to give higher access rights – superiority to higher priority calls in a saturated system. The third quasi-orthogonal parameter of calls we consider – called *degradation tolerance* – is the maximum degradation individual calls can tolerate. This allows the system to degrade calls with more than one channel while satisfying the prenegotiated constraints at the time of call set up. We then introduce new call arrivals and handoff arrivals/terminations into our framework, extending

[1] A revised HSCSD will probably play an important role in GSM's third generation called the *Universal Mobile Telecommunication System* (UMTS) [22].

it with the ability to handle the *dynamics* of a wireless mobile system. The total network revenue equation that we derive will have to satisfy some constraints (e.g., channels used by calls should not exceed channel capacity of cell) to be able to determine the calls to be admitted. The complexity and dependency of these equations and constraints lead us to use special optimization tools to look for sub-optimal solutions. Among these tools, we choose *simulated annealing* because of its quite fast convergence and well suited random behavior to achieve sub-optimal solutions for call admission and maximal revenue generation. With the help of a discrete event simulation tool developed by us, we show interesting features (e.g., influence on call blocking probabilities) of the proposed framework.

The rest of the paper is organized as follows. Section 2 describes our call degradation framework in the static case followed by the description of the dynamic admission scenario. In this section we also address how to incorporate non-real-time calls into our framework. Section 3 introduces a simulation tool implementing our system and presents experimental results. Section 4 concludes the paper.

2. Call degradation framework

Let us now describe our approach to call degradation and admission. Since different multimedia applications have different bandwidth requirements, to be able to transmit multimedia data flows in cellular networks, more than one channel has to be assigned to these calls simultaneously. Since calls can employ more than one channel, significantly less calls may make the cell overloaded by claiming all available channels. If a cell becomes saturated (i.e., no more available channels/slots), the system's policy can be: either to

- reject/block all incoming calls, or
- degrade (take away one or more channels from) ongoing calls in order to be able to accommodate more calls from the available call mix.

In the extreme of the second scenario, the system may try to accommodate as many calls as possible by degrading all calls to use only one channel. This might lead to too much degradation of calls, which in turn could result in a poor QoS and hence reduced revenue generated for the system provider. Although by negotiating a minimum bandwidth/channel requirement for each call at the call set up time, the maximum degradation can be controlled, it is not assured that this maximum degradation would result in the highest revenue. Also, the selection of the calls to be degraded in a less overloaded situation requires some serious considerations. This problem can be solved by formulating a cost (or revenue) function for the cell. By maximizing this function, the respective degradation values could be evaluated. Another goal towards formalizing the framework is to select an optimal call mix that the system can accommodate considering the available calls. The calls in this optimal call mix will thus become those, which get admitted – resulting in a call admission policy.

In the remainder of this section, we introduce the call parameters and build a framework which is *static* in the sense that it only considers a momentary state of a cell. Assuming that the call mix is already available we derive the best degradation policy. Next we extend our framework to dynamically handle incoming and terminating calls (both handoffs and new arrivals), and have the system select the best call mix (admission) and hence the most profitable degradation policy. Finally, we discuss some issues related to non-real-time calls into saturated systems.

2.1. Call degradation with a static call mix

We characterize each call with a 3-tuple: $\langle$*traffic class, priority class, degradation tolerance*$\rangle = \langle i, j, m\rangle$, where $1 \leqslant i \leqslant K$, $1 \leqslant j \leqslant P$ and $0 \leqslant m \leqslant i-1$. The traffic class i means that i is the preferred (and maximum) number of channels a call is requesting, where K is a constant for the entire network. In the priority classes, a higher priority call is more expensive but it will be more likely to have that call admitted perhaps with less degradation; the value P is also a constant for the entire cellular network and represents the highest priority. The third attribute of each call defines the maximum number of channel degradation individual calls can tolerate (meaning that the minimum requested number of channels by the call is $i-m$), thus defining a set of degradation on calls without major quality sacrifice. Each released channel – due to the degradation – goes to a common pool to be reallocated to other calls.

The number of ongoing class-i, priority-j, and tolerance-m calls is denoted as $n_{i,j,m}$. Also let $n_{i,j} = \sum_{m=0}^{i-1} n_{i,j,m}$ be the total number of class-i, priority-j calls. A *bandwidth degradation policy*,

$$D_{\mathrm{bw}} = \{y_{2,1,1}; y_{2,1,2}; \ldots; y_{K,P,K-1}\},$$

specifies the number $y_{i,j,d}$ of the ongoing class-i, priority-j calls that are degraded by d channels, where $2 \leqslant i \leqslant K$, $1 \leqslant j \leqslant P$ and $1 \leqslant d \leqslant i-1$, since no call can tolerate to lose all of its assigned channels. Let us now take a look at the case where the cell is already saturated. The revenue generated, Φ_{g}, due to carried traffic in the cell is given by

$$\Phi_{\mathrm{g}} = \sum_{i=1}^{K}\sum_{j=1}^{P} T_{i,j} \cdot n_{i,j}, \tag{1}$$

where $T_{i,j}$ denotes the revenue gain for one class-i priority-j call with no degradation.

