Journal of King Saud University – Computer and Information Sciences (2017) **29**, 103–115

King Saud University

Journal of King Saud University –
Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Journal of
King Saud University –
Computer and
Information Sciences

Users' classification-based call admission control with adaptive resource reservation for LTE-A networks

Salman Ali AlQahtani *

Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

Received 27 July 2015; revised 7 December 2015; accepted 31 December 2015
Available online 20 April 2016

KEYWORDS

Adaptive resource allocation;
Admission control;
Delay aware;
LTE-Advanced user categorization

Abstract In this paper, we introduce the user's privileges and traffic maximum delay tolerance as additional dimensions in the call admission control processes to efficiently control the utilization of LTE-A network resources. Based on this idea, we propose an efficient call admission control scheme named "delay aware and user categorizing-based CAC with adaptive resource reservation (DA–UC-ARR)", where the user priority is adjusted dynamically based on the current network conditions and the users' categorizations and traffic delay tolerances, to increase the network's resource utilization and at the same time to maximize the operators' revenue. In this proposed scheme, the users are classified into Golden users and Silver users, and the type of service per user is classified as real time (RT) and non-real time (NRT) services. We compare the performance of the proposed scheme with the corresponding results of previous schemes, referred to as the adaptive resource reservation-based call admission control (ARR-CAC) (Andrews et al., 2010; AlQahtani, 2014), where user categorization and delay were not taken into consideration in the call admission control process. Simulation results indicate the superiority of the proposed scheme because it is able to achieve a better balance between system utilization, users' privileges provided by network operators and QoS provisioning compared to the ARR-CAC scheme.

---

* Tel.: +966 505225172.
E-mail address: salmanq@ksu.edu.sa.

Peer review under responsibility of King Saud University.

ELSEVIER **Production and hosting by Elsevier**

http://dx.doi.org/10.1016/j.jksuci.2015.12.002

1. Introduction

The Third Generation Partnership Project (3GPP) has been working on various aspects to improve the performance of the Long Term Evolution (LTE) and LTE-Advanced (LTE-A) as new radio access technology and network architecture at a reduced cost. These aspects include Orthogonal Frequency Division Multiple Access (OFDMA), higher order multiple inputs multiple output (MIMO), carrier aggregation, and heterogeneous networks (relays, picos and femtos) (3GPP TR 36.814, 2010; 3GPP, 2010; Chadchan and Akki, 2010). LTE-A has been put in place to support diverse IP-based traffic, such as voice, data and multimedia. In addition to these different supported traffic services and from the network operator's point of view, users' categorization (differentiation) is becoming an important issue because quality of service (QoS) and users' privileges issues vary among users and their traffic classes.

The main objective of the network operators is to maximize the revenue earned by having maximum network utilization, at the same time satisfying the communication needs of their users by developing various programs and benefits. These programs are designed to offer high-class services and privileges to distinguished users. Mostly, the users are classified into a number of categories based on their annual billings. Currently, some network operators categorize their users into a number of different types based on the frequency of their use of the network resources or based on the amount of their annual billing. For, example, an operator can categorize their users as Golden and Silver users, where the Golden category has higher annual billing than the Silver one. Therefore, the network operators give the Golden users higher priority than the Silver users.

Contemporary and future wireless networks are required to serve not just different user categories but also different traffic types, which are classified by standardization bodies. Some of them require guaranteed bit rate (GBR), and some applications do not require guaranteed bit rate (Chadchan and Akki, 2010). The QoS parameters of the different traffic per user category can be significantly different as well. Normally, the real-time services are categorized as GBR applications, and non-real-time services are categorized as non-guaranteed bit rate (NGBR) applications. A GBR bearer confirms the value of QoS parameters associated with it, and the corresponding service assumes that congestion-related packet losses do not occur. A non-GBR bearer does not confirm bearer QoS values, and the corresponding service should be prepared for congestion-related packet losses. Based on users' categories and traffic services types for each user, we can have four types of bearers in LTE: Golden users with real time services (G-RT), Silver users with real time services (S-RT), Golden users with non-real time services (G-NRT) and Silver users with non-real time services (S-NRT) bearers.

Call admission control is located at layer 3 (network layer) (L3) in the evolved Node B (eNB) and is used both for setup of a new bearer and for handover candidates (Chadchan and Akki, 2010). In the LTE-A networks, it is still a challenge for an evolved Node B (eNB) to perform resource block (RB) allocation and scheduling. The task of CAC is to admit or reject the establishment requests for new radio bearers. To do this, it is very important to design an efficient CAC that can take into account the overall resource situation, the user's priority based on its category level and the provided QoS of the in-progress request and the QoS requirements of the new radio bearer request. An important aspect for achieving these objectives is to design an effective CAC scheme with users' varying categories and QoS requirements. This paper aims to design an efficient CAC that takes this issue into account.

The rest of this paper is organized as follows. In Section 2, the literature review is presented. The system model and assumptions are explained in Section 3. In Section 4, details of the proposed scheme's model are presented. The simulation results are presented in Section 5. The paper concludes with Section 6.

2. Literature review and contributions

Call admission controls for the satisfaction of utilization and QoS requirements in integrated networks and LTE-A cellular networks have been an active area of research (AlQahtani and Mahmoud, 2006, 2007; Andrews et al., 2010; AlQahtani, 2014; Kaur and Selvamuthu, 2014; Qian et al., 2009; Lie et al., 2008; Kosta et al., 2010; Niu et al., 2013; Hegazy and Nasr, 2015; Yang et al., 2015). Typical CAC schemes for multi-service are complete sharing (CS), complete partitioning (CP) and threshold (AlQahtani and Mahmoud, 2006, 2007; Andrews et al., 2010; AlQahtani, 2014). These papers briefly discussed the CS, CP and hybrid resource sharing approaches and their shortcomings. To overcome these shortcomings, a number of schemes were proposed for multiple integrated networks by controlling the shared resources in an efficient and adaptive way. In AlQahtani and Mahmoud (2006, 2007), a new CAC admission control with dynamic resource allocation for two types of services in a 3G wireless network was proposed and analyzed. This CAC combines the advantages of CS and CP between real time and non-real time services in an efficient and adaptive way using dynamic priority. The scheme presented in AlQahtani and Mahmoud (2006, 2007) was extended to multiple integrated networks. In Andrews et al. (2010), an adaptive resource call admission control for integrated WiMAX/WiFi networks is presented. The author developed a combined CS and virtual partitioning (VP) resource allocation to control the connection requests in an adaptive way. Two types of traffic were assumed: WiMAX traffic and WiFi traffic. In AlQahtani (2014), a similar scheme was used to adaptively control the connection request in LTE-A with Type I Relay Nodes. In this case, the types of traffic are LTE direct traffic and LTE indirect aggregated traffic (through a relay node). However, in AlQahtani and Mahmoud (2006, 2007), Andrews et al. (2010), AlQahtani (2014), the users' categorizations and their maximum delay tolerance were not addressed. Mostly, the resources were virtually apportioned between the two types of requests and the waiting requests were serviced using a dynamic priority that was calculated based on the predefined VP of each request and the current network usage. In addition, the QoS of service requirements such as delay was not included in calculating the dynamic priority.

Some of the CAC schemes have been modified and implemented in the networks that have integrated services. A number of these implemented schemes are discussed in Kaur and

Selvamuthu (2014) and Kosta et al. (2010), but their discussions are not specific to user categorizations. A CAC and resource scheduling schemes for LTE networks with heterogeneous services are proposed in Kaur and Selvamuthu (2014). The proposed allocation algorithm gives high priority to the RT service packets approaching the delay deadline by using a transmission guard interval. The proposed CAC can adjust the threshold according to the network conditions in an adaptive way. The results show that this proposed CAC can not only balance the in-service connections of different classes of traffic but can also easily reserve the RBs and potentially support user handover. In Qian et al. (2009), the author proposed a two-stage CAC scheme that uses the received signal quality as an indicator to estimate the number of RBs of each request to improve the performance of the network overall. The first stage of the proposed CAC uses a SINR with a dynamic threshold for admitting new arriving requests. At the second stage, the utilization of resources issue plays a major role in maintaining a minimum QoS level for each new request. Simulation results show that the proposed CAC scheme contributes to improving the performance of non-GBR services and maintaining quite well the QoS of delay-sensitive services and finally achieves a reduction in the Packet Drop Ratio for the network.

