Utility-based Adaptive Resource Allocation in Hybrid Wireless Networks *†

Calin Curescu, Simin Nadjm-Tehrani, [calcu/simin@ida.liu.se]
Department of Computer and Information Science, Linköping University

Bing Cao and Teresa A. Dahlberg [bcao/tdahlber@uncc.edu]
Department of Computer Science, University of North Carolina at Charlotte

Abstract

Service availability in wireless networks is highly dependent on efficient resource allocation and guaranteed Quality of Service (QoS) amid overloads and failures. This paper addresses optimal bandwidth allocation in a hybrid network (cellular and ad hoc), where added reach through an ad hoc overlay is combined with the stability and essential services of a cellular network. The paper builds on a near optimal approach in which Resource-Utility functions are used as a means of adaptive delivery of QoS, user differentiation, and maximisation of system level utility. It distinguishes between non-adaptive, semi-adaptive, and fully adaptive applications. First, the global cellular bandwidth allocation (in the presence of multiple routes through ad hoc relays) is cast in terms of a Linear Programming problem. Second, a heuristic algorithm that has far lower computational overhead and accrues at worse 12% less than the utility of the optimal solution is presented. Both algorithms are implemented within a model of a hybrid network on top of the J-Sim simulation environment. Comparative studies are made to show effective load balancing and crash tolerance in the presence of a high traffic overload.

1. Introduction

In Future Generation wireless networks, diverse wireless technologies such as Cellular, WLAN, and Bluetooth will proliferate in different edges of the Internet and complement each other to provide untethered multimedia services and seamless visits to the IP-core network. Most wireless access technologies are deployed in either infrastructure based cellular mode or infrastructure-less ad hoc mode. While each access mode was initially designed with distinct characteristics, many recent efforts are underway to define hybrid networks, that combine the advantages of both access modes [1, 21, 2, 9]. These approaches to hybrid networks can be classified as either "ad hoc over cellular" [9, 24, 28, 25] or "cellular over ad hoc" [27, 3].

In our work, we focus on "ad hoc over cellular" approach that aims to "stretch" the reach of cellular networks, and integrate high speed access, global coverage and roaming support into a single seamless system. These concepts motivated the ODMA option in 3GPP [1] or the next generation A-GSM [2].

Among the challenges in the hybrid wireless networks, is optimal resource management of diverse radio resources, from the perspectives of both the users and the service provider. Hybrid network radio resources often include "cellular" capacity (licensed frequency spectrum centered around fixed base stations) in addition to "ad hoc" capacity (unlicensed frequency spectrum limited by interference local to each mobile). Furthermore, hybrid network models, that employ user equipment to serve as mobile relays, must include resulting usage costs into resource management.

In earlier work [4, 5], resource-utility (R-U) functions were employed and a Time-Aware Resource Allocation (TARA) scheme was proposed for Quality of Service (QoS) resource allocation in cellular networks. TARA maximizes the accumulated utility of the whole cell (over time) in an adaptive way. In this paper, we continue to use utility functions to characterise the bandwidth of the cellular link (as the bottleneck resource). Moreover, we extend the approach to model additional resources at the ad hoc nodes. That is, we consider non-critical resource usage (such as power usage, processing capacity and bandwidth of the ad hoc link) at the relaying user equipment and model it as a cost function. Two resource allocation algorithms result from this study.

First, a centralized optimal allocation algorithm based on linear programming is formulated. Second, a distributed heuristic algorithm is formulated that attempts to perform close to the optimal solution with considerably lower runtime complexity. The simulation analysis, using the J-Sim simulator, illustrates the performance gains in "ad hoc over cellular" hybrid networks. It demonstrates the capability of the proposed heuristic algorithm to efficiently utilize resources in the hybrid radio context and provide benefits such as load-balancing and fault tolerance.

The paper is organised as follows. In the next section we review related work on resource allocation for wireless net-

†This work was supported by CUGS (the National Graduate School in Computer Science, Sweden). The cooperation between Linköping University and UNCC was facilitated by an NSF supported travel grant.

works and utility-based solutions. Section 3 presents background information about TARA and utility maximisation. In Section 4, we present our system model together with two different bandwidth allocation algorithms. Sections 5 and 6 present the evaluation setup and compare the simulation results for the different algorithms. Section 7 concludes the paper.

2. Related works

Several "ad hoc over cellular" approaches such as MCN [9] and iCAR [24] use relays to overcome cellular shortcomings, such as limited spatial coverage, low bit rates, and a high bit cost for data services. Relays, being either static infrastructure or other mobile stations (MS), form a virtual overlay for congestion mitigation and alternate routing to extend and improve coverage of the cellular base stations (BS) [2, 8, 16]. In these hybrid networks, admission control (AC) and bandwidth allocation (BA) schemes for resource management are necessary to ensure QoS guarantees.

Work on resource management for cellular networks often focuses on management of licensed frequency spectrum local to each base station. For example, the authors in [14] a novel adaptive bandwidth allocation scheme by dynamic local estimation of changing traffic parameters, and a probabilistic control policy for high channel utilization. The work in [6, 17] employs bandwidth borrowing and degradation as part of AC with each connection request submitting acceptable max and min resource requirement. In [20] an AC algorithm is proposed that uses controlled QoS degradation of on-going calls to manages a tradeoff between resource allocation of on-going calls and new calls. A similar tradeoff is managed by an AC algorithm proposed in [19] using Guard Channel policies. AC schemes proposed in [7] consider both "non-prioritized" schemes in which the BS made no distinction between new and HO calls, and two "priority oriented" schemes that allow queuing of handover calls.