The system will lose revenue proportionately to the number of degraded calls, and to the level of degradation tolerance. The revenue loss (damage) function, Φ_{d}, can be formulated as follows:

$$\Phi_{\mathrm{d}} = \sum_{i=2}^{K}\sum_{j=1}^{P}\sum_{d=1}^{i-1} L_{i,j,d} \cdot y_{i,j,d}, \tag{2}$$

where $L_{i,j,d}$ denotes the revenue loss in order to degrade a class-i, priority-j call by d channels.

Thus, the *effective revenue*, Φ, earned by the system with the degradation policy D_{bw} is given by

$$\Phi = \Phi_{\mathrm{g}} - \Phi_{\mathrm{d}} = \sum_{i=1}^{K}\sum_{j=1}^{P} T_{i,j}\cdot n_{i,j} - \sum_{i=2}^{K}\sum_{j=1}^{P}\sum_{d=1}^{i-1} L_{i,j,d}\cdot y_{i,j,d}. \quad (3)$$

To compute the optimal $y_{i,j,m}$ values for the system (i.e., the optimal D_{bw} policy), the function Φ has to be maximized. Using simple calculus, it is easy to show that the number of variables to be optimized is $(KP/2)(K-1)$. The constraints to satisfy are

$$\{0 \leqslant y_{i,j,d}\} \quad (\forall i,j,d \mid 2 \leqslant i \leqslant K,\ 1 \leqslant j \leqslant P,\ 1 \leqslant d \leqslant i-1), \quad (4)$$

$$\left\{0 \leqslant \sum_{d=m}^{i-1}(n_{i,j,d} - y_{i,j,d})\right\} \quad (\forall i,j,d \mid 2 \leqslant i \leqslant K,\ 1 \leqslant j \leqslant P,\ 1 \leqslant m \leqslant i-1), \quad (5)$$

$$\sum_{i=1}^{K}\sum_{j=1}^{P}\sum_{m=0}^{i-1} n_{i,j,m} - \sum_{i=2}^{K}\sum_{j=1}^{P}\sum_{d=1}^{i-1} d\cdot y_{i,j,d} \leqslant C, \quad (6)$$

where C is the total number of channels available in the cell. The total number of constraints is $KP(K-1)+1$.

Constraints in (4) formalize the trivial fact that there can be no calls that are degraded by a negative amount of channels (no call can be assigned more channels than requested). Constraints in (5) correspond to the fact that calls that specify d as their degradation tolerance, may only be degraded at most d channels. To formalize an inequality satisfying that calls are not degraded more than they are allowed by their degradation tolerance, we need to compute the number of released channels for all the calls that have a d value greater than or equal to the actual channel value we are calculating for. Later on, this constraint will become significantly more complicated, but the idea behind it will remain the same. Finally, the constraint in (6) assures that the cell controller maintains the used number of channels at or under the channel capacity C of the cell.

We can observe that the complexity of the revenue function in equation (3) is growing quadratically with the number of classes K and linearly with the number of priorities P, and that the revenue is a linear function of $y_{i,j,d}$. Because of the non-trivial dependency of the variables to be optimized, advanced optimization techniques such as integer programming, simulated annealing, or genetic algorithms may be used to compute an optimal or sub-optimal set D_{bw}.

2.2. Call degradation with dynamic call mixes

Let us now consider a dynamic system, which is recursive in the sense that in each step it collects predictive and/or substantive information on the future call mix based on the past call mixes. In each step, $n_{i,j}$ can be seen as the call mix from the previous step. Introducing mobility into the system, there are two types of incoming calls:

- arriving handoff calls, denoted as $h_{i,j,m}$, which is the number of arriving class-i, priority-j, tolerance-m handoff calls and
- new incoming calls ($b_{i,j,m}$).

There are also two types of dissolving calls:

- terminating calls ($t_{i,j,m}$), and
- departing handoff calls ($l_{i,j,m}$).

We assume the availability of predicted handoff data partially coming from the neighboring base stations; at this time we do not deal with how such information is obtained.

The total effective revenue Φ in the dynamic scenario can be calculated as follows:

$$\Phi = \sum_{i=1}^{K}\sum_{j=1}^{P} T_{i,j}(n_{i,j} - t_{i,j} - l_{i,j} + b_{i,j} + h_{i,j}) - \sum_{i=2}^{K}\sum_{j=1}^{P}\sum_{d=1}^{i-1} L_{i,j,d}\cdot y_{i,j,d}, \quad (7)$$

assuming that

(1) all incoming calls can be accommodated, and

(2) the revenue earned by the scheme will be higher than that with rejecting some calls.

We will show that these assumptions can be relaxed easily. To introduce an optimal-mix call admission control to the system, let us add two more sets of variables $\beta_{i,j,m}$ and $\eta_{i,j,m}$ to denote respectively the number of class-i, priority-j, tolerance-m new and handoff calls to be admitted such that $0 \leqslant \beta_{i,j,m} \leqslant b_{i,j,m}$ and $0 \leqslant \eta_{i,j,m} \leqslant h_{i,j,m}$.

