# Auction-based Resource Reservation in 2.5/3G Networks

Manos Dramitinos, George D. Stamoulis, and Costas Courcoubetis

Network Economics and Services Group (N.E.S.),
Department of Informatics, Athens University of Economics and Business
76 Patision Str., Athens, GR 10434, Greece
Tel: +30 210 8203693, Fax: +30 210 8203686
Email: {mdramit, gstamoul, courcou}@aueb.gr

## Abstract

We consider UMTS networks in which users request services other than telephony that last for long time intervals: e.g., video clips that last for several minutes. The duration of network time-slots over which resource units are allocated is much shorter. This complicates consistent reservation of resources over longer time scales, where consistent reservation is required to ensure that service quality is constant throughout the entire service session. In this paper, we define an auction-based mechanism for nearly consistent reservation of the resources of a UMTS (or GPRS) network by the users that value them the most, in order to satisfy the longer time scale requirements of their service sessions. Each of these sessions has a fixed target bit-rate. The mechanism is based on a series of Generalized Vickrey Auctions and a set of predefined user utility functions that we propose. Bidding is performed automatically on behalf of the users on the basis of each user's selection of one of these utility functions and his declaration of a total willingness to pay. We argue that under our mechanism the user does not have a clear incentive of not performing a truthful selection of a bidding function according to his own utility. The utility functions we define express appropriately the preferences of the users with respect to the resource allocation pattern in the cases where perfectly consistent allocation cannot be attained. We also provide a mapping of these functions to the UMTS service classes. The effectiveness of our resource reservation mechanism is demonstrated by means of experiments. It appears that most of the users either are served very satisfactorily or essentially are not served at all. The mechanism is implemented at the network base station, and is applicable in practical cases of networks with large numbers of users whose sessions last for many slots.

**Keywords:** Auction, efficiency, resource reservation, UMTS, utility.

## 1 Introduction

Multi-unit auctions have recently received considerable attention as an economic mechanism for resource reservation and price discovery in networks. The case where users compete for reserving resources consistently (i.e., without undesirable fluctuations of the bit-rate) for *large* *time scales* remains an open research topic. This is of particular interest for many practical cases such as the provision of network services with high duration. A prominent case is that of UMTS [5], [11]: except for voice, the duration of services that users generally request is significantly longer than a single time slot; e.g., news downloading or video streaming of a certain bit-rate. Note that the duration of such sessions depends on the application and the content: the duration of video jokes can be one minute (or even less), while that of a football match is approximately 90 minutes. In any case, however, the duration of such sessions is much longer than that of network slots, over which resource units can be reserved. Apart from UMTS, the problem of consistent resource allocation also applies to GPRS technology including its enhanced version EDGE[1] [9].

In this paper, we deal with how to satisfy the requirements of large time scale services, each having a *fixed* target bit-rate, by allocating resources *nearly consistently* to those users who value them the most. Each user succeeding this enjoys an almost constant level of quality of service throughout the duration of his session. We assume that the population of users generally varies over time, which further complicates the problem. Our objective is to allocate resources *efficiently*. That is, attain a high social welfare (i.e. total value induced to the users), while providing proper incentives for rational usage of resources. Note that demand

[1] The unit of resource allocation and the definition of a time-slot depends on the network technology. In UMTS, a 10msec UTRAN frame constitutes a slot, while resources are allocated in bits; see Appendix. In GPRS and EDGE, the unit of resource allocation is the radio block.

fluctuations make it impossible for a provider to publish fixed prices for the resources, which would have been the way to attain efficiency under steady demand. Since the mobile industry is extremely competitive and spectrum is a scarce resource, efficient exploitation of spectrum is crucial for commercial providers. In this paper, we propose and evaluate a mechanism attaining to a significant extent nearly consistent reservation of resources. This mechanism comprises: (a) a series of repeated Generalized Vickrey Auctions, one per slot, and (b) certain bidding functions for the user to choose from.

The repetition of auctions has been analyzed by economists. Nevertheless, most of the related work focuses on comparisons of sequential and simultaneous auctions and the problems that auction repetition imposes on incentive compatibility (see [2] and [3]). Closely related to our problem is that studied by Crémer and Hariton in [1], who also recognize that certain Internet applications require guaranteed capacity over longer time periods than others. Thus, [1] deals analytically with the properties of a Vickrey-type mechanism for accommodating such applications' demand for a pipe. However, the assumptions in the models analyzed in [1] are considerably different than ours. It is taken that each application lasts for either one or two time slots, and that usage of the pipe is dedicated entirely to a single application (as opposed to being shared at portions to be determined) by means of a Vickrey-type auction in which at most two users participate. Under these assumptions, the authors of [1] analyze the users' bidding strategies, the revenues associated with the auction mechanism, and for a special case they derive the optimal revenues. Hence, it remains an open problem how to reserve resources nearly consistently throughout the service duration by means of auctions, particularly in practical cases with large numbers of users whose sessions last for many slots. It should be noted that Lazar and Semret [6] have employed the Vickrey auction in bandwidth trading. However, they deal with "one-shot" trading of bandwidth, although the process comprises a series of Vickrey auctions until convergence is reached. Also, Varian and MacKie-Mason [7] employ a Vickrey auction in order to determine the price per packet in the Internet.

In this paper, we develop a new auction-based resource reservation mechanism that is applicable for 2.5/3G networks, i.e. networks based on technologies such as UMTS, GPRS and EDGE. For simplicity reasons, we henceforth restrict our discussion to UMTS, but we employ the general terms "slot" and "unit" (of resource allocation) so as to keep the presentation of the material as general as possible; see Appendix. Due to the very short duration of each slot, it would be practically impossible for each user to place a new bid per slot. Thus, our mechanism comprises bidding functions; at the start of his service session, each user selects such a function and declares his total willingness to pay (for perfectly consistent allocation of resources). The network performs bidding on *behalf of the user* according to the aforementioned declarations. Thus, all computation is performed at the network base stations, so that there is no need for extra communication with the user terminals. The bidding functions of our mechanism are directly related to certain utility functions that we introduce. These appropriately express the preferences of the users with respect to the resource allocation pattern in the cases where perfectly consistent allocation is not attained. We also provide a mapping of our utility functions to the UMTS service classes. Furthermore, when dealing with auction mechanisms, *user incentives* is an important issue. We argue that, under our mechanism, the user does not have a clear incentive for not making truthful declarations when selecting his bidding function. This property also provides users with the incentive for rational usage of resources, according to their actual needs. The definition of bidding functions and the fact that users have to choose one of these constitutes an integral part of our mechanism for nearly consistent resource reservation. The effectiveness of our mechanism is demonstrated experimentally by analyzing the resulting resource allocation patterns. It is seen that users with competitive bids are served very satisfactorily while they are allocated patterns that are in very good accordance to their respective preferences. On the other hand, users with non-competitive bids are allocated very limited quantities of resources (if at all) at a low charge.