Work on resource management for ad hoc networks often focuses on managing the interference, generated by consumption of unlicensed frequencies, local to each MS involved in an ad hoc path between a source and destination node. For example, in [26] a contention-aware AC is proposed which attempts to support QoS guarantees by limiting the number of connections allowed within a neighborhood of nodes. A distributed AC algorithm is introduced in [23] that is based on the concept of a "service curve" to reflect the status of the network (number of active nodes, activity index and contention status). An ad hoc node wanting to establish a new connection must compare the "service curve" with a predefined universal performance threshold curve for QoS purpose.

Variable resource needs and differentiated importance levels of most of the new services decrease the relevance of traditional performance metrics such as blocking/dropping probabilities. Thus, the user-perceived utility might be more suited as performance criterion. Chen Lee et al. [13] use resource-utility functions in a QoS management framework with the goal to maximise the total utility of the system. They propose two approximation algorithms, and compare the run-times and solution quality with an optimal solution based on dynamic programming. As opposed to maximising the total utility of the system, Rui-Feng Liao et al. [15] provide "utility fair allocation". Their algorithm extends "max-min fair allocation", with utility replacing bandwidth as the fairness criterion.

3. Background

The utility model used in this paper is the same as the one used in the Time-Aware Resource Allocation scheme (TARA) [5], and enables the allocation algorithm to differentiate between different needs for resource guarantees. This section presents the main concepts of TARA.

One of the main concepts in TARA is the usage of resource-utility (R-U) functions. The utility of an application (and its associated connection) represents the value assigned by the user to the quality of the application's results. The accrued utility, at any time point, depends on the allocated resource, which in our case is the bandwidth of the cellular link. For the ease of implementation, and to keep complexity low, it is necessary to quantise the utility functions using a small set of parameters. Thus, the utility function for a connection can be represented by a list of bandwidth-utility pairs: $u_i = \left(\binom{U_{i1}}{B_{i1}}, ..., \binom{U_{ik}}{B_{ik}} \right)$ where k is the number of utility levels attainable by the connection. We can regard the k levels as allocation segments, and an important parameter of each segment is its efficiency (slope), calculated as the utility increase divided by resource requirement. Thus, for segment k, $e_{ik} = \frac{U_{ik} - U_{ik-1}}{B_{ik} - B_{ik-1}}$. For an allocated bandwidth x_i, the accrued utility is denoted by $u_i(x_i)$.

In order to keep an optimal allocation in such a dynamic system, with new connections and handovers constantly appearing, we run our allocation algorithm periodically. This means connections might have their allocated resource changed during their lifetime. Now, applications react differently to changes in their resource allocation (especially resource degradations). To take this into consideration we have identified three connection classes depending on their adaptability to bandwidth reallocation. Thus for accounting utility TARA uses a new parameter, u_i^a, which is the utility accumulated over time for the connection i.

Class I represents non-adaptive connections. Once accepted, the resource amount cannot be re-negotiated. If the initial resource amount cannot be assured at any time point, the connection should be dropped and no utility is gained for the whole duration of running. Otherwise, the utility accumulated over the duration of the connection is: $u_i^a = u_i^{init} \times duration$. Examples are real-time control data, and real-time data streams.

Class II represents semi-adaptive connections. Here the lowest utility experienced during its lifetime determines the utility for the whole duration: $u_i^a = u_i^{min} \times duration$. Examples are streams of sensor readings with different accuracy, and different types of streaming multi-media.

Class III represents fully-adaptive connections. These

have no real-time requirements, and can therefore adapt to both increases and decreases of the bandwidth. $u_i^a = \int_0^{duration} u_i(b_i(t))\,dt$, where $b_i(t)$ describes the amount of allocated bandwidth over time. Examples are fetching e-mail, or different types of file transfer.

Besides the three classes, TARA accounts also for a) the dissatisfaction created when an ongoing application is interrupted, and b) the sensitivity to frequent reallocations. Thus, P_drop_i (drop penalty) is to be applied to u_i^a if the system drops connection i. Also, when the resource level for connection i is changed within t_adapt_i (adaptation time) since the last allocation, the connection will suffer a proportional penalty.

To maximise accrued utility, at each allocation point we modify (scale) the original R-U functions to reflect the influence of the previous parameters on the connections' u_i^a. Basically, by modifying the R-U functions [5] we make connections that are new and old, flexible and inflexible, etc., directly comparable. Then, by using these modified R-U functions as an input to an optimised allocation algorithm, we achieve a maximised accumulated system utility.

The focus of this paper is the allocation algorithm for hybrid networks, so in the reminder of the paper we consider that all the utility functions (u_i), used at a certain allocation time point, were already modified according to the TARA model.

4. Bandwidth allocation in hybrid networks

In this section we explain the system model used for our utility maximisation scheme. We start from a classic cellular network model, where in each cell a base station (BS) services the mobile station (MS) inside the covered area. MSs can connect to the BS using the direct cellular wireless link. In addition, we assume each MS is equipped with a second wireless interface that can be used to connect to other MSs in an ad hoc manner. We consider the two spectra (cellular vs. ad hoc) to be in different bands, the cellular using a narrower, highly regulated band while the ad hoc belongs to a broader, reusable, unregulated band. Thus, there is no interconnection/interference between the two bands.

At a certain point in time, a MS can connect to a BS directly through the cellular link, or relay via ad hoc paths using other MSs, and further through the cellular interface of the last MS in the path. Figure 1 presents an example. The ad hoc network serves only as an extension for the cellular network, with most of the functionality (allocation, security, billing, etc.) located in the nodes of the cellular network.

Regarding the bandwidth allocation problem, we make the following observation: The bandwidth of the ad hoc network is usually more than one order of magnitude greater than the bandwidth of the cellular network. E.g. today the bandwidth of a 3G base station is 2 Mb/s (with 10Mb for HSDPA mode) while the bandwidth of 802.11g is 40Mb/s (with 802.11n $>$ 100 Mb/s). Therefore we consider, the cellular link bandwidth as the bottleneck resource of the hybrid system, which makes the ad hoc links bandwidth virtually unrestricted in comparison.