We can now reformulate the total revenue function as

$$\Phi = \sum_{i=1}^{K}\sum_{j=1}^{P}\big[T_{i,j}\cdot(n_{i,j} - t_{i,j} - l_{i,j} + \beta_{i,j} + \eta_{i,j})\big] - \sum_{i=2}^{K}\sum_{j=1}^{P}\sum_{d=1}^{i-1} L_{i,j,d}\cdot y_{i,j,d}. \quad (8)$$

Also, the constraints must be changed to

$$\sum_{i=1}^{K}\sum_{j=1}^{P}\sum_{m=0}^{i-1}\big[(n_{i,j,m} - t_{i,j,m} - l_{i,j,m} + \beta_{i,j,m} + \eta_{i,j,m})\cdot i\big] - \sum_{i=2}^{K}\sum_{j=1}^{P}\sum_{d=1}^{i-1} d\cdot y_{i,j,d} \leqslant C, \quad (9)$$

$$\{0 \leqslant y_{i,j,d}\} \quad (\forall i,j,d \mid 2 \leqslant i \leqslant K,\ 1 \leqslant j \leqslant P,\ 1 \leqslant d \leqslant i-1), \quad (10)$$

$$\left\{0 \leqslant \sum_{d=m}^{i-1}(n_{i,j,d} - t_{i,j,d} - l_{i,j,d} + \beta_{i,j,d} + \eta_{i,j,d} - y_{i,j,d})\right\} \quad (\forall i,j,d \mid 2 \leqslant i \leqslant K,\ 1 \leqslant j \leqslant P,\ 1 \leqslant m \leqslant i-1), \quad (11)$$

$$\{0 \leqslant \beta_{i,j,d} \leqslant b_{i,j,d}\} \quad (\forall i,j,d \mid 1 \leqslant i \leqslant K,\ 1 \leqslant j \leqslant P,\ 1 \leqslant d \leqslant i-1), \quad (12)$$

$$\{0 \leqslant \eta_{i,j,d} \leqslant h_{i,j,d}\}$$
$$(\forall i, j, d \mid 1 \leqslant i \leqslant K,\ 1 \leqslant j \leqslant P,\ 1 \leqslant d \leqslant i-1). \quad (13)$$

Maximizing the revenue function (8) not only obtains the optimal degradation policy D_{bw}, but also the optimal call mix for the admission control algorithm.

The revenue function Φ has $(PK/2)(3K+1)$ variables and $PK(7K+1)+1$ constraints. It can be observed that the introduction of admission control does not change the rank (or dimensionality) of the optimization problem.

By maximizing $\Phi(\{y_{i,j,d}\}, \{\beta_{i,j,m}\}, \{\eta_{i,j,m}\})$, the cell will reach the maximum effective revenue. However, a closer look reveals that equation (8) does not consider the "superiority" of arriving handoff calls over new incoming calls. With the introduction of priorities, it may not be true that all handoff calls are superior to all new calls. To be able to handle this situation we have to consider two more virtual revenue gain functions: $H_{i,j}$ for the incoming handoffs and $B_{i,j}$ for the new call arrivals. These two functions can be equipped with the characteristic of "superiority". The new "virtual" revenue function is

$$\Phi = \sum_{i=1}^{K}\sum_{j=1}^{P}\big[T_{i,j}\cdot(n_{i,j}-t_{i,j}-l_{i,j}) + B_{i,j}\cdot\beta_{i,j} + H_{i,j}\cdot\eta_{i,j})\big] - \sum_{i=2}^{K}\sum_{j=1}^{P}\sum_{d=1}^{i-1} L_{i,j,d}\cdot y_{i,j,d}. \quad (14)$$

The revenue function in equation (14) is a "virtual" revenue, which helps us distinguish between handoffs and new calls. The actual revenue for a cell will still be calculated using equation (8).

The revenue function in equation (14) has the same constraints as equation (8), i.e., constraints (9)–(13). The choice of the gain functions $T_{i,j}$, $B_{i,j}$, $H_{i,j}$, and the loss function $L_{i,j,d}$ should be based on statistical information collected from existing networks. The gain functions might be in the form of matrices of numerical data instead of explicit formulas, which will increase the computational speed of the optimization. There are some obvious rules which apply to all four of these functions, such as they are all monotone ascendent in both i and j and additionally $L_{i,j,d}$ is monotone ascendent also in d.

2.3. Incorporating non real-time calls

Non-real-time calls do not have to satisfy the strict end-to-end delay constraints that real-time calls have. In general, real-time calls usually have preemptive priority over non-real-time calls, particularly when a system becomes saturated with real-time calls. This idea is implicitly implemented in CDPD [3] and GPRS [6], since non-real-time packet forwarding services utilize the left-over channels of cells.

From the service provider's point of view, it is desirable to provide enough bandwidth for all users at all time, since the more subscribers they can serve, the more the revenue they can earn. At the same time, they do not wish to over engineer cells by adding channels, that are idle most of the time, and thus waste bandwidth and capital. By carefully monitoring, the network service providers can pinpoint the cells which may become frequently saturated. These cells can then be repartitioned into smaller cells using more base stations with a smaller transmission radius (power) – this is called cell splitting. Another approach is to implement a hierarchical cellular system on top of the critical cells (micro- and macro-cells). No matter what is done to avoid the overloading of cells, it may not be possible to get rid of saturated conditions altogether in an effective manner. In the previous section we provided a framework to handle this situation for real-time multimedia calls. Once the network is saturated, the non-real-time traffic is blocked and it will have to wait until free channels become available in the cell. Still, the service times can become too long. Although non-real-time data transfers do not have delay constraints, it is not acceptable to have too long service times either. In other words, a fair scheduling of submitted calls is preferred. One solution approach is to monitor the traffic of each individual call. Since most network multimedia applications employ VBR (variable bit rate) transmission services, there are periods when the transmission medium is idle or does not use all the resources. These resources could then be reallocated temporarily to non-real-time calls, as proposed in [4]. Another solution is to treat non-real-time calls to be low-priority real-time calls so that we can accommodate these non-real-time calls in our dynamic framework.

