Available online at www.sciencedirect.com
ScienceDirect
Engineering Applications of Artificial Intelligence 21 (2008) 619–632
Engineering Applications of
ARTIFICIAL
INTELLIGENCE
www.elsevier.com/locate/engappai

An intelligent-agent approach for congestion management in 3G networks

Soamsiri Chantaraskul*, Laurie Cuthbert

Department of Electronic Engineering, Queen Mary, University of London Mile End Road, London E1 4NS, UK

Received 2 June 2006; received in revised form 20 February 2007; accepted 15 May 2007
Available online 12 July 2007

Abstract

The exploitation of wideband code division multiple access (W-CDMA) technology in third generation (3G) networks gives an inherent flexibility in managing the system capacity, although radio resource management (RRM), including congestion management, is more complicated. To guarantee the quality of service (QoS) provided to customers, the concept of a "service level agreement" (SLA) is introduced and these must be managed by the RRM. This work proposes the application of intelligent agents in SLA-based control in the RRM, essentially for congestion management and demonstrates the ability of intelligent agents to improve and maintain the QoS to meet the required SLA. A particularly novel aspect of this work is the use of learning (case-based reasoning—CBR) to predict the control strategies to be imposed. If there is no congestion, the network operates as provisioned, but, if congestion occurs, it is detected by the agent monitoring process and CBR will be used to provide a suitable policy either by recalling from experience or recalculating the solution from its knowledge. With this approach, the system performance will be monitored at all times and a suitable policy can be applied immediately as the system environment changes, resulting in the QoS being maintained.

Keywords: Case based reasoning; Congestion management; Intelligent agent; Quality of service; Service level agreement; UMTS radio resource management

1. Introduction

With wideband code division multiple access (W-CDMA) being used for third generation cellular networks (3G networks), the system capacity becomes more flexible since all users share the same spectrum allocation and use codes to identify themselves from others. Hence, the whole bandwidth can be reused in every cell and the system capacity is limited by the total interference that occurs from other users, other base stations and the background noise. This provides the flexible, higher bandwidth services and maintains the best system capacity, but also leads to more complexity in radio resource management (RRM).

The concept of a service level agreement (SLA) is becoming one of the main interests in 3G networks as it allows network providers (NPs) to offer different levels of service guarantees to different customers (who are paying different rates). In other words, a customer's quality of service (QoS) is guaranteed according to the SLA they have made with the provider.

This work proposes the exploitation of intelligent agents to offer SLA-based control policies (here, they are RRM parameters), to maintain the QoS as guaranteed especially when congestion occurs. Employing a multilayer agent approach offers different decision timescales for resource management. In a stable traffic situation, the decision can be made immediately, but as the circumstances change, a learning approach is used as part of the agent system to identify the congestion pattern and find the best policy (RRM configuration).

The novelty of the work proposed in this paper goes beyond the previous publications of the authors. In that previous work, the congestion pattern can be recognised as it occurs and the agent efficiently determines the best policy to maintain the QoS; here, the agent is further developed to deal with congestion patterns that are similar to those in

*Corresponding author. Tel.: +44 20 7882 5333; fax: +44 20 7882 7997.
E-mail addresses: soamsiri.chantaraskul@elec.qmul.ac.uk (S. Chantaraskul), laurie.cuthbert@elec.qmul.ac.uk (L. Cuthbert).

0952-1976/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2007.05.004

the library but not the same. It also uses the knowledge it has learnt to respond to unfamiliar congestion patterns by implementing a rule-based algorithm. The learning process of the agent is demonstrated here.

In the following sections, 3G networks and RRM are introduced and the proposed agent control system is described followed by the simulation model, monitoring process and congestion pattern recognition; results of applying this technique are also presented.

2. Intelligent agent in SLA-based control system

2.1. 3G The network and the RRM

The 3G cellular system (also called the universal mobile telecommunication system (UMTS)) uses W-CDMA as a multiple access technology in the frequency band around 2 GHz. The system architecture consists of three main functions: user equipment (UE), universal terrestrial radio access network (UTRAN) and core network (Holma and Toskala, 2002; Kaaranen et al., 2001). Fig. 1 shows the general architecture of UMTS.

UE is the 3G network terminal, which contains two separate parts: mobile equipment (ME) and universal subscriber identity module (USIM). The core network covers all the network elements needed for switching and subscriber control. It maps the end-to-end QoS requirements to the UMTS bearer service and also maps onto the available external bearer service when inter-connecting to the other networks (3GPP TS 25, 2001).

UTRAN is the subsystem controlling the wideband radio access, W-CDMA. Radio access bearers (RAB) for the communication between UE and CN are created and maintained in the UTRAN. The UTRAN consists of a set of radio network subsystems (RNS) connected to the core network through the Iu interface and connects to a UE through the Uu interface, which is W-CDMA radio interface. The RNS consists of a radio network controller (RNC) and one or more Node B, essentially the UMTS base station. Each RNS is responsible for the resource of its set of cells. A Node B is responsible for the transmission in a cell or a number of cells.

UTRAN is the focus of this work as RRM is one of the most important entities in the UTRAN. RRM is responsible for maximising the system capacity and delivering the required QoS, given the finite radio resources at the air interface. In general, the RRM algorithms can be divided into five functions: power control (Baker and Moulsley, 2000), handover control (Laiho et al., 2002), admission control (Holma and Toskala, 2002), congestion control (load control) (Muckenheim and Bernhard, 2001) and packet scheduling (Laiho et al., 2002).

2.2. SLA and agent in the RRM

In 3G networks, users are likely to have access to a widespread choice of service providers (SPs) or NPs. They will be able to choose which SP they want to connect to, according to their perception of the QoS (in the general sense) they are receiving or in terms of the service offerings available. The SP will then be able to choose which network to use to carry the service requested (Cuthbert et al., 2001).

In setting up such an approach the SP at least will require some sort of SLA with the network operators and will use these SLAs as part of the decision-making process when allocating traffic. Corporate customers will also want SLAs between themselves and the SP; the SP will also monitor the service performance it delivers to all its customers.

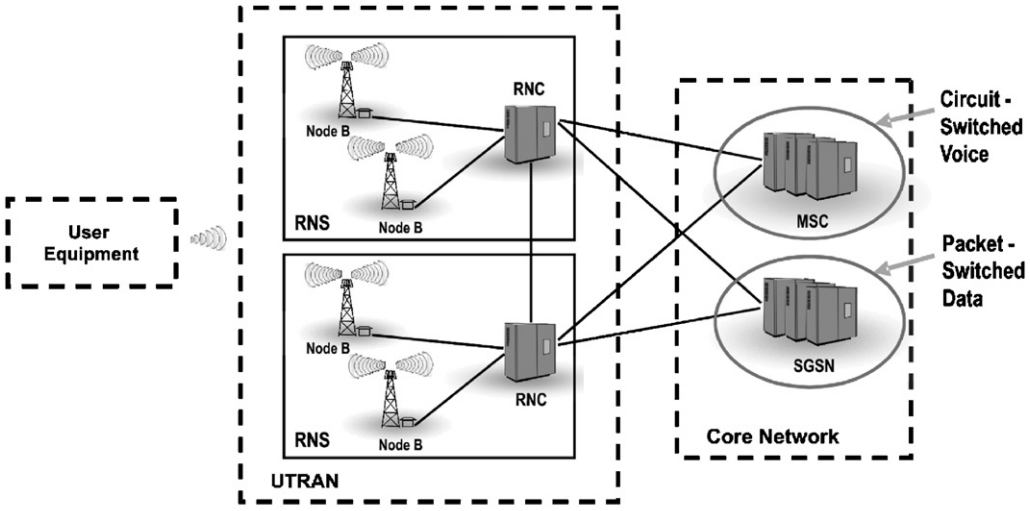


Fig. 1. Overall UMTS architecture.

Referring to the definition by Lewis and Ray (1999) and Marilly et al. (2002), an SLA is a contract between SPs and NPs, or between SPs and customers, that specifies (usually in measurable terms) what services the SP will furnish (the supporting services, service parameters, acceptable/unacceptable service levels, liabilities and action to be taken in specific circumstance) and what penalties the SP will pay if it cannot meet the committed goals. Hence, SLAs allow SPs to differentiate themselves from their competitors and allow them to offer different levels of service guarantees.

Resource management is a very crucial aspect as it aims to maximise the bandwidth utilisation while maintaining SLAs. It can be very complicated as one of the main features in wireless network is that the users are moving and the traffic pattern can change rapidly.

The first work on using intelligent agents to control mobile networks was by Bodanese (2000) (Bodanese and Cuthbert, 2000a, b). This resulted in a distributed resource allocation scheme for first generation mobile networks using intelligent agents that offers an efficient solution for resource allocation under moderate and heavy loads.