In Lie et al. (2008), the author proposed delay-aware CAC (DACAC) schemes to guarantee the required QoS for packet delay in the LTE network. The DACAC scheme accepts or rejects the requests based on packet delay measurements for each service class and the current RB utilization of the LTE-A network. The DACAC scheme uses two thresholds for RB utilization. These two thresholds' values are determined as measured RB utilization and packet delay. Those measurements are achieved by adopting the moving-window average method based on the newly defined window structure. In Kosta et al. (2010), a combined admission control (AC) scheme for controlling multiclass QoS and Grade of Service (GoS) in the uplink LTE network was proposed, where the combined AC scheme assigns RBs in a fair way such that the delay and throughput are dynamically adjusted according to the traffic load. Numerical results show that it is possible to satisfy the QoS constraints of all accepted requests using a priority scheme without any network capacity loss. In Niu et al. (2013), Hegazy and Nasr (2015) and Yang et al. (2015), the users are categorized from instantaneous users' conditions and traffic classes. The user classifications from operators' perspective were not considered. In Niu et al. (2013) the users are categorized based on mobility speed. In Hegazy and Nasr (2015) and Yang et al. (2015), the users are categorized based on users' speed and types of traffic.

In all of the above, several works (AlQahtani and Mahmoud, 2006, 2007; Andrews et al., 2010; AlQahtani, 2014; Kaur and Selvamuthu, 2014; Qian et al., 2009; Lie et al., 2008; Kosta et al., 2010) mainly considered the integrated networks (Andrews et al., 2010; AlQahtani, 2014) or the 3G and LTE-A network with multi-services (AlQahtani and Mahmoud, 2006, 2007; Andrews et al., 2010; Kaur and Selvamuthu, 2014; Qian et al., 2009; Lie et al., 2008; Kosta et al., 2010) but rarely considered allocation of resources to different user categorizations from operators perspective in addition to traffic classifications. In Niu et al. (2013), Hegazy and Nasr (2015) and Yang et al. (2015), the classifications of users are based on users' behavior only. Therefore, most of the previous works focus on maximizing the system throughput or achieving fairness among traffic services. However, allocating RBs without considering both the users' priority level (user category) (G or S) and service QoS would make the demand and supplies not balance. This paper aims to design an efficient CAC that takes this issue into account. This paper introduces a CAC strategy for the LTE-A supporting multimedia services with different classes of traffic and diverse user categories. To address this, we propose an efficient resource allocation scheme named "user categorizations and delay aware adaptive resource reservation-based call admission control (DA–UC-ARR)," which aims at allocating RBs adaptively based on the user's priority granted by the network operator based on its category and the QoS required by the supported integrated services. Unlike most previous studies, this work takes into account both the varying users' categories and the QoS needs of the users' traffic. In addition to traffic classifications and dynamic resource allocation proposed in Andrews et al. (2010), AlQahtani (2014), the proposed scheme introduces a new dimension of categorizations, which is Golden (G) and Silver (S) users, and adds one new QoS metric for calculating the dynamic priority.

3. System model and assumptions

A single cell uplink LTE-A network is considered. In this cell of interest, there is a set of users' equipment devices (UEs) and one LTE-A base station (eNB). The users are uniformly distributed in the cell, and each user can be categorized into different types based on their QoS requirements and user category. In this network, we have a new dimension of categorization, which is golden (G) and silver (S) user classes, denoted in general by j user type (where G type is $j = 1$ and S type is $j = 2$). Every user's class can be classified to serve different traffic classes, which are classified by standardization to real-time (RT) services and non-real-time (NRT) services, denoted in general by i service class (where RT class $i = 1$, and NRT class $i = 2$). Therefore, we will have four different request types (request ij, where $i = 1, 2$ and $j = 1, 2$). These types are Golden user class of real time service class (G-RT), Silver user class of real time service class (S-RT), Golden user class of non-real time service class (G-NRT), and Silver user class of non-real time service class (S-NRT) requests. We assume that the requests of G-RT users are of highest priority; next is S-RT. S-NRT user requests are given lowest priority. Users with higher priority factors are subject to increased subscription rates in return for prioritized network access and QoS. Two ratios are also introduced, which are Beta (β) and Alpha (α) and are used to measure the priority of each connection type. β and $(1 - \beta)$ ratios are used in the calculation of the instantaneous priority level of Golden and Silver users, respectively, while α and $(1 - \alpha)$ ratios are used in the calculation of the instantaneous priority level of RT and NRT traffic classes within each user's category, respectively. A higher ratio will lead to higher priority.

The general architecture of a single cell LTE network with different classes of UE is presented in Fig. 1.

All connection requests are assumed to follow a Poisson arrival process with mean rates of $\lambda_{i,j}$ ($\lambda_{1,1}, \lambda_{1,2}, \lambda_{2,1}$ and $\lambda_{2,2}$, respectively). Each connection request ij, where we have four types of aggregated connection requests, has its own queue

(Q_{ij}), and the queuing time limit of each connection request type is a constant value $D_{\max}$ defined based on the requested QoS requirements. The total arrival rate into the system is $\lambda = \sum_{\forall i,j} \lambda_{i,j}$. Each connection request ij has a negative exponential service time distribution (connection time) with mean rates of $1/\mu_{i,j}$. Using the definitions above, the generated load of each request type i,j, denoted by $\rho_{i,j}$, is given by $\rho_{i,j} = \mu_{i,j}^{-1} \lambda_{i,j}$. The total offered load intensity for the system, denoted by ρ, is given by $\rho = \sum_{\forall i,j} \mu_{i,j}^{-1}.\lambda_{i,j}$.

The frame structure of the LTE-A network is organized into 10-ms duration radio frames (Chadchan and Akki, 2010) as shown in Fig. 2. Each frame of 10-ms duration is divided into ten 1-ms subframes. Two types of frame structures are supported by LTE-A. A Type 1 frame is for Frequency Division Duplex (FDD) transmissions, and Type 2 is for Time Division Duplex (TDD) transmissions. For FDD, in each radio frame, 10 subframes are available for uplink and 10 subframes are available for downlink transmission, and these for downlink/uplink transmissions directions are separated in the frequency domain. The LTE Transmission Time Interval (TTI) is defined as one subframe duration.

In LTE, transmission resources are called Resource Blocks (RB) and are assigned to physical channels in time–frequency units. The smallest time–frequency unit is called a resource unit or resource element (RE) and is used for downlink/uplink transmissions. An RE is defined as one subcarrier over one symbol. An RB is formed by a group of 12 subcarriers contiguous in frequency over one slot (0.5 ms) in the time domain, for both TDD and FDD systems. Each subcarrier has a spacing of 15 kHz. The total bandwidth that one RB occupies is 180 kHz for 12 subcarriers. A physical channel covers a frequency band containing one or more contiguous RBs. The physical channel bandwidth is a multiple of 180 kHz. All of the RBs in the available bandwidth constitute a resource grid. Two RBs (180 kHz × 1 ms) is the minimal allocated resource unit.

We assume that each radio frame consists of a set N of RBs for data and control transmission. The value of N depends on the LTE bandwidth, B, selected. Some of these N RBs are used for control transmission and signaling overhead and denoted by N_{co}. The RBs used by Random Access Channel (RACH) and Physical Uplink Control Channel (PUCCH) for controlling the transmission are time-variable, and we have assumed a constant number of RBs for data transmission and control transmission and signaling overhead in this research. Therefore, the maximum LTE-cell capacity (in terms of RBs) that is available and can be used for data transmission is the total number of RBs minus the number of RBs used for control transmission and signaling overhead and is denoted as $N - N_{\mathrm{co}}$.