Even though we consider the bandwidth of the BS as bottleneck, in an optimised allocation we have to consider the effects of using the ad hoc paths. First, there is increased resource consumption on the relaying MS such as battery energy and processing power. We assume that users would appreciate some incentive for letting other connections use their MS. Second, there is the problem of the weaker QoS offered by the ad hoc route. Delay increases with hop count. Moreover, an ad hoc path might get disconnected due to mobility. We model these relay costs and QoS losses with the help of a path dependent *cost* that is proportional to the number of ad hoc hops and the amount of traffic sent on that path. From a pricing viewpoint, we can regard the utility functions as proportional to the rates the user is willing to pay for a certain connection. In the same manner, incentives proportional to the per hop costs could be regarded as reimbursements to the owner of the MS used as relay.

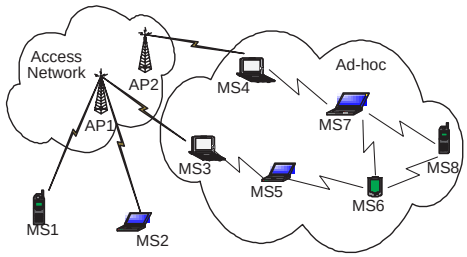

Figure 1. Hybrid network

In order to establish an end-to-end connection the algorithm must choose among a set of reachable BSs, and for each BS there might be several ad hoc paths available. Taking into account the above cost model, it is obvious that the shortest (in the number of hops) ad hoc route to a BS has also the lowest costs. Thus, the path choice in our scheme consists of two phases. First, a shortest path first (SPF) routing algorithm (such as AODV [18]) is employed to find the best paths from an MS to a set of near BSs. A BS is considered near if a shortest path exists, given that the hop-count does not exceed a certain threshold value. Second, the allocation algorithm will use this set of paths in the optimised allocation.

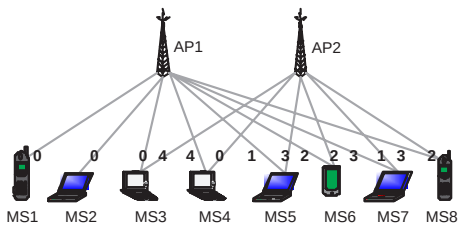

Figure 2. Hybrid network - condensed

Thus the model of Figure 1 can be condensed to the one presented in Figure 2, where the links represent potential paths from the MSs to the BSs. The numbers on the links in Figure 2 represent the number of relays, to which the cost/bit of that path is proportional. A direct cellular connection has zero relays.

4.1. Optimal bandwidth allocation

Assuming each connections has an attached utility function u_i, let's calculate the system-wide utility we obtain with a certain bandwidth allocation (at a certain time point). In this paper, when we refer to a connection we imply an end-to-end, OSI transport layer connection. In the optimal allocation, the packets for a connection (e.g. between an application on a MS and its server in the core network), could be sent over several paths through several BSs. Let there be n active MSs that connect to m BSs. Assume X_{ij} the amount of bandwidth allocated to a connection from MS i over the ij path (if such a path exists) to BS j. The cost the system incurs over a path ij is modelled as $C_{ij} = c \times h_{ij} \times X_{ij}$, where c is a cost constant that represents the cost/bit/hop and h_{ij} is the number of hops. If a direct connection over cellular link is possible, then $h_{ij} = 0$. The existence of a path ij and its hop-count, h_{ij}, is given by the underlying SPF routing algorithm that finds the shortest paths between MS i and the set of near BSs j.

The total utility of the system at this moment is the sum of the utility generated by all connections minus the sum of the costs over all paths.

$$U = \sum_{i=1}^{n} u_i(\sum_{j=1}^{m} X_{ij}) - \sum_{i,j=1,1}^{n,m} c \times h_{ij} \times X_{ij} \qquad (1)$$

Therefore, to derive the optimal value for bandwidth (now represented by a variable x_{ij}) we have to solve the following maximisation problem:

$$Maximise\ U = \sum_{i=1}^{n} \left(u_i(\sum_{j=1}^{m} x_{ij}) - \sum_{j=1}^{m} c \times h_{ij} \times x_{ij} \right) \qquad (2)$$

$$subject\ to: \sum_{i}^{n} x_{ij} \leq X_j^{max} \qquad (3)$$

$$x_{ij} \geq 0 \qquad (4)$$

where X_j^{max} is the maximum bandwidth available in the cell j. If there are no paths between a MS and a BS then the corresponding term is excluded from all the j-indexed sums.

4.2. Linear programming formulation

Lee et al [13] show that maximising $\sum_i u_i(x_i)$ subject to $\sum_i x_i < X^{max}$ (single resource pool), where u_i is a discrete R-U function and X^{max} is the maximum available resource, is an NP-hard problem (closely related to bin packing). They also show that by approximating utility functions with their convex hull frontier, a low complexity algorithm can be used that yields results close tot he optimal solution (1%). If we denote x_i to be the allocated bandwidth to connection i over all possible paths, $x_i = \sum_{j=1}^{m} x_{ij}$, and we set $j = 1$ and $c = 0$ in equation (2) we observe that their problem is an instance of our problem, which makes our maximisation problem also NP-hard. To make the problem tractable, we also approximate u_i with its convex-hull frontier u_i'.

