RESEARCH   **Open Access**

QoS priority-based dynamic frequency band allocation algorithm for load balancing and interference avoidance in 3GPP LTE HetNet

Zeeshan Kaleem, Bing Hui and KyungHi Chang[*]

Abstract

A 3rd Generation Partnership Project (3GPP) Long Term Evolution (LTE) system uses the concept of two-tier heterogeneous networks (HetNets), where low-power and short-range femtocells are laid under macrocells to fulfill the quality of service (QoS) requirements of users and to boost overall network capacity. However, co-channel interference is one of the major issues that need to be resolved for the successful deployment of HetNets. To overcome this problem, fractional frequency reuse (FFR) schemes have been proposed that can efficiently utilize the available spectrum. Nevertheless, these schemes waste limited frequency resources owing to their static allocation and lack of following QoS requirements, network loading conditions, and service priority of users. In this paper, a QoS-based dynamic FFR (QoS-DFFR) scheme is proposed that efficiently allocates the non-occupied center-zone frequency bands, i.e., bonus bandwidth (BBW), to cell-edge users by considering their QoS requirements. Consequently, the proposed QoS-DFFR scheme can optimize cell-edge user throughput and sector throughput and reduce co-channel interference by dynamically allocating the BBW to the most demanding cell-edge users. The proposed QoS-DFFR scheme improves performance because of its ability to dynamically allocate the limited portion of the frequency bands based on the service priorities of users. The system-level simulation results show that the proposed QoS-DFFR scheme performs remarkably well in a HetNet environment. Compared with the usual FFR schemes, the proposed scheme almost doubles the cell-edge user's throughput and reduces the user's packet loss rate.

Keywords: Fractional frequency reuse; Load balancing; Quality of service; Heterogeneous networks; 3GPP LTE

1 Introduction

Conventional cellular systems use a planned homogeneous macrocell-based network architecture where the macrocells, i.e., evolved NodeBs or eNBs, provide services to the users in the network. In a homogeneous environment, the eNBs have the same transmit power levels, antenna patterns, receiver noise floors, and backhaul connectivity. This allows the network to offer a similar quality of service (QoS) to the user equipment (UE) across all cells [1,2]. However, such a deployment is not suitable for cell-edge users because they have weak received signal power from the base stations (BSs). This degrades the QoS and throughput performance of the users.

To improve QoS and throughput performance for the users, two-tier heterogeneous networks (HetNets) are

considered [3]. These networks have conventional eNBs in the first tier, overlaid with low-power LTE-Advanced (LTE-A) femtocells. In LTE-A, femtocells are commonly referred to as home evolved NodeBs (HeNBs). A recent study shows that 114 million mobile users will be accessing mobile networks through HeNBs by 2014 [4]. In HetNets that are composed of eNB and HeNB environments, two classes of users are considered. Users who are connected to eNB are called macro users (MUE), while those connected to HeNBs are referred as femto users (FUE). In two-tier HetNets, the HeNBs have the advantages of small form factor, low cost, and smaller coverage. These advantages support reuse of the same licensed band multiple times within the second tier of the HetNets. That is, the employment of HeNB improves QoS for users, spectral efficiency, and hence the capacity of the overall network. However, the use of the same frequency bands between MUE and FUE leads to several

* Correspondence: khchang@inha.ac.kr
Department of Electronic Engineering, Inha University, Incheon 402-751,
South Korea

kinds of co-channel interference. The major interference scenario involves downlink (DL) interference from HeNBs to MUEs.

In order to reduce co-channel interference, various kinds of cooperation techniques have been considered in recent literature, including collaborative frequency scheduling among eNBs and HeNBs [5], group formation of HeNBs by exchanging information among neighboring HeNBs [6,7], power control approaches [8-10], and intelligent spectrum access [11]. Moreover, many uncoordinated schemes such as classical fractional frequency reuse (FFR) [12] have been proposed; a cellular network using a frequency reuse factor of 3 (FRF3) would be an example. This type of network reduces co-channel interference and is particularly helpful in improving cell-edge users' throughput by proper cell planning or spectrum management between the eNB and HeNBs, at the cost of system throughput. In order to balance the cell-edge users' throughput and thus the overall throughput, a modified FFR scheme where the cell is divided into two zones, i.e., full-reuse (FR) and the partial-reuse (PR) zones, as shown in Figure 1, has been proposed [13]. By applying this modified FFR scheme, much research has been conducted to optimize the system throughput. The optimal static FFR (OSFFR) scheme is proposed in [14], where a bigger reuse factor of 6 is employed to improve the cell-edge user throughput. Notice that in OSFFR, the available frequency bands are equally allocated to PR zones without considering the network loading conditions.