An alternative to performing auctions sequentially (one per slot) would be to run combinatorial auctions, thus allowing users to bid for resources spanning several consecutive slots.

Such an approach should not be adopted due to the following reasons: i) the computational overhead is excessive, ii) guaranteeing resources for the entire duration of the sessions currently served would render it impossible for the network to serve users with higher value that would arrive late, thus resulting in high connectivity delay, poor revenue and inefficiency.

The remainder of this paper is organized as follows: In Section 2, we present our auction-based mechanism. In Section 3, we define user utility functions capable to express certain types of user preferences with respect to patterns of inconsistent resource allocation. In Section 4, we experimentally assess the effectiveness of our mechanism. In Section 5, we discuss issues on user incentives. In Section 6 we present certain extensions of our mechanism. Finally, in Section 7, we provide some concluding remarks.

## 2 ATHENA: A new Resource Reservation Mechanism[2]

Most of the complexity of the problem of consistent resource reservation in UMTS lies in the fact that users demand sessions spanning partly overlapping intervals with different durations, which in general are much larger than the time scale $t_a$ of a slot. The approach that we propose is to conduct a sequence of "mini-auctions", each concerning reservation of resources within one slot. Each mini-auction is a sealed-bid Generalized Vickrey Auction (GVA) [4], with atomic bids (i.e., bids that are either fully satisfied or rejected) of the type $(p, q)$, where $q$ is

[2] ATHENA: Auction-based THird gEneration Networks resource reservAtion

the quantity of resource units sought in the present slot and $p$ is the price proposed for each such unit. Henceforth, unless otherwise specified, we restrict attention to services requiring constant bit-rates. For UMTS, since resources are allocated in bits, if the service involves traffic of a specific bit-rate $m$, then we have $q = m \cdot t_a$.(For each such service session, one bid is placed per mini-auction; see Section 3.) As explained in [4], in a GVA, users with elastic utility functions have the incentive to bid truthfully their willingness to pay. This is a very attractive property, motivating our selection of running a GVA in each mini-auction. Indeed, most often efficiency comes together with incentive compatibility (i.e. the incentive for truthful bidding), since it is hard to allocate the goods auctioned to the users that value them the most if their bids do not reveal their true values for these goods.

The rules of the Generalized Vickrey Auction [4] prescribe that: i) each user reports his valuation for a subset or for all points of his demand function for the good auctioned, ii) units are allocated to the highest bids until demand exhausts supply, iii) each user is charged according to the social opportunity cost that his presence entails. Formally, user's $i$ charge equals $\mu_i(\theta) = SW_{-i}(0, \theta_{-i}) - SW_{-i}(\theta)$, where $SW_{-i}$ is the social welfare of bidders other than $i$, $\theta$ is the set of the users' reported valuations and $(0, \theta_{-i})$ is the efficient outcome if $i$'s reported value were 0 while the reports by other users remain unchanged. Note that winners' charges are less than their respective bids. We illustrate this below, by means of an example relevant to our case, namely with a single atomic bid per user: Assume that six units of a good are auctioned to five users, namely $A$, $B$, $C$, $D$ and $E$. Users $A$ and $B$ are interested in being awarded exactly four units, while $C$, $D$ and $E$ claim exactly two. The vector of the per unit (of good) valuation of users $A$, $B$, $C$, $D$ and $E$ respectively is $< 9, 8, 6, 7, 3 >$. As already mentioned, users have the incentive to bid truthfully. The corresponding bids are $b_A = (9, 4)$, $b_B = (8, 4)$, $b_C = (6, 2)$, $b_D = (7, 2)$ and $b_E = (2, 2)$. By ranking them in decreasing order of price, it follows that $A$ is awarded four units, while $D$ is awarded the remaining two units. Note that $B$ is not awarded any units since his bid cannot be fully satisfied. Under our mechanism, bids are atomic. Hence, if a bid is high enough to be "winning" but the quantity demanded exceeds the residual supply, it is declared as losing; the remaining units are allocated to the lower bids that can be fully satisfied. Recall now that the charge for each user equals the social opportunity cost his presence entails. Thus, $A$ pays according to the prices of the highest losing bids after excluding his own bids, paying as many such bids as the units he won. Hence, $A$ is charged $8 \cdot 4 = 32$ and $D$ is charged $6 \cdot 2 = 12$.

Our mechanism consists of a series of GVAs, one per slot. However, in a realistic case of a UMTS network, these mini-auctions will need to be run so frequently that it would not be feasible for users to participate in all these mini-auctions, neither manually nor automatically by means of an agent running in their respective terminal. Thus, since the user cannot give his bid on a per mini-auction basis, we define utility functions, pertaining to the various services. These functions are provided by the network operator as bidding functions for the user to choose from; they are scaled by the user's total willingness to pay, which is to be given by the user himself (as part of his service request). This approach is similar to that of 3GPP [12] regarding predefined QoS profiles. Then, the network runs all mini-auctions by bidding on behalf of each user, according to his respective selection of bidding function. As explained later, the user does not have a clear incentive for not selecting truthfully his bidding function.

Furthermore, note that since users do not bid themselves, the details of the physical layer and the auction are hidden from them: A user demanding a service selects among the predefined bidding functions the one that better expresses his preferences and declares a total willingness to pay $U_s$. A session that lasts for time $t_s$ is then created. Each user aims at achieving constantly the bit-rate $m$ pertaining to the selected service by participating in a large number $K_s$ of mini-auctions, where $K_s = \frac{t_s}{t_a}$. (Recall, however, that the network is bidding on each user's behalf.) For example, if the user wishes to watch his favorite music video clip lasting for 4 minutes, all he needs to declare is the title of the video, a total willingness to pay $U_s$, the desired quality level, and the utility function type, as shown in Figure 1. Requests are sent from the user's terminal to the UMTS base station over the Random Access CHannel (RACH). The parameters $t_s$, $K_s$ and $m$ = 2Mbps are computed automatically by the network and are transparent to the user. Note that a successful choice of $U_s$ is very important for the resources to be allocated to the user. Such a choice should be based on the user's urgency to receive the service, on the extent to which the content to be transported is interesting to him, on a rough estimate of the duration of the service, as well as on the user's past experience from participation in ATHENA mechanism. It is plausible, that future terminals may be equipped with software providing some related assistance or recommendation to the user.