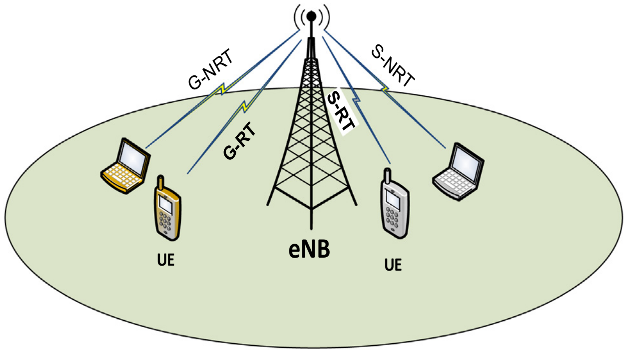

Figure 1 Architecture of a single cell LTE network with different classes of UE.

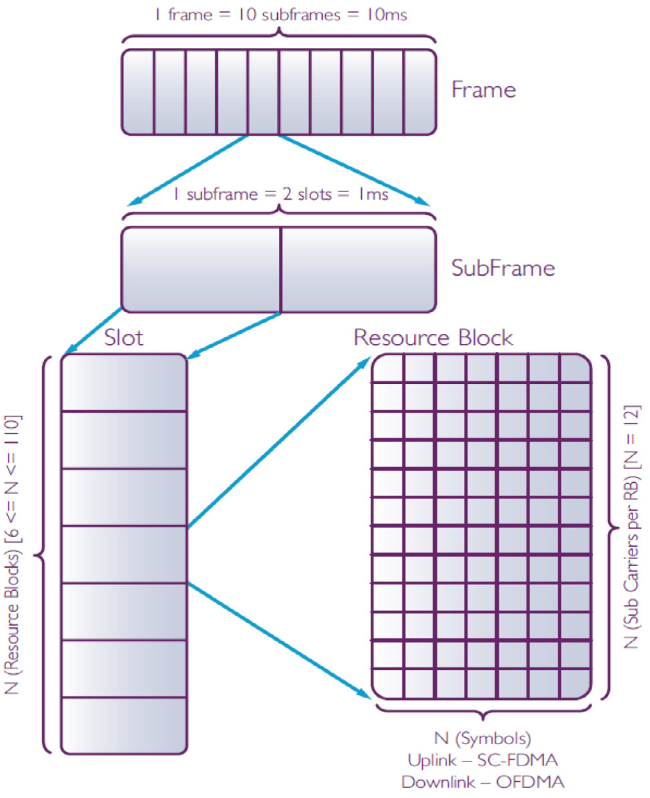

Figure 2 Architecture of a single cell LTE network with different classes of UE.

In this network model, the N RBs are virtually partitioned into four partitions: N_{11}, N_{12}, N_{21} and N_{22}. N_{11} RBs are for G-RT, N_{12} RBs are for G-NRT, N_{21} RBs are for S-RT, and N_{22} RBs are for S-NRT, where $N_{11} + N_{12} + N_{21} + N_{22}$ is equal to $N - N_{co}$. Each connection's virtual resource partition is predefined dynamically based on its β and α ratios.

4. Call admission control strategies

CAC is used in general to control the number of users in the LTE-A network and must be designed to guarantee the QoS requirements for both traffic services and user types. In the LTE-A network, CAC denotes the process of making a decision on a new connection request based on the number of available RBs versus users' types and QoS requirements for different traffic services. Before the CAC makes the decision regarding the arriving connection requests, it first estimates the total RBs currently in use, N_c, (i.e., current cell loads) and the number of RBs required by the new arrival connection request ij (new load increment, ΔN_{ij}), and then applies them in the decision-making process of accepting or rejecting new connection requests. The estimation process of the new load increment required by the new arrival request ij and the current cell loads, N_c, follows.

The data rate requirement for each request type can be defined in terms of its required RBs. Assuming that the average data rate requirement for each connection request *ij* is R_{ij}, then the required bandwidth (B_{ij}) to guarantee the required data rates can be expressed as (AlQahtani, 2014; TR 36.942 Annex A, 2010; TR 36211, 2010)

$$B_{ij} = \frac{R_{ij}}{\log(1 + \mathrm{SNR})}$$

Considering the resources on the uplink and assuming that there are four different connection request types, the resource increment, ΔN_{ij} , required by newly arrived connection request *ij* in terms of the required RBs can be calculated as

$$\Delta N_{ij} = S \cdot \left\lceil \frac{B_{ij}}{W_{RB}} \right\rceil \tag{1}$$

where W_{RB} is the bandwidth of one resource block, S is the number of sub-frames (in LTE-A, $S = 10$), and $\lceil x \rceil$ denotes the nearest integer greater than or equal to x. At any point in time, we will have k_{ij} current active connections from each request type *ij*. Based on this information and using (1), the current instantaneous total number of RBs in use (i.e., the total current cell load), N_c can be calculated as follows using the following criterion:

$$N_c = \sum_{\forall ij} k_{ij} \cdot \Delta N_{ij} \leqslant N - N_{co} \tag{2}$$

In the next subsections, we present one of the current existing related CAC schemes, known as adaptive resource allocation CAC (ARR-CAC) (AlQahtani, 2014), and we adapted it to our system to use it as a reference to compare its performance with the performance of the proposed schemes. Then, the herein proposed user categorizing and delay-aware based CAC with adaptive resource reservation is presented and explained.

4.1. ARR-CAC

In this section, we will adapt the idea of schemes that were proposed in AlQahtani and Mahmoud (2006), Andrews et al. (2010), AlQahtani (2014), known as ARR-CAC, to our system, as shown in Fig. 3. To do so and use the ARR-CAC scheme, all active connections can be grouped into two different types based on traffic service types, one for RT connections (Golden & Silver users) and the other for NRT connections (Golden & Silver users). In ARR-CAC, each request type has its own queue, and all request types share the available RBs. At the same time, each request type has a predefined virtual resource partition defined as N_1 RBs for RT services (Golden & Silver users) and N_2 for NRT services (Golden & Silver users) such that $N_1 + N_2 = N - N_{con}$. The required RBs for a newly arrived request i ($i = 1$ for RT, $i = 2$ for NRT), ΔN_i, can be calculated as follows,

$$\Delta N_i = S \cdot \left\lceil \frac{B_i}{W_{RB}} \right\rceil \tag{3}$$

Therefore, the ARR-CAC scheme decides to admit or reject a new connection request i based on the following criteria:

$$N_c + \Delta N_i \leqslant N - N_{co} \tag{4}$$

The main steps of ARR-CAC can be defined as follows (AlQahtani, 2014). An incoming connection request i is served as long as the criterion (4) is satisfied. When all RBs are unavailable, the arrived request i is buffered in its queue or blocked if its queue is full. Any connection request is dropped from its queue if it exceeds its queuing time limit, $D_{\max\text{-}i}$. Upon an RB release, the adaptive priority (AP) value is calculated for all request types that have non-empty queues, and then the queue with higher priority (i.e., with least AP value) is served first based on the FIFO policy.

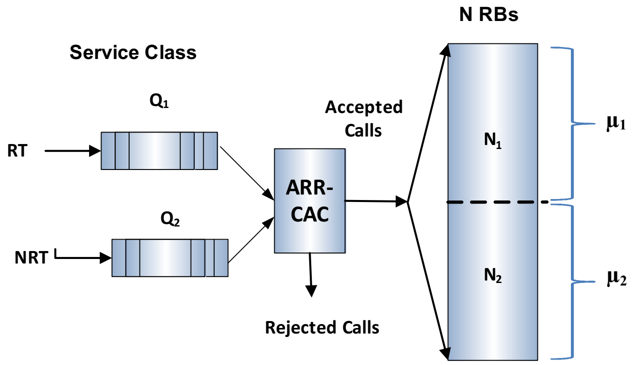

Figure 3 ARR-CAC system model.

