Dynamic Bandwidth Allocation Algorithms for High-Speed Data Wireless Networks

Matthew Andrews, Sem Borst, Francis Dominique, Predrag Jelenkovic,
Krishnan Kumaran, K.G. Ramakrishnan, Phil Whiting
Bell Labs, Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974

Abstract

Next-generation wireless networks are expected to support a wide range of high-speed data services, Web browsing being one of the major applications. Although high data rates have been shown feasible in a single-user setting, the resource allocation issues that arise in a multiple-user context remain extremely challenging. As compared with voice, data traffic is typically more bursty, while the users are less sensitive to delay. These characteristics require resource allocation strategies to operate in a fundamentally different manner in order to achieve efficient spectrum utilization.

The paper proposes several algorithms for scheduling the transmission of data to multiple users in an efficient manner. As a novel feature, the various schemes exploit knowledge of the buffer contents in order to achieve high throughput, while maintaining fairness by providing quality-of-service to individual users. The proposed algorithms are backward compatible with existing cellular and PCS standards such as IS-136, and provide a powerful approach to improve spectrum efficiency in forthcoming high-speed data complement services to existing cellular systems. Extensive simulation experiments demonstrate that the algorithms significantly outperform conventional schemes.

1 Introduction

Present mobile cellular systems only allow a limited capability for the transmission of data, see for instance Alanko *et al.* [1] and Mathias [12]. GSM and IS-95 standards for example permit transmission only up to a rate of approximately 10 kbps. This rate is too low to satisfy the requirements of many applications. Here we will present algorithms for dynamically allocating carriers to mobiles so as to carry data with peak rates of the order of 400 kbps. These algorithms can be implemented in conjunction with a group of narrow band Time Division Multiple Access (TDMA) carriers, such as the North American IS-136 system or the European GSM system. Indeed, they can be utilized in existing networks and operated within modest bandwidth requirements of a few MHz.

There are two important distinctions to be made between voice and data traffic. The first is that data traffic is much more bursty than voice. The second is that data traffic has quite different Quality-of-Service (QoS) requirements from voice. On the one hand, it is much more tolerant to delay than voice; response times of several seconds are acceptable for data, whereas voice signals can only be delayed a fraction of a second. On the other hand, data is much more sensitive to bit errors, requiring error rates of 10^{-6}, whereas voice can be satisfactorily transmitted at 10^{-3}.

These distinctions lead to very different approaches for bandwidth allocation. In particular, the allocation of a dedicated connection to a data user is less reasonable. If a low-speed connection is provided, then delays are unacceptably high; providing a high-speed connection reduces delay, but cannot be used efficiently because of the burstiness. However, the fact that data applications are tolerant of delay provides the possibility of coordinating base stations and scheduling transmissions. Such an approach has also been suggested by Linnartz [11], who, however, made a proposal involving tight synchronization at the level of time slots to be used in conjunction with a reservation scheme. An alternative collision resolution approach is also described in [11].

Of course, sophisticated scheduling would not be necessary if Fixed Channel Assignment (FCA) provided an adequate QoS at a cost acceptable to network operators. This is not the case however. In FCA, the base stations are partitioned into re-use groups, i.e., groups of mutually non-interfering cells, and the carriers are statically divided among the re-use groups. Since the carriers are allocated on a fixed basis, the approach lacks flexibility. Once the number of mobiles has reached the maximum which can be supported on the carriers allocated to a base station, further requests for service must be rejected. In reality, of course, a simple reallocation of carriers might have enabled these extra mobiles to have been supported with little cost to other network users. A similar difficulty arises if there is a build-up of packets at a particular base station, which can occur even if the mobiles are evenly distributed because of the burstiness of data traffic. Once again, extra carriers cannot be allocated, since the number at each base station is fixed, even though few packets are being presented for transmission at base stations nearby. This is in contrast to voice systems, where the peak bandwidth requirements are fixed at a relatively low level. Thus, the inefficiency of FCA in terms of spectrum utilization precludes its use by operators.

The algorithms we present here are for wireless networks operated according to a frequency re-use plan, but do allow for Dynamic Channel Assignment (DCA). In DCA, carriers are not statically assigned to base stations, and may be diverted in a dynamic fashion from cells with traffic lulls to cells with traffic peaks. A frequency re-use plan offers the advantage that interference between co-channel users is limited by geographic distance according to worst-case conditions. There is no need to obtain on-line propagation or interference measurements. The sole problem remaining is one of scheduling subject to these re-use constraints. Unfortunately, the optimal solution to such a scheduling problem is not known. Furthermore, even if it were known, it would be of such complexity that it would be impractical to implement. Our approach therefore has been to devise a number of heuristics. Each determines carrier allocations to base stations for a fixed period on the basis of the number of packets awaiting transmission. Once allocated, a common algorithm determines how the carriers are used to transmit the mobile packets themselves. The work by Harris [8] addresses the related and easier question of how to schedule a given set of packets for transmission onto carriers with an already determined allocation of carriers to base stations. The model in [8] includes the possibility of mobiles being able to receive packets from more than one base station, which is not considered here. Of course, the problem of how to assign carriers to base stations and the problem of then assigning these carriers to mobiles are closely coupled in the circumstances described in [8].

The remainder of the paper is organized as follows. In Section 2, we present a detailed model description, and then discuss the issues surrounding the traffic model. The next three sections propose a number of heuristic carrier allocation algorithms. In Section 6, we evaluate the communication requirements and computational complexity of the various algorithms. In Section 7, we describe a series of simulation experiments, in which we examine the impact of the burstiness and the duration of the scheduling interval on the delay and throughput performance. The conclusions are presented in Section 8.

2 Model description

We consider a wireless network of N base stations supporting M users. Packets destined for the users arrive at the serving base station, where they are queued until transmission. (Throughout, we focus on the stream of packets from the base station down to the users. We do not consider the flow of packets from the users up to the base station.)

The base stations share a pool of K orthogonal carriers for transmitting packets down to the users. Transmission occurs in a slotted fashion, with exactly one packet transmitted on a single carrier during a time slot.

Base stations may re-use carriers subject to certain interference constraints. We assume the re-use constraints may be described in the form of an *interference graph* with the vertices corresponding to the base stations and the edges representing the pairs of interfering base stations. Thus, if the interference graph contains the edge $\{j_1, j_2\}$, then base stations j_1 and j_2 are barred from transmitting on the same carrier in the same time slot. Such models of networks of queues with service constraints determined by a graph also appear in the context of input-queued switches, see for instance Tassiulas [14].

The scheduling policy determines the assignment of carriers to users in each time slot. Let $X_{jkt} \in \{0, 1\}$ indicate whether or not carrier k is allocated to base station j in the t-th time slot. The re-use constraints require $X_{j_1kt} + X_{j_2kt} \leq 1$ for all edges $\{j_1, j_2\}$ in the interference graph in order for the assignment to be feasible.

Let $Y_{ikt} \in \{0, 1\}$ indicate whether or not carrier k is assigned to user i in the t-th time slot. Then $\sum_{i \in \mathcal{M}_j} Y_{ikt} \leq X_{jkt}$ for all $j = 1, \ldots, N$, with $\mathcal{M}_j$ representing the set of all users served by base station j. Define $Z_{it} := \sum_{k=1}^{K} Y_{ikt}$ as the number of carriers assigned to user i in the t-th time slot. A user may not be assigned more than L carriers simultaneously, i.e., $Z_{it} \leq L$.

Denote by Q_{it} the queue length for user i at the start of the t-th time slot, i.e., the number of packets for user i waiting for transmission. The queue lengths evolve over time as

$$Q_{i(t+1)} = A_{it} + [Q_{it} - Z_{it}]^+,$$

where $[z]^+ := \max\{z, 0\}$ and A_{it} is the number of packets that arrive for user i during the t-th time slot.

Suppose that a packet for user i arrives during the t'-th time slot into a queue of length $Q_{it'}$. The delay experienced by that packet, measured in time slots, is then defined to be $t'' - t'$, with

$$t'' := \min\{t : \sum_{t=t'+1}^{t} Z_{it} \geq Q_{it'}\}.$$

Traffic model The most basic statistical model that that is able of capturing data 'burstiness' and complex dependency structures is the so-called On-Off model, first investigated by Cohen [5] and Rubinovitch [13]. For two-state Markovian On-Off models, Anick, Mitra, & Sondhi [2] analyzed the impact of the burstiness on queueing performance. Subsequent studies explored more general Markov models with finite state space. These led to the Equivalent-Bandwidth theory for Markovian arrival processes, see for instance Elwalid *et al.* [6] and Glynn & Whitt [7].