## 3 User Utility Functions

Next, we define the utility functions offered by ATHENA mechanism. Each user selects one such function as his bidding function to be employed by the network. In Section 5, we argue that the user does not have a clear incentive not to select his actual utility function as his bidding function.

In order to simplify bidding, the utility functions we define below are expressed as sums: we assume that the user's utility (value) $u_s$ for receiving a service is the sum of the sub-utilities $u_i(x_i; h_{i-1})$ obtained in each slot $i$, each of which depends on the target bit-rate $x_i$ and the history $h_{i-1}$ of resource reservation for this service session up to the present slot; that is,

$$u_s = \sum_{i=1}^{K_s} u_i(x_i; h_{i-1}) \ . \tag{1}$$

The dependence on the *history* $h_{i-1}$ implies that when there are gaps in the resource allocation pattern, not only the quantity of resources but also the *way* these are allocated can possibly make considerable difference to the degree of user satisfaction. Thus, by selecting one of the predefined user utility functions as his bidding function, each user declares his preferred form of allocation pattern for the cases where perfectly consistent resource allocation is not possible. We consider three cases, for all of which each user is only interested in attaining a certain fixed bit-rate $m$ in each slot. (Different users may have different values of $m$.) Thus, as explained in Section 2, the corresponding quantity $q$ of resources required is $q = m \cdot t_a$, for which the network offers (on the user's behalf) an amount of money $u_i(m; h_{i-1})$, i.e. the bid $(p, q)$ placed is the pair $[u_i(m; h_{i-1})/(m \cdot t_a), m \cdot t_a]$.

**Type 1:** *Users indifferent to the allocation pattern.* This is the simplest type, and pertains to "volume-oriented" users, i.e. users whose utility solely depends on the quantity of resources allocated, as opposed to the allocation pattern. Hence the user's total value $U_s$ is equally apportioned among the various slots. That is,

$$u_i(x_i; h_{i-1}) = \mathbf{1}(x_i = m) \cdot \frac{U_s}{K_s}\ , \tag{2}$$

where $\mathbf{1}(\cdot)$ is the indicator function. Therefore, the lack of information transfer due to the gaps in the resource allocation pattern, results in proportional loss of user satisfaction. An appropriate example of users belonging to this type is those accessing news with a certain target bit-rate. Also, users downloading information by means of FTP can be considered as users of this type, provided that the FTP session is capable of resuming download when time-outs occur. This utility function is suitable for the UMTS Background Class services with $m$ pertaining to the Maximum Bit-rate parameter of this class [12]. Note that since users are "volume-oriented", a rate less than $m$ would also be useful to them. That is, if a user has the last winning bid and he can be awarded just a fraction of the target bit-rate, it is both meaningful and efficient to do so. This can be attained if the assumption of atomic bids is relaxed. On the contrary, it is obviously inefficient to assign to the higher value users a bit rate less than their maximum.

**Type 2:** *Users sensitive to service continuity.* This type pertains, for example, to users that prefer watching consistently half of a football match rather than watching several shorter periods of the total duration 45min. In general, it is applicable to services where the degree of user satisfaction depends heavily on both the volume of information transmitted and the delay observed. Thus, users of this type prefer the allocation pattern of Figure 2(a) to that of Figure 2(b), although the total quantity of resources of the two patterns is the same. In order to express this preference, we define the sub-utility function as

$$u_i(x_i; h_{i-1}) = \mathbf{1}(x_i = m)\frac{U_s}{K_s} \cdot \alpha^d \ , \tag{3}$$

where $d$ is the distance between the current and the previous slots during which this user achieved reservations and $\alpha$ is a discount factor such that $0 < \alpha < 1$. Recall that $h_{i-1}$ is the history of resource reservation for this service session, and influences $u(x_i; h_{i-1})$ through the value of $d$, which is monitored by ATHENA module. This utility function is suitable for the UMTS Streaming Class services [12].

**Type 3:** *Users sensitive to the regularity of the allocation pattern.* This type pertains to users of services such as stock-market information. Such users prefer allocation pattern Figure 2(b) to that of Figure 2(a). Indeed, the former pattern is more regular than the latter, in the sense that the total amount of resources allocated in small time windows has smaller variation. In order to express this preference, we define the sub-utility function as

$$u_i(x_i; h_{i-1}) = \mathbf{1}(x_i = m)\frac{U_s}{K_s} \cdot \alpha^{\max\{0,\Delta d\}} \,, \tag{4}$$

where $\Delta d$ is the difference of the present and the previous values of the distance defined above, and $\alpha \in (0,1)$ is again a discount factor. Note that $\alpha^{\max\{0,\Delta d\}}$ equals 1 if $\Delta d \leq 0$ and thus the received quality of service improves or remains constant, and it is less than 1 if the distance increases and hence the quality deteriorates. This utility function is suitable for users of the UMTS Interactive Class services [12].

Both Type 2 and 3 utility functions have certain features in common: $\frac{U_s}{K_s}$ expresses the additional user satisfaction per slot in cases of reservations in consecutive slots, while the discount factor $\alpha$ and its powers express the dis-satisfaction resulting from the gaps in the allocation pattern and thus the diminishing willingness to pay of the user for a new allocation.

For a user whose utility function is indeed that of Type 1 or Type 2 or Type 3, consistent allocation of resources would result in utility $U_s$. Also, for a user whose bidding is performed according to one of these functions, neither the total amount of money offered nor the total charge to be paid can ever exceed $U_s$, even in the case of perfectly consistent allocation.

The aforementioned utility functions are not the only appropriate ones to express the user satisfaction with respect to the allocation pattern attained. What is important, is the fact that the values of these utilities reflect correctly the preferences of each type of user. Thus, it can be easily seen that for a user of Type 2 the utility value obtained from equation (3) exceeds that obtained from equation (4), despite the fact that the same total quantity of resources is allocated in the two cases. Moreover, for a user of Type 3, the reverse inequality applies. Experimental assessment reveals the effectiveness of the utility functions presented above with respect to the resulting resource allocation patterns; see Section 4.