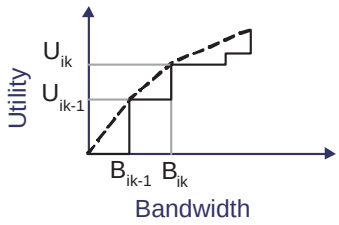

Figure 3. Segment of a R-U function

In Figure 3 we present a utility function (solid line), its convex hull (dashed line) and highlight the $k = 2$ segment of the convex hull. While u_i' is not linear, it consists of several segments $(1, ..., l)$ that are piecewise linear, e.g. three in Figure 3. For a segment k of u_i' let x_{ik} be the allocated bandwidth. The efficiency of the segment being $e_{ik} = \frac{U_{ik} - U_{ik-1}}{B_{ik} - B_{ik-1}}$, its corresponding incremental utility would be $e_{ik} * x_{ik}$. Then, for a certain allocation, we have

$$u_i'(x_i) = \sum_{k=1}^{l} e_{ik} * x_{ik} \qquad (5)$$

where $x_i = \sum_{k=1}^{l} x_{ik}$. To transform our maximisation problem into a linear programming form, we would like to replace $u_i'(x_i)$ with $\sum_{k=1}^{l} e_{ik} * x_{ik}$. However, if we regard the left and right sides of equation (5) as functions, the function on the right hand side is less constrained. Therefore, we need to add the following two constraints. First, the bandwidth of a segment is restricted by its maximum as specified in the utility function, that is $x_{ik} \leq X_{ik}^{max}$, where $X_{ik}^{max} = B_{ik} - B_{ik-1}$. Second, a higher level (segment) cannot be used if the earlier levels are not fully used. That is, if $x_{ik} > 0$ then for all $k' < k,\ x_{ik'} = X_{ik'}^{max}$.

Having expressed x_i with utility-segment related terms x_{ik}, we can go back to the multi-path formulation in equation (2). The bandwidth allocated to x_{ik} can be distributed over a number of paths. Thus, $x_{ik} = \sum_{j=1}^{m} x_{ikj}$ and x_{ikj} is the bandwidth allocated to the utility segment k of connection i on path ij.

Now, the maximisation problem based on u_i' can be for-

mulated as:

$$Maximise\ U' = \sum_{i=1}^{n}\left(u_i'(\sum_{j=1}^{m} x_{ij}) - \sum_{j=1}^{m} c \times h_{ij} \times x_{ij}\right) =$$

$$\sum_{i=1}^{n}\left(\sum_{k=1}^{l}(e_{ik} \times \sum_{j=1}^{m} x_{ikj}) - \sum_{k,j=1,1}^{l,m} c \times h_{ij} \times x_{ikj}\right) =$$

$$\sum_{i,k,j=1,1,1}^{n,l,m} (e_{ik} - c \times h_{ij}) \times x_{ikj} \quad (6)$$

$$subject\ to: \quad \sum_{i,k=1,1}^{n,l} x_{ikj} \leq X_j^{max} \quad (7)$$

$$x_{ikj} \geq 0 \quad (8)$$

$$\sum_{j=1}^{m} x_{ikj} \leq X_{ik}^{max} \quad (9)$$

$$\forall ikk' \text{ if } \sum_{j=1}^{m} x_{ikj} > 0 \text{ and } k' < k \text{ then } \sum_{j=1}^{m} x_{ik'j} = X_{ik'}^{max} \quad (10)$$

Proposition 4.1 *Any solution to equation (6) subject to conditions (7 - 9) respects condition (10).*

Proof Let's assume the opposite, which means that in such a solution there are two utility-segments $k < k'$ of a connection i where $\sum_{j=1}^{m} x_{ikj} < X_{ik}^{max}$ and $\sum_{j=1}^{m} x_{ik'j} > 0$. We know that $x_{ik} = \sum_{j=1}^{m} x_{ikj}$. Let $y = min(X_{ik}^{max} - x_{ik}, x_{ik'})$. The utility generated by segments k and k' is $U_{kk'} = e_{ik} \times x_{ik} + e_{ik'} \times x_{ik'}$. We then take y from segment k' and allocate it to segment k. Then $U'_{kk'} = e_{ik} \times (x_{ik} + y) - e_{ik'} \times (x_{ik'} - y) > U_{kk'}$ because $e_{ik} > e_{ik'}$ for a concave function such as the convex hull. The allocations for other segments being equal, this means that the original allocation was not maximal, so we arrive at a contradiction.

Thus, linear programming can be used to solve the global bandwidth allocation problem optimally and we adopt the LP approach, as a baseline in our comparisons. However, we are aware of the drawbacks of the approach that justify looking for a better solution. These are:

- Centralised allocation. The LP algorithm need to know the state of all the MSs and BS and paths in the whole network to reach an optimal allocation. This is unrealistic for a large network.
- Time complexity. As will be presented in Section 6.3 the LP algorithm is prohibitively computationally intensive for an online allocation.
- Signalling/control overhead. The LP solution might spread the allocation of a connection over several paths through different BSs, which can increase both the logistical overhead and the contention on the MAC-layer.

Note that the topology of the network changes in time, ongoing connections end, and new ones are created. Old resource allocations can break or become suboptimal. To address this, we run our (re)allocation algorithms periodically. This bounds the reallocation rate in the system, even if the rate of events (traffic and topology changes) is much higher. The only disadvantage is that new and rerouted connections must wait until the next allocation time, to receive new resources. Choosing an appropriate period will imply tradeoff between a) utility optimisation and reducing the delay of path establishment and b) the computational and signalling overhead of an allocation round.

4.3. Hybrid-heuristic algorithm

To solve the hybrid resource allocation problem in a distributed manner with less complexity (and overhead) we have devised the following heuristic algorithm. The algorithm can be divided in two parts.

