# Fair Resource Management in Diverse Cellular Systems

James Irvine[†], Gwenaël Le Bodic[†], Robert Atkinson[†] and Dilshan Weerakoon[‡]

[†] Dept. of Electronic and Electrical Engineering
University of Strathclyde,
Glasgow G1 1XW, Scotland
j.m.irvine@strath.ac.uk

[‡] Centre for Telecommunications Research
Kings College London, The Strand,
London WC2R 2LS, England
Dilshan@iee.org

## Abstract

This paper describes the work undertaken in Core 1 of the UK Mobile Virtual Centre for Excellence (MVCE) programme on Resource Management. MVCE Core 1 was a 56 man year project involving a collaboration of 7 UK universities and 24 companies. The work involved the design of a flexible resource management system which can be used in diverse mobile systems involving multiple standards, air interfaces, networks and operators.

## 1. Introduction

Mobile radio systems are becoming increasingly complex, and user requirements increasingly varied. To obtain the guaranteed quality of service users demand, a simple method is to partition the resources in a fixed manner service by service or system by system. However, in this case unused resources are not available for other services, which may then be blocked. The problem is to share resources for efficiency, but maintain QoS guarantees for different systems and services in a practical manner.

Third generation technologies are resulting in an increasing number of air interfaces and network standards. While standardisation efforts have concentrated on reducing the number of different competing standards, the vast increase in the range of multi-media and data service requirements when compared to predominantly voice 2nd generation services has meant that different solutions are required for different situations.

Pressure has come from regulatory authorities. In many countries, while mobile technologies matured, these authorities have been content to allow a small number of players to control the market in return for the considerable expenditure required to deploy a network. However, recent moves have been towards diversifying the mobile market, not only with regard to service providers, but also with regard to network operators.

In the future, therefore, it is anticipated that networks will become more diverse, with multiple providers, multiple operators, and multiple air interface standards; both cellular and cordless. Macro resource management, where frequency bands are assigned to network operators on long term licenses may well be replaced by more dynamic allocation to allow resources to be assigned where and when necessary.

In addition to this high level diversity between networks, there is an increasing trend towards a lower level diversity of provision within individual networks. This is due to an increasing use of adaptive techniques tailoring the air interface in terms of modulation, coding, bandwidth, etc to the exact requirements of the particular user service and communication condition. Managing a large number of techniques within a network is complex, requiring distributed solutions.

In the light of these trends, the UK Mobile Virtual Centre for Excellence consortium has been conducting research into novel resource management techniques. The central theme of this work has been to develop techniques which will allow resources to be allocated in a 'fair' manner. By fair, we mean allocations which conform to previously agreed service specifications.

## 2. Resource Management Requirements

In order to implement such functions, two levels of interactions can be identified. The first are high level inter-system or inter-network interactions.

A key issue is that of fair access to scarce radio resources. The explosion in cellular subscription has increased demand for indoor coverage. Cordless systems designed for such environments have the advantage of optimised air interfaces and can support high data rates, but bring the problem of increased system diversity and the problem of the management of many small, possibly unco-operative, cells.

Other interactions take place at a lower level between services within the same network. The goal of a fair allocation scheme is to ensure that the guaranteed service quality is maintained while making the most efficient use of resources. The number of services involved make this difficult, as does the translation of requirements between different networks and air interfaces.

Note that a fair allocation may in fact be non-optimal when the system is considered as a whole (i.e. total traffic carried over all users). However, if a user is promised a particular service, that is the service that he or she expects. Users should not be asked to give up quality they have paid for due to the action of other users.

## 3. MVCE Resource Management Architecture

In order to provide a framework for the resource management structure, a new hierarchical architecture is introduced. The architecture is designed to work over a number of different network types to achieve efficient control of resources in varied and possibly distributed systems. The architecture has four layers (see Figure 1), a *digital marketplace*, a *flow controller*, a *service contract manager* and a *radio resource manager*. Each of these layers is similar in that it contracts to the layer above to provide *quality* for a particular *cost* and at a particular level of *commitment*, while using the layer below to provide this. Each layer also chooses between the best contract based on its quality:cost:commitment offering from the entities in the layer below, thus providing devolved management of resources.