Statistical analysis has provided increasing evidence that the traffic streams in modern broadband networks exhibit long-tailed (subexponential) characteristics. For the case of Ethernet traffic, such findings were reported by Willinger *et al.* [15]. These statistical observations have stimulated research in queueing analysis under heavy-tailed (non-Cramér) assumptions. We refer to Jelenkovic & Lazar [9] for recent results on multiplexing On-Off sources with subexponential characteristics and an extensive list of references on subexponential models.

It is difficult to predict the predominant applications and statistical characteristics that will govern future networks. Consequently, we have conducted our experiments using both exponential and subexponential (long-range dependent) models.

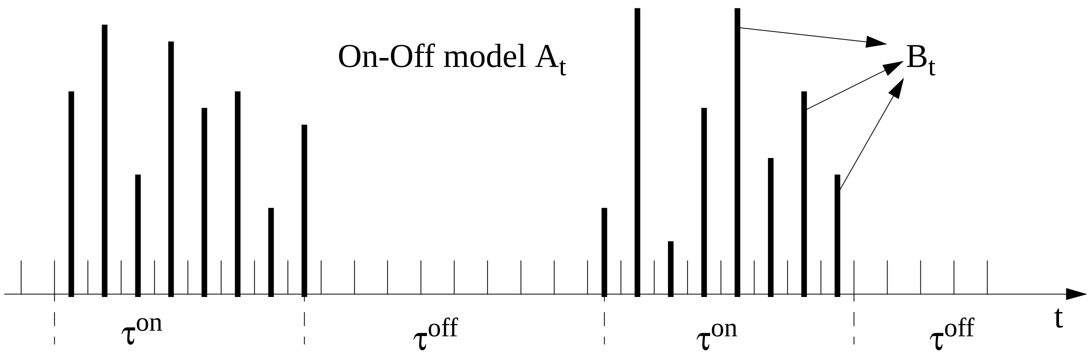

Figure 1: Sample path of an On-Off process.

τ^{off}	τ^{on}	B_t
Exponential	Exponential	Poisson
Exponential	Pareto	Poisson

Table 1: Combinations of the distributions that will be used in the experiments.

Here, we present a formal definition of an On-Off process. Consider two independent sequences of i.i.d. random variables $\{\tau_n^{off}, n \geq 0\}$, $\{\tau_n^{on}, n \geq 0\}$, $\tau_0^{off} = \tau_0^{on} = 0$. Define a point process $T_n^{off} := \sum_{i=0}^{n}(\tau_i^{off} + \tau_i^{on})$, $n \geq 0$; this process will be interpreted as representing the beginnings of Off periods in an On-Off process. Next, let $\{B_t, t = 0, 1, \ldots\}$ be a sequence of i.i.d. random variables. Then, a discrete time On-Off process A_t is defined as

$$A_t = B_t \text{ if } T_n^{off} - \tau_n^{on} \leq t < T_n^{off}, \quad n \geq 1,$$

and $A_t = 0$, otherwise. Figure 1 illustrates the sample path of an On-Off process.

Table 1 presents the combinations of the various distributions function that will be used to model τ^{off}, τ^{on}, and B_t. The reason that we choose only the exponential distribution for the Off periods is because large queue build-ups are insensitive to the distribution of Off periods (for a rigorous treatment of this insensitivity phenomenon see [9]). In contrast, the queueing behavior is strongly affected by the duration of the On times, with a clear dichotomy between exponential and subexponential distributions. The fluctuations in the peak data rate are described by a Poisson distribution. In general, for all different choices of distributions and their parameters, it will be shown that the DCA schemes significantly outperform the FCA approach.

3 Preliminaries

We now proceed to describe various schemes for scheduling the transmission of packets to the users. In view of the cell-by-cell re-use constraints, the problem may be decomposed into two parts: (i) allocation of carriers to base stations; (ii) assignment of carriers to individual users, given the number of carriers allocated to the base stations.

In the present paper, we focus on the first problem; in the next two sections, we will describe various schemes which all differ in the way they allocate carriers to base stations. The approach to the second problem is common to all these schemes; we assume that the assignment of carriers to users is proportional to the queue lengths, subject to the constraint that no user may be assigned more than L carriers.

Specifically, suppose that base station j has been allocated K_j carriers. First, we assign $L_i = \min\{L, \lfloor K_j Q_i / \sum_{m \in \mathcal{M}_j} Q_m \rfloor\}$ carriers to each user $i \in \mathcal{M}_j$ served by base station j. In case there are any carriers remaining, i.e., $\sum_{i \in \mathcal{M}_j} L_i < K_j$, we assign them to the user $i \in \mathcal{M}_j$ with the largest ratio of Q_i/L_i among those with $L_i < L$, We repeat the above procedure until either there are no carriers remaining, or $L_i = L$ for all users $i \in \mathcal{M}_j$.

There are of course numerous alternative approaches for assigning carriers to users. Since no user may be assigned more than L carriers, capacity may potentially be wasted if there are only a few users with non-empty queues. Thus, from the point of view of throughput, it seems attractive to assign carriers to the user with the longest queue. From a fairness perspective, however, it appears more reasonable to assign a comparable number of carriers to each user. The proportional assignment scheme described above may be viewed as an intermediate strategy with regard to these two extremes. Among the class of work-conserving schemes, we do not expect any significant variation in *average-delay* performance as long as the value of L is not extremely small.

Since the above procedure for assigning carriers to users does not require any exchange of information or complex calculations, we assume that the assignment is done every time slot. In contrast, we assume that the allocation of carriers to base stations is done only once every S time slots, for some parameter S, since the process may involve communication overhead and extensive computations. We refer to such a scheduling interval of S time slots as a *superframe*. The parameter S may be different for different schemes, as it will depend on the communication and computational complexity involved. A large value of S provides for a more elaborate allocation process, while a small value of S allows for a quicker response to congestion conditions.

Each scheme uses a superframe structure, and so there must be global synchronization of the base stations. This need not be very accurate if superframes are comparatively long (i.e. if S is large). Long superframes introduce some latency, but this does not increase delay significantly as long as the mean burst period is longer than the superframe duration.

All schemes for assigning carriers to base stations must face up to the engineering trade-off between the amount of communication and computation required and the quality of the assignment that is made. Thus, it may be advantageous to design a scheme which makes assignments of carriers to base stations much more rapidly, but in a less efficient way than a scheme which does the reverse.

Average-delay bounds Determining the delay-minimizing allocation scheme is prohibitively demanding for all but the simplest cases. Hence, a bound for the achievable average delay is instrumental in evaluating the performance of heuristics.

A simple but effective delay bound may be derived as follows. Consider a clique of base stations in the interference graph, i.e., a group of base stations that all mutually interfere. By definition, carriers cannot be re-used in the clique. Thus, the aggregate queue length in the clique cannot be any smaller than in case the aggregate packet arrival stream were served by K dedicated carriers in isolation. Hence, by Little's law, the average packet delay in the clique is also bounded below by the average delay if the aggregate packet arrival stream were served by K isolated carriers.

Now suppose that the interference graph is symmetric, and that the maximum clique size is F. In that case, the network-average delay is bounded below by the average delay if the superposition of F arrival streams were served by K isolated carriers. We refer to the latter delay bound as the *pooling bound*. Tighter bounds may be derived along similar lines, see Borst & Whiting [3].

4 Distributed carrier assignment

We now outline two distributed assignment schemes, which only require local exchange of information between base stations for allocating carriers. In the next section, we will present a centralized allocation scheme, which uses global knowledge in assigning carriers to base stations.

Both schemes use the concept of re-use groups. We assume that the base stations are partitioned into F re-use groups $\mathcal{N}_1, \ldots, \mathcal{N}_F$. A re-use group is an independent set in the interference graph, i.e., a group of mutually non-interfering base stations.