The main reason for using intelligent agents is to give greater autonomy to the base stations. The autonomy gives an increase in flexibility to deal with new situations in traffic load and to decrease the information load (the messaging resulting from taking, or determining control actions) on the network.

The most direct previous work relevant to this paper is that of the IST Project SHUFFLE (IST-1999-11014) (SHUFFLE, 2002). In that project, the work of Bodanese was extended to 3G networks. However, the project only offered the hypothesis that the agents could control SLAs: there was no detailed study as to how to implement such control, nor were there any results on SLA management.

In Chantaraskul and Cuthbert (2004), initial investigation is presented to support the idea of improving the congestion situation by changing the SLA-based control policy. Chantaraskul and Cuthbert (2005) give introduction of the implementation of intelligent agent in the control system together with results for the performance evaluation for several congestion scenarios. Here, the system has been further developed to be able to manage the congestion patterns not just the one it experienced. The system is also possible to use its knowledge and offer the suitable control policy to deal with unfamiliar cases and the results showing the system performance in such a situation are provided.

2.3. Agent system and architecture

In current mobile networks, users make connection to a particular SP, who buys network capacity from an NP. Resources are provided by the NP, who owns the physical network. The user will then be restricted to that particular provider until the contract ends or is terminated.

In any business environment, all parties have their own interests, so that there is interest in a more flexible model whereby more choice is allowed depending on performance.

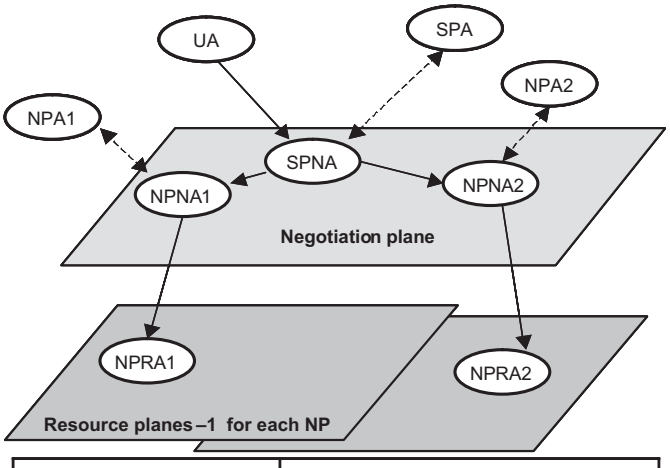
SLA as a concept has been proposed as a contractual arrangement between parties in order to maintain offered QoS and to meet the objectives. Here, a multi-agent system is introduced to give more flexibility of provider selection and in resource utilisation for the best resource management in 3G networks.

The functional architecture used here comes from SHUFFLE (Cuthbert et al., 2001). By using agents, it is possible to allow selection of SP by offering on price, QoS, or by value added service type; it is also possible for SPs to choose an NP on a similar wide range of criteria. The outline functional architecture to achieve this is shown in Fig. 2.

The control concept from SHUFFLE is split into two functional places in the network: the *negotiation* plane and the *resource* plane:

- The negotiation plane is the place where all the interactions between the customers, the SPs and the NPs occur. The communication between different entities' agents will take place here. As users request services, SPs will handle the requests and negotiate with NPs.
- The resource plane is where NPs manage their network resources both across and within individual radio cells to meet the SLAs they have with SPs. This plane is the domain of the NP. In this work, the resource plane will be focused on as the network provider resource agent (NPRA) is a crucial agent that manages the resource within its network.

Both of these planes use intelligent agents to manage their area of responsibility. Detailed explanation on



UA: User Agent	SPNA: Service Provider Negotiation Agent
SPA: Service Provider Agent	NPNA: Network Provider Negotiation Agent
NPA: Network Provider Agent	NPRA: Network Provider Resource Agent

Fig. 2. Illustration of functional architecture adopted by SHUFFLE (from Cuthbert et al., 2001).

individual agents can be found from Chantaraskul (2005). In this work, the resource plane is the focus as it is a crucial part in managing the resource within the network; the important agent here is the NPRA. This agent is the main interest in this work and will be mentioned in more detail.

The basic agent structure is also taken from Cuthbert et al. (2001); each agent follows the concept of Bodanese (Bodanese and Cuthbert, 2000b) who used three layers taking action and decisions on different timescales: reactive, local planning and co-operation.

As an individual connection must have the decision made in real-time, the reactive layer is designed for a very fast response. More complex, and slower acting, functions are implemented in the planning layers. Generally the local planning layer is concerned with long-term actions within its own instance, whereas the co-operative layer is concerned with long-term actions between peer agents, or with other types of agent.

The reactive layer is, therefore, very simple, implementing policies being passed down by the higher layer. It also monitors its actions and feeds back this status information to the planning layers so that they can monitor the effectiveness of their actions.

Fig. 3 illustrates the agent architecture and the detailed structure of the NPRA, which is modified in this work to use case-based reasoning (CBR) in the local planning layer. For the NPRA, the reactive layer is designed to be fast, performing the same function that would be in a conventional RNC, assigning the connection to a Node B, and performing connection admission control (CAC) but it does this according to the policies assigned by the planning layer.

The connection request (containing information about the SP, QoS, type of connection) is first considered for assignment to a Node B using an algorithm or set of rules passed down from the planning layer.

The assigned Node B is the one that is allowed to perform power control, and which subsequently accepts or rejects the call. The assignment and CAC scheme can be passed down as a policy, as could a new scheme. Since these can be changed dynamically, the planning layer can respond to the local resource issues by changing the assignment and CAC strategy. Hence, a suitable policy can be chosen to match the current situation, reporting to the planning layer from the reactive layer in order to maintain the system performance.

The local planning layer of the NPRA is responsible for setting the policies. As explained previously, this can change the reactive layer assignment strategy, or it can change the QoS allowed on the connections within its control.

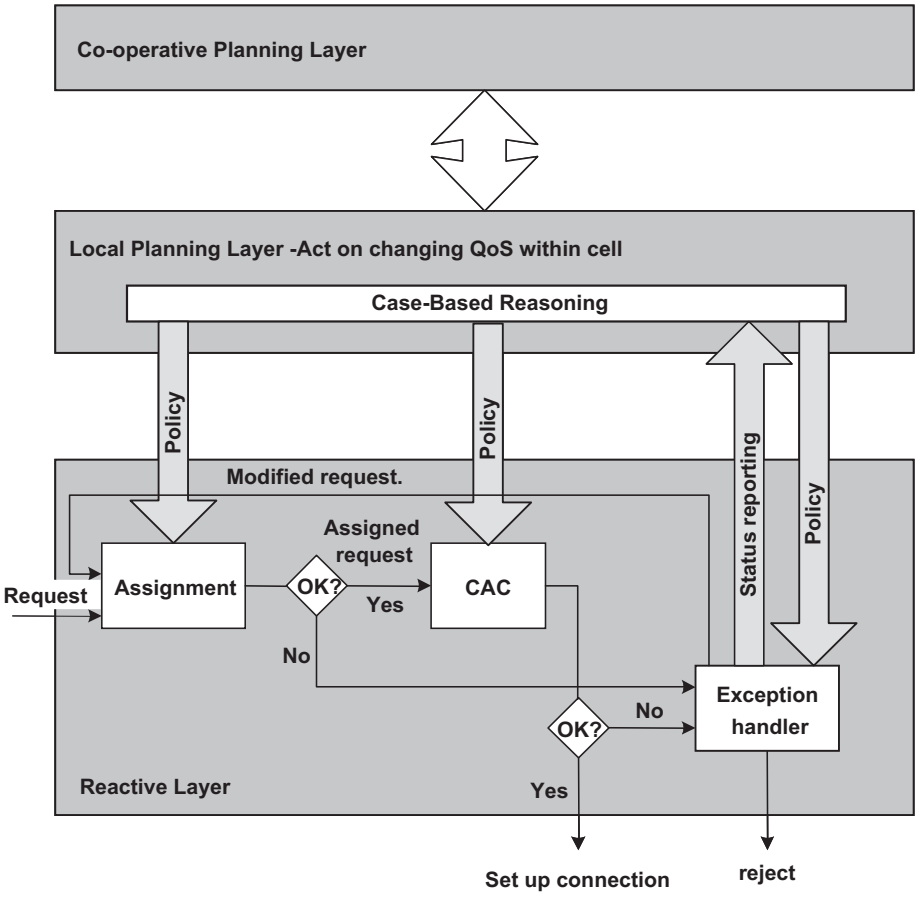


Fig. 3. NPRA (with CBR) internal architecture.