We consider each cell having a priority queue system for non-real-time calls. Arriving non-real-time calls are placed into a sub-queue in the queue system corresponding to their respective priority. The queue itself will have a measure l_q, which is the weighted sum of the lengths of its sub-queues. This measure is then used to decide on placing non-real-time calls into the call mix of the decision process, thereby letting the system decide on accepting or rejecting those calls. This model is strongly built upon a non-real-time call distribution system, which is equipped with segmentation functions for user data like the one suggested in [18]. This would help speed up non-real-time data delivery while sacrificing only an insignificant amount of "real-time" bandwidth. The reformulated revenue function is

$$\Phi = Q\cdot l_q\cdot q + \sum_{i=1}^{K}\sum_{j=1}^{P}\big[T_{i,j}\cdot(n_{i,j}-t_{i,j}-l_{i,j}) + B_{i,j}\cdot\beta_{i,j} + H_{i,j}\cdot\eta_{i,j})\big] - \sum_{i=2}^{K}\sum_{j=1}^{P}\sum_{d=1}^{i-1} L_{i,j,d}\cdot y_{i,j,d}, \quad (15)$$

where Q is the virtual revenue gain caused by using one channel for non-real-time calls, and $q \geqslant 0$ is the number of channels to be allocated to non-real-time calls. Constraint (9) has to be reformulated as well, to reflect the number of channels removed from real-time calls, as follows:

$$q + \sum_{i=1}^{K}\sum_{j=1}^{P}\sum_{m=0}^{i-1}\big[(n_{i,j,m} - t_{i,j,m} - l_{i,j,m} + \beta_{i,j,m} + \eta_{i,j,m}) \cdot i\big] - \sum_{i=2}^{K}\sum_{j=1}^{P}\sum_{d=1}^{i-1} d \cdot y_{i,j,d} \leqslant C. \tag{16}$$

3. Simulation experiments

The evaluation of our integrated (call degradation and admission) framework by hand or by deterministic optimization methods seems impossible, considering its complexity and more importantly the strong dependence of constraints on each other and on the revenue function. To validate our dynamic framework, we were seeking a heuristic optimization method and tool that is intelligent enough to cope with the dependences and complexity.

Our search led through evaluating the most well known heuristic optimization methods for our purpose. Our final choices were genetic algorithms (GA) and simulated annealing (SA). Eventually we chose the *simulated annealing* optimization approach. Although it seems that GA can be also used to find sub-optimal solutions for our problem, it cannot be easily modified to cope with variable dependence due to the simple genetic operators like crossover and mutation. On the other hand, in SA, after randomly selecting the variable to be modified in the first step, we can select other variables that depend on the selected variable, thus assuring that in most of the cases the solution remains feasible.

We developed a custom-made discrete simulation engine to emulate mobile nodes interacting with the cell controller, thus creating a cellular environment. In the following subsections we first describe the customized SA function used to determine degradation-admission, followed by the description of the simulation environment. Finally we will present some experimental results and discussion.

3.1. Simulated annealing (SA)

The idea of simulated annealing to simulate the cooling process of material in a heat bath [12] has been applied to discrete function optimizations since the 1980s [8,16]. The basic terms of the original application (e.g., temperature, cooling factor, etc.) were also adapted by the discrete optimization community. While running SA algorithms, a variable called the *temperature* (t) is decreased periodically by employing a monotone descendent cooling function $a(t)$. When the temperature reaches the final temperature defined by the user, the SA algorithm should have a sub-optimal solution for the revenue function. For each temperature, several repetitions are done. In each iteration, a randomly selected variable of the revenue function is selected and changed. For these new values the revenue function and the constraints are reevaluated. If the result for the new solution is better than the previous one, the new solution is accepted for the next iteration. If the new solution is worse than the previous one, it still can be accepted based on a probabilistic function operating on the current temperature. Although SA can easily handle constraints – by evaluating them to check whether the new solution is feasible – it cannot cope with strong dependencies by itself. Thus, to use SA for our purpose we had to change several pseudo-randomly selected variables at the same time. To decide on the future call mix and degradation in our case, the revenue function had to be maximized. The SA function is implemented with approximately $L = 90$ temperature steps, each repeated 100 times. In each of this 9000 steps, the SA function selects one behavior from the following list:

- Admitting a randomly selected handoff call by degrading or removing arbitrary calls admitted in the same simulation step.
- Admitting a randomly selected new call by degrading or removing arbitrary calls admitted in the same simulation step.
- Removing a recently admitted new call from the future call mix.
- Removing a recently admitted handoff call from the future call mix.
- Keeping the current call mix but arbitrarily reconfiguring the degradation values of calls in the mix.