Quality is the standard throughput, corruption and delay. The cost could be simply the number of resources required to support the call, but it is possible for the network provider to vary this quantity based on more economic factors such as the effect of other network users or the remaining resource level. The commitment is the probability that the contracted quality will be delivered. In a mobile system with moving terminals and variable channel quality, 100% commitment is impossible to guarantee, and expensive to approach, so the commitment allows lower priority services to trade off availability against cost. It is also possible to define multi-mode contracts which offer different qualities for different proportions of the call. This would be suitable for services which can adapt to varying channel quality, and may therefore be able to secure a lower-cost contract, since the network can instruct the application to switch to a less resource demanding mode if necessary.

The lowest management layer, the Radio Resource Manager (RRM), operates on top of the MAC, and controls radio resources on a specific air interface. The Service Contract Manager (SCM) is a network entity controlling resources at a particular location (i.e. base station of group of interacting base stations). The SCM would be connected to the RRMs of the different air interfaces of that network in that location (cordless and cellular, for example), and can therefore control resource allocation between them and choose the most efficient (lowest cost) interface for a particular call given its quality and commitment requirements.

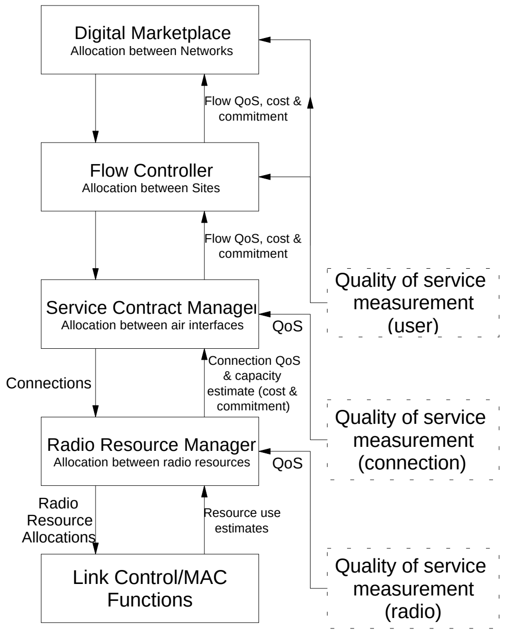
Digital Marketplace
Allocation between Networks
Flow QoS, cost & commitment
Flow Controller
Allocation between Sites
Flow QoS, cost & commitment
Service Contract Manage
Allocation between air interfaces
QoS
Quality of service measurement (user)
Connections
Connection QoS & capacity estimate (cost & commitment)
Radio Resource Manager
Allocation between radio resources
QoS
Quality of service measurement (connection)
Radio Resource Allocations
Resource use estimates
Link Control/MAC Functions
Quality of service measurement (radio)

**Figure 1: Layered Resource Management Hierarchy**

The Flow Controller (FC) is the highest level of resource management within a particular network and is responsible for a call whenever it is in the network. It would switch calls between SCM as the terminal changed location, for example. Again, it is possible that several SCMs could service a call (macrodiversity), and the FC would choose the best for the call's requirements. The highest level, which would not be present in a resource management system for a single network, allows trading of calls between networks. Through a system of contracted control channels, a user or their service provider tender a service contract in an electronic marketplace at the call admission request and the various FCs registered in that marketplace make offers of quality, cost and commitment if they are in a position to support the call. This means that a user can achieve the lowest cost for a particular call without being tied to a specific network operator. It also allows service providers to offer enhanced services by making use of the capabilities of more than one network operator, perhaps by sending different flows of a multimedia call through different networks. For example, the audio portion of a call could be sent through a network with high levels of coverage (and therefore commitment), but to reduce costs the video may be sent through a microcellular network resulting in breaks in that service but a much lower cost. Complex variations are possible, limited more by marketing concerns than the capabilities of the system. A key