In the first scheme, the carriers are also partitioned into F sets $\mathcal{K}_1, \ldots, \mathcal{K}_F$, each associated with a particular re-use group. The carriers are designated, but not restricted, for use by the base stations in the corresponding re-use group. On a request basis, base stations may borrow carriers that are currently not needed by the designated owners. This scheme will henceforth be referred to as *distributed carrier requesting*.

In the second scheme, the carriers are not associated with any particular group of base stations, but freely floating. At the start of every superframe, the base stations in a particular re-use group grab all the carriers that are currently not used by any interfering base stations. This scheme will be termed as *distributed carrier raking*.

Of course, there are also hybrid schemes conceivable where some carriers are designated while others are completely floating. For ease of presentation, however, we restrict the attention to the two extreme schemes mentioned above.

Both schemes also use the concept of *demand* for carriers by the base stations. At the start of every superframe, base station j calculates the demand for carriers in the next superframe as

$$D_j = \lfloor \sum_{i \in \mathcal{M}_j} \min \left\{ L, \frac{Q_i}{S} \right\} \rfloor.$$

The demand is the number of carriers that base station j could acquire without risking any capacity to go unused in the next superframe.

We now proceed to describe the two proposed schemes in detail.

4.1 Distributed carrier requesting

In this scheme, base stations may borrow carriers that are currently not needed by the designated owners. The scheme consists of the following three steps, performed at the start of every superframe.

Restricted retainment of designated carriers Base station j acquires $\min\{D_j, K/F\}$ designated carriers, with K/F representing the number of carriers associated with each re-use group.

Re-use groups

Carrier sets	N_1	N_2	N_3	N_4
K_1	A D_0 R	B_3 D_3	B_2 D_2	B_1 D_1
K_2	B_1 D_1	A D_0 R	B_3 D_3	B_2 D_2
K_3	B_2 D_2	B_1 D_1	A D_0 R	B_3 D_3
K_4	B_3 D_3	B_2 D_2	B_1 D_1	A D_0 R

Figure 2: Distributed carrier requesting process; A: acquisition of designated carriers; D: declaration of used carriers; B: borrowing unused carriers; R: reclaiming designated carriers.

Restricted acquisition of non-designated carriers Next, base stations may borrow carriers that are not already claimed by the designated owners. The process is organized in $F - 1$ rounds. In the n-th round, the base stations in re-use group $\mathcal{N}_f$ concurrently acquire the carriers in the set $\mathcal{K}_{f+n}$ that have not yet been claimed by any interfering base stations in re-use groups $\mathcal{N}_{f+1}, \ldots, \mathcal{N}_{f+n}$. However, base stations do not acquire more carriers than their computed demand. By definition, the base-stations in a re-use group are all mutually non-interfering. Thus, the assignment of carriers is guaranteed to remain feasible throughout. Of course, the order in which the base stations are allowed to acquire non-designated carriers could be periodically changed or dynamically adjusted in response to congestion conditions to enhance performance.

Unrestricted acquisition of designated carriers Since base stations do not acquire more carriers than their computed demand, it is possible that some carriers remain unclaimed. Therefore, the base stations finally reclaim their designated carriers that have not been claimed by any of the interfering base stations.

The distributed carrier requesting process is exemplified in Figure 2 for a scenario with $F = 4$ re-use groups. The letter A corresponds to the first stage, the restricted retainment of carriers by the designated base stations. Then, in D_0, the base stations declare which of their designated carriers they wish to retain. The symbol B_n represents the n-th borrowing round during the second phase, the restricted acquisition of carriers by non-designated base stations. The base stations in re-use group $\mathcal{N}_f$ then decide which of the carriers in the set $\mathcal{K}_{f+n}$ they wish to borrow, among those that have not been declared in use by any of the interfering base stations in re-use groups $\mathcal{N}_{f+1}, \ldots, \mathcal{N}_{f+n}$. Then, in D_n, the base stations in re-use group $\mathcal{N}_f$ declare which of those carriers they have decided to borrow. The letter R finally indicates the third stage, the unrestricted acquisition of carriers by the designated base stations.

4.2 Distributed carrier raking

In this scheme, the base stations periodically rake all the carriers that are not currently used by any interfering base stations. In each superframe, only the base stations in one re-use group are active, while the base stations in all the other re-use groups remain passive. The active re-use group alternates in a cyclic manner, i.e., the base stations in re-use group $\mathcal{N}_f$ are active in the $(nF + f)$-th superframe, $n = 1, 2, \ldots$ The scheme consists of the following three steps, again executed at the start of every superframe.

Restricted retainment of carriers by passive base stations First, the passive base stations decide which carriers they wish to keep of the ones which they used in the previous superframe. Suppose that base station j used K_j carriers in the previous superframe. If $D_j < K_j$, then base station j retains D_j of these carriers, and releases the remaining $K_j - D_j$ ones. The base stations must always relinquish carriers after using them for T superframes, for some parameter T. If $D_j \geq K_j$, then base station j holds on to all K_j carriers.

Unrestricted acquisition of carriers by active base stations Next, the active base stations concurrently grab all the carriers that are not already claimed by any interfering base stations. By definition, the base-stations in a re-use group are all mutually non-interfering. Thus, the assignment of carriers is guaranteed to remain feasible.

Unrestricted retainment of carriers by passive base stations Finally, the base stations in the passive re-use group reclaim the carriers which they relinquished in the first stage, but which were not claimed by any of the interfering active base stations.

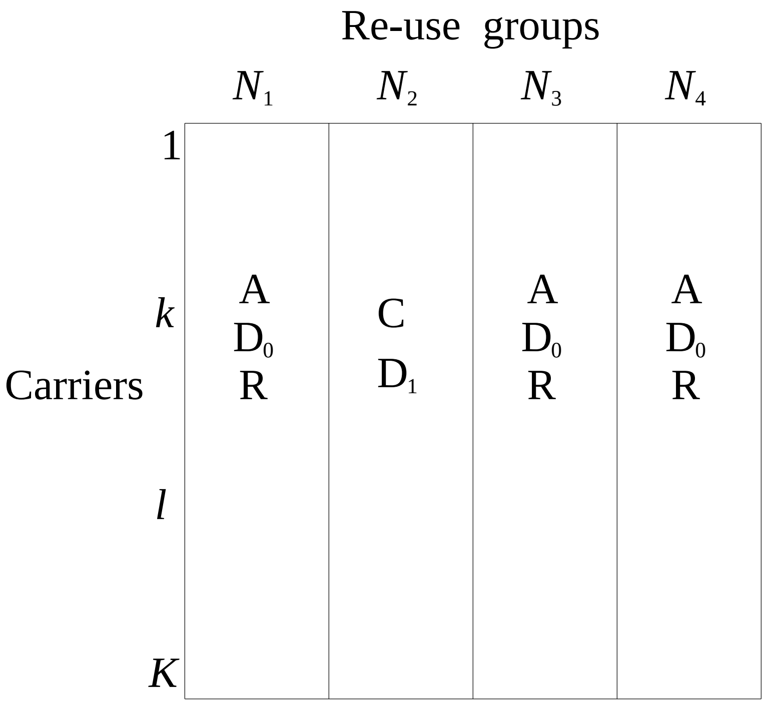

Figure 3: Distributed carrier raking process; A: acquisition of previously used carriers by passive base stations; D_0: declaration of used carriers by passive base stations; C: active base stations grabbing unused carriers; D_1: declaration of used carriers by active base stations; passive base stations reclaiming previously used carriers.

The distributed carrier raking process is illustrated in Figure 3 for a scenario with $F = 4$ re-use groups. We assume that it is the turn of the base stations in the second re-use group to rake carriers. The letter A corresponds to the first stage, the restricted retainment of carriers used in the previous superframe by the passive base stations. Then, in D_0, the passive base stations declare which of the carriers used in the previous superframe they wish to retain. The letter C represents the second phase, the unrestricted acquisition of carriers by the active base stations. Then, in D_1, the active base stations declare which of the carriers they have been able to acquire. The letter R finally indicates the third stage, the unrestricted retainment of carriers by the passive base stations.

5 Centralized carrier allocation

In the previous section, we presented two distributed assignment schemes, which only involve local exchange of information between base stations for allocating carriers. We now outline a centralized allocation scheme, which relies on global knowledge in assigning carriers to base stations.