The priority is adaptively adjusted to prevent all request types from adversely affecting each other. For low to moderate loads, all RBs (i.e., $N - N_{co}$ RBs) are open for all arrived connection request types to enhance the LTE-resource utilization. However, under congestion conditions the AP is used to dynamically differentiate between the types of queued connection requests. Therefore, the adaptive priority (AP_i) value for a type i request is computed using the RBs currently in use by type i requests and the virtual resource partition predefined for type i requests as follows:

$$AP_i = \frac{N_c^i}{N_i} \tag{5}$$

where N_c^i is the number of current RBs in use by active connection requests of type i (k_i) such that $N_c^i = \sum_{\forall i} \Delta N_i k_i \leqslant N$, and the next highest priority connection request type is determined as

$$\begin{cases} RT \text{ If } AP_1 \leqslant AP_2 \\ NRT \text{ If } AP_2 < AP_1 \end{cases} \tag{6}$$

Thus, the request type i with minimum AP_i will be given the highest priority and can be served first. Here, as the total RBs currently occupied by all users of ith request type, N_c^i, drops below the number of virtual reserved RBs for the ith request type, N_i, its AP_i value decreases and hence its priority increases. When two or more request types have the same AP value, then the request type with higher priority will be served first. Using these rules, the unutilized RBs of one user type can be utilized by other user types when needed. Additionally, at high system load, the AP value prevents all users' requests from adversely affecting each other.

4.2. Proposed CAC schemes

The network operators feel that the importance of the user categorizations issues is becoming increasingly important. An important issue in providing differentiated users and services in LTE-A is to design an adaptive and efficient CAC scheme. The limitations with the previous ARR-CAC (AlQahtani and Mahmoud, 2006; Andrews et al., 2010; AlQahtani, 2014) scheme are as follows:

- It differentiates only between two types of connection requests, i.e., RT and NRT (AlQahtani and Mahmoud, 2006) or WiMAX and WiFi (Andrews et al., 2010), or LTE direct connections and indirect connections (through relay) (AlQahtani, 2014), regardless of the user category and its priority level with respect to the operators.
- The adaptive priority value and pre-defined virtual partition are defined in ARR-CAC (Andrews et al., 2010; AlQahtani, 2014) independent of user category and the traffic's QoS in terms of maximum delay limitations.

The proposed delay aware user categorizing-based CAC with adaptive resource reservation takes the previous scheme's limitation into account. In this proposed scheme, a new dimension of classifications, which are Golden and Silver user categories, is defined, and every user category can be classified to have different traffic types, which are classified by standardization to RT services and NRT services. For network operators with prioritized access and QoS, it is found that higher priority users are subject to increased subscription rates in return.

In this proposed scheme, α and β are very important ratios that are defined and estimated by the service provider (operator) according to the target revenue and the importance of the user category and its traffic class, respectively, where $0 < \alpha < 1$ and $0 < \beta < 1$. The α ratio and $(1 - \alpha)$ ratio are used to define the Golden and Silver users' priority levels. The user with the higher ratio has the higher priority. In our case, the Golden user has the higher priority. The β ratio and $(1 - \beta)$ ratio are used to define the RT and NRT priority level for each user category. The user's traffic class with the higher ratio has the higher priority. In our case, RT traffic has the higher priority. Therefore, α is used to define the priority level (importance) of users (Golden has higher value) and β is used to define the priority level (importance) of traffic services for each user category (RT has higher β value).

Using these predefined ratios, LTE RBs can be virtually partitioned among users and traffic classes into four predefined partitions. The number of each predefined virtual RBs for *i* service class of *j* user type (connection request *ij*) can be calculated as follows:

- Number of virtual predefined RBs for *RT* service class of *G* user type

$$\boldsymbol{N}_{1.1} = \alpha * \beta * N \tag{7}$$

- Number of virtual predefined RBs for *RT* service class of *S* user type

$$\boldsymbol{N}_{1.2} = \alpha * (1 - \beta) * N \tag{8}$$

- Number of virtual predefined RBs for *NRT* service class of *G* user type

$$\boldsymbol{N}_{2.1} = (1 - \alpha) * \beta * N \tag{9}$$

- Number of virtual predefined RBs for *NRT* service class of *S* user type

$$\boldsymbol{N}_{2.2} = (1 - \alpha) * (1 - \beta) * N \tag{10}$$

The predefined partition is then defined based on the priority level of each request, which is reflected by the two ratios. In addition, each request *ij* has its own queue $\boldsymbol{Q}_{i,j}$, and all types share the available RBs.

Our proposed scheme can be divided into two sub-schemes. The first one is called "user categorizing-based CAC with adaptive resource reservation (UC-ARR)". This scheme considers the user categorization and traffic services. The second sub-scheme is named "delay aware and user categorizing-based CAC with adaptive resource reservation (DA–UC-ARR)". This sub-scheme takes the maximum delay limitations of each request type into account when calculating the adaptive priority value of each connection request type. These two sub-scheme steps and model are explained in the following sub-sections.

4.2.1. UC-ARR

The system model of this scheme is as shown in Fig. 4.

The UC-ARR scheme procedures can be stated as follows:

1. The arrived *i* service class of *j* users type request is served as long as the criterion in (11) is satisfied.

$$N_c + \Delta N_{i,j} \leqslant N - N_{con} \tag{11}$$

   where $\Delta N_{i,j}$ is the total increment of RBs required by the arrived user of the *i*th service class and the *j*th user type.
2. When all RBs are in use, the arrived connection request *i, j* is inserted in its queue $Q_{i,j}$ if it is not full.
3. A request *ij* is dropped from its queue if it exceeds the queuing time limit.
4. Upon an RB's release, the adaptive priority, $AP_{i,j}$, value is calculated for all connections requests *ij* with non-empty queues. Then, the non-empty $Q_{i,j}$ with the lowest $AP_{i,j}$ value is served first.

The adaptive priority *AP*, used in step 4 of the above procedure, is calculated using:

1. The total number of current RBs occupied by all users of the i^{th} service class and the j^{th} user type $N_c^{i,j}$.
2. The number of virtual reserved RBs for *i*th service class *j*th user type $N_{i,j}$.

Therefore, the *AP* value of UC-ARR for request *i,j* is given by:

$$AP_{i,j} = \frac{N_c^{i,j}}{N_{i,j}} \tag{12}$$

This sub-scheme pseudocode is shown in Table 1.

The request *i,j* with the minimum $AP_{i,j}$ value will be given the highest priority. Here, as the total number of currently occupied RBs $N_c^{i,j}$ drops below the number of virtual reserved RBs $N_{i,j}$, its *AP* value decreases and its priority increases. Using these rules, the unused load of one type can be used by other traffic classes when needed. Additionally, at high network load, the *AP* value prevents the requests from adversely affecting each other. The priority level of each request *ij* depends on the value of its importance ratios α and β. Therefore, the priority is adaptively adjusted taking into account the defined importance level of all users and their service classes, as shown in Table 2.

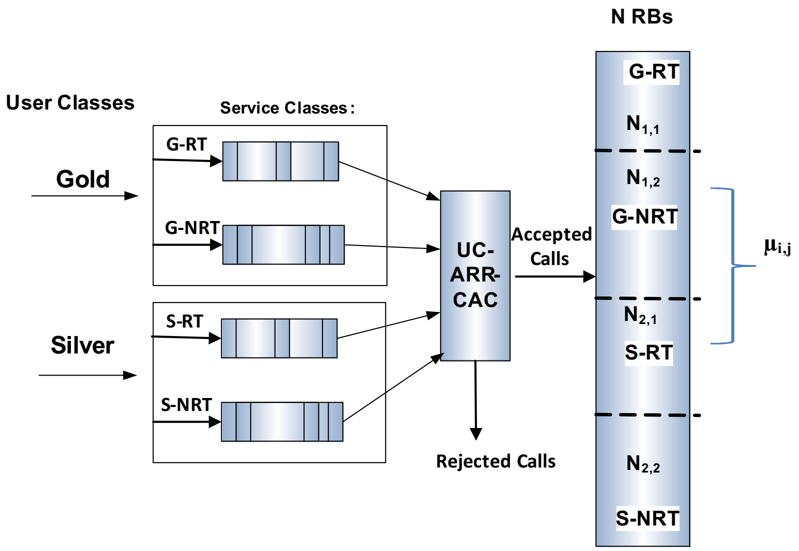

Figure 4 System model for UC-ARR scheme.