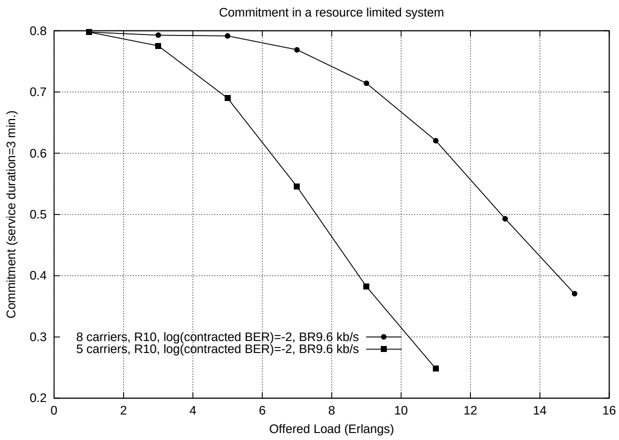

**Figure 2: Trade off between commitment and users in a TETRA system with 10km cells**

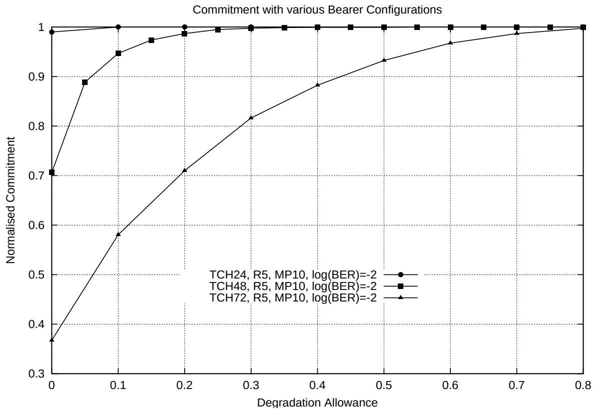

**Figure 3: Commitment and Bearer Configurations**

feature of the marketplace for trading services resides in the ability to develop fairer resource pricing schemes. With these schemes, the price of the radio resource is directly proportional to the ratio between demand and supply of this resource. Innovative services can exploit the dynamics of a marketplace in order to use more efficiently the scarce radio resources. For instance, the download of emails could be performed only when the offered market price drop below a pre-defined threshold (meaning the demand for resources is low in comparison with the associated supply). The marketplace concept is detailed in [1].

## 4. Service Contracts and Commitment

The concept of commitment is of primary important to the resource management system in that it allows a consistent method to trade off quality guarantees against the cost of providing them. In order to see how such a system would work, a comprehensive simulation of a TETRA PMR system has been developed [2]. The system is easily scaleable for future cellular systems with higher bit rates, but TETRA was chosen initially because it is an existing system for which real measurements are available but which has a very diverse set of bearers and QoS requirements which make control difficult. Link adaptation has been added to the TETRA system using the standard bearers. Figure 2 shows the effect of increasing users in a cell for a system with 5 carriers (19 user slots) and 8 carriers (31 user slots). The service chosen, 9.6kb/s data with low delay and 1% residual BER, requires between 2 and 4 slots depending on channel conditions. In both cases, providing a commitment of 75% limits users to roughly half that of a level of 60%. The cost for a commitment of 75% in such a system would therefore be twice that of a commitment of 60%, although economic factors may vary what will be actually charged by the network. Generally speaking, the highest commitment is delivered by reserving the most robust bearer all the time, although this is very expensive. Figure 3 shows what commitment can be offered with various bearer services in the TETRA system. In this scenario, bearer services are the TCH7.2 (7.2kb/s net per slot, no error protection) TCH4.8 (4.8kb/s net per slot, low error protection) and TCH2.4 (2.4kb/s net per slot, high error protection).

Using link adaptation significantly reduces the resource cost, but commitment still reduces due to the time taken to switch between modes but also due to lack of radio resources.

In the mobile environment, consideration must be given to the fact that it is difficult for a network operator to ensure that a certain quality will be maintained for the entire duration of the communication session. Some applications will be dramatically affected by quality degradations of the radio link (like compressed video) whereas the same degradations will not have a significant effect on others (like for voice communications). In order to differentiate the quality offered to different classes of services, the notion of service contract is introduced. The service contract allows a fine-level QoS differentiation. A coarse-level QoS differentiation can also be offered by limiting the possible instantiations of the service contract (to match UMTS service classes for instance). A service contract is specified in generic QoS parameters to make it easily tradable in a digital marketplace. So, mapping functions might be necessary to map the service contract parameters onto network specific parameters. This service contract is composed of several primary performance parameters but also of secondary parameters that allow a quantification of quality degradation. The primary parameters inform on the targeted application requirements and the secondary parameters inform on the application tolerance to non-conformance regarding the primary requirements. For instance, in a circuit-switched