Successively, for each carrier, the scheme attempts to find a subset of mutually non-interfering base stations with the maximum number of packets queued. In reference to the associated interference graph, we term such a set as the *maximum-weight independent set.* The scheme is motivated by the stability results for the maximum-weight independent set established in Kahale & Wright [10]. As before, the allocation is done at the start of each superframe, and the capacity of the carrier is subtracted from the queues to which an assignment has been made.

The maximum-weight independent set problem can be formulated as an Integer Linear Program (ILP), with variables $x_j \in \{0, 1\}$ indicating whether or not base station j belongs to the desired set. In general, the problem is known to be NP-complete, even on very simple graphs. We expect this to be the case on most wireless interference graphs as well. Hence, we *relax* the integrality constraints, and consider the corresponding Linear Program (LP).

We first assign weights $\{w_j\}$ to each of the base stations $j = 1, \ldots, N$ determined by $w_j = \sum_{i \in \mathcal{M}_j} Q_i$, i.e., the aggregate queue length of the users served by base station j. Then, we successively perform the following steps for each of the carriers $k = 1, \ldots, K$.

1. Solve the linear program

$$\begin{array}{lll} \max & \sum_{j=1}^{N} w_j x_j & \\ \text{sub} & \sum_{l \in \mathcal{C}} x_l \leq 1 & \text{for all maximal cliques } \mathcal{C}, \\ & 0 \leq x_j \leq 1 & \text{for all } j = 1, \ldots, N. \end{array}$$

2. Let $\{x_j^*\}$ be the solution to the linear program. Initialize $\mathcal{J} := \{1, \ldots, N\}$ to the set of all base stations. Then perform the following steps.

   (a) Let $\hat{\jmath} := \arg\max_{j \in \mathcal{J}} \{x_j^*\}$. Assign carrier k to base station $\hat{\jmath}$, i.e., $X_{jk} := 1$. Remove base station $\hat{\jmath}$ and all its neighbors from the set $\mathcal{J}$.

   (b) Repeat the above step until $\mathcal{J} = \emptyset$.

3. For each base station that has been assigned carrier k, reduce its weight by S, i.e., $w_j := w_j - S X_{jk}$.

The maximal cliques in step 1 of the algorithm can be constructed by repeatedly appending vertices to a set initialized to some vertex $\{j\}$ so that the new vertex has edges to all the current elements of the set, until no more vertices can be added. Enumerating all maximal cliques is a formidable task on arbitrary graphs, but the local nature of the interference in a wireless network ensures that the number of distinct maximal cliques stays within a constant factor of the number of cells in the network, thus rendering a linear program with a tractable number of constraints. However, determining the true maximum-weight independent set requires us to constrain x_j to be binary, i.e., 0 or 1. In the absence of these constraints, the LP solution only provides an *upper bound* to the ILP solution.

Step 2 is essentially a method of *rounding* the fractional solution of the LP to a feasible assignment of carriers. This produces a *lower bound* to the optimal value, which, in conjunction with the above-mentioned upper bound, makes the LP a useful bounding procedure. In step 3, for each base station that has been assigned carrier k, the corresponding weight is reduced by the number of packets S that can be transmitted during the next superframe.

For well-structured interference graphs and random weights, the LP method provides a reasonable approximation to the true maximum-weight independent set. In experiments carried out with a regular 4-cell re-use pattern, which are discussed in detail in a later section, the LP almost always produced the exact optimal solution. This was evident from the fact that the values of $\{x_j^*\}$ returned by the LP were integral, despite the relaxation.

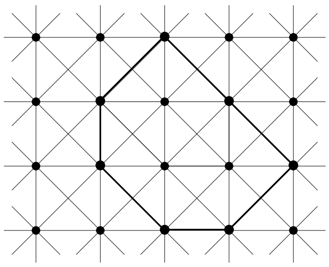

Figure 4: An odd cycle with 7 base stations.

However, there are cases in which the LP can be shown to 'fail', i.e., provide ambiguous values of x_j^* with no natural rounding. The simplest such cases are so-called induced *odd cycles*, see Figure 4. Consider a snapshot of the network where the only non-empty queues are located on such a ring with an odd number of base stations, and the queues are all equal in size. It can then be shown, by straightforward substitution, that the LP solution would be $x_j^* = 0.5$ for all j. The true solution would have $x_j^* = 1$ on p base stations in a cycle of length $2p+1$, with $x_j^* = 0$ elsewhere. Such cases, although pathological, can arise in subnetworks of small size, and must be dealt with by adding constraints to the LP of the form $\sum_{j \in C} x_j \le p$. These constraints rule out the ambiguous solution, which has $\sum_{j \in C} x_j = p + 1/2$. Incorporating all such constraints on an arbitrary graph is prohibitive, but the effort is moderate for small values of p on wireless interference graphs. However, even here we cannot cover all values of p, but this is not likely to be too detrimental in practice, since the effect of large values of p only occurs in the unlikely adversarial circumstance of a large ring of base stations containing the only non-empty queues.

Thus, the LP provides a heuristic for allocating carriers in a stable and efficient manner while avoiding interference. However, the algorithm as described above may be unfair by excessively favoring users or base stations with very long queues. The potential unfairness may be partly ameliorated by appropriately modifying the weights to reflect the desired level of fairness among users, possibly at the expense of throughput. The ultimate solution may involve the integration of the scheduling algorithm with policing and admission control.

6 Complexity

The complexity of the various algorithms is determined by the level of communications required to pass messages between various network elements and the amount of computing effort needed to perform the calculations themselves. Carrier raking and carrier requesting have a similar level of complexity, with the centralized algorithm being significantly more complex in terms of both communication and computation.

We first consider carrier requesting. A base station need only communicate with its neighbors. As shown in Figure 2, just before the start of each round, each base station broadcasts the list of carriers which it has available from those which it could have acquired previously. Thus, in the first round, each base station declares which carriers are remaining from those designated for its use. In the second round, each base station declares which carriers it is using from the group of carriers offered in the first round. Similar declarations are made in the third round and so on.

The scheme may be implemented by using a common control channel. This would then be used for a total of $F - 1$ rounds. If the re-use for signaling is H, then the signaling load for each round is $HFK/F = HK$ bits, where K/F bits are needed to indicate which designated carriers are being used by each base station. The total number of bits which must be transmitted on the common control channel is, therefore, HFK, since there are $F - 1$ rounds, and there is a declaration at the end. For example, if $F = 4$, $H = 12$, and $K = 32$, then 1536 bits must be transmitted per superframe, so that if the capacity is 10 kbit/s, 6 superframes a second might be processed.

The use of a common control channel does require that the base stations are synchronized, so that message and base station identifiers need not be sent. The computing resources needed to calculate the demand and carry out the algorithm are very limited. Note however that the base stations must compute which carriers they will borrow before the next round is started.

In carrier raking, the passive base stations send out a broadcast signal once per superframe, indicating which carriers they are prepared to relinquish. The active base stations respond by declaring the carriers they have taken in a subsequent broadcast message. The signaling load is once again HFK bits per superframe if a common control channel is used. The computational load is similar to carrier requesting. The demand calculations are the same as for carrier requesting, but these only have to be performed once every F superframes.

We finally consider the centralized scheme. If the controller is located at the Mobile Switching Center (MSC), then there will be no need to pass queueing information from the base stations, as far as the down link is concerned. Once the succession of linear programs have been solved, the results must be passed to the base stations. This may be passed as a K-bit field for each base station in the network, and supplied with a base station identifier of length B bits. Therefore, the total information which must be passed per superframe is $(K + B)N$, where N is the number of base stations. This information can be passed through wireline connections between the MSC and the base stations.

There are N variables in the linear program, and there are GN constraints, where G is a small factor depending on the number of maximal cliques. The LP must be solved once per carrier, but obtaining the solutions can be sped up by using the previous solution as the starting point for the next. There are several efficient methods for solving such LP's.

7 Simulation experiments

We now present an overview of the results from the simulation experiments that we have performed to examine the performance of the various algorithms proposed in the previous sections. We also provide results for the pooling bound (see Section 3), which helps us set a lower bound on the achievable delay.

We consider a network of $N = 64$ base stations arranged in a grid as depicted in Figure 5. We assume the grid is 'wrap-around' (i.e. a periodic boundary) to reduce the edge effects.