## 4 Experimental Assessment

In this section we assess experimentally ATHENA mechanism. For simplicity, we restrict attention to the resource allocation patterns of users selecting bidding functions of Types 2 and 3 (see Section 3). We have implemented special software that simulates ATHENA. We have run numerous simulation experiments according to a detailed simulation model, specifying: i) the distributions of user arrivals, departures, and service requests, and ii) the mix of users in terms of the number of users per type and the distribution of their total willingness to pay. For example, a specific scenario ran comprised 500 mini-auctions, 50 "low-value" users of Type 1, 25 "high-value" and 25 "low-value" users of Type 2 and 50 "medium-value" users of Type 3. For each user, the total willingness to pay is randomly selected according to a uniform distribution over the intervals $[1, 100]$, $[101, 200]$ and $[201, 300]$, depending on whether the user is of low, medium or high value respectively. The total quantity of resource units available at each mini-auction also fluctuates (due to the varying allocation of resources to phone calls and SMS/MMS) in the simulation model, and is selected according to the uniform distribution over the interval $[80, 120]$. Finally, in each experiment, the value of the discount factor $\alpha$ in user utility functions [see equations (3) and (4)] was in the range $[0.91, 0.99]$.

Experiments confirm that when a user's bids are always higher (resp. always lower) than each mini-auction's lowest winning bid (i.e., "cut-off" price), then the allocation pattern was perfectly consistent (resp. the user was allocated no resources at all), as expected. Furthermore, experiments have revealed that our mechanism has the important property that the vast majority of the users' resource allocation patterns either are *nearly consistent* or comprise *very limited* quantities of resources. Note also that the congestion level of the network (i.e. total *demand* for resources), affects the percentage of the population of users that receives very satisfactory service. However, the congestion level has very limited effect on the aforementioned bi-modal distribution of the resource allocation patterns. “Intermediate” allocation patterns in general arise rarely. (Such patterns are those that are not nearly consistent and yet comprise a non-negligible percentage of the total quantity of resources that would arise for the same user in the case of consistent allocation.) The distribution of allocation patterns for various congestion levels is depicted in Figures 6, 7 and 8. The total percentage of users being allocated intermediate patterns does not exceed 30%, while this percentage is typically considerably lower[3]. In fact, intermediate patterns mostly arise in cases of low congestion and they are “harmless” in the sense that their associated charge is typically very low. Also, the higher the level of congestion, the higher (resp. lower) the percentage of users allocated almost no resources (resp. the percentage of users that receive very satisfactory service), as depicted by the leftmost (resp. rightmost) bar of each of Figures 6, 7 and 8. Hence, ATHENA serves as a “soft” call admission control mechanism actually blocking users with non-competitive total willingness to pay from receiving service by the network, while leading to nearly consistent allocation for the competitive users. This property is due to the expression of the bidding functions, as explained below.

Next, we discuss the placement of gaps in cases of nearly consistent patterns. The most interesting such patterns arise for users whose bids are often close to the cut-off prices of the

[3] This is the percentage of the resources allocated to each user with respect to the total quantity of resources demanded by *this user* during his service session.

various mini-auctions, thus being vulnerable to entries of new users and to the fluctuations of the total quantity of resources available. Experiments showed the effectiveness of our mechanism for such users with respect to allocating them patterns according to the preferences expressed by their bidding functions. In particular, successful users with Type 2 bidding function receive almost consistent allocation of resources for large time intervals (see Figure 4) while those with Type 3 bidding function receive more fragmented allocation patterns, similar to those depicted in Figure 5. Note also that due to the expressions of these functions [see equations (3) and (4)] when a gap starts in the resource allocation pattern, then the bids of the user are reduced due to the discount factor $\alpha$. If the user is not competitive enough, then this discounting prevails for several consecutive slots, the bids become very low and the user essentially quits service. This is preferable than this user receiving mediocre service with large gaps in the resource allocation pattern. Note that in all experiments the competition encountered by each individual user was a stationary process. This would also apply to cases of real networks with large populations of competing users. Hence, in general, a user is either competitive for almost all of his service session or essentially he is not competitive at all. Due to the expressions of the bidding functions of our mechanism, non-competitive users essentially quit early enough, without insisting in competing almost hopelessly.

## 5 User Incentives

A user participating in ATHENA in order to reserve resources has to select one of the predefined utility functions as his bidding function and declare a total willingness to pay $U_s$, which parameterizes this bidding function. In this section, we deal with the issue of user incentives. Consider a user whose actual preferences are indeed expressed by one of the predefined utility functions. Has such a user the incentive to *truthfully* declare this particular function as well as the *true* value of his total willingness to pay? Below, we provide considerable evidence that this is often the case, although we cannot prove formally that this applies always.

First, consider a user whose utility function is of Type 1 (see Section 3). Then, his sub-utility $u_i(m; h_{i-1})$ from attaining his targeted bit-rate $m$ in a particular slot $i$ equals $\frac{U_s}{K_s}$, and is independent of the rest of the allocation pattern. Thus, by the incentive compatibility property of the Generalized Vickrey Auction, it can be established that the optimal bid for this user is an amount of money equal to $\frac{U_s}{K_s}$. In fact, this is a dominant bidding strategy for the user in order to maximize the expected value of his net benefit. This is defined as the utility resulting by the resource allocation pattern minus the charge paid for it.

Consider now a user whose utility function is either of Type 2 or of Type 3. The reservation of resources in one slot brings both instant value (sub-utility) to the additive user utility function as well as extra value to resources to be reserved in the future by improving the overall allocation pattern. [This improvement is due to the resulting reduction of the values of $d$ and $\Delta d$ in equations (3) and (4) respectively.] Therefore, the value for this user from attaining his targeted bit-rate $m$ in a particular slot $i$ is dependent on the rest of the resource allocation pattern. Clearly, placing a bid less than the corresponding sub-utility in a particular mini-auction is less beneficial for this user than placing a truthful bid, because this only increases the risk of losing without affecting the payment. It could be argued however that a bid higher than the sub-utility should be submitted, due to the extra value that a present resource reservation (which becomes more likely by over-bidding) would bring to future ones. However, using techniques on sequential games, it can be proved that this would *not* be the case for "extremely uncertainty averse" (i.e. extremely conservative) users, who (by definition) in cases of choice/behavior under uncertainty maximize the net benefit of the worst possible outcome. Note that this "maximin" behavior was proposed by [10] for situations of complete uncertainty, and was adopted by several other researchers (e.g. see [2] and [8]). On the other hand, for a user aiming to maximize his expected net benefit, truthful bidding in each slot would not be the optimal strategy. This may lead such a user not to select his own utility as his bidding function, as another of the functions may ex post turn out to be more profitable. However, this does not seem likely to be the case for the reason explained in the next paragraph. Finally, it can be easily seen that if the user selects truthfully his bidding function, then he will also declare truthfully his total willingness to pay.