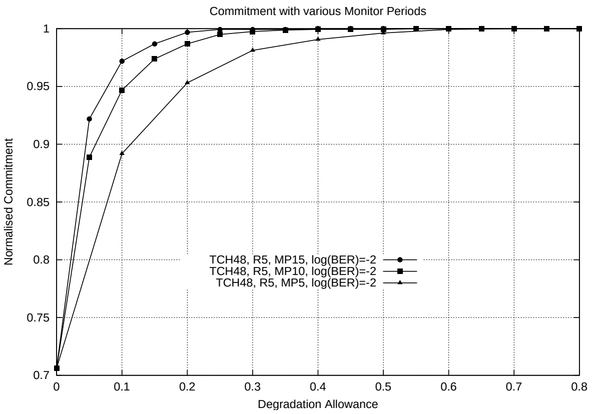

**Figure 4: Commitment and Monitoring Periods**

environment, the three primary performance parameters could be the *BER*, the *bitrate* and the *delay* whereas the secondary parameters would be the *monitoring period*, the *sampling rate* and the *degradation allowance*. Degradation Allowance represents the proportion of quality measures allowed to be non-compliant with the three first parameters over a sliding monitoring period. The sampling rate is the rate at which the system checks what has been delivered against what was contracted. Each parameter of the service contract is fixed or negotiable depending on service adaptability and network capabilities. Tuning the values of the service contract parameters has an effect on the associated resource cost.

Figure 4 shows the effect of the monitoring period length on the commitment probability. The commitment probability is defined as the probability that the associated contract will be committed (the delivered quality meets the contracted degradation tolerance). In this scenario, the degradation allowance varies from 0 to 80% (x-axis). The required BER is $10^{-2}$, the cell radius is 5 km and the bearer configuration is TCH4.8 (Traffic Channel at 4.8kb/s net rate per slots). The monitoring periods considered are 5, 10 and 15 seconds.

From the graph of Figure 4 it can be seen that the system offers higher levels of commitment for long monitoring periods. If short monitoring periods are specified as part of service contracts then the network operator has to be more reactive and/or preventive to network quality degradations. However, it is important that error sensitive applications like video are associated with short monitoring periods whereas longer monitoring periods are acceptable for applications such as voice. When the degradation allowance is 0, meaning non-conformant measures are not allowed, then the monitoring period length does not have any impact on the commitment probability.

The notion of service contract is extended for adaptive applications. In this context, a multi-mode contract specifies the requirement of each mode in which the adaptive application can operate. For instance a video might operate at different frame rates, with different colour depths or frame resolutions. When the RRM is enable to maintain the current operating mode then the SCM can require the associated application to change its operating mode for a one which is less demanding in terms of QoS.

## 5. Resource Management between Services

The programme has also looked at optimal algorithms for the control of resources within each entity. The main concern has been fairness - a user should receive their contracted QoS irrespectively of the service given to other users. This means that globally optimal algorithms which maximise resource use across the system may not be appropriate if to do so they may penalise individual users who have accepted contracts. Much of the research programme has been involved with fair queuing schemes, for example a variation of Self Clocked Fair Queuing for operation within the RRM, and a DCA variant for use within the SCM for arbitrating between different uncoordinated radio ports. When some co-ordination is possible, for example when there is a feedback path to the source, finer control is possible, but one of the problems which becomes quickly apparent is that of measurement, and work has also been undertaken to allow the measurement of local parameters (queue length, delay, etc) and to use these measurements to estimate user QoS and therefore to control the system.

Another research area that investigated in detail within the VCE program is radio resource management schemes for networks which support multiple services with diverse QoS requirements. One of the key resource management schemes proposed is Pre-emptive resource allocation with dynamic partitioning.

Pre-emptive resource allocation with dynamic partitioning: This scheme uses a pre-emptive bandwidth allocation by dynamic partitioning of the total spectrum available. This allows mobile user groups with high priority services to access greater amounts of bandwidth than mobile user groups with low priority services when the network is overloaded. The scheme does not reduce overall trunking efficiency and the network can still guarantee QoS for high priority services under overloaded conditions. Analytical and simulation results show that the QoS for all the services considered is kept closer to the target at the expense of increasing the forced termination of calls already in progress with respect to a whole range of system load.