4.2.2. DA–UC-ARR

This sub-scheme, known as DA–UC-ARR, takes into account the different categorizations of users as in the previous sub-scheme and adds another dimension in the adaptive priority calculations, which is the maximum delay tolerance. It gives the connection request a high priority value when the difference between the pre-defined tolerable maximum delay $Dmax_i$ and the current latency of the user's request for the ith service class, TA_i (i.e., $Dmax_i - TA_i$), becomes very small. The procedure for this scheme is the same as that of the UC-ARR but the adaptive priority AP is calculated using:

1. The total number of RBs currently occupied by all users of the ith service class and the jth user type, $N_c^{i,j}$.
2. The number of virtual reserved RBs for the ith service class and the jth user type, $N_{i,j}$.
3. Tolerable maximum delay for user with the ith service class, $Dmax_i$.
4. Current latency for user of with the ith service class, TA_i.

Table 1 The pseudocode of UC-ARR.

ARRIVAL event of request i,j

```
IF $N_c + \Delta N_{i,j} \leqslant N - N_{con}$
      Admit request $i,j$
ELSE
   If ($Q_{i,j}$ is not full)
      Insert request $i,j$ call into its queue
   Else Reject request $i,j$
END
```

DEPARTURE event of request i,j

```
IF (There are non-empty queues)
   calculate: $AP_{i,j} = \frac{N_c^{i,j}}{N_{i,j}}$
   Select non-empty $Q_{i,j}$ with the lowest $AP_{i,j}$ value
      IF $N_c + \Delta N_{i,j} \leqslant N - N_{con}$
Admit request $i$,j
      Else Release the RB
Else RB
END
```

Table 2 The Adaptive priority *AP* of each user traffic class of UC-ARR.

Category/services	RT	NRT
Golden	$AP_{1,1} = \frac{N_c^{1,1}}{\alpha*\beta*N}$	$AP_{2,1} = \frac{N_c^{2,1}}{(1-\alpha)*\beta*N}$
Silver	$AP_{1,2} = \frac{N_c^{1,2}}{\alpha*(1-\beta)*N}$	$AP_{2,2} = \frac{N_c^{2,2}}{(1-\alpha)*(1-\beta)*N}$

Therefore, the AP of DA–UC-ARR value for request i,j request is given by:

$$AP_{i,j} = \frac{N_c^{i,j}}{N_{i,j}} \times (Dmax_i - TA_i) \tag{13}$$

In (13), as the user of class i service has higher latency in the queue, the difference between the tolerable maximum delay $Dmax_i$ and the user latency of TA_i becomes very small and this will decrease the value of $AP_{i,j}$. Decreasing the priority value will increase the priority of users with service class i. At high system load the delay factor will prevent the classes from adversely affecting each other. The adaptive priority AP of each user traffic class of DA–UC-ARR is shown in Table 3.

This sub-scheme pseudocode is as shown in Table 1, except that the AP is calculated using the equation shown in (13).

5. Performance evaluation and result analysis

This section discusses the investigations and experiments undertaken to assess the proposed CACs' performance. The results of the experiments are then elucidated with the help of a developed NS-2 simulator. It needs to be emphasized here that the simulation experiments were carried out on the developed simulator so that a suitable result is achieved and the diverse features of the LTE uplink performance can be accurately measured within dissimilar and atypical circumstances. The measures used for performance evaluation include system utilization and QoS metrics such as total blocking probability (TBP) and average delay.

Table 3 The adaptive priority AP of each user traffic class of DA–UC-ARR.

Category/services	RT	NRT
Golden	$AP_{1,1} = \frac{N_c^{1,1}}{\alpha * \beta * N} \times (Dmax_1 - TA_1)$	$AP_{2,1} = \frac{N_c^{2,1}}{(1-\alpha) * \beta * N} \times (Dmax_2 - TA_2)$
Silver	$AP_{1,2} = \frac{N_c^{1,2}}{\alpha * (1-\beta) * N} \times (Dmax_1 - TA_1)$	$AP_{2,2} = \frac{N_c^{2,2}}{(1-\alpha) * (1-\beta) * N} \times (Dmax_2 - TA_2)$

5.1. Simulation setup

The proposed schemes have been modeled by developing an NS-2 simulation. The intended simulation model is composed of a single cell set up with an omni-directional antenna using an SC-FDMA uplink based on the 3GPP LTE-A system model. The total bandwidth considered is $B = 20$ MHz, subdivided into 100 RBs of 12 sub-carriers. The simulation has, as a matter of fact, four FIFO queues that hold the value of the arrival time and drop time of each connection request. The important simulation parameters are shown in Table 4. Requests are generated in the system according to a Poisson arrival process. The arrival rate of all types of users is assumed to be the same, i.e., $\lambda_{i,j}$ are equal for all i and j classes and the service rate of all types of users is μ, where $\mu_{i,j} = \mu$ for all i and j service and user types, respectively. The utilization rate is given by λ/μ.

The traffic model configurations are given in Table 5. We compare the performance of our proposed CAC's algorithm with a reference ARR-CAC (AlQahtani and Mahmoud, 2006).

5.2. Performance measures

It needs to be emphasized here that the focal rationale of using the simulations is to evaluate the performance of the proposed CAC schemes in terms of resource utilization, average delay and total blocking probabilities of each traffic class. The below-mentioned metrics were measured to quantify the performance of the proposed system, to assess the performance of the system through the use of different CAC schemes.

5.2.1. Total blocking probabilities

The total blocking probability (TBP_k) of a type k request is the probability that a type k request does not attempt to enter the service. In fact, this probability is the sum of two probabilities: the blocking probability and the time-out probability. The blocking probability (BP_k), of a type k request, is the probability that a type k request is not able to be queued (i.e., is blocked) when its queue is full. It can be defined as follows:

$$BP_k = \frac{Bl_k}{Ar_K}$$

where Bl_k is the total number of type k requests that blocked and Ar_k is the total number of type k requests that have arrived. The time-out probability (TP_k) is the probability that a queued (i.e., unblocked) type k request is deleted from its queue when the mobile (users) exceeds its maximum waiting time before getting a resource. It can be defined as follows:

$$TP_k = \frac{Dr_k}{Ar_K}$$

where Dr_k is the total number of a type k requests dropped and Ar_k is the total number of type k requests that have arrived. Therefore, the total blocking probability of a type k request is the blocking probability plus the time out probability of unblocked requests of the same type, and can be written as follows:

$$TBP_k = TP_k + BP_k$$

Table 4 Simulation parameters.

Parameter	Assumption
System bandwidth	20 MHz (100 PRBs, 180 kHz per PRB)
TTI	1 ms
Number of RBs for data transmission	96
Number of RBs for control transmission	4
User arrival	Poisson process
Connections duration	Exponential 120 s
Queue size	10
α ratio	65%
β ratio	60%

Table 5 Requirements of QoS.

User class	Service class	Delay budget	Rate budget
Golden/Silver	RT	100 ms	64 Kbps
Golden/Silver	NRT	300 ms	128 Kbps

5.2.2. System utilization (U)

The system utilization can be defined as the ratio of the average number of total current cell occupied RBs (N_c) of interest to the total number of RBs in one radio frame (N). So, the system (cell) utilization can be expressed as:

$$U = \frac{N_c}{N}$$

5.2.3. Average waiting time (W)

The average queuing time delay for the k request type, W_k, is the average time a new arrival user (call) waits before it gets service or it is blocked.

$$W_k = \frac{\Sigma \text{waiting time of request } k}{\text{total } \# \text{ of arrived request } k}$$

5.3. Results and discussion

We are going to compare the performance of the proposed schemes with ARR-CAC (Andrews et al., 2010; AlQahtani,

2014) under the same conditions in terms of system utilization and other QoS metrics, such as TBP and resource utilization, for RT and NRT. In the case of fixing channel holding times for all users, the increase in request arrival rates (i.e., offered traffic loads) corresponds to the increase in traffic intensity. Consequently, most of the performance measures are plotted against the total offered load measured in Erlang (Erl.). To study the full impact and importance of the comparisons made in the following sections, and to arrive at the results thereof, a number of scenarios are studied.