After the final temperature T_f has been reached, the SA function returns with the sub-optimal call mix to be used for the simulation engine in the next step. Figure 1 shows a flow chart depicting the SA function.

3.2. Simulation environment

The simulation engine we developed simulates a large number of mobile nodes or mobile stations (MS) moving in an area. The cell in which our degradation framework is evaluated is placed into the center of this area. This central cell (CC) has a channel capacity of $C = 50$ channels. For each MS, calls are generated according to a Poisson process. We changed the parameters of the Poisson process to generate different loads in the cell. Each MS with an ongoing call entering the service area of CC will initiate a handoff, thus changing the $\{h_{i,j,m}\}$ variable set. If the admission policy decides to admit this handoff call, then it will be added to the future call mix, otherwise it will be forced to terminate.

Those MSs whose calls are terminated by any unexpected manner, may try to reestablish these calls according to a predefined probability P_r (i.e., a call that is forced to terminate may generate a new call request with probability P_r). The call-mix selection algorithm is run every second, to admit the new arriving calls. The total moving area for nodes in the simulations performed is a square with a side length of 1000 units. The CC has a service radius of 125 units. The mobility of the MSs was based on a random walk mobility model of a continuous-time stochastic process [11]. For each node,

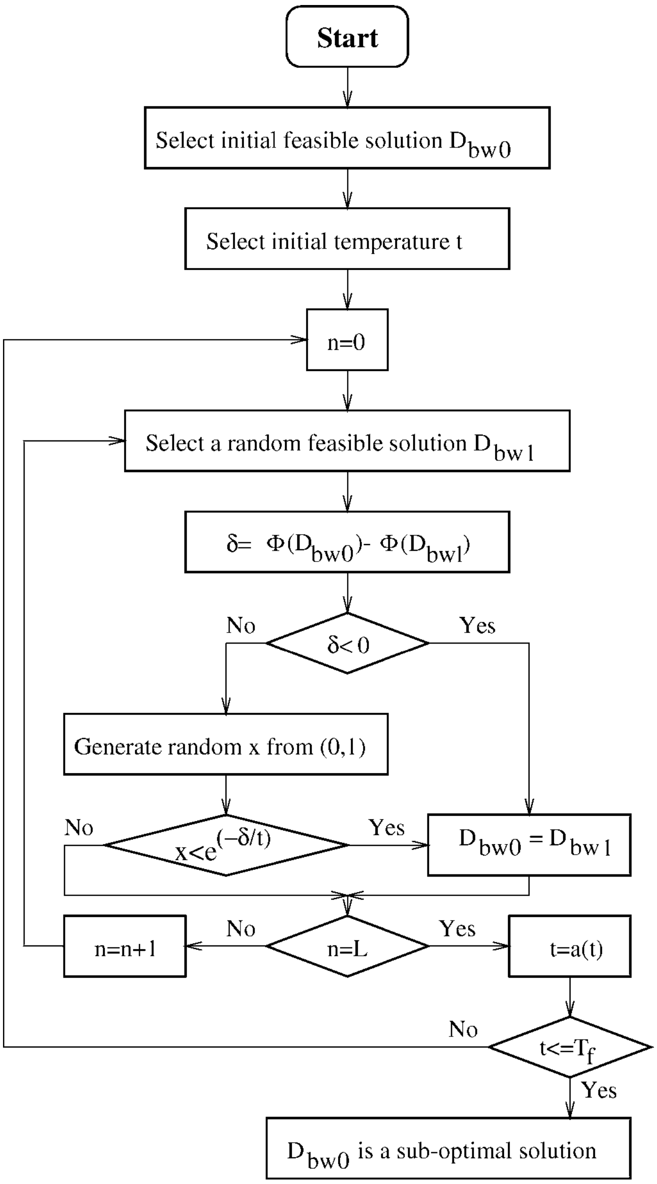

Figure 1. Flow chart for simulated annealing.

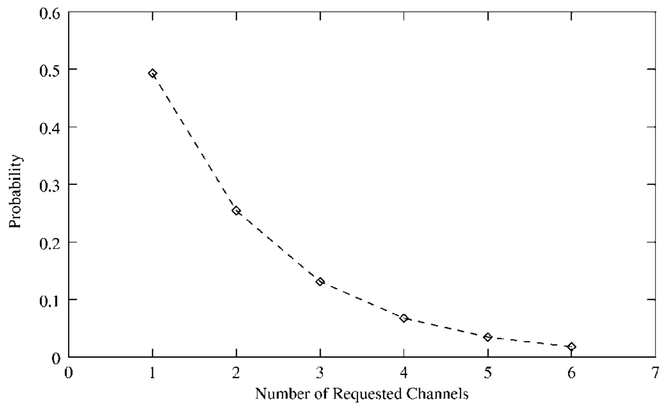

Figure 2. Channel distribution of calls.