## 5.1 System Description

Consider a network, which supports multiple types of service with different QoS requirements. The mobile users supported by the network are classified according to their QoS requirements. Users with different types of service are given a different priority [3]. This prioritisation may be based on many characteristics, for example (a). Revenue Optimization: The service providers associate revenue earned for users in each class. Priorities can be based on the revenue structure of the particular service provider, (b) Transmission Rate: The priorities can be assigned on the basis of the transmission rate required by each class or (c) Delay Tolerance: Service classes that have a higher delay tolerance can be buffered longer before rejection, such classes can be awarded a lower priority. This produces n priority groups denoted by $G_1$ to $G_n$, $G_1$ being the service with the highest priority and $G_n$ the service with least priority. Figure 5.(a) shows the frequency spectrum occupied by mobile users in the case of a Frequency Division Multiple Access (FDMA) Scheme (the pre-emptive scheme is also valid for both Time Division Multiple Access (TDMA) and Code Division Multiple Access (CDMA)). The system divides the available bandwidth $C_t$ into n partitions, with each partition having a number of channels that corresponds to the call arrival rate, channel occupancy distribution and QoS requirements of each corresponding priority group $G_j$. Figure 5.b shows the logical bandwidth partitioning The total bandwidth is divided into n logical partitions with the jth partition having $B(t)_j$ channels. $B(t)_j$ is the size of jth logical partition, i.e., the amount of bandwidth group $G_j$ is allowed to have access to with pre-emptive priority over other user groups when the network is overloaded. The number of channels within each partition is changeable as the QoS requirements of mobile user group changes with time. This will allow the system to adapt dynamically according to the traffic load of each service. On the other hand, $U(t)_j$ is the number of channels occupied by each group $G_j$ at time t.

## 5.2 System Operation

The operation of the proposed scheme is described below.

1. The network will classify every new call request into one of the priority levels based on different criteria as below.
2. At the time of a bandwidth request, if there are sufficient resources available, they will be allocated to that user regardless of the priority level they belong to. However, if at the time of the bandwidth request the network is overloaded and all the channels are occupied, the following system operation will be performed.

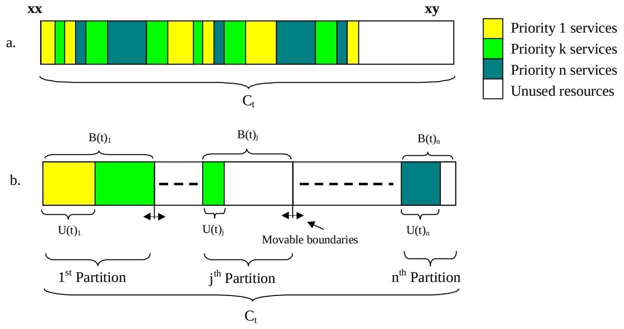

**Figure 5 : (a). Physical channel occupancy by multiple services; (b). Logical channel occupancy by multiple services**

3. If the user requesting bandwidth belongs to the priority group $G_j$, the network will calculate the bandwidth occupied by all the users belonging to that group. Then, the following equation is checked
$$U(t)_j > B(t)_j$$
4. If the above inequality is satisfied, i.e., group $G_j$ occupies more bandwidth than that available in $B(t)_j$, then the mobile user requesting the bandwidth will be denied access to the network.
5. In the case where the above inequality is not satisfied, i.e., users in group $G_j$ occupy less bandwidth than granted in their $B(t)_j$ logical partition, the system will perform the following steps.
6. The system checks if any other priority group occupies more bandwidth than its corresponding pre-emptive bandwidth limit. This search is carried out from the lowest priority group to the highest, i.e. from $G_n$ to $G_1$. If any group occupying more bandwidth than the corresponding pre-emptive bandwidth limit is found, then a user from that group (randomly selected) will be forced to terminate and the released channel will be allocated to the user requesting a channel. If all the groups occupy within their corresponding pre-emptive bandwidth limit, the user requesting bandwidth will be denied access.

## 5.3 Partition size estimation

In this scheme, the amount of bandwidth within each partitioning is adjusted in response to changes in instantaneous call arrival rate [4], channel occupancy distribution and variation of QoS requirements of higher priority services. The aim is to adapt the channel allocation to the traffic variations, by minimally disturbing the existing allocation of channels to other partitions. This adjustment to partition size will allows scheme to perform best. However, under stationary traffic load where the amount of traffic load from each priority group remains constant, there will not be any significant improvement to overall performance. But, under non-stationary traffic conditions, the scheme enables the higher priority services to stay closer to required blocking probability target without significantly reducing the overall performance.