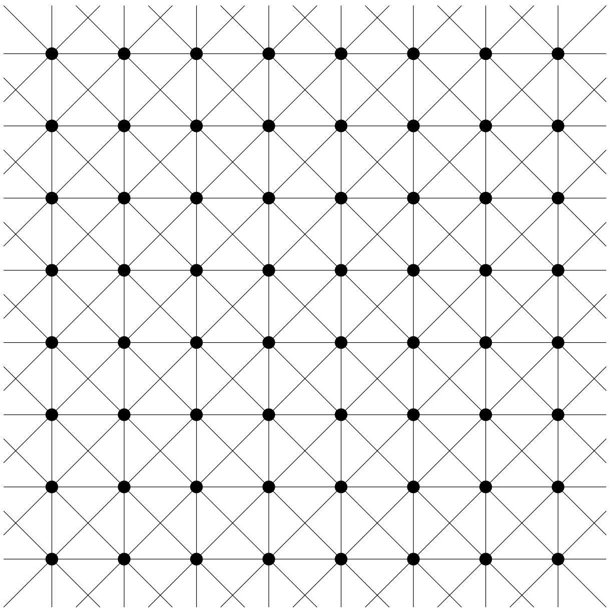

Figure 5: The interference graph.

The total number of carriers in the system is $K = 32$, and the maximum number of carriers that can be assigned to any particular mobile is $L = 8$. The bandwidth per carrier is 48 kbps, and the packet size is 1920 bits so that for example $S = 25$ packets can be transmitted during 1 second.

For the first four experiments, we used a superframe of 1 second for carrier requesting and the LP-based algorithm. For carrier raking, we used a superframe of 0.24 seconds, so that each base station has the opportunity to acquire carriers about once every second.

In all experiments, we adopted the On-Off traffic model presented in Section 2. This model was used to describe the arrival of packets which need to be transmitted to a particu-

	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5
On period (s)	varying	varying	4	4	4
Off period (s)	varying	10 * On period	36	36	36
On + Off (s)	40	varying	40	40	40
Peak rate (kbps)	varying	384	384	384	384
Burst size (Mb)	1.536	varying	1.536	1.536	1.536
Mobiles per BS	4	4	varying	non-uniform	4
Superframe (s)	.24/1.00	.24/1.00	.24/1.00	.24/1.00	varying

Table 2: Simulation parameters

lar mobile. The key parameters in the traffic model are the average On period, the average Off period, the mean rate, the peak rate, and the burst size. These quantities are related through

$$Mean\ rate = \frac{On\ period \times Peak\ rate}{On\ period + Off\ period}, \qquad Burst\ size = On\ period \times Peak\ rate.$$

To investigate the impact of the traffic characteristics on the performance of the various algorithms, we varied these parameters as summarized in Table 2 above. The mean rate is always taken to be 38.4 kbps. The number of packet arrivals per slot is assumed to have a Poisson distribution with as parameter the peak rate divided by the packet size. We now proceed to describe the results from the simulation experiments in greater detail.

7.1 Experiment 1 – Fixed burst size

For this experiment, we had 4 mobiles per base station. We constructed the traffic model so that the 'bursts of packets' would have a fixed mean size. However, we altered the peak arrival rate so that these bursts would have different shapes.

Figures 6-9 show the average delay as a function of the On period. (Note that an increasing On period corresponds to a decreasing peak rate and decreasing burstiness.) As expected, the average delay decreases as the burstiness decreases. Furthermore, the schemes are more set apart when the traffic is more bursty. This illustrates the responsiveness of the dynamic schemes. However, when the burstiness is reduced, both carrier raking and LP exhibit a significant residual latency, caused by the continuing response to queue fluctuations, despite the near-constant traffic.

7.2 Experiment 2 – Fixed peak rate

For this experiment, we also had 4 mobiles per base station. However, the peak rate was fixed during the On period, whose length was altered to vary the burst size. The length of the Off period was also changed to keep the load fixed. Because the peak rate and load are fixed, the burstiness is not changing in this case.

As Figures 10, 11 show, the delay worsens somewhat with the mean length of the On period. This reflects the fact that more packets are queued up in a burst as its duration increases.

7.3 Experiment 3 – Varying number of mobiles per base station

We kept the burst size and peak rate fixed, and varied the number of mobiles per base station. This has the effect of varying the total load on the system.

Figures 12, 13 indicate the gains in throughput from using DCA. If we assume a mean delay requirement of a few seconds, for example, then we can attain improvements in throughput of around 30-70%. Carrier requesting and LP have similar throughput, with carrier raking being a little worse. The gains in throughput over FCA are greater when the traffic is more bursty, see Figures 14, 15.

7.4 Experiment 4 – Non-uniform distribution of mobiles

This experiment was the same as experiment 3, except that we had a non-uniform distribution of mobiles. If the average number of mobiles per base station was m, then we assigned to each base station a number of mobiles that was chosen with equal probability from $\{m-2,\ldots,m+2\}$.

Figures 16, 17 show the impact of the non-uniformity on performance. Here FCA delay performance worsens significantly as the mismatch between the number of carriers allocated and the actual number of mobiles at a base station increases. Carrier raking also does badly, because of the possibility of getting 'locked' into a poor match between carriers and load, causing the base stations to hold on to their carriers.

7.5 Experiment 5 – Varying superframe duration

For the DCA schemes, we altered the length of the superframes so as to examine the effect that the scheduling interval in the carrier allocation process has on performance.

Figures 18, 19 show that for all the schemes, the performance improves as the superframe duration is reduced, as expected. As we can see, a superframe duration of the order a fraction of a second is very desirable. The complexity calculations in the previous section indicate that this should be achievable for both carrier raking and carrier requesting. The performance of carrier requesting degrades gradually as the duration of the superframe increases. The average delay approaches that of FCA as the superframe duration grows large, because the queue length becomes negligibly small compared to the superframe capacity, reducing the computed demand to zero. In contrast, the performance of carrier raking degrades severely as the superframe duration increases, causing excessive delays for superframes longer than a second.

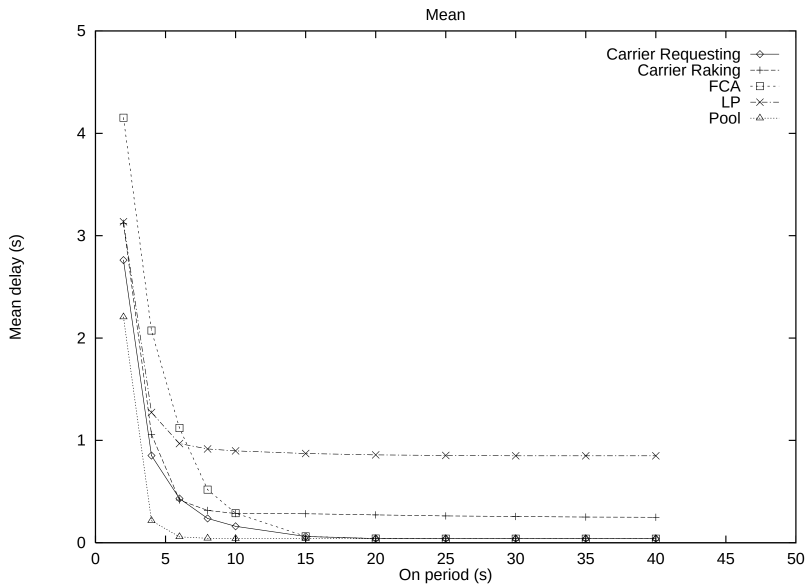

Figure 6: Mean packet delay – Fixed burst size (Exponential On period)