To start with, it is very important to investigate and evaluate the general behavior of the proposed scheme and to compare it with the previous scheme. The second situation to study, compare and explicate is the proposed scheme's behavior in terms of QoS provision and TBP for different system configurations. For the third scenario, a study has to be made with respect to the effect of the importance ratios (α and β) in the case of UC-ARR and DA–UC-ARR. What is important in relation to the study of all three scenarios is that it helps us to study and consider a wide range of offered traffic from a low value of 20 Erlangs to a larger value of 240 Erlangs, rather than to consider a light load. This comprehensive study of the scenarios is undertaken to capture the behavior of the proposed systems in all traffic load conditions (i.e., at low, moderate and high traffic loads), especially in the cases using the queuing techniques.

5.3.1. Performance comparisons of ARR-CAC and UC-ARR

Using the parameters in Tables 4 and 5 unless specified otherwise, the performance comparison of ARR-CAC (AlQahtani and Mahmoud, 2006; Andrews et al., 2010; AlQahtani, 2014) and the proposed UC-ARR is investigated. Using the α and β parameter values shown in Table 4, the Golden users are of the highest priority and Silver users have the lowest priority. The priority of each user category and its virtual resource partition are adaptively adjusted by applying the UC-ARR formula shown in (12). The performance measures of each user and traffic class are shown as a function of total system traffic loads. The comparisons between these two schemes in terms of TBP and total system utilization for RT and NRT services versus traffic loads are shown in Figs. 5–8.

The behavior of Golden and Silver users using the normal ARR-CAC scheme (AlQahtani and Mahmoud, 2006; Andrews et al., 2010; AlQahtani, 2014) is exact because the scheme differentiates only between traffic classes and there is no differentiation between users' categories. When we pay attention to the performance observation of the comparison between the previous scheme and the proposed scheme, as shown in Figs. 5–8, we discover that for low to moderate loads (<100 Erl.), there is no difference between the schemes' results because of resource availability. However, at heavy loads, the users' categorization criteria protect user types from each other by applying the *AP* formula (12), which results in preventing the Golden user high load from starving the other class (i.e., Silver users), and driving their QoS performance down in both traffic classes. In addition to that, the results of UC-ARR are found to be quite optimistic and clearly indicate that high priority user classes have lower call blocking probability when compared to lower priority user categories because of virtual resource reservation of each user's category that were predefined, as shown above in (7–11). The UC-ARR provides the pleasure of high priority users (Golden) in terms of low TBP and high resource utilizations, generating at the same time more revenue to network service providers.

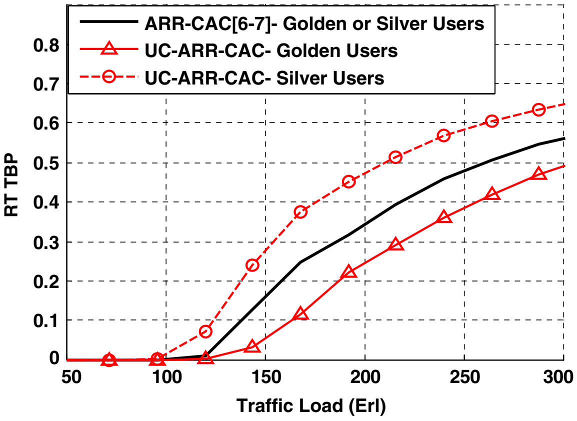

Figure 5 TBP of RT versus total offered traffic loads.

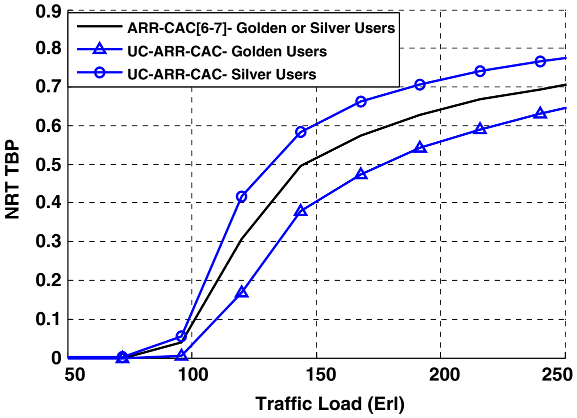

Figure 6 TBP of NRT versus total offered traffic loads.

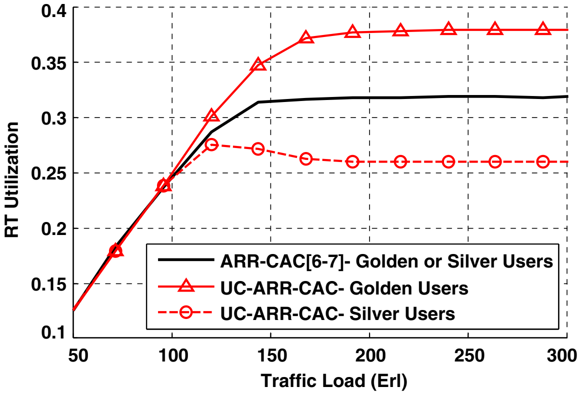

Figure 7 Utilization of RT versus total offered traffic loads.

From an overall system utilization point of view, all three schemes provide the best system utilizations as shown in Fig. 9. This is because of applying the complete sharing concept in an adaptive way by all three schemes. However, the resource distributions between all connection types based on

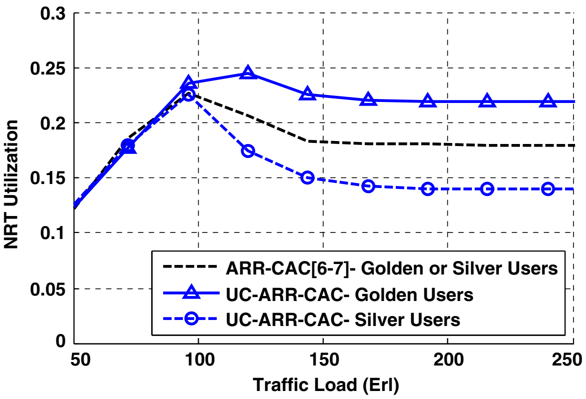

Figure 8 Utilization of NRT versus total offered traffic loads.

their priority and QoS can be provided in an adequate way by using the proposed schemes.

5.3.2. Performance study of the proposed schemes

In this section, we study the performance of the proposed schemes for different configurations. First, we compare the performance of UC-ARR and DA–UC-ARR using the parameters shown in Tables 4 and 5, where we study the effect of applying the delayed awareness in (13). Second, we investigate the proposed schemes' performance at different configurations using different values of α and β.

The performance measures of each user's traffic classes are shown as a function of total system traffic loads. In this context, we also observe and examine the comparisons between DA–UC-ARR and the UC-ARR. The results for TBP, utilizations and average delay of RT and NRT services versus traffic loads are shown in Figs. 10–14. The performance of Golden and Silver users using both proposed schemes are compared. As an observation, for low to moderate loads ($<$100 Erl.), there is no effect by adding the delay aware factor. However, at high system loads, the lower priority users' category and lower traffic class will experience a higher value of TA_i waiting for services due to resource unavailability. Depending on the maximum tolerance delay, $Dmax_i$, of each traffic class, the AP value of each traffic class decreases as the value of $(Dmax_i - TA_i)$ decreases. In general, when any user starts having less AP value, its priority increases and it will be given better treatment, as shown in (13), which will result in better TBP and better average delay.

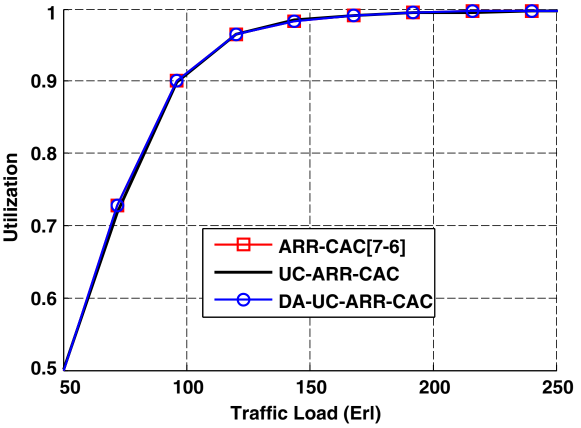

Figure 9 Total system utilization versus total offered traffic loads.