So far, we have dealt with the incentives of users aiming to optimize either their expected net benefit or the net benefit of the worst possible outcome. Consider now a user whose utility function is of either Type 2 or Type 3 and is interested in getting the best possible service without exceeding his actual total willingness to pay. According to the experimental results of Section 4, if this user is indeed allocated a non-negligible quantity of resources, then he will most likely be allocated a nearly consistent pattern whose gaps will be in accordance to the pattern preferences expressed by his selected bidding function. Therefore, this user will be better off by truthfully revealing his utility function and total willingness to pay.

## 6 Extensions of ATHENA mechanism

In Section 3, we have defined utility functions suitable for users seeking for strict guarantees, i.e. users whose bit-rate has a single target-value. Below, we deal with extending the utility function of Type 2 for the case of two alternative bit-rates; Type 3 can be extended similarly.

**Type 4:** *Extension of Type 2 for two alternative bit-rates.* We consider a user who is willing to watch video either with the "high quality" bit-rate $r_{\text{high}}$, or with just the "good quality" bit-rate $r_{\text{good}}$, whenever the former is not feasible. Watching the video consistently with either "good quality" or "high quality" results in different degrees of user satisfaction; hence, the total willingness to pay respectively equals $V_{\text{good}}$ and $V_{\text{high}} = V_{\text{good}} + \Delta V$, where $\Delta V > 0$.

Figure 3 depicts a typical allocation pattern arising when consistent resource allocation with bit-rate $r_{\text{high}}$ is not possible. This pattern can be viewed as the superposition of two allocation sub-patterns: one with bit-rate equal to either 0 or $r_{\text{good}}$ and another sub-pattern with bit-rate equal to either 0 or $r_{\text{high}} - r_{\text{good}}$. Since the user is still sensitive to service continuity, but there are now two possible bit-rates, we can extend the user utility function of equation (3) as follows:

$$u(x_i; h_{i-1}) = \mathbf{1}(x_i \geq r_{\text{good}}) \frac{V_{\text{good}}}{K_s} \cdot \alpha^{d_1} + \mathbf{1}(x_i = r_{\text{high}}) \frac{\Delta V}{K_s} \cdot \alpha^{d_2} \ , \tag{5}$$

where $d_1$ and $d_2$ are defined with respect to the length of the gaps incurred in the two aforementioned sub-patterns. This utility function is suitable for the UMTS Conversational Class [12], with $r_{\text{good}}$ and $r_{\text{high}}$ corresponding to the Guaranteed and the Maximum Bit-rates of this class; however, the lower bit-rate $r_{\text{good}}$ is not guaranteed by ATHENA. Furthermore, for each user of Type 4, there should be placed two *summable* bids, i.e. $(p_{\text{good}}, q_{\text{good}})$ and $(p_{\text{extra}}, q_{\text{extra}})$, expressing his willingness to pay (per unit) for the basic bit-rate $r_{\text{good}}$ and his extra willingness to pay for the extra bit-rate $r_{\text{high}} - r_{\text{good}}$, respectively. However, the latter bid is taken into account only when the former is a winning bid. In general, the number of atomic bids to be given on behalf of each user in each mini-auction equals the number of alternative bit rates, which in the general case may be higher than two.

We have explained that ATHENA cannot provide strict guarantees of quality of service

(QoS) on an end-to-end basis. It might be desirable though to do so for a limited number of high-value users, to be referred to as "gold" users. This can be accomplished by means of allowing some ATHENA users to declare an "infinite", i.e. extremely high, willingness to pay for the service requested. This option could be made available for a limited number of users that are both insensitive regarding their charges, and highly demanding regarding the QoS they experience. Apart from their usage-based charge[4], the provider should charge these users an additional *high* monthly premium in order for them to subscribe to this service of strictly guaranteed quality. Since ATHENA bidding module will be submitting the highest bids for these users, they will always win at all the mini-auctions they participate and receive perfect quality at their terminals. However, this option should be offered to very few users, the exact number of which depends on the minimum total quantity of resources available in all slots. Note also that the introduction of this service option results in *service differentiation* for the users, according to their type (namely "gold" or "normal"). This can be generalized by introducing different prespecified options for the willingness to pay, each corresponding to a certain class of users. Hence, ATHENA would serve as a *price discrimination* mechanism enabling DiffServ.

It is also possible to modify ATHENA so that it *favors* the users already receiving service

[4] Note that for these users too the usage-based charge equals the social opportunity cost as computed in the various auctions, which however may be very high in times of congestion.

versus those initiating service sessions. This should be done in a way that the primary objective of efficiency is not considerably affected. The idea is for each user to gradually increase the value of $\alpha$ (towards 1) as the user's time in the system increases. This way it becomes harder for new users to displace those that have already received part of their service.

Finally, note that throughout the paper we have assumed that a service session is served within the same cell for its entire duration. However, *handovers* can also be easily accommodated: the selected bidding function of the session should be passed to the new cell, together with the associated parameters and the current values of distance $d$ (for Type 2), $d$ and $\Delta d$ (for Type 3) and $d_1$, $d_2$ (for Type 4).

## 7 Concluding Remarks

In this paper we presented ATHENA, an effective auction-based mechanism for nearly consistent reservation of the resources of a UMTS network by the users that value them the most. The mechanism can deal successfully with long time scale data, audio and video services in practical cases of networks with large numbers of competing users. It is based on a series of Generalized Vickrey Auctions and a set of new user utility functions expressing the user's preferences with respect to the resource allocation patterns. We also provide a mapping of these functions to the UMTS service classes. Finally, we have outlined several extensions of our mechanism, mostly regarding the guarantees provided to the users. Assessment of these extensions constitutes an interesting direction for future research.

## References

[1] J. Crémer and C. Hariton, “The pricing of critical applications in the Internet”, Journal of the Japanese and International Economies, 13:4, 281-310, December 1999.

[2] J. Dow and S. R. da Costa Werlang, “Uncertainty Aversion, Risk Aversion, and the Optimal Choice of Portfolio”, Econometrica, 60:1, 197-204, January 1992.

[3] D. Friedman, “Evolutionary Games in Economics”, Econometrica, 59:3, 637-666, May 1991.