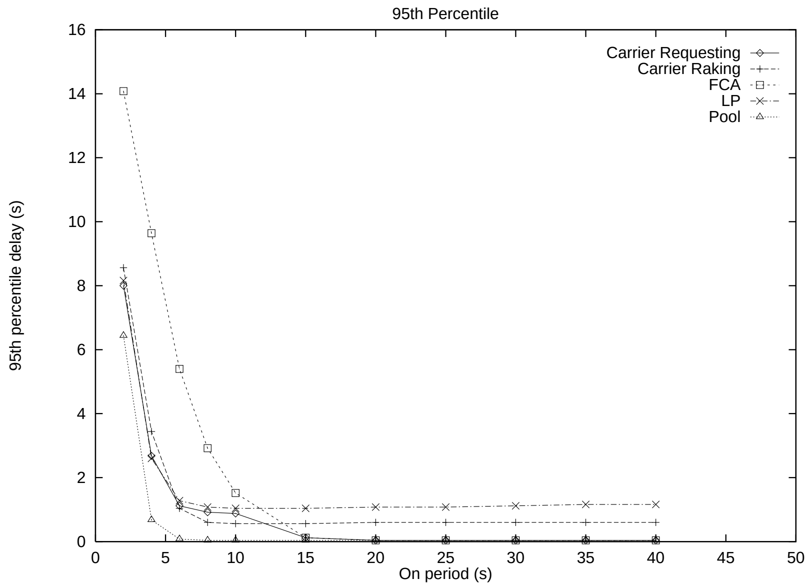

Figure 7: 95th Percentile packet delay – Fixed burst size (Exponential On period)

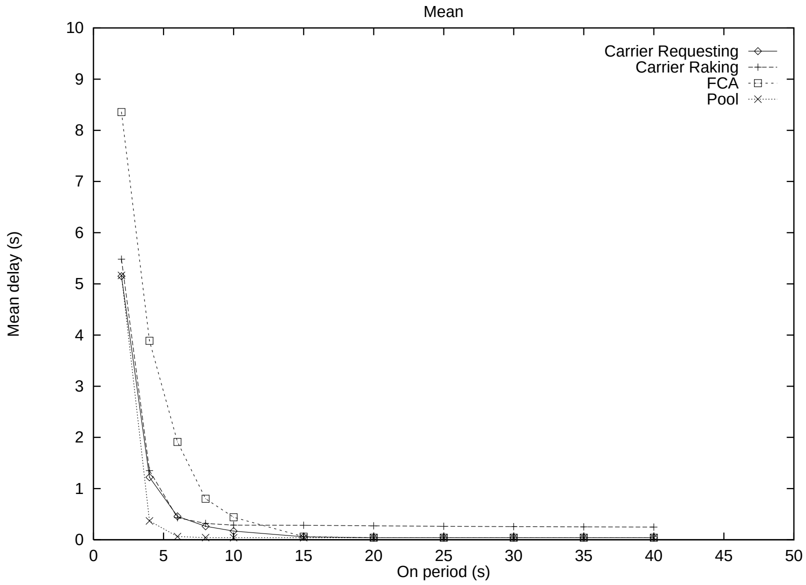

Figure 8: Mean packet delay – Fixed burst size (Pareto On period)

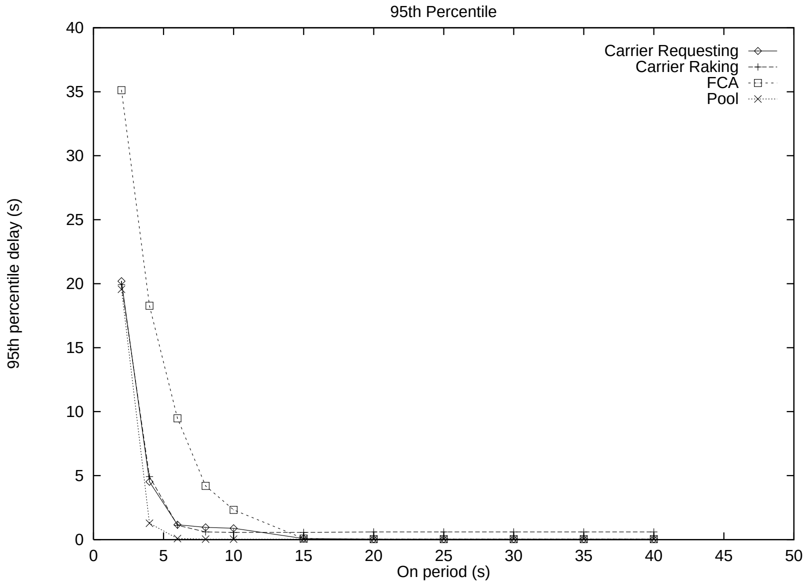

Figure 9: 95th Percentile packet delay – Fixed burst size (Pareto On period)

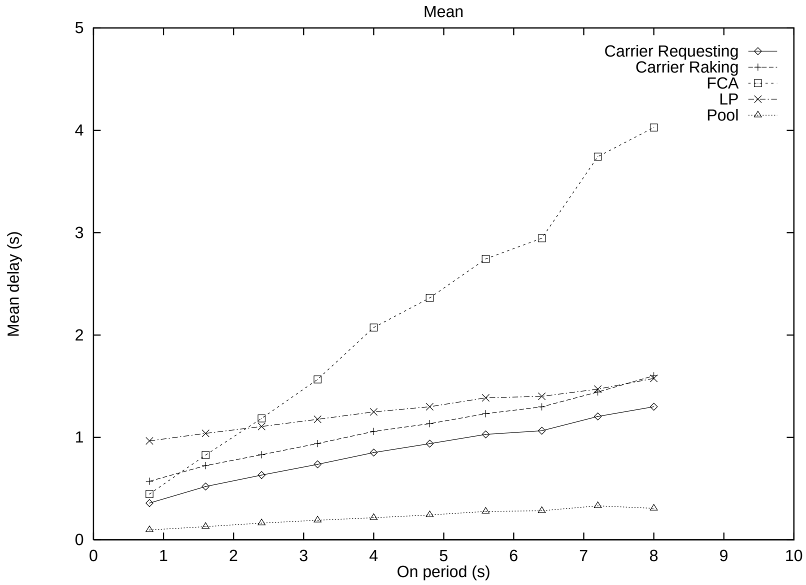

Figure 10: Mean packet delay – Fixed peak rate (Exponential On period)

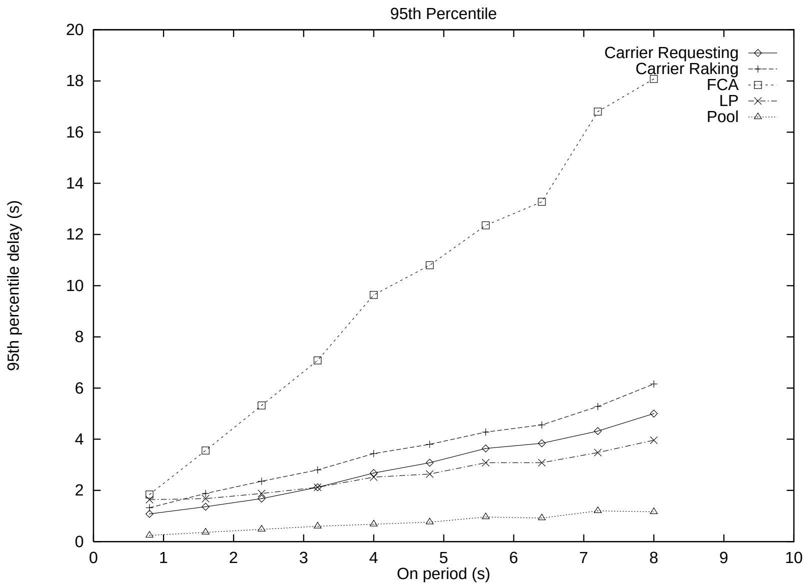

Figure 11: 95th Percentile packet delay – Fixed peak rate (Exponential On period)

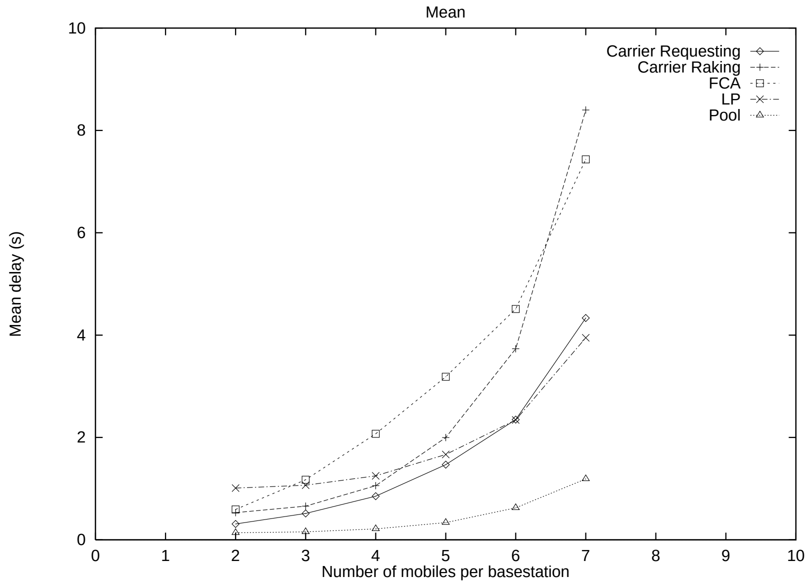

Figure 12: Mean packet delay – Varying number of mobiles (Exponential On period)