Higher-layer control (at the co-operative planning layer) depends on co-ordinated action between groups of cells and hence is likely to need handling by negotiation between several NPRAs (Du et al., 2003).

2.4. CBR approach in SLA-based control system

As explained earlier, intelligent agents are attractive in controlling communication networks because they can decide for themselves what they need to do to meet their objective, both by using knowledge programmed into the agent and from knowledge they have learnt.

Learning is not only about remembering successful strategies, but also learning from scenarios that did not work in order to prevent such failures to occur again.

Here, the attraction in using learning is to be able to jump straight to a policy (that would be downloaded to the reactive layer) that worked in a similar situation previously. By using such a learning method, the response time will be much faster because the planning layers would not need to approach each situation *ab initio*, but would be able to recall a solution that matches, or is closer to, the situation at hand.

CBR is one AI approach that allows the agent to learn from past successes. It is a method that finds the solution to the new problem by analysing previously solved problems, called *cases*, or adapting old solutions to meet new demands. From the background study done in this work (Chantaraskul, 2005), CBR has been chosen as different traffic patterns can be studied, and they together with the solution can be indexed in the case library to use for solving future problems.

The process model of CBR used here is based on the CBR cycle in Aamodt and Plaza (1994). The CBR process starts when there is a new problem or new case happening. The first step is case retrieval, which uses the characterising indexes of the event to find the best-match solved case(s) from the case library. The solution from the retrieved case(s) will be reused.

However, the solution might need to be modified to fit the new situation as the new situation will rarely match the old one exactly: this step is called 'revising'. Once the new solution is proposed, the next step is to test it with the real environment. The result is either success or failure. If the solution fails, a monitoring process will analyse the failure, repair the working solution and test again. If the solution succeeds, this new solution will be indexed and retained in the case library to use for future problem solving.

The work reported in Caulier and Houriez (1995) gives an example of using CBR in network traffic control by using it to control traffic flow in the standard public switched telephone network of the Ile de France. In other work in Hassanein et al. (2001), CBR is used to correct the error estimation of the required bandwidth computed by conventional CAC schemes.

In the work described here, CBR is introduced to find the solution for the particular congestion pattern recognised in the network.

3. Simulation model

In order to test the system performance, the simulation model is developed. Here, a system with nine hexagonal cells is used, with 1 km radius; each cell has its own base station with an omni-directional antenna placed at the centre of the cell, although sectoring could easily be included. There are three different classes of users: gold, silver and bronze. A number of mobiles have been generated randomly according to input traffic pattern with 50% of users being bronze, 30% silver and 20% gold. It is assumed that the gold customers will pay the highest service charge followed by silver and bronze customers, so that the gold customer is paying for the best service and more flexibility (changing RRM parameters in order to ascertain QoS) than the others.

The model consists of two traffic types, voice and video. For voice traffic, the approach of Kuri and Mermelstein (1999) is used with an activity factor of 0.45. The transmission rate for voice traffic is assumed to be 8 kbit/s and mean holding time is 180 s. For video traffic, an activity factor of 1 has been used (Angelou et al., 2002; Kuri and Mermelstein, 1999). The transmission rate for video traffic is assumed to be 64, 144 or 384 kbit/s and the mean holding time is 300 s. With this approach of using activity factor, voice traffic is assumed to be present (counted as interference to other users) 45% of time and 100% (throughout the connection time) for video traffic.

The equation below shows the relationship between the transmitted power and received power as being used in the radio propagation model (from Liu and Zarki, 1994):

$$P(r) = 10^{\xi/10} \cdot r^{-\alpha} \cdot P_0, \tag{1}$$

where $P(r)$ is the received power, P_0 is transmitted power, r is the distance from the base station to mobile, ξ in decibels has a normal distribution with zero mean and standard deviation of σ (typical value of 8 dB) and α^2 represents the gain (typical values of α in a cellular environment are 2.7–4.0).

For the uplink capacity limited, the SIR of each transmission is calculated at the base station and it can be expressed as follows (based on Capone and Redana, 2001):

$$\mathrm{SIR}_i = \left(\frac{W}{R_i}\right) \cdot \frac{\mathrm{Pr}_i}{(I_{\mathrm{intra}} + I_{\mathrm{inter}}) \cdot \mathrm{AF} + N_{\mathrm{thermal}}}, \tag{2}$$

where (W/R) is the processing gain, Pr is the received signal strength, I_{intra} is the sum of the received signal powers of other transmissions within the same cell, I_{inter} is the sum of the received signal powers from the other cells, AF is activity factor and N_{thermal} is the thermal noise power.

Power control in UMTS consists of three main functions: (i) open-loop power control (ii) inner-loop power control, and (iii) outer-loop power control (Laiho et al., 2002). As the simulation focuses on the uplink-limited capacity, power control for the uplink is applied. In this work, the first two types are applied in the simulation since they have the major effect on the simulation result. Without outer-loop power control, the target SIR has to be fixed; here, it is assumed to be 6 dB and the threshold is 4 dB (Kuri and Mermelstein, 1999). The power control step is assumed to be 1 dB at each power control cycle (Baker and Moulsley, 2000; Thong and Bigham, 2002).

SIR-based CAC is chosen here for the assignment and admission scheme with uplink capacity limitation (which means the signal to interference of the received signal from mobile to base station is calculated). It has been modified to make sure none of the existing connections will be dropped when accepting a new connection request.

In summary, the system environment can be controlled by generating connection requests accordingly. The admission of the new connection request is done by calculating its SIR as shown in Eqs. (1) and (2). The new connection request is accepted if the calculated SIR is not less than the target SIR. For the existing connections, power control is used to regulate the transmitted power.

According to Kolodner (1993) there are several proposed schemes of organisational structure and retrieval algorithms for CBR. In this work, the hierarchical memory with parallel search is used as it provides an efficient retrieval that is less time consuming, as the matching and retrieving happens in one step, which also gives less complexity.

All cases in the library are organised according to the hierarchical memory scheme. Here, the cases that share the same feature will be clustered together and this method has been mentioned as a shared-feature network. The most important feature will be used on the top to differentiate all the cases and followed by less important ones. This refers to the diagram shown in Fig. 4.

As the parallel search has been chosen for the CBR model, the whole library will be searched for each characterising index in one step. Together with the hierarchical memory, the search will start from the most important feature, followed by the next important one and so on. In this way, the number of matched cases will get smaller and smaller until the best match can be achieved.

If the new case is to be retained in the library, the library index has to be re-sorted according to the priority of the characterising index of the new case.

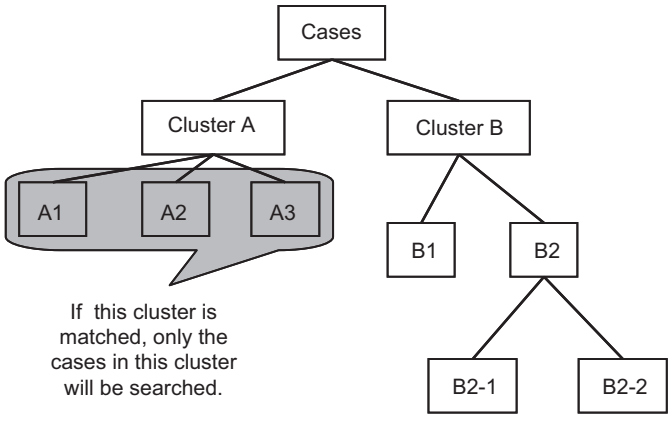


Fig. 4. A shared-feature network.

4. Monitoring and congestion pattern recognition

4.1. Simulation scenarios and SLA assumption

The system environment is controlled to give different types of congestion patterns. This section introduces the cases (or congestion patterns) in the initial library for the CBR models implemented here. There are two main scenarios being tested here: the random overload cases and the hotspot cases. These initial cases represent the common congestion scenarios. Fig. 5 illustrates the initial case library structure with the shared-feature network. The most important parameter, which is used to differentiate the two main clusters, in this case is the congestion parameter for each cell. More details are given in Section 4.2.

For the random overload cases, the whole system traffic load is increased from a normal traffic level[1] when the simulation reaches stability. Stability and verification have been carefully considered in setting up this simulation. There are three subcases for the random overload scenario differentiated by the accumulative value of call blocking rate for gold and silver customers, because it is not generally short-term violations that are important in SLAs. An SLA might specify, for instance, that the blocking rate must not exceed a certain value during a day or months. In other words, this specification is intended to guarantee that the customer will not be blocked again and again in a certain period of time as that could cause frustration, which might lead to the customer changing network operator.