- The *core allocation* algorithm is used to independently allocate resources for each BS. It compares all the different connections requesting resources at the given BS, and the most utility-efficient connections are chosen, taking into account also the incurred relaying costs. That is, at BS j, for each utility-segment k of a connection i a new core-efficiency, e_{ikj} is computed by subtracting the relaying costs of path ij from the original efficiency e_{ik}. Thus, $e_{ikj} = e_{ik} - c \times h_{ij}$. Then bandwidth is allocated in decreasing order of e_{ikj}. Let $comp_j$ be the lowest core-efficiency of an accepted (i.e. a non-zero allocation) utility-segment of a connection, $comp_j = min(e_{ikj} \mid x_{ikj} > 0)$. This parameter characterises the level of the competitiveness of the connections requiring bandwidth at BS j, and will be used by a MS to choose the least competitive BS from its point of view. Note that $comp_j$ depends on both BS capacity and the importance of the contending connections. Similar to the LP algorithm, the "core allocation" part is invoked periodically, to keep the allocation updated.
- The *path choice* algorithm compares the paths to different BSs returned by the SPF routing algorithm, and chooses only one to carry the entire connection. The path choice algorithm will choose the connection to the BS where it assumes it has the highest chance to be accepted. To achieve this, it asks all near BSs about their $comp_j$ parameter. Intuitively, the BS with the lowest competitiveness during last allocation round should give the highest chance of accepting the connection. Nevertheless, the efficiency of the connection will be diminished by the path cost to the respective BS. Therefore among all the paths to possible BSs, the algorithm chooses path ij with $min_{j=1}^{p}\ (comp_j + c \times h_{ij})$, where p represents the number of BSs near to MS i. The "path choice" algorithm is event-triggered: a) at the arrival of a new connection, b) when the current path cannot be sustained anymore (due to mobility, fading, failures). In

this new context of ad hoc paths, handovers can be of two types: the traditional, when the MS uses the direct cellular link and moves out of the BS reach area, or, when one of the relays in the ad hoc path moves out of range. In either case, handovers are dealt with as new connections, only that the already accrued utility is taken into account.

The path cost related component in the "path choice" algorithm also prevents oscillations in the system. By oscillations we mean that at a certain point in time a cell is the target of most of the new connections/handovers, while at the next point all the load is directed to another cell. Due to the cost differences of different paths however, the MSs tend to connect to closer BS if the competitiveness factors are roughly the same.

5. Evaluation setup

To evaluate the behaviour of our hybrid network resource allocation scheme we use a traffic mix that is representative for a future mobile communication network, also used in earlier works [6, 17, 5]. Connections can belong to one of the six application groups presented in Table 1. To create a diverse traffic mix, the maximum required bandwidth and connection duration are not fixed values, but follow a geometric distribution with the given minimum, maximum and mean values (columns 2 and 3). The flexibility of the application with respect to bandwidth reallocations is given in the second column from the right and represents its TARA class.

Table 1. Traffic mix used in the experiments

Applic. Group	Bandwidth Requirement (Kbps)			Connection Duration (sec)			Examples	TARA class	Utility scaling factor
	min	max	avg	min	max	avg			
1	30	30	30	60	600	180	Voice Service & Audio Phone	I	1
2	256	256	256	60	1800	300	Video -phone & Video -conference	II	1/3
3	1000	6000	10000	300	18000	600	Interact. Multimedia & Video on Demand	II	1/10
4	5	20	10	10	120	30	E-Mail, Paging, & Fax	III	3
5	64	512	256	30	36000	180	Remote Login & Data on Demand	III	1/5
6	1000	10000	5000	30	1200	120	File Transfer & Retrieval S ervice	III	1/7

Each of the six application groups have an associated R-U function with a different base shape. All the R-U functions used in the experiments follow the minimum and maximum bandwidth requirements specified in Table 1.

To complete the utility specification, a relative importance has to be associated with each application group. For example, even though one might be ready to pay roughly three times more for a video-phone conversation (bandwidth demand of 256 Kbps), the utility per bit is almost three times higher for an audio-phone application (which requires only 30 Kbps). This information is shown in the rightmost column of Table 1. It represents the utility per bit associated with the maximum required bandwidth (e.g if the maximum required bandwidth of a connection in application group 3 is $4{,}000$ Kbps then the utility for this bandwidth is $4{,}000{,}000 \times 1/10 = 400{,}000$). All the other utility values of the R-U functions are calculated following the given basic shapes.

Our simulations were performed in a simulation environment described by Jonasson [12] and built on top of J-Sim, a component-based, simulation environment developed at Ohio State University [22, 11]. For the linear programming part, we have used the java package from the operation research objects collection (OR-Objects) [10].

Connections arrive on the MSs following an exponentially distributed inter-arrival time with a mean of 15 minutes. All the 6 application groups arrive with equal probability. Mobility is modelled in the following way: the time at which a user moves in a new geographical cell follows a geometric distribution starting from 60 sec and mean 300 sec, with equal probability to move in any of the neighbouring cells. If the MS uses an ad hoc path, a handover will be triggered when the current path gets disconnected. To simulate this, we employ a simple but efficient ad hoc path generation mechanism. Following a geometric distribution with the mean of 300 sec it triggers in a MS a path renewal that discards the old paths and creates p new paths to randomly chosen neighbouring BSs. To each of the paths, a random hop count between 1 and max_hop is attached. The experiments have been conducted with $p = 3$ and $max\_hop = 4$. The cost/bit/hop of using an ad hoc path has been set to $c = 0.02$ (unless otherwise stated).

We have simulated go-around world model to preserve uniformity in our gird. Each cell has a capacity of 30 Mbp/s. For all experiments the bandwidth allocation/reallocation has been performed with a period of 2 seconds. Further QoS controlling parameters have been kept unchanged from our previous TARA experiments [1] [5].

6. Experimental results

In this section we test the performance of the hybrid-heuristic algorithm, that has a low complexity and works in a distributed manner, to show how close it comes to the optimal LP algorithm. Furthermore, we compare with the optimal allocation when using only the direct cellular link (TARA scheme). The comparisons are performed using three key scenarios to expose the characteristics of the hybrid network. The first scenario simulates uneven loaded cells (hot-spots) and we test the load balancing features of the hybrid network. The second scenario exposes the fault tolerant capabilities as we simulate a BS failure, while in the last scenario we test a uniformly balanced setting, where the hybrid network should not perform better than a pure cellular setting. We then go on to show QoS differentiation properties and algorithm timing overheads.