## 6. Resource Management between Sites

This research adopted a complementary approach to that for Resource Management between Services. Rather the issue here is fair access to radio resources from a base station perspective. Much research on fair access to radio spectrum has focussed on techniques for providing fair access for multiple users on a given band of spectrum to the base station. In this case the base station, as the central entity, can coordinate resource usage in a fair manner. Within the context of this research, fairness is concerned with fair access to a common carrier pool by base stations.

The explosive growth of cellular subscribers has lead to operators and manufacturers investigating techniques aimed at increasing capacity. One approach is to introduce smaller cells (micro cells), either to replace or operate in tandem with existing macro cells. This idea can be extended by the introduction of pico cells, forming a hierarchical cellular structure. Pico cells may primarily be utilized in the indoor arena.

Employing a fixed channel allocation (FCA) strategy in the indoor environment is not feasible for a number of reasons.

First is the penalty of reduced trunking efficiency caused by dividing the available channels between indoor stations that are located sufficiently close to cause interference. The problem is exacerbated by the services likely to be used by indoor users. It is widely anticipated that slow-moving and indoor users are more likely to demand higher bit-rate services. If indoor stations are deployed in great numbers adjacent to one another then the number channels available to each will be small. Operators may place exterior stations in locations that allow for a regular frequency re-use plan. This flexibility is not afforded in the indoor arena where base stations are likely to be deployed within the building that they will provide with coverage. After all, it is unreasonable to assume that the proprietor of an office will be willing to house the base station for another business, aside from the obvious practical reasons there are also security implications. The lack of location flexibility combined with FCA may produce large cluster sizes, resulting in inefficient spectrum utilization due to the reduction in trunking efficiency .

The second problem related to FCA is that of financial burden. The financial burden of continual frequency re-planning the macro cellular network currently represents a significant cost to the network operator. If indoor stations are deployed in the numbers that many predict then the financial burden of re-planning may prove prohibitive.

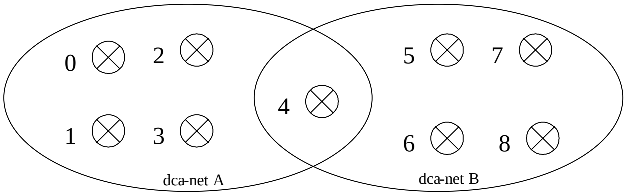

**Figure 6 : DCA-nets**

The solution is to allow base station to co-operate in sharing the available radio resources. That is, a collective of indoor base stations operate a form of dynamic channel allocation (DCA) in a fully distributed manner.

Freed from the constraints imposed by FCA, indoor base stations could be deployed in an uncoordinated fashion. This would transfer responsibility for installation and maintenance of the base stations to the customer or third party companies. The idea is that the indoor base stations operate as cordless/cellular hybrids. Cellular technologies are used, and handover to the exterior network is permitted. However, handover to another customer's base station is not permitted, and channel allocation is performed by the base stations at a local level.

Simulation have revealed that although radio signals from indoor base stations have the potential to illuminate adjacent building and surrounding streets [5]. They provide only minimal interference in other buildings located more than one city-block away. This gives rise to isolated dca-nets, small groups of base stations that are isolated from one another which operate a DCA algorithm. Due to the isolation of the dca-nets, a pure DCA algorithm can be employed, where all base stations have equal access to carriers from a common pool on a demand basis.

Research has shown that this approach increases overall capacity due to the greater trunking efficiency, for both homogeneous and non-homogeneous traffic loads. However, it was shown that considerable unfairness could result. In this context unfairness refers to poorer QoS on some base stations. The poor QoS is a consequence of resource starvation on base stations which are members of more than one dca-net. For example, base station 4, in Figure 6, will be subject to poorer QoS (in this instance, higher blocking probability) since it must share radio channels with members of both dca-nets.

By exchanging QoS information base stations can operate a benevolent policy, whereby base stations that are already satisfying their GoS requirements (e.g. 5% blocking probability) can limit call admissions. The base station thereby frees resources to be used by base stations that are not meeting their GoS requirements. The effect of this benevolent policy, for the previous example, is shown in Figure 7.