[4] T. Groves and J. Ledyard, “Optimal Allocation of Public Goods: A Solution to the ‘Free Rider’ Problem”, Econometrica, 45:4, 85-96, May 1997.

[5] H. Holma and A. Toskala, “WCDMA for UMTS: Radio Access for Third Generation Mobile Communications”, John Wiley, ISBN 0-471-72051-8, September 2000.

[6] A. Lazar and N. Semret, “Design, Analysis and Simulation of the Progressive Second Price Auction for Network Bandwidth Sharing”, Columbia University, April 1998.

[7] J. MacKie-Mason, “A Smart Market for Resource Reservation in a Multiple Quality of Service Information Network”, Technical Report, University of Michigan, September 1997, URL: http://www-personal.umich.edu/ jmm/research.html.

[8] J. Rawls, “A Theory of Justice”, Cambridge: Harvard University Press, 1971.

[9] S. Soursos, C. Courcoubetis, and G.C. Polyzos, “Differentiated Services in the GPRS Wireless Access Environment”, Proceedings of IWDC 2001, Evolutionary Trends of the Internet, September 17-20, 2001, Italy.

[10] A. Wald, Statistical Decision Functions, New York: John Wiley, 1950.

[11] The 3rd Generation Partnership Project (3GPP), URL: http://www.3gpp.org/.

[12] 3GPP, Technical Specification Group Services and System Aspects, QoS Concept (3G TR 23.107 version 5.6.0), URL: http://www.3gpp.org/.

# Appendix: Resources in UMTS and GPRS

In this Appendix we define the *slots* and *units* of allocation of our auction-based resource reservation mechanism when applied in UMTS and GPRS networks. More related information can be found in [5], [9], [11] and [12]. Figure 9 illustrates the 10msec WCDMA frame in UMTS. The entity that allocates the network resources within each frame among user connections is the Resource Manager. In GPRS (and EDGE), user data services compete for reserving a number of *time slots*, with the unit of allocation being the *Radio Block*. Users attempt to reserve these resources in order to accommodate their large time scale data traffic; this is transferred in bursts, by means of creating Temporary Block Flows within the PDP context, as depicted in Figure 10.

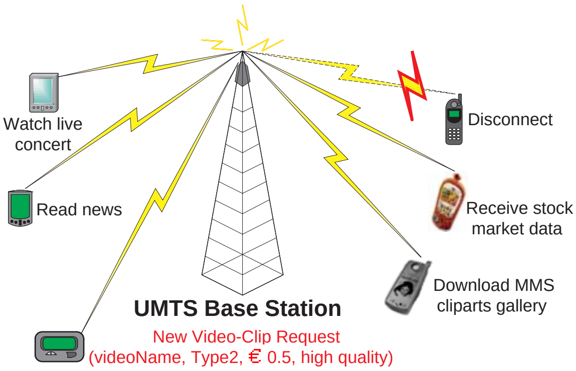

Watch live concert
Disconnect
Read news
Receive stock market data
Download MMS cliparts gallery
UMTS Base Station
New Video-Clip Request
(videoName, Type2, € 0.5, high quality)

Figure 1: A UMTS cell and its served users.

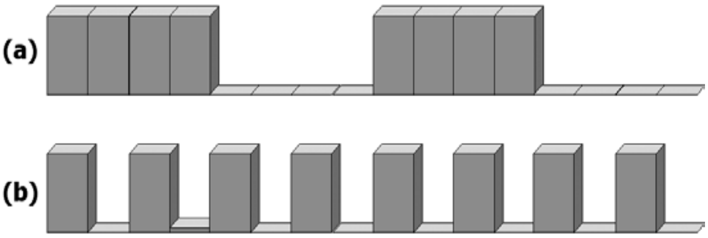


Figure 2: Inconsistent resource allocation patterns.

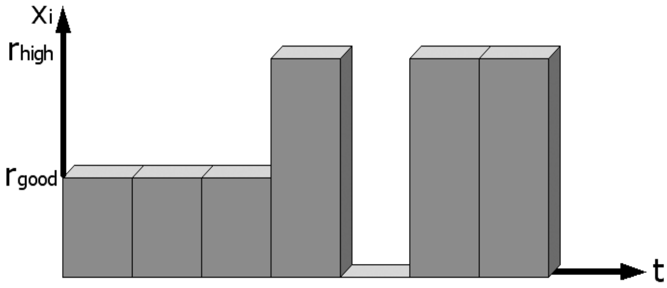
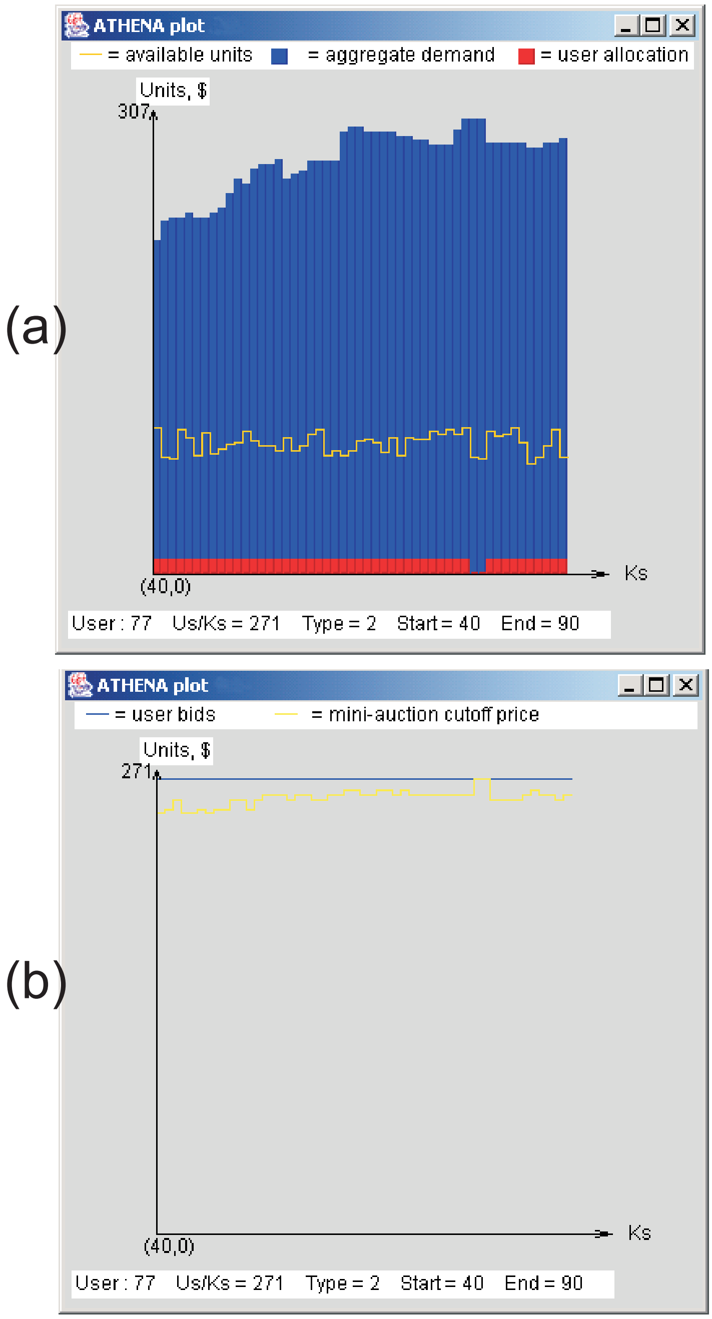