Figs. 10, 11 and 14 show that the TBP and average delay of Golden users' traffic are always better than the TBP and average delay of Silver users' traffic using both schemes. However, the DA–UC-ARR scheme provides fair and better treatment for Silver users' traffic (S-RT and S-NRT) compared to the UC-ARR scheme. Therefore, the DA–UC-ARR prevents the Silver users from being overwhelmed by Golden users and at the same time guarantees the higher performance (lower TBP and average delay) for Golden users. In addition and as expected, when we decrease the TBP of Silver users' traffic, its resource utilization will increase, as shown in Figs. 11 and 12. This implies that adding the maximum delay tolerance factor, $(Dmax_i - TA_i)$, plays an additional role in improving TBP and average delay as well as the satisfaction of the QoS requirements for Silver users and prevents the Silver users' traffic from being overwhelmed by higher priority Golden traffic. As more explanation, as the latency, $TA_{i,}$ of users' traffic approaches its maximum delay, $Dmax_i$, the value of $(Dmax_i - TA_i)$ will become very small and will result in a very low AP value, which will increase its service priority.

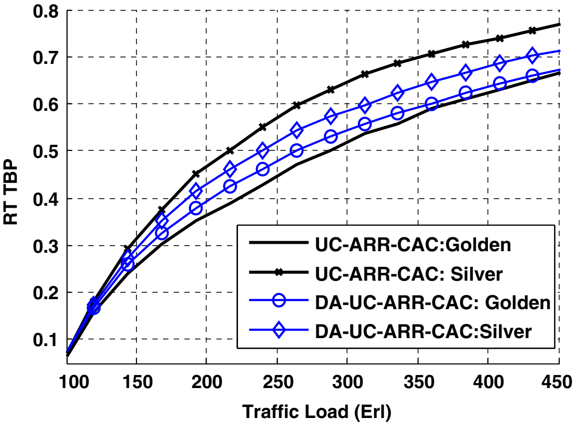

Figure 10 TBP of RT versus total offered traffic loads.

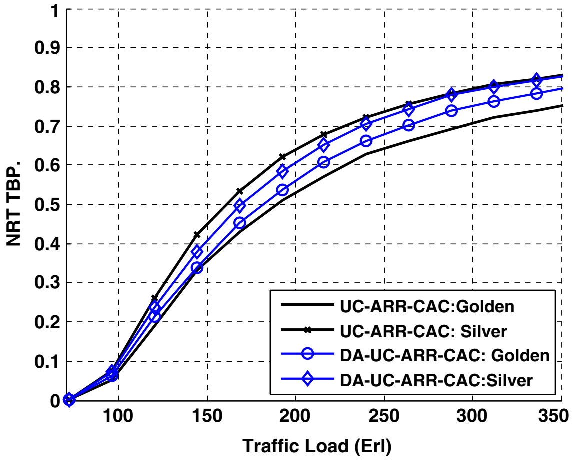

Figure 11 TBP of NRT versus total offered traffic loads.

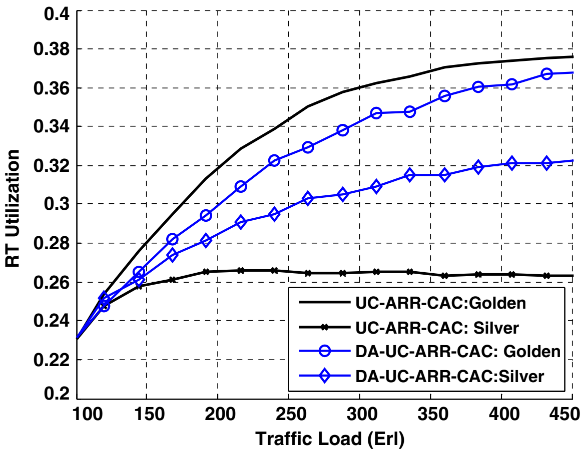

Figure 12 Utilization of RT versus total offered traffic loads.

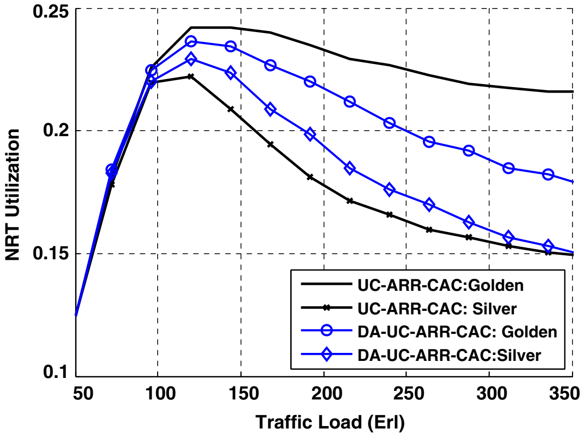

Figure 13 Utilization of NRT versus total offered traffic loads.

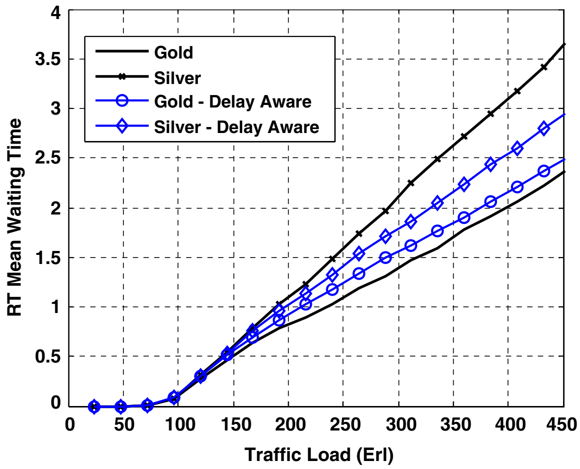

Figure 14 Average Delay of RT of both users' categories versus total offered traffic loads.

The enhancements provided for Silver users by the DA–UC-ARR scheme will be at the expense of lowering the performances of Golden users' services, as shown in Figs. 10, 12 and 14. However, this improvement does not lower the performance measures of Golden users below the Silver users' performance; the performance measures of Golden users' traffic is still better than that of Silver users' traffic using both schemes. Therefore, from Figs. 10–14, we can conclude that the proposed DA–UC-AAR algorithm is adequate to provide good performance for both user categories as well as the satisfaction of the QoS requirements in terms of maximum delay in an adaptive and balanced way. The level of improvement for each user can be adjusted well using the appropriate configurations of each users category traffic in terms of maximum delay, $Dmax_i$, and the right ratio value in terms of α and β.

In fact, the parameter values chosen for both importance factors (α and β) and $Dmax_i$ will play the major role in providing performance results required by each users' services. To investigate this issue more, we compared the impact of changing the importance ratio Alpha (α) in both DA–UC-ARR and UC-ARR. In this investigation the α ratio was adjusted to give one user category better treatment than the other. The parameters in Table 4 and Table 5 are used unless otherwise specified. The TBP for each user's category class is shown when α has been assigned the values of 20%, 50% and 80%. The TBP of any user's category traffic class is measured as the number of blocked or dropped connection requests of that user's traffic class over the total number of connection requests of that user's traffic class. Figs. 15 and 16 show the TBP of the G-RT and S-NRT connection requests, respectively, as a function of the offered load at different alpha (α) ratio. The other users' category traffic classes behave in a similar way.