The three random overload cases consist of

- Random overload case with accumulative blocking rate of gold customer exceeding the limit.
- Random overload case with accumulative blocking rate of gold and silver customers exceeding the limit.
- Random overload case with accumulative blocking rate of silver customer exceeding the limit.

For the hotspot cases, the traffic load is increased to create a congestion environment as the simulation reaches stability in a particular cell or a particular area within that hotspot cell. In order to identify the area of congestion

[1]Here, normal traffic situation means system is under acceptable traffic load as it utilises the default policy and QoS is maintained according to SLAs.

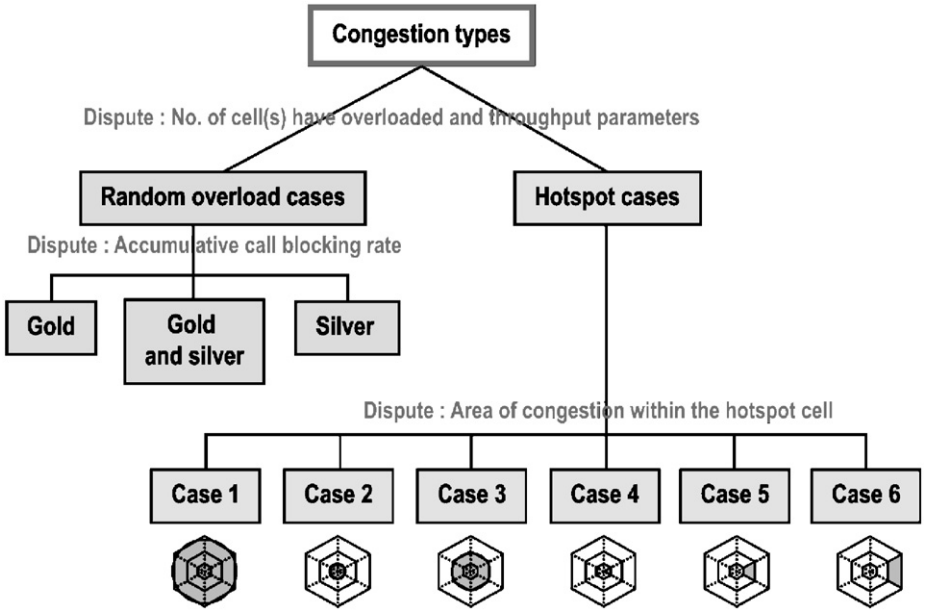


Fig. 5. Case library structure according to shared-feature network.

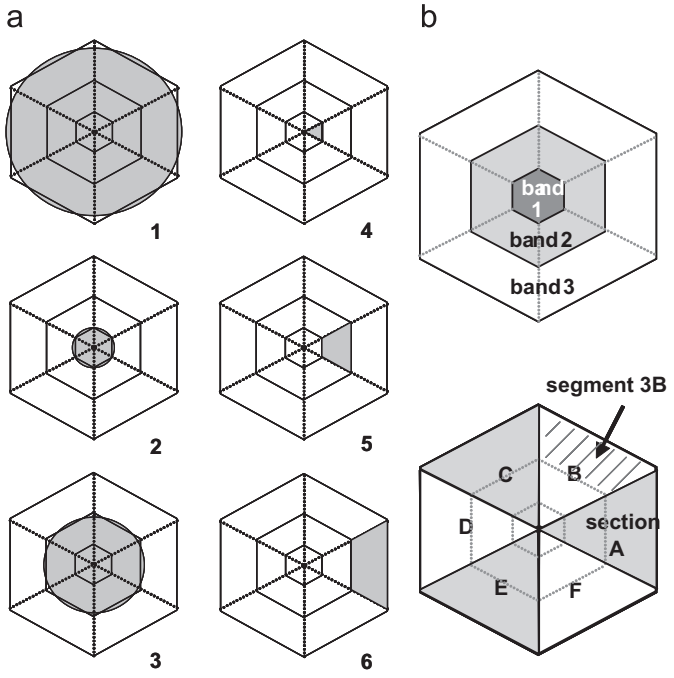


Fig. 6. (a) Hotspot cases illustrated by hotspot cell layouts. (b) Hotspot cell monitoring areas.

within the hotspot cell, the hexagonal cell is partitioned as shown in each hexagon in Fig. 6(a). As the simulation runs, traffic will be monitored and data are collected for each subsection of the hotspot cell. A particular sector where congestion occurs can be identified by using the users' exact locations to calculate the offered throughput, which represents the amount of traffic requested, in each sector then comparing against the set threshold. In order to position the mobile terminal, positioning data from three or more separate base stations is required. Ahonen and Eskelinen (2003) give an overview on terminal positioning in UMTS. There are six subcases being initially tested for the hotspot environment as shown by six hotspot cell layouts in Fig. 6(a). Fig. 6(b) illustrates in detail the hotspot cell monitoring areas, which first divides the cell into bands 1–3 then subdivides each band into six segments.

These hotspot cases cover a variety of patterns as in reality the hotspot area could be large or small and also it could be in any specific part of the cell, not only around the centre.

It has been stated previously that congestion is determined by the deviation of monitoring parameters from the SLA. As mentioned before, SLA is an agreement between provider and provider or provider and customer to guarantee their service quality according to the contract and the price they pay for. Hence, each SLA would have its own individual specifications according to the satisfaction of the parties.

Here, the SLA assumptions made for the maximum acceptable level of call blocking rate are:

- Maximum acceptable rate for gold: 0.03.
- Maximum acceptable rate for silver: 0.05.

Note that the SLA assumption for the maximum acceptable call blocking rate for bronze customer is not considered in this work. Experiments were done to confirm that the chosen connection and admission control would maintain the call dropping rate as congestion occurs and affects the call blocking rate more.

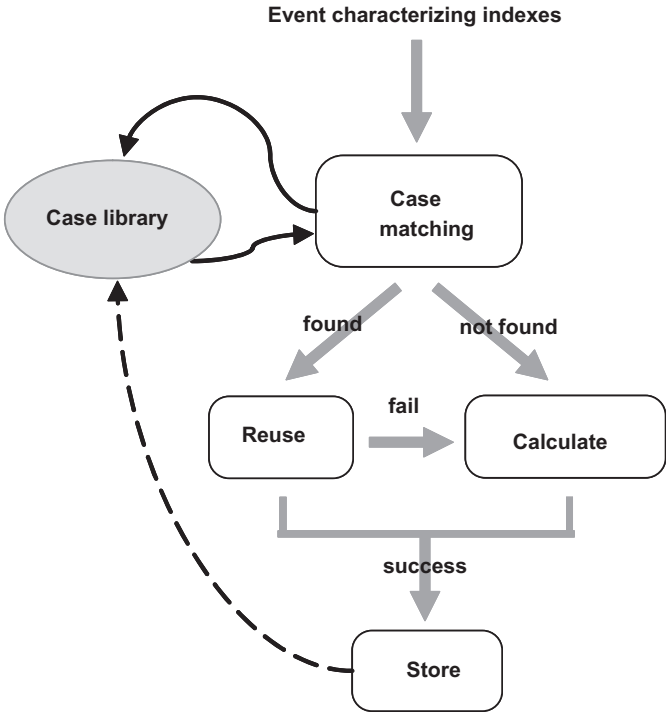


Fig. 7. Case matching process using CBR approach.

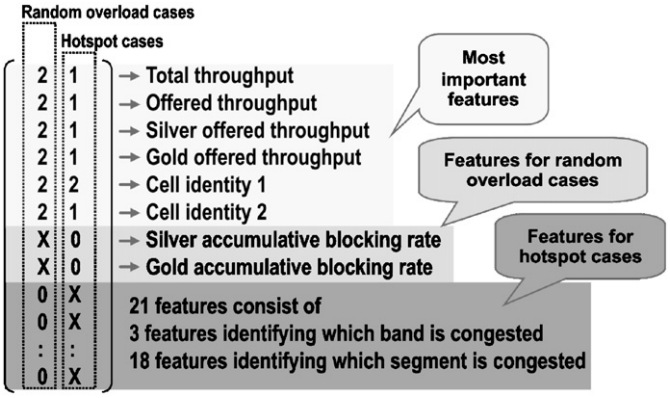


Fig. 8. Case library structure for the case matching.

4.2. Monitoring and case matching process for congestion pattern recognition

Monitoring parameters are collected periodically and sent to the local planning layer of an agent where the CBR model is located as shown in Fig. 3. The parameters will then be compared with the SLA requirements and any deviation from the SLA can be reported. The CBR model will then be used to find the best solution for the situation.

Bases on the CBR model described earlier, Fig. 7 shows the CBR model as being adapted to use in this simulation.