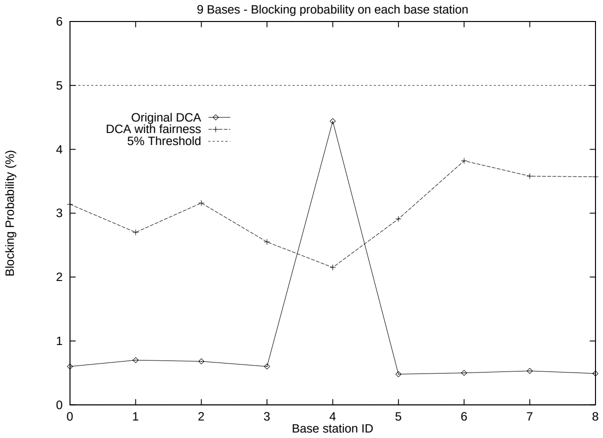

**Figure 7 : Effect of benevolent policy**

The plot shows call blocking on each of the nine base stations, with and without the benevolent call admissions policy. When DCA is applied on it own, users on base station 4 clearly have a poorer QoS, than those on all others. However, with the benefit of the call admissions policy, the GoS becomes more uniform, hence providing a fairer QoS to all members of the dca-nets.

Further research in this area is ongoing, with particular emphasis on the interaction of the DCA scheme and a packet-based speech system.

## 7. Conclusions

This paper has described the Resource Management work undertaken within Core I of the Mobile VCE research programme. The aim of this research was to devise novel resource management techniques allowing fair and efficient allocation of resources within and between increasing complex cellular systems.

At the core of the work has been the development of a management architecture. This has simplified the definitions of resource management algorithms. In addition, the paper has shown how the resource management architecture could be integrated with a concept of marketplace where network operators and service providers can trade electronically communications services. In this context, the notion of service contract is of primary importance. A generic service contract has been specified for and is associated with the possibility to quantify degradation tolerance. These concepts have been illustrated with a number of simulation results for the TETRA system.

The proposed pre-emptive resource allocation scheme allows mobile user groups with high priority services to access greater amounts of bandwidth than mobile user groups with low priority services when the network is overloaded. This scheme does not reduce overall trunking efficiency and the network can still guarantee QoS for high priority services under overloaded conditions. Analytical and simulation results show that the QoS for all the services considered is kept at the expense of increasing the forced termination of calls already in progress. However, the forced termination probability can be reduced significantly at the expense of end-to-end delay by introducing a queuing system. Also, this scheme maintains low blocking probability with respect to the whole range of system load. Another significant advantage is the ability to dynamically change the number of channels allocated for each service as the traffic volume of different service classes changes over long periods.

It has been shown that a distributed DCA algorithm operating on its own may lead to unfair QoS for users on other base stations. Unfairness arises when base stations are members of multiple dca-nets and is the result of resource starvation. A benevolent call admissions policy has been devised which has shown promise in providing a more uniform QoS across participating base stations.

## Acknowledgements

The authors would like to acknowledge the contributions of their colleagues in the programme working on the resource management task : Hamid Aghvami, Osama Baldo, Babak Jafarian (Kings College), Nikolaos Bavetsas, Rahim Tafavolli (University of Surrey), John Dunlop, Demessie Girma (University of Strathclyde) and Mark Rayne (Simoco). The work reported in this paper has been part of the Networks Work Area of the Core I Research Programme of the Virtual Centre of Excellence in Mobile & Personal Communications, Mobile VCE, www.mobilevce.co.uk, whose funding support is gratefully acknowledged. More detailed information on this research is available to Industrial Members of Mobile VCE.

## References

[1] G. Le Bodic, D. Girma, J. Irvine and J. Dunlop, "Dynamic 3G Network Selection for Increasing the Competition in the Mobile Communications Market", *these proceedings*.

[2] G. Le Bodic, J. Irvine, D. Girma and J.Dunlop, "QoS Management With Dynamic Bearer Selection Schemes", to be published *in Proc. of European Wireless 2000*, Dresden (Germany), September 2000.

[3] D. Ayyagari, A. Ephremides, "Admission control with priorities: Approaches for multi-rate wireless systems", *Mobile Networks and Applications* (4), Baltzer Science Publishers, 1999.

[4] O.T.W. Yu, V.C.M. Leung, "Adaptive Resource Allocation for Prioritised Call Admission over an ATM-Based Wireless PCN", *IEEE JSAC*, Vol.15, No.7, 1997

[5] R. Atkinson & J. Dunlop, "Potential for Mobile Coverage Improvement in Urban Areas", *Proc. ICUPC '98*, Florence, Italy, October 1998, pp. 9-13