6.1. Accumulated utility as performance

As our main performance metric we use the time-accumulated system utility, generated by all the connections

[1] The drop penalty was set using the following formula $P\_drop_i = 20\% \times u_i^{req} \times avg\_dur$, where u_i^{req} is the maximum required bandwidth, and avg_dur is the average duration, see Table 1. Adaptation time was set to 5 seconds

in the system. We show the behaviour of the system when subjected to increased traffic loads, as marked on the x-axis. The numbers represent the offered traffic load compared to system capacity (e.g. 2.42 means that the offered load traffic is on average 2.42 times the systems capacity). For each of the offered loads we conducted five different experiments (by changing the seed of the various distributions) and plotted the average. The coefficient of variance (standard deviation / average) was less than 0.07 in most cases. Note that an offered traffic overload does not mean the system is in a congested state, since the allocation/admission control mechanism ensures that the system is not accepting more that it can handle (connections can be accepted with less than their maximum requirements).

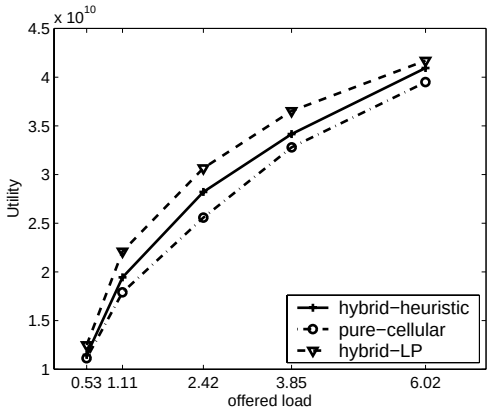

Figure 4. Utility performance for the hot-spots scenario

6.1.1. Unbalanced load. In the first scenario we simulate the effects of a hot-spot area. We simulated a checkered pattern, where half of the cells are subject to a three times higher offered load than the others.

Compared to the hybrid-LP algorithm that represents the optimal allocation, and the pure-cellular which represents the optimal allocation without using ad hoc paths, the performance of the hybrid-heuristic is roughly in the middle, as shown in Figure 4. That is, the heuristic is around 10% better than the pure-cellular, and the LP is around 12% better than the heuristic. This is for moderate overloads such as 1.11 - 2.42. At heavy loads the "lighter" loaded cells are themselves quite overloaded and the algorithms tend to converge.

A more dramatic change can be observed in Figure 5. Here we plotted the utility generated by all connections originating from the MSs located in one of the heavy loaded cell. Being able to connect to the neighbouring cells, allows this set of MSs to generate 30% more utility with the hybrid-heuristic algorithm than in the pure-cellular setting. On the x-axis of the graph we have the half-hour simulated. The offered load in this setup is 2.42.

So, what explains the increase of service for the overload cell by 30% while the overall gain is only around 10%? This is because the higher importance connections that are accepted thanks to the ad hoc paths will replace less important connections in the surrounding cells, so the absolute gain in utility is the difference between the accepted and the replaced. A direct effect of this load-balancing is that the degree of "QoS inversions" has been diminished. By "QoS inversions" we mean that connections/users with a higher importance are rejected in the overloaded cell while less important ones are serviced in the cells around.

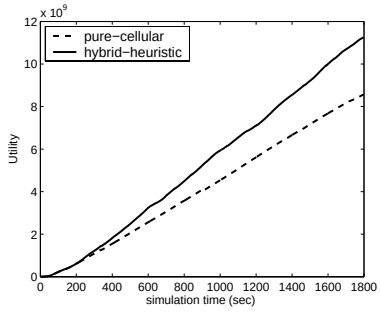

Figure 5. Utility of all connections originating in an overloaded cell

6.1.2. BS failure and dead-spots. In the second scenario, Figure 6, we simulate the extreme case of a BS failure in one of four cells. The difference between the hybrid algorithms and the pure-cellular one increases greatly, as the pure cellular network is clearly handicapped by not being able to use alternative paths to the direct cellular links. A similar situation arises in the handling of dead-spots. Dead spots are areas that are not covered by a BS due to obstacles or interference or big distances. The cell of a crashed BS becomes uncovered area, a big dead-spot. Therefore, these results should be proportionally applicable to dead-spot situations.

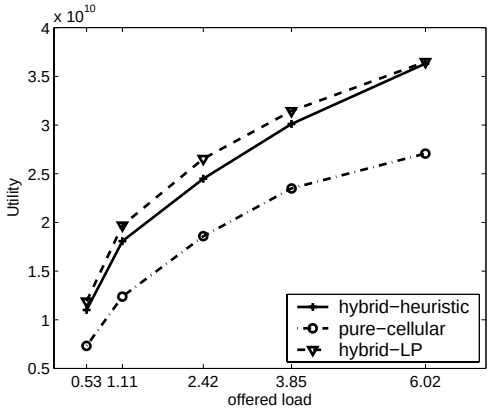

Figure 6. Utility performance for the BS crash scenario

6.1.3. Balanced load. The third scenario presents a balanced load, which means that the MSs located in each cell generate on average the same amount of traffic. This is the

scenario in which the hybrid network behaves most similar to a pure cellular network. When the load and type of the offered traffic is equally distributed, there should be no gain in sending connections to other BSs, especially considering the cost over the ad hoc paths. Nevertheless, at 1.11 (111%) offered load the LP can take advantage of the global knowledge of load and ad hoc paths and shows a 10% improvement over the pure cellular setting, see Figure 7. The hybrid-heuristic relays only on local knowledge to choose the target BS, and acts close to the pure cellular, which is what we expected from such a scenario. As an overall trend we can observe that the highest benefits of using the ad hoc paths are gained for light to moderate overload. At underload and very high loads, the algorithms tend to converge.