As congestion starts to be reported, the characterising indexes of the event will be sent to the CBR model, which will try to retrieve a base-match solved case from the case library. If the match can be found, the solution will then be proposed to apply in the system; if a near match is found then that will be used, but the performance will be monitored and if necessary the case will be *revised*. A rule-based algorithm is used to compute the solution if there is no match or if the near match fails. Fig. 8 shows the case library structure for the case matching process implemented here. In order to identify the best match, tolerance ranges for each characterising index are set; these tolerance ranges are used to identify the threshold of the matching process. The characterising indexes are filtered using the tolerance ranges into small integer numbers as shown in Fig. 8. These new characterising indexes are then used to find the matched case by retrieving only the exact match from the case library.

The first set of characterising indexes (first six parameters shown in Fig. 8) describing reported congestion scenario is used to describe which main congestion pattern the reported situation falls into. By doing this, the system can differentiate the random overload case from the hotspot one.

In the case of random overload, there are two extra parameters used to identify which subcase is occurring as follows:

- Accumulative blocking rate for silver class.
- Accumulative blocking rate for gold class.

The situation is more complicated when the hotspot case is detected. A two-step hotspot pattern identifying method has then been introduced. The process is to identify which area (referring to bands 1–3 in the hotspot layout in Fig. 6(b)) has congestion in the first step. The second step will look into that congested band(s) and find out the subarea(s) (or segment) that has/have congestion.

The method utilises thresholds for the offered traffic in each band. Fig. 9 illustrates the mechanism by presenting the identification process of an example of congestion pattern.

The first step, which classifies the congestion in bands 1–3, is depicted in Fig. 9(a). Each dotted line represents the limit set for each band. From the first step, bands 1 and 3 are determined to be overloaded. Fig. 9(b) and (c) illustrate the second step. Each histogram represents the offered traffic plot for bands 1 and 3 (where congestion is detected) and expands into the six subareas (segments A–F in Fig. 6(b)). Again, the dotted line in each histogram is set as a threshold. The histogram for band 1 shows that all subareas are under high traffic load while the histogram for band 3 shows that only segment A is under congestion. As a result, the congestion area in this hotspot cell can be taken as shown in the layout on the right-hand side of Fig. 9.

4.3. Rule-based algorithm

To complete the CBR model, a simple rule-based algorithm is set up to show that cases that are not matched can be dealt with. More detail will be given along with the simulation results in the next section.

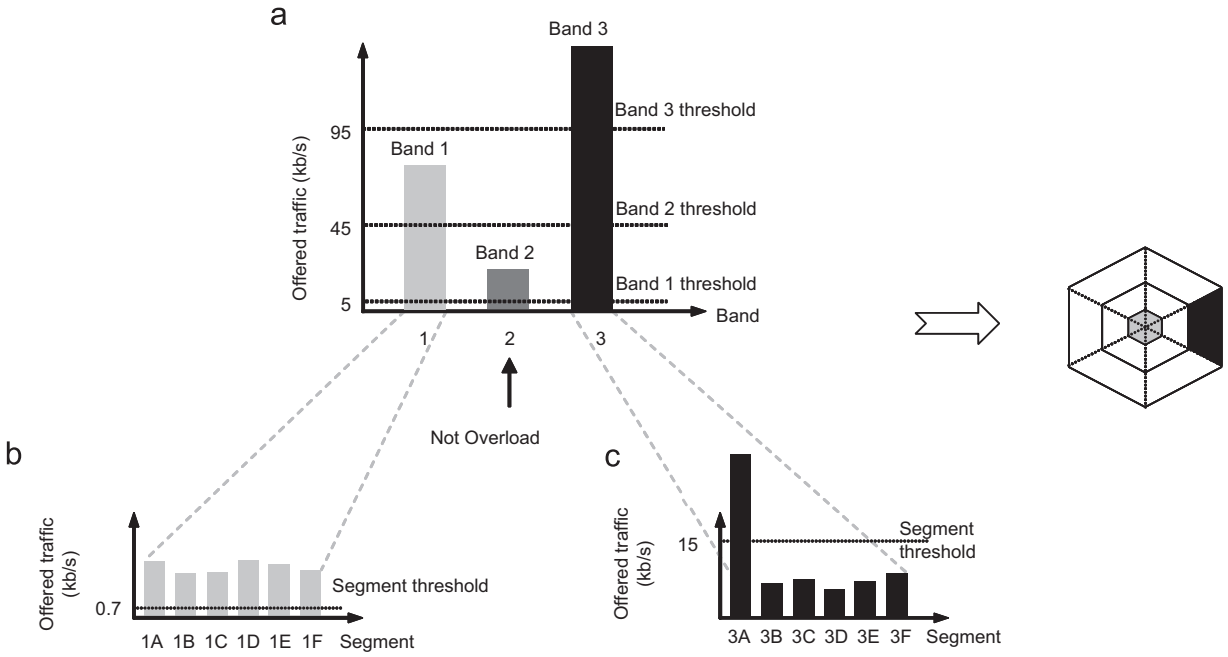


Fig. 9. A two-step hotspot pattern identifying method.

5. Numerical results and discussions

5.1. Generating cases

In this section, the generation of cases and their solutions for the CBR model is the focus. Several methods have been taken into account and the configurations used to recover from congestion have been chosen based on experiments, so the solution for each case is achieved by selecting the method and parameters that provided the best result. The results are categorised into random overload cases and hotspot cases, which show the system performance when the selected solution for each congestion pattern is implemented.

The system performance is generally monitored as *call blocking* and *call dropping rate*. The results shown here, however, are for the call blocking rate. The reason is that the assignment and admission control used here is designed to favour existing connections over the new ones, hence having a greater effect on call blocking; favouring established connections is a well-known approach in mobile networks.

5.1.1. Random overload cases

Fig. 10 shows the results from the random overload cases. Fig. 10(a) demonstrates the comparison between the call blocking rates across the simulation time for the conventional system (dashed lines) and the SLA-based control system with CBR approach (solid lines) as the traffic load increases for the first random overload case, where the whole system is randomly overloaded with the accumulative blocking rate of gold exceeding the limit. The conventional system means the ordinary system that does not include SLA-based control (no change in policy).

Without the controlled system, the call blocking rate is increased for all types of customer and exceeds the limit set for gold and silver customers. With CBR, the pattern of the congestion can be recognised by the control system. The technique used here to control the congestion and give adequate service to gold customers is to adjust the buffering time for each customer class so that a connection request that cannot be served immediately is put in the queue in case resource becomes available. In this case, gold has the longest buffering time (4 s) and lower value for silver (2 s) with that for bronze still at 0. It can be seen that the call blocking rate for gold is maintained at the expense of both silver and bronze.

Fig. 10(b) shows the results from the control system of the next random overload case, where both gold and silver accumulative blocking rates exceed the limit. In this case, both silver and gold QoS need to be handled. By giving highest buffering time to silver and slightly lower for gold, the blocking for both can be kept within the range. As the buffer in this implementation uses the priority arrangement, gold customers are always at the top of the queue. Therefore, in order to also give priority to silver customers, their buffering time has to be higher (the buffering times for gold, silver and bronze customer have been set as 2, 4, and 0 s, respectively). The result for the last random overload case is shown in Fig. 10(c). The situation is that the long-term value for gold customers has been met, but that for silver is at the limit. Therefore, silver customers have to be given priority in order that their long-term blocking is not exceeded, but gold customers can be allowed to have worse service since there is still "slack" in