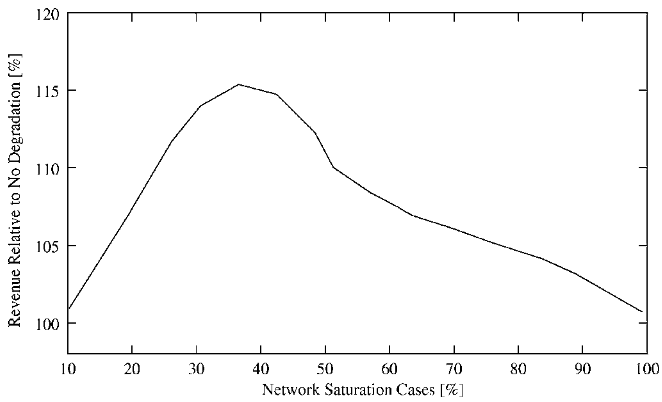

Figure 3. Relative revenue earned.

random length intervals are generated. During these intervals each node is moving into a constant direction with a constant speed. These directions and speeds are randomly generated at the beginning of the intervals. The maximum speed in our simulations was 16 units/s (vehicular speed) with an IID uniform distribution, while the intervals were IID exponentially distributed random variables with a mean of 200 s. The movement directions for each interval are IID uniform distributed. The maximum number of channels used by calls has been determined according to IID geometrically distributed random variables concentrated and normalized onto the closed interval $[1, 6]$, implying that the maximum number of channels for calls used was 6. This distribution is depicted in figure 2. As it can be seen, we assumed that approximately 50% of the traffic is generated by simple voice calls. The priority of calls is determined according to the same distribution as the number of channels but with a maximum value of $P = 4$. The degradation tolerance is generated according to a Poisson distribution concentrated and normalized onto $[0, i - 1]$, with a mean of $i/2$, where i is the assigned number of channels. The revenue generated by a call is a linear growing function with the number of assigned channels and with the priority. The revenue loss (damage) function is a linear increasing function of the gain function (with a multiplier constant of 0.1), it also grows linearly with the degradation parameter.

For each seed (of the random generator to achieve an acceptable confidence interval) and load, two simulations have been completed. The first simulation did not incorporate the degradation but it employed the SA function for call admission (all degradation parameters are set to 0). The second simulation has been run with the degradation features. The duration of the simulations was 1000 s each. The cases when the network becomes saturated were counted in the simulations and this measure was used to relate the obtained results to the parameter along X axis. A 100% stands for a cell that is always saturated.

3.3. Summary of results

We measured the revenue generated, the call handoff blocking and the new call establishment rejections with and without degradations allowed. We then calculated the proportion of the results in those two systems, which shows the benefits of the admission with degradation framework. In figure 3

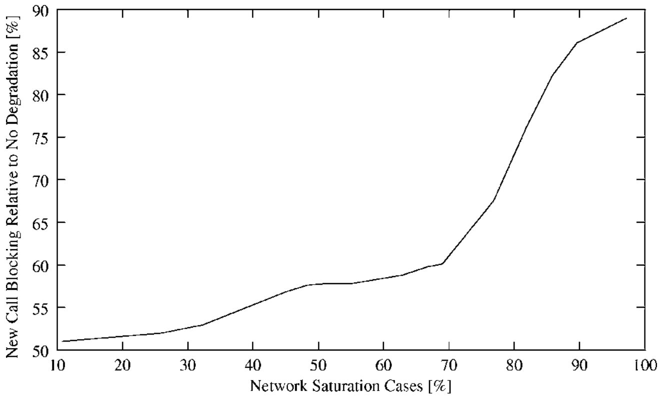

Figure 4. Relative new call rejection.

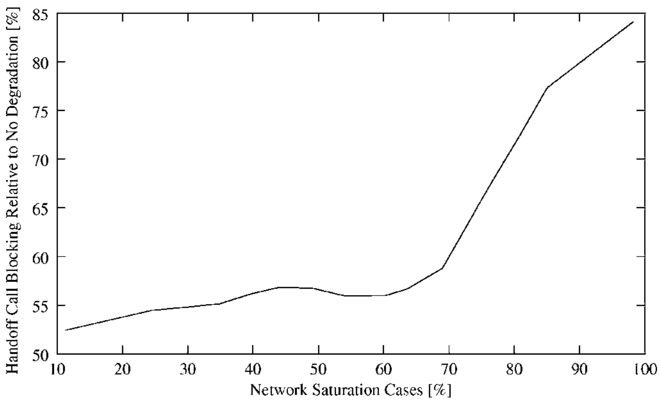

Figure 5. Relative call handoff rejection.

the revenue comparison of the two systems can be observed, while figures 4 and 5 show the new call rejection and the handoff rejection differences (respectively) between the two systems.

Analyzing the revenue results we observe that at very low and very high overloads, the degradation algorithm hardly outperforms a simple admission policy for the following reasons. At low overloads there are not enough high bandwidth requesting calls on the average to make the cell saturated, thus creating a difference. On the other hand, at high overloads the non-degradation system has such a rich call mix to select from that it can select the calls providing better revenues too.

The call rejection figures show interesting results too. At low overloads the non-degradation policy rejects almost twice as many handoffs and new calls than the degradation policy. At higher overloads they become more comparable since most of the calls have to be rejected in both cases and the call mix contains more expensive calls, so degradation may not result in a better call blocking probability at high overloads.