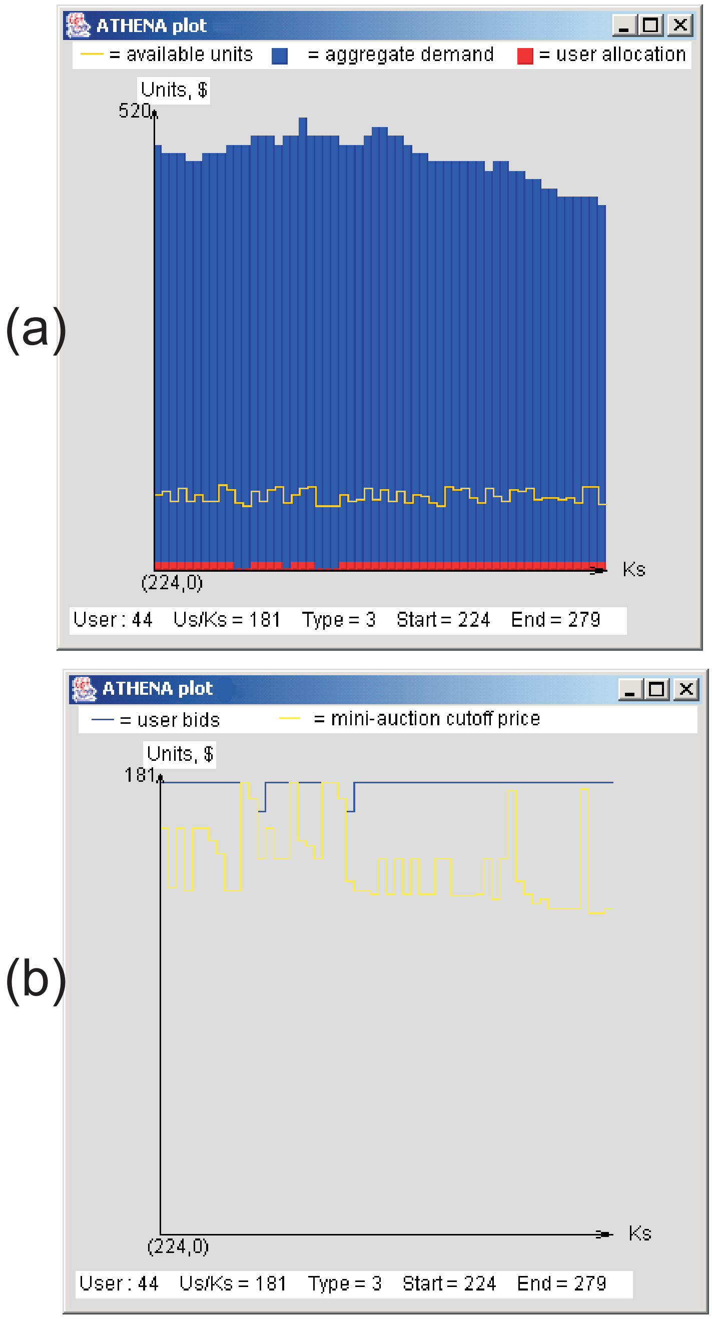
Figure 3: A resource allocation pattern with two acceptable bit- rates.



User : 77 Us/Ks = 271 Type = 2 Start = 40 End = 90

User : 77 Us/Ks = 271 Type = 2 Start = 40 End = 90

Figure 4: Screenshot (a) depicts the total available resource units (highest bars), the aggregate demand for resources (line), and the allocation pattern (lowest bars) of a user with Type 2 bidding function whose bids are in general competitive. Screenshot (b) depicts this user's bids (darker line) and the mini-auctions' cut-off prices throughout his service session.



User : 44 Us/Ks = 181 Type = 3 Start = 224 End = 279

User : 44 Us/Ks = 181 Type = 3 Start = 224 End = 279

Figure 5: Screenshots (a) and (b) depict the same type of information as in Figure 4 for a user with Type 3 bidding function.

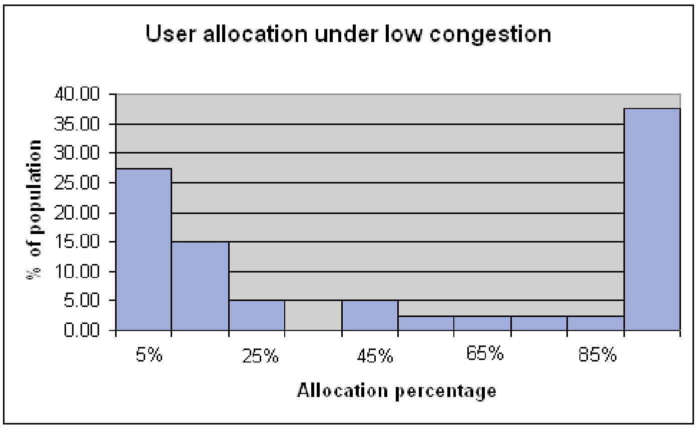


Figure 6: Distribution of resource allocation patterns with respect to the percentage of the total targeted resources that was attained under low network congestion.