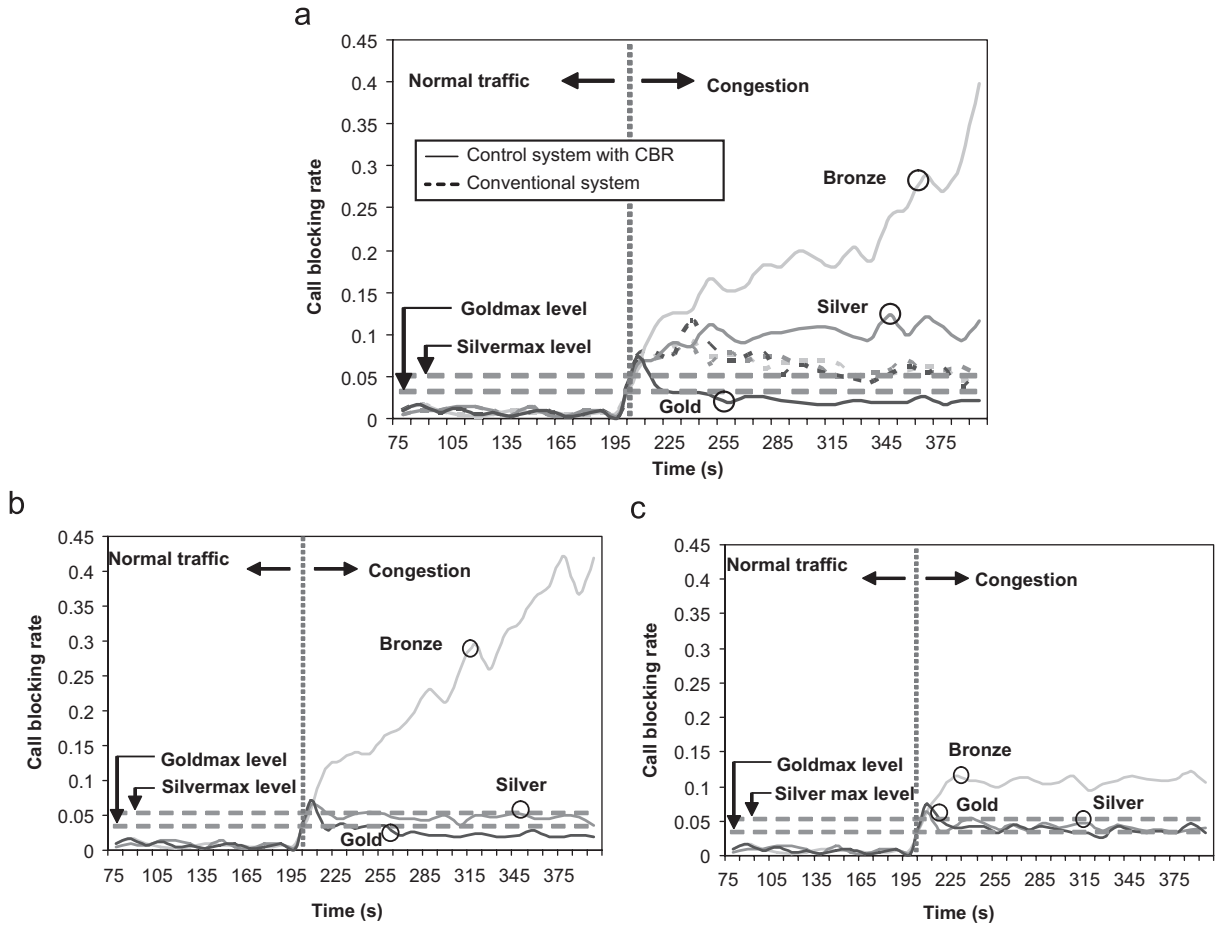


Fig. 10. Simulation results from the random overload cases. (a) Comparison between the results from conventional system and control system for the first random overload case. (b) Result from control system for the second random overload case. (c) Result from control system for the third random overload case.

their SLA. In this case, the buffering times for gold, silver and bronze customer have been set as 0, 1, and 10 s, respectively. It should be noted that the results from the conventional system were not included in Fig. 10(b) and (c) as they have similar trend as the dashed lines appearing in Fig. 10(a).

5.1.2. *Hotspot cases*

The six cases have been tested and provided with the best solution for each case in the case library. The following result is an example of results achieved from the experiments.

Fig. 11 shows the results from the hotspot case 1 (refer to Fig. 6(a) for the number of hotspot cases), where the whole hotspot cell is overloaded. The result from the conventional system is illustrated in Fig. 11(a) and the one from the controlled system is in Fig. 11(c). The solid lines are the call blocking rate across the simulation time as congestion occurs for each class of customer in the hotspot cell and the dotted lines are the result from normal traffic cell.

From Fig. 11(a), it can be seen that as congestion occurs in the hotspot cell the gold customer has already exceeded the limit while all blocking rates from normal traffic cells are in acceptable range. As congestion is detected by the controlled system, the CBR engine proposes a new policy, this time by shrinking the centre cell and transferring the users near the edge to neighbouring cells; it also only accepts new connection requests within this area until the situation is back to normal. Here, the hotspot cell is shrunk to a radius of 400 m.

Fig. 11(b) demonstrates the shrinking method. It can be seen from the result in Fig. 11(c) that the call blocking rate for all customer classes is kept within the limit as displayed by the solid lines with a small compensation from the neighbouring cells represented by the dashed lines. Comparing with the dashed lines in Fig. 11(a), the values are higher but still acceptable.

5.2. *System performance with similar cases*

A further case study is presented in this section by examining the system performance with different (but similar) congestion patterns to the existing cases (hotspot cases only). As mentioned, the CBR engine retrieves a case and solution for either an exact or close match. Here, an

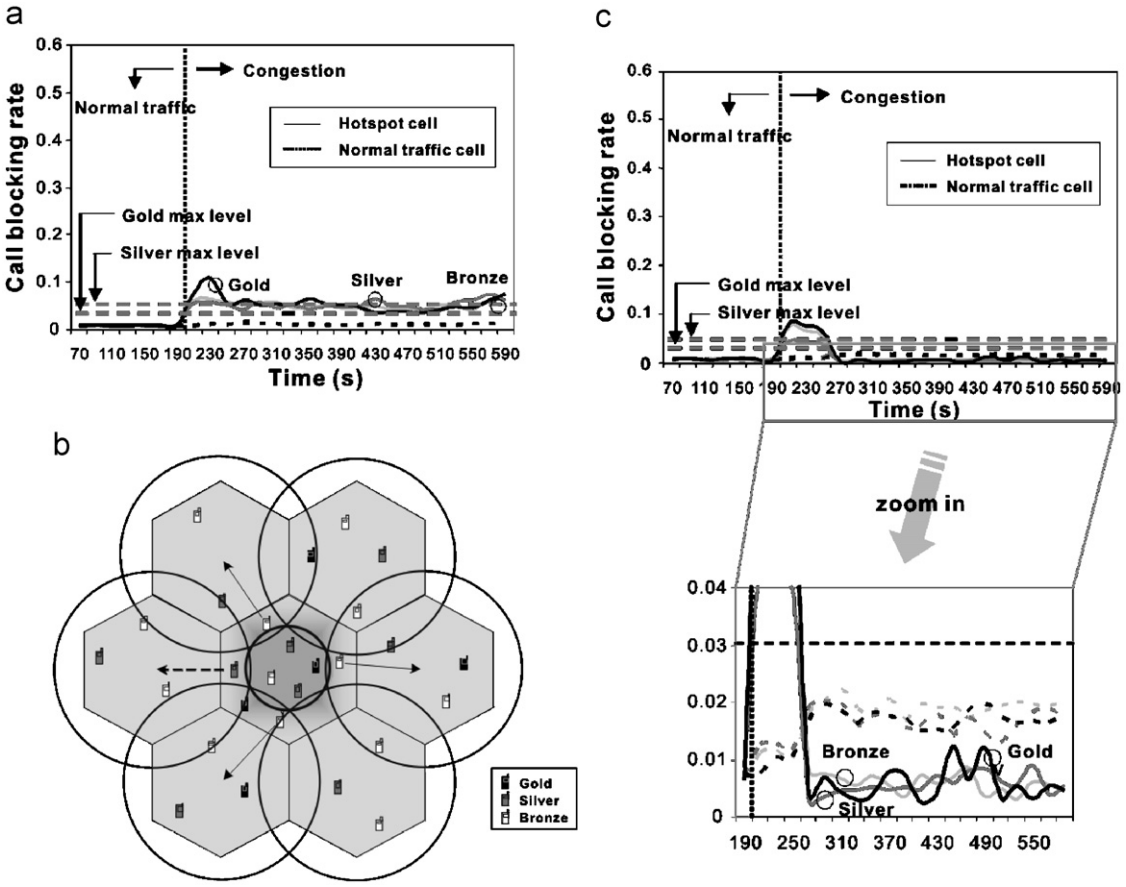


Fig. 11. Simulation results from the hotspot case 1. (a) Result from the conventional system. (b) Cell shrinking method. (c) Result from the control system.