From these figures we can observe the following main findings. The TBP of all users' category traffic using both schemes increases as the traffic load increases, which is intuitive. However, the status of the TBP of each user's traffic class depends on the effect of the alpha (α) ratio when calculating the AP (12), (13). Based on the AP equations shown in Tables 1 and 2 for both schemes, it is clear that when the alpha (α) ratio increases, the AP value of RT traffic decreases while the AP value of NRT traffic increases. As the AP of any user's traffic class decreases, the service priority of that traffic class increases, and hence its TBP decreases and vice versa. However, the DA–UC-AAR scheme performs better than UC-AAR in the case of RT traffic at the expense of lowering the performance (Higher TBP) of NRT services at low to medium

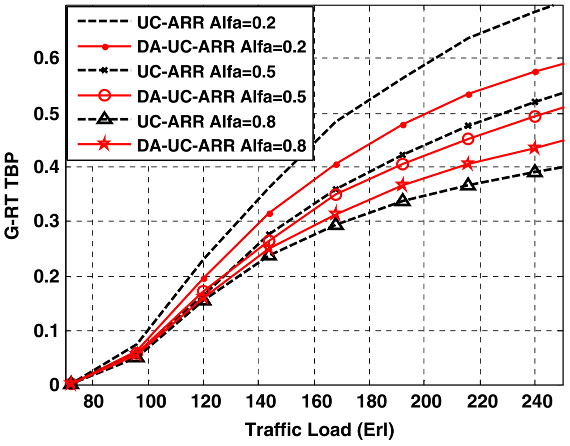

Figure 15 G-RT's TBP versus offered traffic loads at different α ratio.

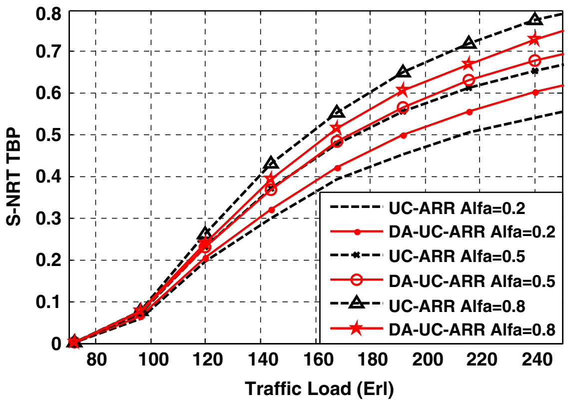

Figure 16 TBP of S-NRT versus total offered traffic loads at different α ratio.

alpha (α) ratio. At high alpha (α) ratios, the DA–UC-AAR scheme performs better than UC-AAR in the case of NRT traffic at expense of lowering the performance (Higher TBP) of RT services, but this lowering does not drop below that of NRT; the RT TBP is still better than that of NRT. Therefore, in all cases, the DA–UC-AAR scheme achieves better TBP for RT services compared to NRT by adding the maximum tolerance delay effects to the AP factor in (12), (13). Because RT has lower maximum delay than NRT, as RT approaches its maximum delay, the AP value decreases, and this will result in better TBP. However, when we have a very high alpha (α) ratio, it becomes the dominant factor in decreasing the AP value and the effect of the delay aware factor is negligible. Figs. 15 and 16 compare the TBP of G-RT and S-NRT connection requests, respectively, as a function of the offered load at different alpha (α) ratios for both schemes. The performance is similar to what we explained because the alpha (α) ratio affects the AP value in a similar way.

Overall, an interesting observation is that the AD–UC-AAR performance gain is better than the UC-AAR scheme for sensitive delay services. Reducing the time-out probability reduces the TBP and makes the network efficient in terms of providing optimized QoS. Therefore, at high system load, the delay factor will prevent the traffic classes of each Golden and Silver user from adversely affecting each other while keeping the best performance for Golden users. Finally, based on these observations, we can safely conclude that DA–UC-ARR can be used to prevent the delay insensitive service class high load from starving the RT service class when the Beta or Alpha ratio has a sharp decline without disturbing the overall system utilization. In addition, both schemes prevent the Silver users' high traffic from starving the traffic of Golden users. The DA–UC-ARR succeeds in providing an adequate and good performance that balances between the high priority of Golden users and the maximum delay tolerance of different traffic.

6. Conclusion

In this paper, we propose an efficient resource allocation scheme named "delay aware user classification and adaptive resource reservation-based call admission control (DA–UC-ARR)" that aims at allocating RBs adaptively based on the users' category level and the agreed-upon QoS required by the users' traffic services in terms of maximum delay. In this proposed scheme, the users' requests are classified into Golden users' requests and Silver users' requests, and the type of services per user are classified as real time (RT) and non-real time (NRT). Its performance is compared with previous resource allocation, referred to as the adaptive resource reservation-based call admission control (ARR-CAC) (Andrews et al., 2010; AlQahtani, 2014). Simulation results indicate the superiority of the proposed scheme because it is able to achieve a better balance between system utilization, users' privileges provided by network operators and QoS provisioning compared to the ARR-CAC scheme. These performance results verify the considerable improvement that can be achieved by the integration of user and traffic classifications in the call admission control process in LTE-A networks. The proposed scheme allows the mobile operators to control the quality spacing between users based predefined fixed features (Golden Or Silver) independent of traffic class which may not be desirable when the operators have roaming users, or when the operators have no such user categorizations strategy or when the minimum QoS guarantees of the users are not being met.

Benefiting from the fixed user classification predefined by the operator, the computational complexity of the proposed scheme is less than that of the conventional resource allocation algorithm. Furthermore, the system throughput of the proposed scheme is close to that of the optimal method (open sharing with dynamic priority) when the traffic intensity is larger, but the computational complexity of the proposed scheme will be much less than that of the optimal solution. Therefore, the proposed scheme can reach throughput maximization and QoS satisfaction with a lower computational complexity.

References

3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2, TS 36.300, 3rd Generation Partnership Project (3GPP), Dec. 2010.

3GPP TR 36.814. Requirements for further advancements for Evolved UTRA (EUTRA Release 9), physical layer aspects version 9.0.0., March (2010). Available: <www.3gpp.org> (accessed 1.07.2015).

AlQahtani, S.A., 2014. Analysis of an adaptive priority-based resource allocation control for LTE-advanced communications with type I relay nodes. Wireless Pers. Commun. SPRINGERS 77 (4), 2699–2722.

AlQahtani, S.A., Mahmoud, A.S., 2006. Dynamic radio resource allocation for 3G and beyond mobile wireless networks. Comput. Commun. 30 (1), 41–51.

AlQahtani, Salman A., Mahmoud, Ashraf S., 2007. Performance evaluation and analytical modeling of a novel dynamic call admission control scheme for 3G WCDMA. In: ACS/IEEE International Conference on Computer Systems and Applications, AICCSA 2007, Amman, Jordan.

Andrews, Nirmal, Kondareddy, Yogesh, Agrawal, Prathima, 2010. Prioritized resource sharing in WiMAX and Wi-Fi integrated networks. In: IEEE Wireless Communications and Networking Conference (WCNC), Sydney, Australia.

Chadchan, S.M., Akki, C.B., 2010. 3GPP LTE/SAE: an overview. Int. J. Comput. Electr. Eng. 2 (5).

Hegazy, R., Nasr, O., 2015. A user behavior based handover optimization algorithm for LTE networks. In: 2015 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1255–1260.

Kaur, S., Selvamuthu, D., 2014. Adaptive joint call admission control scheme in LTE-UMTS networks. In: IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), pp. 63–70.

Kosta, C., Sodunke, T., Shateri, M., Tafazolli, R., 2010. Two-stage call admission control policy for LTE systems. In: IWCMC '10.

Lie, H. et al, 2008. Adaptive connection admission control algorithm for LTE systems. In: IEEE, Vehicular Technology Conference, pp. 2336–2340.

Niu, Jinping, Lee, Daewon, Su, Tao, Li, G.Y., 2013. User classification and scheduling in LTE downlink systems with heterogeneous user mobilities. IEEE Trans. Wireless Commun. 12 (12), 6205–6213.

Qian, M. et al, 2009. A novel radio admission control scheme for multiclass service in LTE systems. In: Proceedings IEEE Globecom, pp. 1–6.

TR. 36.942 Annex A, Radio frequency (RF) system scenarios (release 10), December 2010, v10.2.0. Available: < www.3gpp.org >.

TR 36211, Physical channels and modulation (Release 10), December 2010, v10.0.0. Available: < www.3gpp.org >.

Yang, Jie, Qiao, Yuanyuan, Zhang, Xinyu, He, Haiyang, Liu, Fang, Cheng, Gang, 2015. Characterizing user behavior in mobile internet. IEEE Trans. Emerg. Top. Comput. 3 (1), 95–106.