4. Conclusions

In this paper, we presented a framework for optimal call mix selection in wireless cellular networks, where call degradation is possible. By degradation we meant that a pre-negotiated amount of channels can be taken away from calls that are assigned more than one channel. Our framework includes priorities of calls, thus making the admission selection biased towards higher priority calls while biasing the degradation towards lower priority calls. We first presented a static case, where the optimal call mix selection was based on the existence of the call mix to be selected from. Next, in the dynamic case, the framework was enhanced with the ability to select the optimal call mix using incoming and departing handoffs, new calls, and terminations in a recursive way. We have shown ways to introduce non-real-time calls interacting with our framework. To validate our degradation framework, we have developed a discrete event simulation tool based on a customized simulated annealing optimization function. This function was used to obtain sub-optimal solutions for admission and degradation. Simulation results validated that our degradation framework not only raises the total revenue earned by a cell, but can also help in reducing handoff and new call blocking probabilities.

References

[1] A.S. Acampora and M. Naghshineh, QoS provisioning in micro-cellular networks supporting multimedia traffic, in: *Proceedings of IEEE Infocom'95*, Boston, USA (April 1995) pp. 1075–1084.

[2] V. Bharghavan, K. Lee, S. Lu, S. Ha, J. Li and D. Dwyer, The TIMELY adaptive resource management architecture, IEEE Personal Communications Magazine 5(8) (August 1998) 20–31.

[3] CDPD Forum, Cellular Digital Packet Data system specification: Release 1.1, Technical Report, CDPD Forum, Inc. (January 1995).

[4] S.K. Das, R. Jayaram, N.K. Kakani and S.K. Sen, Call admission and control for Quality-of-Service (QoS) provisioning in next generation wireless networks, Wireless Networks 6 (2000) 17–30.

[5] European Telecommunications Standards Institute, GSM 03.34, High Speed Circuit Switched Data (HSCSD) – Stage 2 (1999).

[6] European Telecommunications Standards Institute, GSM 03.60, General Packet Radio Service (GPRS) – Stage 2 (1998).

[7] R. Jain and E.W. Knightly, A framework for design and evaluation of admission control algorithms in multi-service mobile networks, in: *Proceedings of IEEE Infocom'99*, New York, USA, Vol. 3 (March 1999) pp. 1027–1035.

[8] S. Kirkpatrick, C.D. Gelatt, Jr. and M.P. Vecchi, Optimization by simulated annealing, Science 220(4598) (May 13, 1983) 671–680.

[9] J. Kurose, Open issues and challenges in providing Quality of Service guarantees in high-speed networks, Computer Communication Review 23(1) (January 1993) 6–15.

[10] T. Kwon, S.K. Das, I. Park and Y. Choi, Bandwidth adaptation algorithms with multi-objectives for adaptive multimedia services in wireless/mobile networks, in: *Proceedings of the Second ACM International Workshop on Wireless Mobile Multimedia 1999 (WoWMoM'99)*, Seattle, WA (August 1999) pp. 62–70.

[11] A.B. McDonald and T. Znati, A path availability model for wireless ad hoc networks, in: *Proceedings of IEEE WCNC 1999*, Vol. 1, New Orleans, LA (September 1999) pp. 35–40.

[12] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, Equation of state calculation by fast computing machines, Journal of Chemical Physics 45 (1953) 1087–1091.

[13] Nortel News Releases, Nortel Networks breaks own 'Land Speed Record' using light – redefines speed of Internet and networking (October 12, 1999) http://www.nortelnetworks.com/corporate/news/newsreleases/1999d/10_12_9999633_80gigabit.html

[14] C. Oliveira, J.B. Kim and T. Suda, An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks, IEEE Journal on Selected Areas in Communications 16(6) (August 1998) 858–874.

[15] S. Rappaport and C. Purzynski, Prioritized resource assignment for mobile cellular communication systems with mixed services and platform types, IEEE Transactions on Vehicular Technology 45(3) (August 1996) 443–457.

[16] C.R. Reeves (ed.), *Modern Heuristic Techniques for Combinatorial Problems* (Wiley, New York, 1993).

[17] K. Seal and S. Singh, Loss profiles: A Quality of Service measure in mobile computing, Wireless Networks 2(1) (March 1996) 45–61.

[18] S.K. Sen, S.K. Das, K. Basu and J. Jawanda, Quality-of-Service degradation strategies in multimedia wireless networks, in: *Proceedings of IEEE 48th Vehicular Technology Conference (VTC'98)*, Ottawa, Canada (May 1998) pp. 1884–1888.

[19] A.K. Talukdar, B.R. Badrinath and A. Acharya, Rate adaptation schemes in networks with mobile hosts, in: *Proceedings of the ACM/IEEE MobiCom'98*, Dallas, TX (October 1998) pp. 169–180.

[20] X. Tian and C. Ji, Bounding the performance of dynamic channel allocation with QoS provisioning for distributed admission control in wireless networks, in: *Proceedings of IEEE Infocom'99*, New York, USA (March 1999) pp. 1356–1363.

[21] G.V. Záruba, I. Chlamtac and S.K. Das, An integrated admission–degradation framework for optimizing real-time call mix in wireless cellular networks, in: *Proceedings of the ACM Modeling, Analysis and Simulation of Wireless and Mobile Systems 2000 (MSWiM'00)*, Boston, MA (August 2000) pp. 44–51.

[22] Z. Zvonar, P. Jung and K. Kammerlander, *GSM – Evolution Towards 3rd Generation Systems* (Kluwer Academic, Boston, 1999).