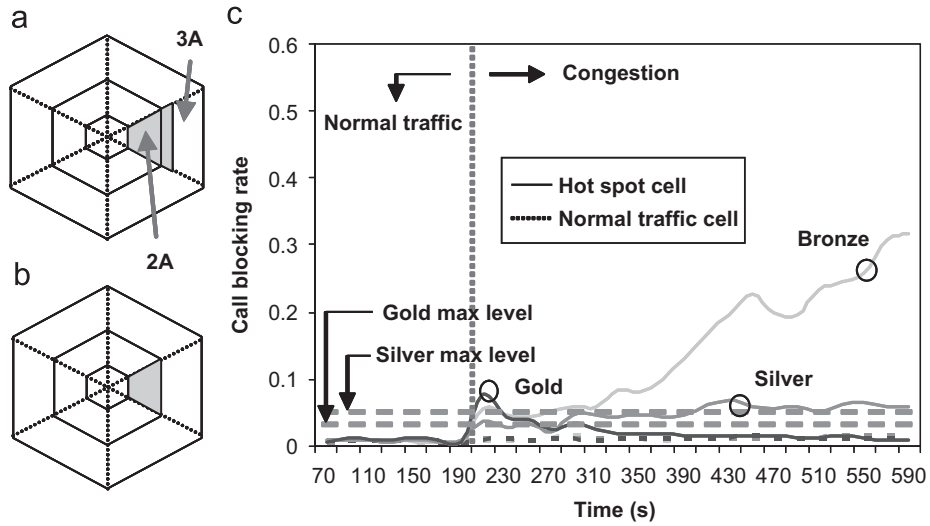


Fig. 12. Simulation result for similar case. (a) Hotspot cell layout. (b) Hotspot cell layout of the matched case. (c) Result from the control system.

example of result is given from one of the experiments, which is designed to observe the system in this aspect.

The congestion pattern tested here is shown by the cell layout in Fig. 12(a). As can be seen, the congestion pattern is close to that existing in the case library (case 5 in Fig. 6(a)), shown in Fig. 12(b). This illustrates the case matching process of Section 4.2 and shows that, by using the tolerance ranges, the close match is retrieved: the hotspot case 5 in Fig. 6(a). Fig. 12(c) shows the simulation result as this congestion pattern was applied. The result from the conventional system is not included here since it has similar trend as the result in Fig. 11(a).

In this case, the solution of case 5 is offered by shrinking the centre cell to cell radius 600 m and altering the

buffering time for gold and silver customers to 4 s and 6 s, respectively. The reason is because the main congestion area (segment 2A according to Fig. 12(a)) was recognised by the monitoring process while the other congestion area (part of segment 3A) was not significant enough and, therefore, did not cause the offered throughput in band 2 to exceed the threshold set at the CBR model.

The simulation result shows that the proposed method from the closest matched case solves the problem by transferring parts of connections to neighbouring cells and also prioritising the rest of customers according to their classes.

5.3. System performance with unfamiliar cases

This section presents the results for a further case study observing the system performance with unfamiliar cases. In order to test the system when faced with cases very different from those in the case library, a rule-based algorithm was developed consisting of a simple set of rules based on the studies done while generating the case library and also from the solutions of all existing cases. Fig. 13 summarises the rule-based algorithm for the CBR.

As mentioned in Section 4.2, the tolerance ranges are used to identify the threshold of the matching process. Therefore, the congestion pattern is defined as an unfamiliar case if the case retrieval process could not offer the matched case. When there is no matched case the system needs to calculate a solution and to demonstrate this a simple set of rules was developed as part of the CBR model. Note that the aim here is to demonstrate the capability of CBR in this application, not to refine the rules, so that the rules have been made deliberately simple, yet effective enough to provide the system with a suitable policy.

Figs. 14 and 15 show examples of results from unfamiliar hotspot cases. In this test, congestion patterns that are different from the ones in the initial case library were implemented. As the system detected the congestion, CBR model was called. In this case because the match cannot be found, the CBR model used integrated rule-based algorithm mentioned earlier to automatically find the solution.

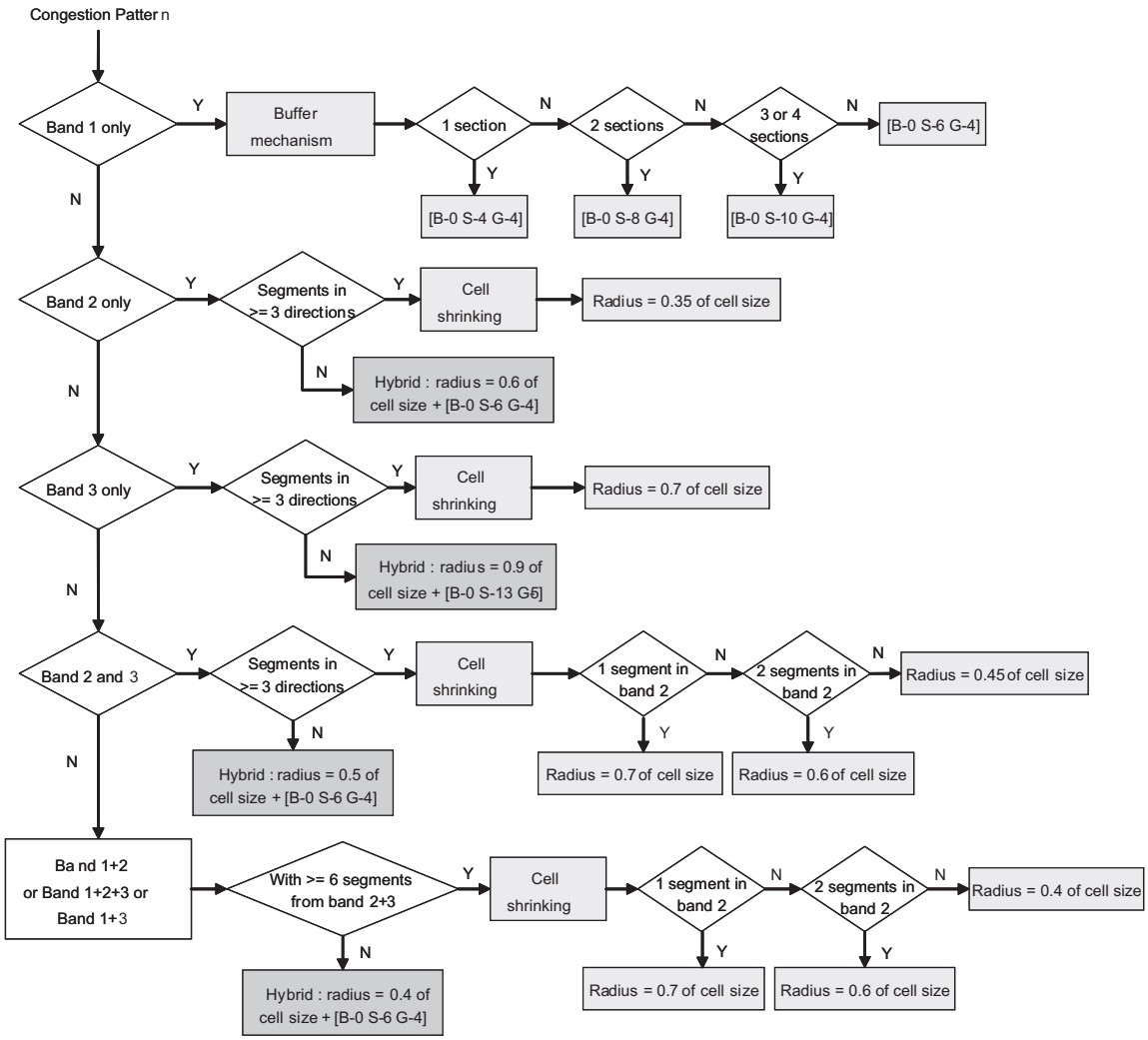


Fig. 13. Summary of the rule-based algorithm.

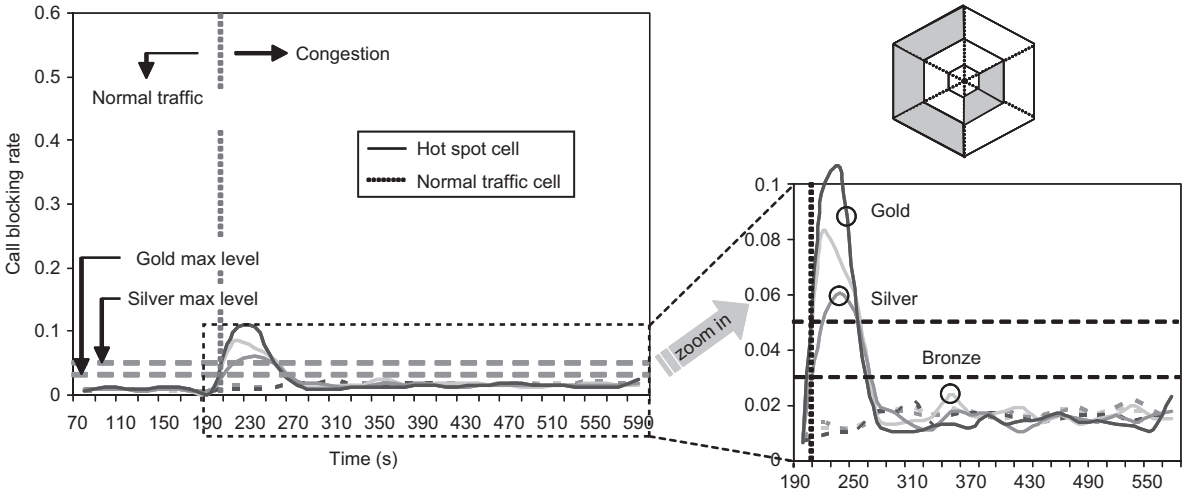


Fig. 14. Unfamiliar case 1—hotspot cell layout and the simulation result.