The scheme of capacity density optimization by fractional frequency partitioning, proposed in [15], maximizes per-user throughput instead of system throughput. However, during the optimization of user throughput per cell, QoS and network loading conditions are not considered. Hence, the capacity density optimization scheme may result in lower spectral efficiency because of the static allocation of frequency bands.

To solve the issues of QoS and network loading conditions, we propose a QoS-based dynamic FFR (QoS-DFFR) scheme for frequency band allocation to fulfill users' QoS requirements and to maximize per-user and per-sector throughput. In the proposed QoS-DFFR scheme, the concept of bonus bandwidth (BBW) is introduced to dynamically allocate the frequency bands to the PR zones according to QoS requirements and service priorities of the users. In this way, the QoS of the users can be guaranteed, and the per-user and per-sector throughput can be optimized, according to network loading conditions.

The rest of the paper is organized as follows: Section 2 describes the system model that includes the network setup, channel models, and physical (PHY) layer. In Section 3, the necessary components of bearer models, schedulers, and traffic models that have their own role in QoS provisioning are covered in detail. Section 4 explains the proposed QoS-DFFR scheme, capacity optimization steps, BBW allocation procedure, and several advantages of the proposed algorithm. In Section 5, the performance of the proposed

Figure 1 Fractional frequency reuse in 3GPP LTE-A HetNet.
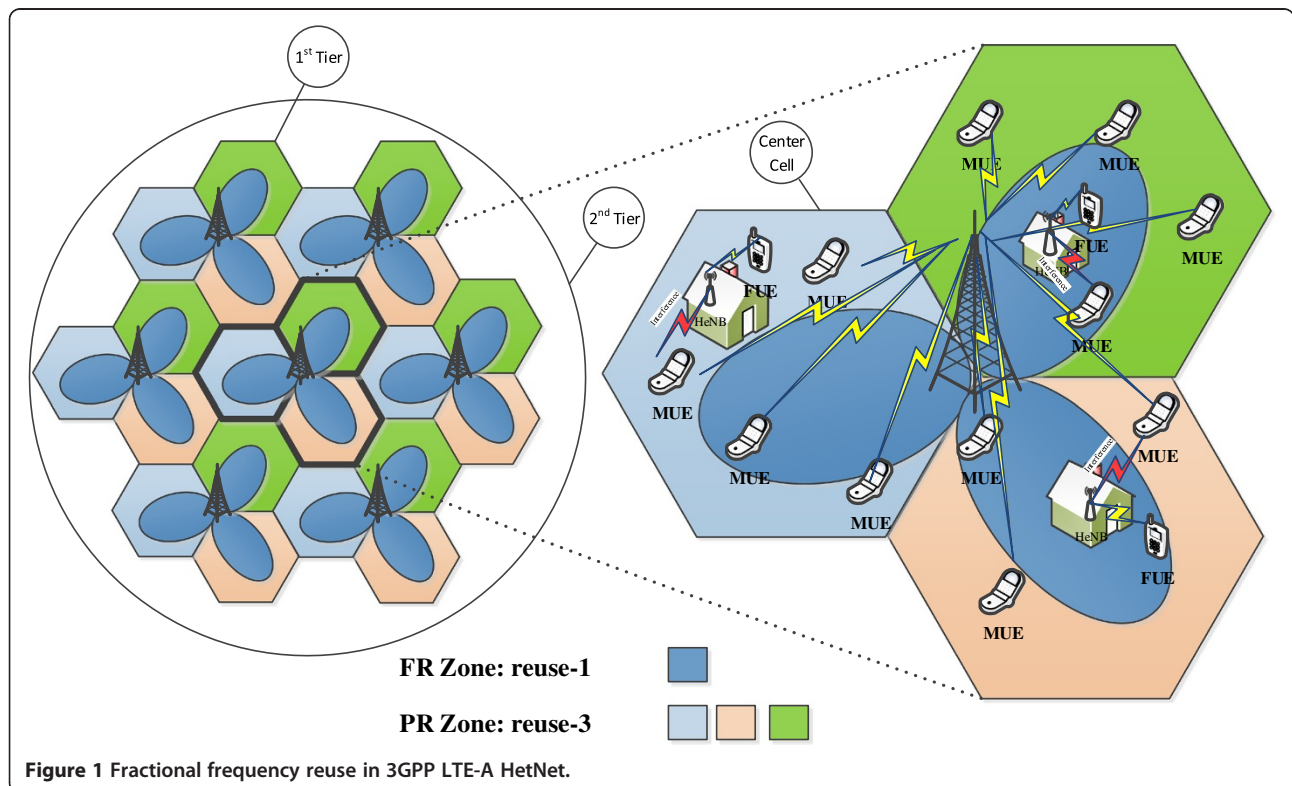
QoS-DFFR scheme is evaluated using system-level simulations. Finally, in Section 6, we draw conclusions.