Gergely V. Záruba received the M.Sc. degree (with honors) in computer engineering from the Technical University of Budapest, Department of Telecommunications and Telematics, in 1997. Since 1997 he is pursuing research towards a Ph.D. degree. From 1995 to 1998 he was a member of Hungary's leading academic telecommunication research laboratory, the High Speed Networks Laboratory at the Technical University of Budapest. In 1998 he joined the Center for Advanced Telecommunications Systems and Services (CATSS) at the University of Texas at Dallas, where he is currently a Research Assistant and a Ph.D. candidate. His research interests include MAC protocol issues in ad hoc networks, admission control in cellular networks, and Bluetooth scatternet formation problems. He is a member of the IEEE.
E-mail: zaruba@utdallas.edu

Imrich Chlamtac holds a Ph.D. degree in computer science from the University of Minnesota, and B.Sc. and M.Sc. degrees awarded with the Highest Distinction from Tel-Aviv University. Since 1997 he holds the Distinguished Chair in Telecommunications at the University of Texas at Dallas, where he is also the Director of CATSS, the Center for Advanced Telecommunications Systems and Services. Prior to joining UTD, Dr. Chlamtac was on faculty at Technion – The Israel Institute of Technology and the University of Massachusetts in Amherst. Dr. Chlamtac is also the recipient of the titles of Senior Professor at Tel-Aviv University, The Bruno Kessler Honorary Professor at the University of Trento, Italy, and of University Professor at the Technical University of Budapest. Dr. Chlamtac is a Fellow of the IEEE and ACM societies for numerous contributions to networking, including the concept of lightpath or wavelength routing and contention free channel access methods. He is also an IEEE Distinguished Lecturer and was a Fulbright Scholar. Dr. Chlamtac published over two hundred and fifty papers in refereed journals and conferences. He is also the co-author of the first textbook on *Local Area Networks* and of the recent book on *Mobile and Wireless Networks Protocols and Services* (Wiley). Dr. Chlamtac serves as the founding Editor-in-Chief of the ACM/URSI/Kluwer Wireless Networks (WINE), Mobile Networks and Applications (MONET) journals, and the SPIE/Kluwer Optical Networks (ONM) Magazine and has served on the editorial boards of most major publications in telecommunications. Dr. Chlamtac is the founder of ACM/IEEE MobiCom and SPIE/IEEE/ACM OptiComm conferences.
E-mail: chlamtac@utdallas.edu

Sajal K. Das received the B.Tech. degree in 1983 from Calcutta University, the M.S. degree in 1984 from the Indian Institute of Science at Bangalore, and the Ph.D. degree in 1988 from the University of Central Florida at Orlando, all in Computer Science. Currently he is a Full Professor of Computer Science and Engineering and also the Founding Director of the Center for Research in Wireless Mobility and Networking (CReWMaN) at the University of Texas at Arlington. During 1988–1999, he was on the Faculty of Computer Science at the University of North Texas, Denton, where he founded the Center for Research in Wireless Computing (CReW) in 1997. During 1995–1997, he was the director of the Center for Research in Parallel and Distributed Computing (CRPDC) at UNT. Dr. Das is a recipient of the Honor Professor Award from UNT in 1991 and 1997 for best teaching and scholarly research, and UNT's Developing Scholars Award in 1996 for outstanding research. He has been a Visiting Scientist at the Council of National Research in Pisa, Italy, and the Slovak Academy of Sciences in Bratislava, and also a Visiting Professor at the Indian Statistical Institute, Calcutta. He has visited numerous universities and research organizations worldwide for collaborative research and seminar talks. He is frequently invited as a speaker at international conferences and symposia. His current research interests include resource management in wireless networks, mobile computing, QoS provisioning, wireless Internet, Wireless multimedia for 3G networks, distributed/parallel computing, performance modeling and simulation. He has published over 170 research papers in these areas and directed several funded projects. He received the Best Paper Awards in the ACM/IEEE Fifth International Conference on Mobile Computing and Networking (MobiCom'99), the Third ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM 2000), and ACM/IEEE International Workshop on Parallel and Distributed Simulation (PADS'97). Dr. Das serves on the Editorial Boards of the Journal of Parallel and Distributed Computing (as the subject area editor of mobile computing), Parallel Processing Letters, and and the Journal of Parallel Algorithms and Applications. In 1997–1999, he was the Co-Editor of the IEEE TCPP Newsletter. He has guest-edited special issues of ACM Wireless Networks (WINE), Journal of Parallel and Distributed Computing and IEEE Transactions on Computers. Dr. Das has served on the program committees of numerous conferences, including IEEE IPDPS, ICPP, IEEE INFOCOM, and ACM MobiCom; was the General Vice-Chair of MobiCom 2000, General Co-Chair of the Fourth IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS'98), General Chair of the Third ACM International Workshop on Wireless Mobile Multimedia (WoWMoM-2000), General Vice-Chair of the IEEE International Conference on High Performance Computing (HiPC2000); Program Vice Chair of HiPC'99; and the Founding Program Chair of WoWMoM'98 and WoWMoM'99. Dr. Das also serves on the ACM SIGMOBILE and IEEE TCPP Executive Committees. He is a member of the IEEE and ACM.
E-mail: das@cse.uta.edu