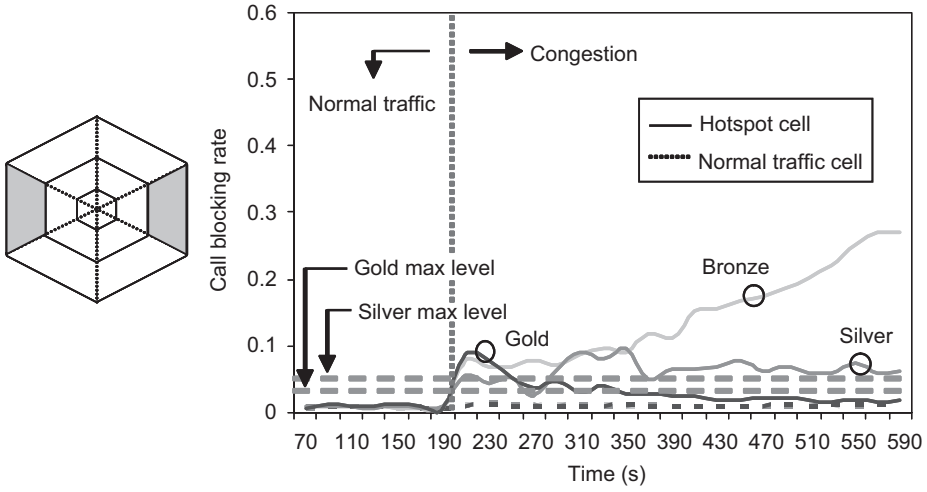


Fig. 15. Unfamiliar case 2—hotspot cell layout and simulation result.

Fig. 14 illustrates the congestion pattern as a hotspot cell layout and the simulation result. In this case, as congestion is detected, the system could not find the match and so results to the rule-based algorithm. The cell shrinking method is proposed with the new hotspot cell radius of 600 m. It can be seen that system QoS is maintained after the implementation of the policy by keeping the call blocking rate for all classes of customers within an acceptable range.

The second example is shown in Fig. 15, including the hotspot cell layout and simulation result. In this case, the hybrid method is proposed by shrinking the hotspot cell to a radius of 900 m and setting the buffering time of 5 and 13 s for gold and silver customers, respectively.

From the success of the proposed solution, these new cases are retained in the case library. As mentioned earlier for the hierarchical memory structure, the library index has to be re-sorted according to the cluster the new cases fall into and their priorities. At this stage, the retaining process is performed offline due to the long period of simulation. However, the learning process is performed leading to the growth of the case library as the system experiences new congestion scenarios.

6. Conclusion

This work has shown that there is value in using CBR to manage the reactive layer policies in an agent-based RRM system. Because the target of the management system can be expressed in any form the system is extremely flexible. The results here have been applied to both instantaneous call blocking and accumulative call blocking, but any measurable key performance indicator could be used.

Moreover, any form of control strategy could be used—the two here (cell shrinking and connection request buffering) were two examples that demonstrated that different schemes could be applied in different types of congestion: because CBR uses matching of previously solved cases from a library, it automatically selects the best approach (or combination of approaches) for a particular situation.

The results also show that unfamiliar cases can also be dealt with and the system as implemented in the simulator can make use of the previous cases to generate a strategy, even though the situation being dealt with does not appear in the case library.

References

3GPP TS 25, 2001. UTRAN Overall Description, 4013GPP Technical Specification, version 5.1.0.

Aamodt, A., Plaza, E., 1994. Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Communications: The European Journal of Artificial Intelligence 7 (1), 39–59.

Ahonen, S., Eskelinen, P., 2003. Mobile terminal location for UMTS. Aerospace and Electronic Systems Magazine, IEEE 18 (2), 23–27.

Angelou, E.S., Koutsokeras, N.Th., Kanatas, A.G., Constantinou, Ph., 2002. SIR-based uplink terrestrial call admission control scheme with handoff for mixed traffic W-CDMA networks. In: Proceedings of the Fourth International Workshop on Mobile and Wireless Communications Network, pp. 83–87.

Baker, M.P.J., Moulsley, T.J., 2000. Power Control in UMTS Release'99. 3G Mobile Telecommunication Technologies. In: Proceedings of the IEE Conference Publication No. 471.

Bodanese, E.L., 2000. A distributed channel allocation scheme for cellular networks using intelligent software agents. Queen Mary, University of London.

Bodanese, E.L., Cuthbert, L.G., 2000a. Application of intelligent agents in channel allocation strategies for mobile networks. In: Proceedings of the ICC 2000, New Orleans, LA, pp. 181–185.

Bodanese, E.L., Cuthbert, L.G., 2000b. A multi-agent channel allocation scheme for cellular mobile networks. In: Proceedings of the ICMAS, Boston, MA, pp. 63–70.

Capone, A., Redana, S., 2001. Call admission control techniques for UMTS. In: Proceedings of the IEEE Vehicular Technology Conference, pp. 925–929.

Caulier, P., Houriez, B., 1995. A case-based reasoning approach in network traffic control. In: Proceedings of the IEEE International Conference on Systems Man and Cybernetics, vol. 2, pp. 1430–1435.

Chantaraskul, S., 2005. An intelligent-agent approach for managing congestion in W-CDMA networks. Queen Mary, University of London.

Chantaraskul, S., Cuthbert, L.G., 2004. Introducing case-based reasoning in SLA control for congestion management in 3G networks. In: Proceedings of the IEEE International Conference on Wireless Communications and Networking, vol. 4, pp. 2498–2503.

Chantaraskul, S., Cuthbert, L.G., 2005. Congestion pattern matching in case-based reasoning control for 3G networks. In: Proceedings of the IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, vol. 2, pp. 134–141.

Cuthbert, L.G., Ryan, D., Tokarchuk, L., Bigham, J., Bodanese, E., 2001. Using intelligent agents to manage resource in 3G networks. Journal of IBTE 2 (Part 4).

Du, L., Bigham, J., Cuthbert, L.G., 2003. Towards intelligent geographic load balancing for mobile cellular networks. IEEE Transactions on Systems Man and Cybernetics, Part C 33 (4), 480–491.

Hassanein, H., Al-Monayyes, A., Al-Zubi, M., 2001. Improving call admission control in ATM networks using case-based reasoning. In: Proceedings of the IEEE International Conference on Performance, Computing, and Communications, pp. 120–127.

Holma, H., Toskala, A., 2002. WCDMA for UMTS: Radio Access for Third Generation Mobile Communication. Wiley, New York.

Kaaranen, H., Ahtiainen, A., Laitinen, L., Naghian, S., Niemi, V., 2001. UMTS Networks Architecture. Mobility and Services. Wiley, New York.

Kolodner, J., 1993. Case-Based Reasoning. Morgan Kaufmann, Los Altos, CA.

Kuri, J., Mermelstein, P., 1999. Call admission on the uplink of a CDMA system based on total received power. In: Proceedings of the IEEE International Conference on Communications, vol. 3, pp. 1431–1436.

Laiho, J., Wacker, A., Novosad, T., 2002. Radio Network Planning and Optimisation for UMTS. Wiley, New York.

Lewis, L., Ray, P., 1999. Service level agreement definition architecture and research challenges. In: Proceedings of the Global Telecommunications Conference, vol. 3, pp. 1974–1978.

Liu, Z., Zarki, M.E., 1994. SIR based call admission control for DS-CDMA cellular system. IEEE Journal on Selected Areas in Communications 12 (4), 638–644.

Marilly, E., Martinot, O., Papini, H., Goderis, D., 2002. Service level agreements: a main challenge for next generation networks. In: Proceedings of the European Conference on Universal Multiservice Network. Avon Books, New York, pp. 297–304.

Muckenheim, J., Bernhard, U., 2001. A framework for load control in 3rd generation CDMA networks. In: Proceedings of the IEEE Global Telecommunications Conference, vol. 6, pp. 3738–3742.

SHUFFLE, 2002. IST Project SHUFFLE (IST-1999-11014). ⟨http://www.elec.qmul.ac.uk/research/projects/shuffle.html⟩.

Thong, W.S., Bigham, J., 2002. Hierarchical management of CDMA network resources. In: Proceedings of the Third International Conference on 3G Mobile Communication Technologies (Publication No. 489), pp. 216–220.