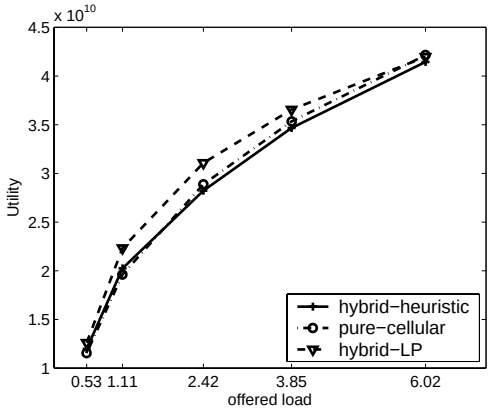

Figure 7. Utility performance for the balanced load scenario

6.2. QoS preservation

So far we have presented the results only from the perspective of the total system utility. Utility is also a good measure for how the QoS of the connections is respected, since each "QoS breach" is penalised (e.g. drop penalty). A more detailed view for the hybrid-heuristic algorithm at load of 2.42 is presented in Table 2. The application groups refer to those in Table 1. We can observe that only connections that have the lowest utility efficiency are blocked (new connections) or dropped (ongoing connections). The allocation algorithms do not treat "ongoing connections" and "handovers" differently, both have a drop penalty attached, so if necessary, the connection with the lowest efficiency will be dropped first. Since application group 6 is a class III connection, it can accept zero allocation situations, so there are no ongoing connections dropped in that case.

Nevertheless it is important to note that the main goal of the system is to generate the highest utility and not to minimise the number of rejected/dropped connections. As the utility for the different bandwidth allocation possibilities can be specified in detail with the help of the R-U functions, it ensures that resources are allocated strictly by importance.

Table 2. Statistics per application group at load 2.42

application group	1	2	3	4	5	6
accepted new	202	191	162	215	216	194
rejected new	0	0	58	0	0	19
rejected ongoing	0	0	22	0	0	0

6.3. Time complexity

Table 3 presents the running time (in seconds) of the simulator for different allocation algorithms, at increasing traffic loads. While there is no difference between the hybrid-heuristic and the pure cellular algorithm, the centralised linear programming algorithm is increasingly slower ($30-200$ times) and definitely not suited for an online allocation.

Table 3. Computational complexity of the algorithms

offered load	0.53	1.11	2.42	3.85	6.02
hybrid heuristic	3	6	9	15	25
pure cellular	3	6	9	15	25
hybrid -LP	92	475	1196	2619	6894

6.4. Cost influence

Using the ad hoc paths means using the equipment of other users in the area, and this translates further into a system-wide cost/hop/bit, c that reduces the gained utility. Until now, we used $c = 0.02$, that is 20% of the efficiency of application group 3. Now depending on how big we assume this cost constant to be, the usage of ad hoc paths will be more or less encouraged. In Figure 8 we plot the dependency of the hybrid-heuristic algorithm on the cost/hop/bit. The baseline is the pure-cellular allocation. At zero cost, the overlay heuristic has a 20% advantage, however, as cost increases the advantage diminishes, and is on par with the pure-cellular at cost 0.1. This is a very high per-hop cost, since in our traffic mix, the efficiency of application group 3 is 0.1. That is, if we send it over a 1-hop path with cost 0.1, no utility would be gained. Thus we can conclude that for reasonable costs the hybrid-heuristic performs as desired. For very high costs $(0.15-0.18)$ the "cost part" of the "path choice algorithm" becomes dominant and the algorithm will use only direct cellular links.

7. Conclusions

To satisfy increasing diverse access requirements and to improve availability, flexibility and higher data rates, today's cellular networks can be foreseen to be replaced by hybrid cellular and ad hoc solutions. Moreover, services and applications with different QoS requirements will compete for the resources of such a network.

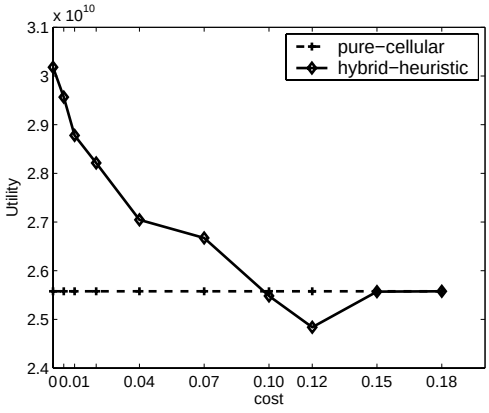

Figure 8. Dependency of the hybrid-heuristic algorithm on cost parameter

Since old performance metrics such as connection blocking/dropping probabilities do not give an adequate measure of relative importance and application requirements (e.g. bandwidth, adaptability), we propose the total system utility as metric. We presented algorithms for resource allocation and admission control that accept detailed QoS specifications in the form of utility functions, and shown their superior behaviour with respect to maximum system utility.

We showed that under reasonable assumptions, the bandwidth allocation problem can be formulated as a linear programming maximisation problem and thus optimally solved. While delivering optimal allocation, the running costs of the LP algorithm make it unsuitable for online allocations. Therefore, we proposed a low-cost heuristic algorithm that aims to send the connection to the BS where it has the highest chance to be accepted.

The experiments show that the hybrid-heuristic algorithm has at worse 12% lower performance than the optimal one. In scenarios dealing with uneven traffic overload or when extending coverage is a necessity, the algorithm showed that it can take advantage of the hybrid setting and performed consistently better when compared to a cellular optimal allocation (TARA). As opposed to the LP algorithm, the hybrid heuristic algorithm works in a distributed manner, being several orders of magnitude faster at runtime.