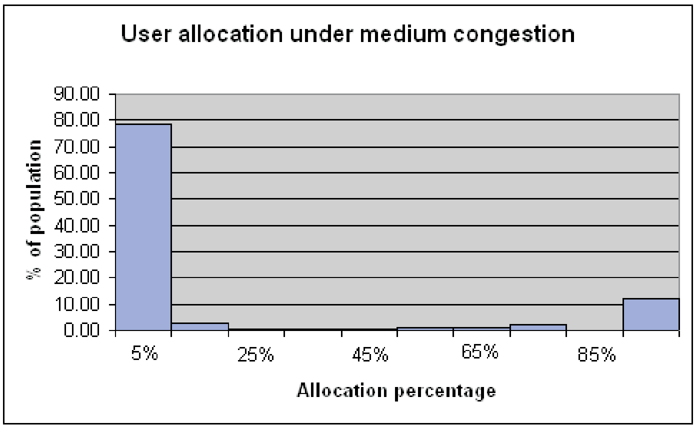


Figure 7: Distribution of resource allocation patterns with respect to the percentage of the total targeted resources that was attained under medium network congestion.

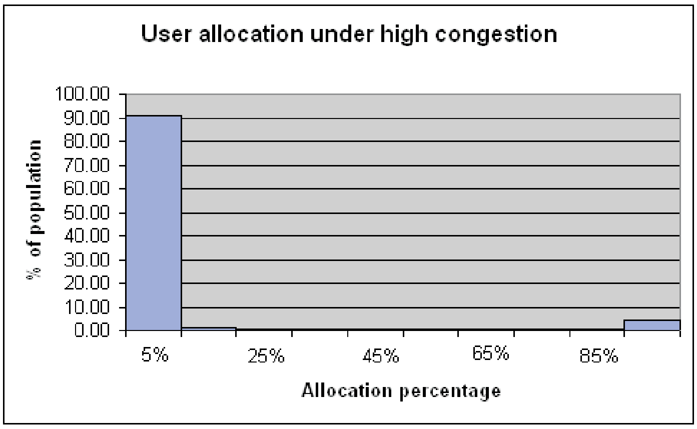


Figure 8: Distribution of resource allocation patterns with respect to the percentage of the total targeted resources that was attained under high network congestion.

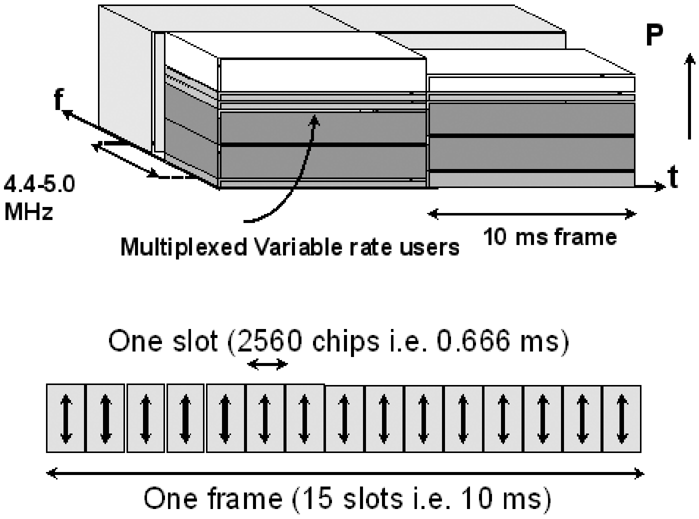

P
f
t
4.4-5.0
MHz
Multiplexed Variable rate users
10 ms frame
One slot (2560 chips i.e. 0.666 ms)
One frame (15 slots i.e. 10 ms)

Figure 9: Allocation of resources in UMTS [5].

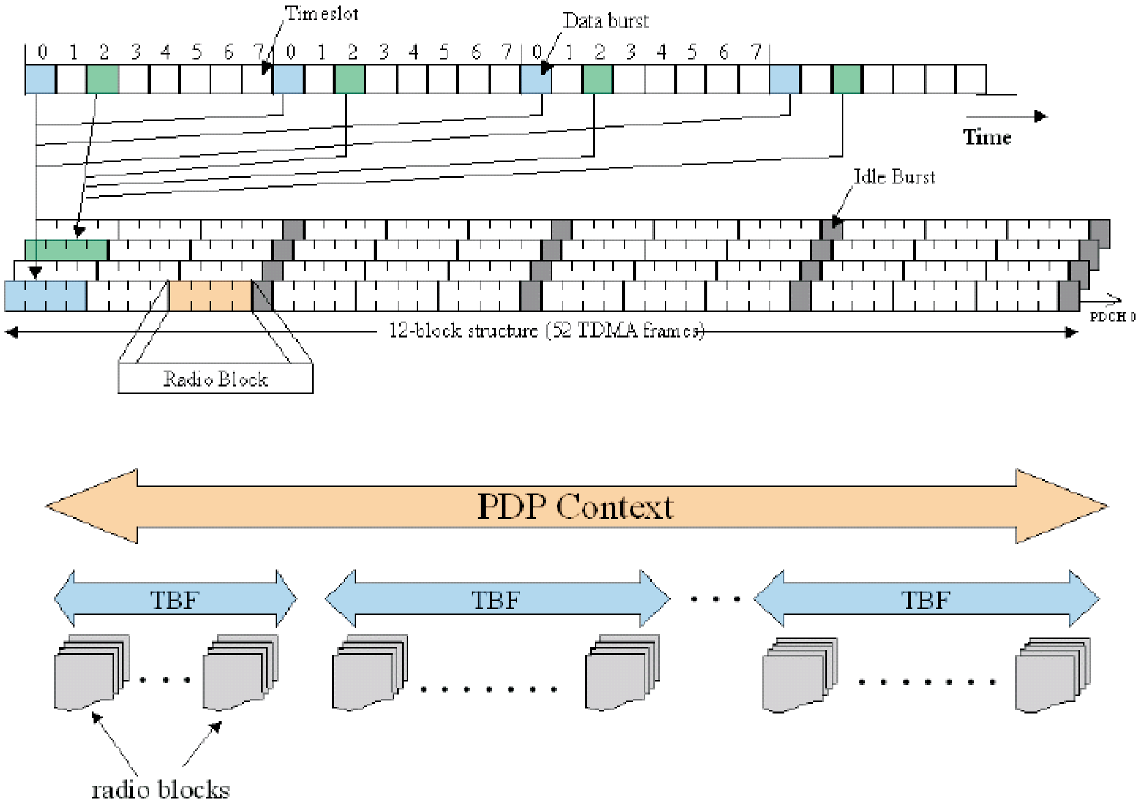


Figure 10: Allocation of resources in GPRS [9].