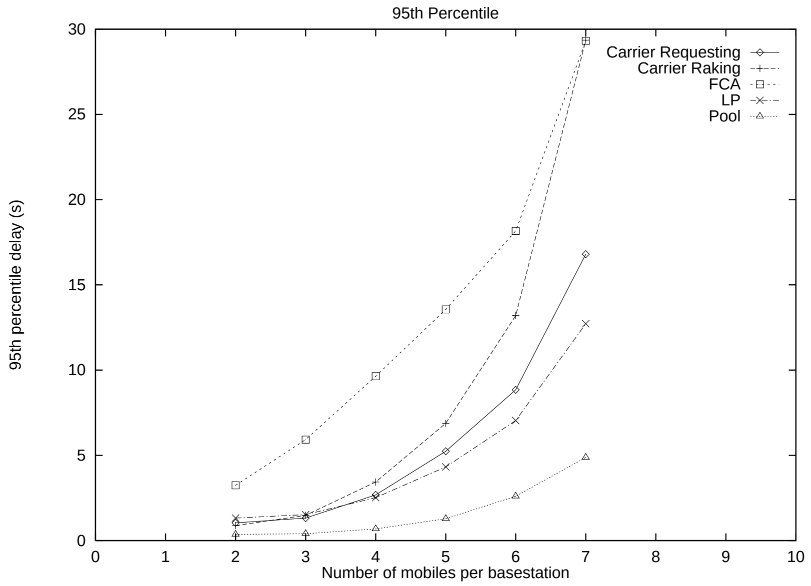

Figure 13: 95th Percentile packet delay – Varying number of mobiles (Exponential On period)

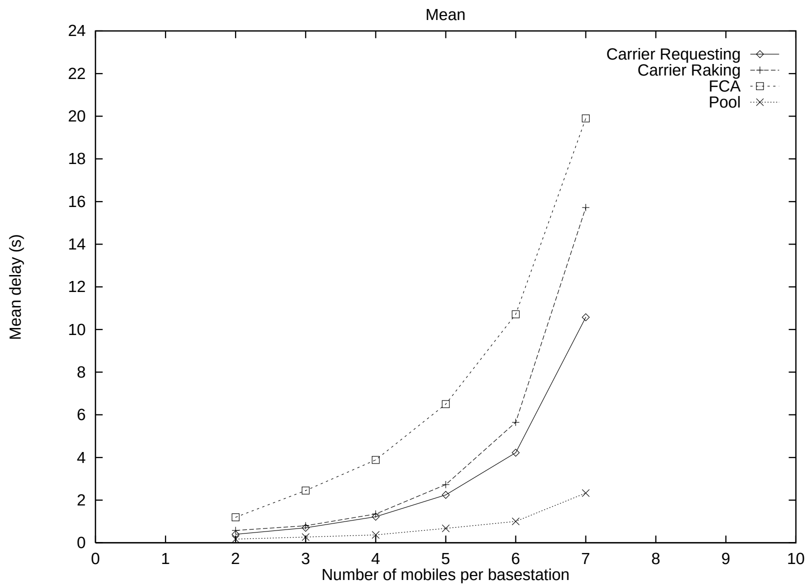

Figure 14: Mean packet delay – Varying number of mobiles (Pareto On period)

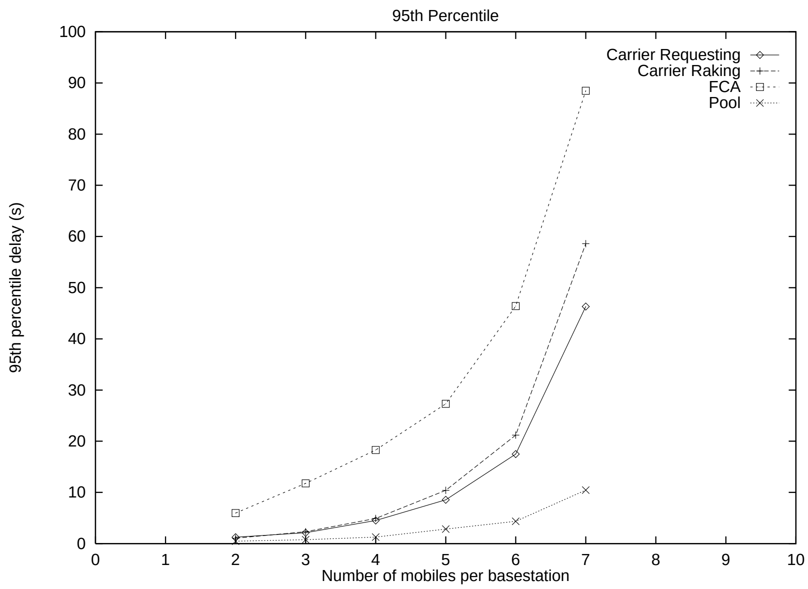

Figure 15: 95th Percentile packet delay – Varying number of mobiles (Pareto On period)

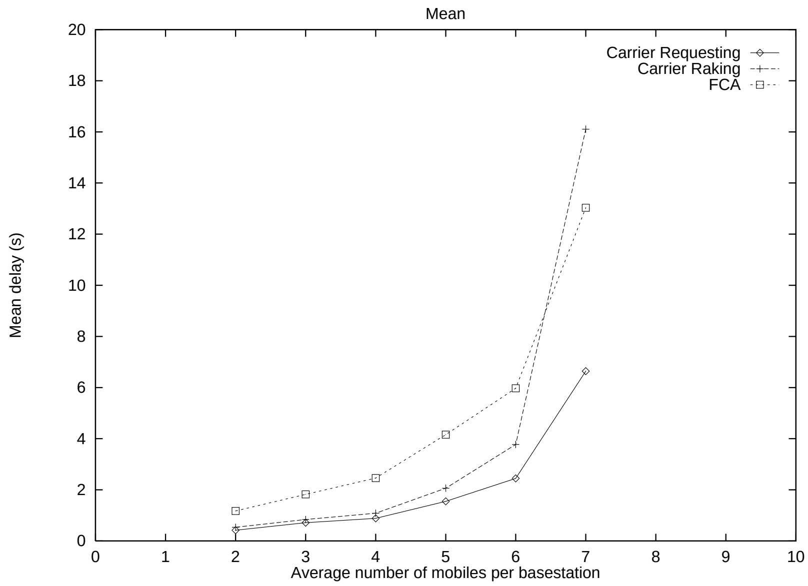

Figure 16: Mean packet delay – Varying (non-uniform) number of mobiles (Exponential On period)

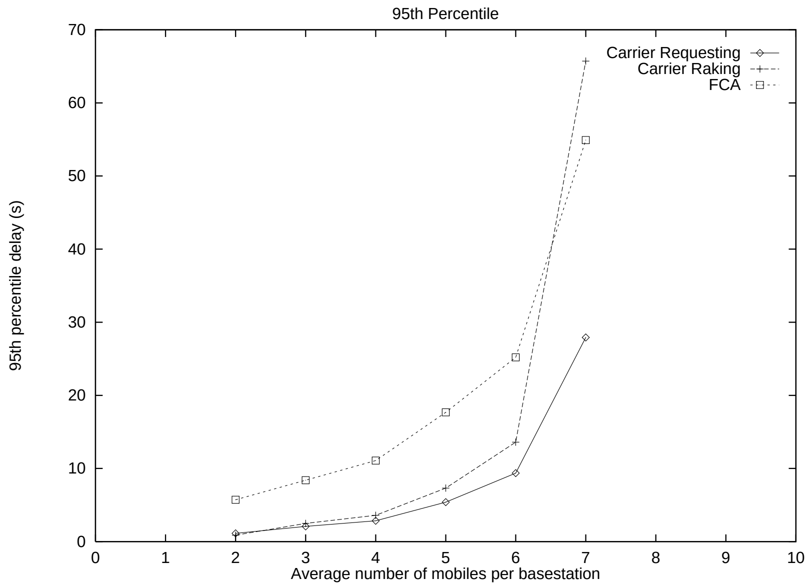

Figure 17: 95th Percentile – Varying (non-uniform) number of mobiles (Exponential On period)