2 System model

2.1 Network setup

In this paper, the DL of a 3rd generation partnership project (3GPP) LTE-A system is considered. Each user is allocated a number of so-called physical resource blocks (PRBs) in the time-frequency grid. The PRBs correspond to the smallest radio resource that can be assigned to UE for data transmission. A resource block has a size of 180 kHz in the frequency domain and 0.5 ms in the time domain. Due to the fixed size of PRBs, the number of PRBs varies according to the system bandwidth (BW) configuration, e. g., 50 and 100 PRBs for system BWs of 10 and 20 MHz, respectively. In this paper, to avoid co-channel interference, both the eNBs and HeNBs utilize the fractional frequency bands instead of the full system BW.

2.1.1 HetNet deployment scenario

We consider a two-tier HetNet environment with 19 eNB sites. Each site is composed of three hexagonal sectors, as shown in Figure 1. In the deployed scenario, the center cell and first-tier cells are the area of interest, while second-tier cells consist of only eNBs to generate interference to the users. The area of interest consists of eNBs, HeNBs, MUEs, and FUEs. In each hexagonal sector, a 5×5 grid model is considered for HeNB deployment.

The 5×5 grid model is a simple model for HeNB urban deployment. It is composed of 25 apartments, as shown in Figure 2. Each apartment is of size 10 m × 10 m, and the apartments as placed next to each other. The number of HeNBs per grid depends upon the deployment ratio, while the percentage of an active HeNB in a single apartment is decided by the activation ratio [16]. MUEs are dropped randomly throughout the indoor and outdoor macro site areas using a uniform distribution pattern. Hence,

it is possible that some MUEs will be dropped into the HeNB area, as shown in Figure 2. It is assumed that at least one FUE is connected per HeNB. In this paper, we set all the HeNBs as active.

2.1.2 General notations

The general notations of the considered network are given below.

- M is the set of eNBs.
- F is the set of HeNBs.
- U is the set of UEs separated into two subsets U_m and U_f, where each subset represents MUE and FUE, respectively.
- N_m is the set of PRBs used in eNB m. Each PRB n_m, with 12 subcarriers, has an equal transmission power for each subcarrier.
- N_f is the set of PRBs used in HeNB f. Each PRB n_f, with 12 subcarriers, has an equal transmission power for each subcarrier.
- $G_{u_m}^{n_m}$ is the channel gain between the eNB m and the MUE u_m on PRB n_m.
- $G_{u_f}^{n_f}$ is the channel gain between the HeNB f and the FUE u_f on PRB n_f.
- $I_{u_m}^{n_m}$ is the aggregate interference at MUE u_m, composed of interference from eNB and HeNB on PRB n_m.
- $I_{u_m}^{n_f}$ is the aggregate interference at FUE u_f composed of interference from eNB and HeNB on PRB n_f.
- η is the thermal noise, including the UE noise figure, per PRB.
- $\text{SINR}_{u_m}^{n_m}$ is the received signal-to-interference-plus-noise ratio (SINR) observed by MUE u_m on PRB n_m.
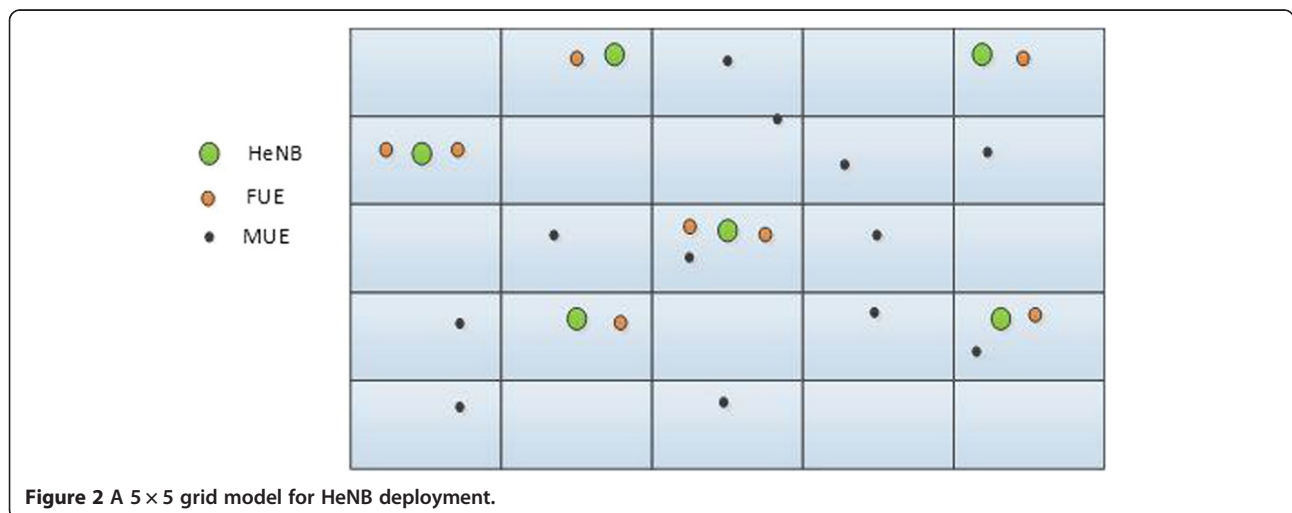- $\text{SINR}_{u_f}^{n_f}$ is the received SINR observed by FUE u_f on PRB n_f.