Future works include relaxing the "virtually unrestricted" assumption on ad hoc network resources and considering a combined routing and resource allocation scheme for hybrid networks.

References

[1] 3GPP. Opportunity driven multiple access. *3GPP TR 25.924 v1.0.0*, Dec. 1999.

[2] G. Aggelou and R. Tafazolli. On the relaying capability of next generation gsm cellular network. *IEEE Personal Communications Magazine: Special Issue on Advances in Mobile Ad Hoc Networking*, 8(1), Feb. 2001.

[3] B. Bhargava, X. Wu, Y. Lu, and W. Wang. Integrating heterogeneous wireless technologies: A cellular aided mobile ad hoc network (cama). *Mobile Networks and Applications (MONET)*, (9):393–408, 2004.

[4] C. Curescu and S. Nadjm-Tehrani. Time-aware utility-based qos optimisation. In *Proceedings of the 15th Euromicro Conference on Real-time Systems*, pages 83–93, Porto, Portugal, July 2003.

[5] C. Curescu and S. Nadjm-Tehrani. Time-aware utility-based resource allocation in wireless networks. *IEEE Transactions on Parallel and Distributed Systems*, 16(7):624–636, July 2005.

[6] M. El-Kadi, S. Olariu, and H. Abdel-Wahab. A rate-based borrowing scheme for qos provisioning in multimedia wireless networks. *IEEE Transactions on Parallel and Distributed Systems*, 13(2):156–167, Feb. 2002.

[7] D. Hong and S. S. Rappaport. Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. *IEEE Transactions on Vehicular Technology*, 35:77–92, Aug. 1986.

[8] H. Y. Hsieh and R. Sivakumar. On using the ad-hoc network model in cellular packet data networks. In *ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC)*, Lausanne, Switzerland, June 2002.

[9] Y.-C. Hsu and Y.-D. Lin. Multihop cellular: A novel architecture for wireless data communications. *Journal of Communications and Networks*, 14(1):30–39, Mar. 2002.

[10] http://opsresearch.com/OR Objects/index.html. Or-objects homepage.

[11] http://www.j sim.org/. J-sim homepage.

[12] R. Jonasson. Simulator for resource allocation in future mobile networks. Master's thesis, Linköping University, Oct. 2002.

[13] C. Lee, J. Lehoczky, R. Rajkumar, and D. Siewiorek. On quality of service optimization with discrete qos options. In *Proceedings of the IEEE Real-time Technology and Applications Symposium*, June 1999.

[14] B. Li, L. Yin, K. Y. M. Wong, and S. Wu. An efficient and adaptive bandwidth allocation scheme for mobile wireless networks using an on-line local estimation technique. *Wireless Networks*, 2:107–116, 2001.

[15] R. R.-F. Liao and A. T. Campbell. A utility-based approach for quantitative adaptation in wireless packet networks. *Wireless Networks*, 7:541–557, Sept. 2001.

[16] H. Luo, R. Ramjee, P. Sinha, L. Li, and S. Lu. Ucan: A unified cellular and ad-hoc network architecture. In *ACM MOBICOM 2003*, San Diego, CA, 2003.

[17] C. Oliveira, J. B. Kim, and T. Suda. An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks. *IEEE Journal on Selected Areas in Communications*, 16:858–878, Aug. 1998.

[18] C. E. Perkins and E. M. Royer. Ad-hoc on-demand distance vector routing. In *WMCSA '99: Proceedings of the Second IEEE Workshop on Mobile Computer Systems and Applications*, page 90. IEEE Computer Society, 1999.

[19] R. Ramjee, D. F. Towsley, and R. Nagarajan. On optimal call admission control in cellular networks. *Wireless Networks*, 3(1):29–41, 1997.

[20] S. K. Sen, S. K. Das, K. Basu, and J. Jawanda. Quality-of-service degradation strategies in multimedia wireless networks. In *Proceedings of IEEE 48th Vehicular Technology Conference*, pages 1884–1888, Ottawa, Canada, May 1998.

[21] A. G. Spilling, A. R. Nix, M. A. Beach, and T. J. Harrold. Self-organization in future mobile communications. *Electronics and Communication Engineering Journal*, 2:133–147, June 2000.

[22] H.-Y. Tyan and C.-J. Hou. Javasim : A component-based compositional network simulation environment. In *Western Simulation Multiconference - Communication Networks and Distributed System Modeling and Simulation*, June 2001.

[23] S. Valaee and B. Li. Distributed call admission control in wireless ad hoc networks. In *Proceedings of IEEE Vehicular Technology Conference (VTC 2002)*, pages 1244–1248, Vancouver, British Columbia, Sept. 2002.

[24] H. Wu, C. Qiao, S. De, and O. Tonguz. An integrated cellular and ad hoc relaying system: icar. *IEEE Journal on Selected Areas in Communications (JSAC)*, 19(10):2105–2115, Oct. 2001.

[25] X. Wu, S.-H. Chan, and B. Mukherjee. Madf: A novel approach to add an ad-hoc overlay on a fixed cellular infrastructure. In *Proceedings of IEEE Wireless Communications and Networking(WCNC)*, Chicago, IL, USA, Sept. 2000.

[26] Y. Yang and R. Kravets. Contention-aware admission control for ad hoc networks. Technical Report UIUCDCS-R-2003-2337, Apr. 2003.

[27] Z. Ye, K. S.V., and S. Tripathi. A framework for reliable routing in mobile ad hoc networks. In *IEEE Infocom 2003*, San Francisco, CA, Mar. 2003.

[28] J. J. Zhou and Y. R. Yang. Parcels - pervasive ad-hoc relaying for cellular systems. In *Proceedings of Med-Hoc-Net*, Sardegna, Italy, Sept. 2002.