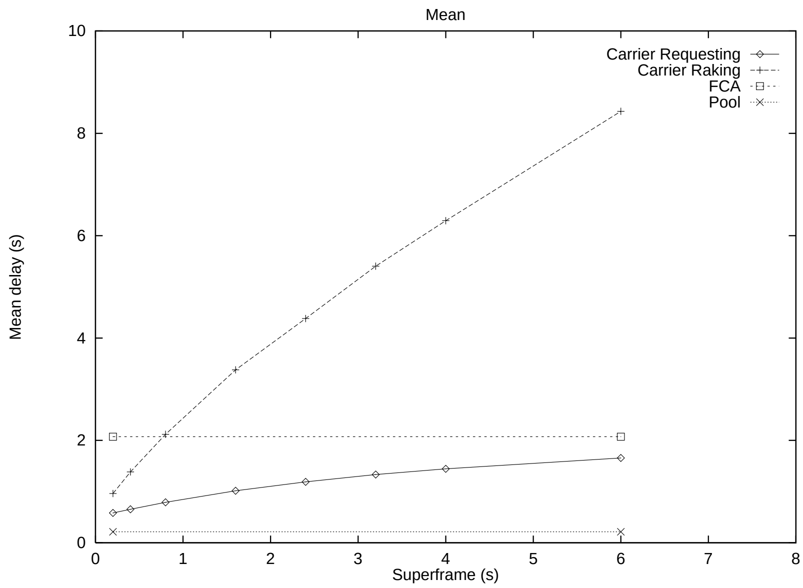

Figure 18: Mean packet delay – Varying superframe duration (Exponential On period)

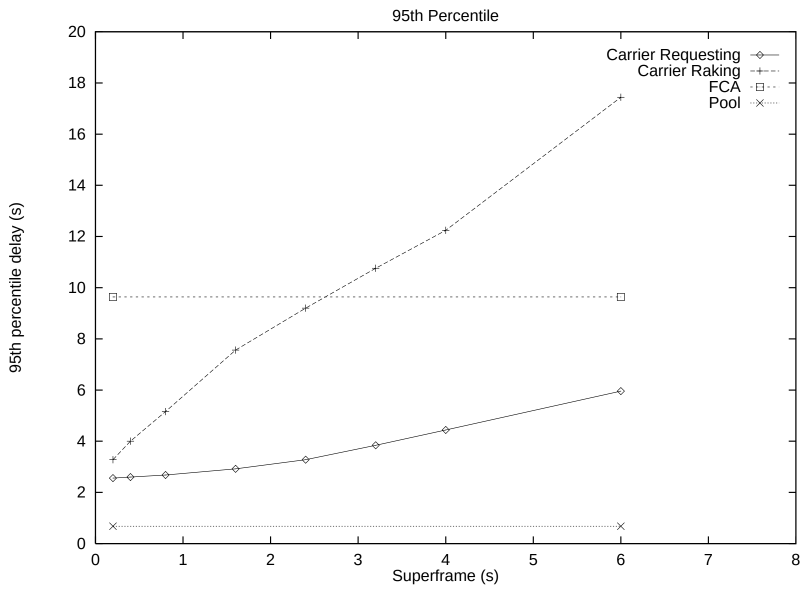

Figure 19: 95th Percentile – Varying superframe duration (Exponential On period)

8 Conclusion

Next-generation wireless networks are expected to support a wide range of high-speed data services, Web browsing being one of the major applications. As compared with voice, data traffic is typically more bursty, while the users are less sensitive to delay. These characteristics require resource allocation strategies to operate in a fundamentally different manner in order to achieve efficient spectrum utilization. In particular, the allocation of dedicated bandwidth to data applications is less reasonable. If a low-bandwidth connection is provided, then the delay and loss of packets will be excessive because of the burstiness. If a high-bandwidth connection is established, then the delay and loss will be less serious, but the utilization will be poor. The delay tolerance of data applications allows however for the possibility of coordinating packet transmissions among base stations. Thus, the efficient operation of high-speed data wireless networks requires the use of dynamic bandwidth allocation algorithms.

We have proposed several such algorithms for scheduling the packet transmissions among base stations in a coordinated fashion. As a novel feature, the various schemes exploit knowledge of the buffer contents in order to achieve high throughput, while maintaining fairness by providing quality-of-service to individual users. The proposed algorithms are backward compatible with existing cellular and PCS standards such as IS-136, and provide a powerful approach to improve spectrum efficiency in forthcoming high-speed data complement services to existing systems.

We have conducted extensive simulation experiments to demonstrate the efficiency of the algorithms. Not surprisingly, there is a trade-off between the throughput and delay performance of the algorithms and the communication and computation overhead involved. Since in practice the available resources for these tasks are limited, the theoretically optimal strategies are not necessarily the most adequate algorithms from a practical perspective.

The algorithms that we considered adopt the concept of a frequency re-use plan. Strategies that use signal measurements to determine the positions of individual mobiles may achieve an even higher spectrum efficiency, at the expense of larger operational complexity, see for instance Chuang & Sollenberger [4].

References

[1] Alanko, I., Kojo M., Laamanen, H., Liljeberg, M., Moilanen, M., Raatikainen, E. (1994). Measured performance of data transmission over cellular telephone networks. *Computer Communications Review (USA)* **24**, 24-44.

[2] Anick, D., Mitra, D., Sondhi, M.M. (1982). Stochastic theory of a data handling system with multiple sources. *Bell System Technical Journal* **61**, 1871-1894.

[3] Borst, S.C., Whiting, P.A. (1998). Achievable performance of dynamic bandwidth allocation algorithms in high-speed data wireless networks. Technical Memorandum, Bell Labs, Lucent Technologies. Submitted for publication.

[4] Chuang, J.C., Sollenberger, N. (1997). Dynamic packet assignment for advanced Internet cellular services. In: *Proc. IEEE GLOBECOM '97*, Phoenix AZ, 1596-1600.

[5] Cohen, J.W. (1974). Superimposed renewal processes and storage with gradual input. *Stochastic Processes and their Applications* **2**, 31-58.

[6] Elwalid, A., Heyman, D., Lakshman, T.V., Mitra, D., Weiss, A. (1995). Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing. *IEEE Journal on Selected Areas in Communications* **13**, 1004-1016.

[7] Glynn, P.V., Whitt, W. (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability* **31**A, 131-156.

[8] Harris, J.M., Kumar, S.P. (1995). An algorithm for minimizing queueing delay of packets in cellular data networks. In: *Proc. 33rd Annual Allerton Conference on Communication, Control & Computing*, Urbana-Champaign IL, 945-954.

[9] Jelenković, P.R., Lazar, A.A. (1998). Asymptotic results for multiplexing subexponential on-off processes. *Advances in Applied Probability*, to appear.

[10] Kahale, N., Wright, P.E. (1997). Dynamic global packet routing in wireless networks. In: *Proc. IEEE INFOCOM '97*, Kobe, Japan, 1416-1423.

[11] Linnartz, J.-M.P.G. (1995). On the performance of packet-switched cellular networks for wireless data communications. *Wireless Networks* **1**, 129-138.

[12] Mathias, C.J. (1997). Wireless data-what's real? What's wrong? *Business Communications Review (USA)* **27**, 52-54.

[13] Rubinovitch, M. (1973). The output of a buffered data communication system. *Stochastic Processes and their Applications* **1**, 375-380.

[14] Tassiulas, L. (1998). Linear complexity algorithms for maximum throughput in radio networks and input queued switches. In: *Proc. IEEE INFOCOM '98*, San Fransisco CA, 553-559.

[15] Willinger, W., Leland, W.E., Taqqu, M.S., Wilson, D.V. (1993). On the self-similar nature of Ethernet traffic. In: *Proc. SIGCOMM '93*, San Francisco CA, 183-193.