Figure 2 A 5×5 grid model for HeNB deployment.

The SINRs observed on PRBs n_m and n_f by MUE u_m and FUE u_f, respectively, are described as

$$\begin{cases} \text{SINR}_{u_m}^{n_m} = \dfrac{G_{u_m}^{n_m} p_{u_m}^{n_m}}{I_{u_m}^{n_m} + \eta} \\ \text{SINR}_{u_f}^{n_f} = \dfrac{G_{u_f}^{n_f} p_{u_f}^{n_f}}{I_{u_f}^{n_f} + \eta} \end{cases}, \tag{1}$$

where $p_{u_m}^{n_m}$ and $p_{u_f}^{n_f}$ is the transmit power from eNB m and HeNB f to MUE u_m and FUE u_f on PRBs n_m and n_f, respectively. The aggregate interferences $I_{u_m}^{n_m}$ and $I_{u_f}^{n_f}$ received by MUE u_m and FUE u_f, respectively, are given as

$$\begin{cases} I_{u_m}^{n_m} = \displaystyle\sum_{i \in M, i \neq m} G_{u_i}^{n_m} p_i^{n_m} + \sum_{j \in F} G_{u_j}^{n_m} p_j^{n_m} \\ I_{u_f}^{n_f} = \displaystyle\sum_{i \in M} G_{u_i}^{n_f} p_i^{n_f} + \sum_{j \in F, j \neq f} G_{u_j}^{n_f} p_j^{n_f} \end{cases}. \tag{2}$$

Equation 2 describes the total interference received by MUE and FUE from their surrounding eNBs and HeNBs. The number of eNBs and HeNBs providing interference to MUE or FUE depends upon the specific frequency reuse scheme being used.

2.2 Channel model and antenna patterns

The channel model represents the propagation loss that occurs when the signal travels from the transmitter to the receiver. The general equation to model the channel gain is given as

$$G \text{ (dB)} = \text{Antenna gain–Path loss–Shadowing} \atop \text{–Fading–Penetration.} \tag{3}$$

2.2.1 Path loss

To calculate the path loss (PL) of MUEs residing outside a building and connected to an eNB, the urban area model [17] is considered as

$$\text{PL (dB)} = 40 \cdot \left(1 \times 10^{-3} \text{ to } 4 \times 10^{-3} \cdot \text{Dhb}\right) \cdot \log_{10}(R) \atop -18 \cdot \log_{10}(\text{Dhb}) + 21 \cdot \log_{10}(f) + 80, \tag{4}$$

where R is the distance between the eNB and MUE in kilometers, f is the carrier frequency in MHz, and Dhb is the BS antenna height, in meters, measured from the average rooftop level. For an MUE located inside a building and served by an eNB, the path loss calculation considers an additional attenuation factor of 20 dB due to the presence of an external wall.

For FUEs connected to HeNBs, the indoor propagation model specified in the LTE-A evaluation methodology [16] for a dense urban deployment of HeNBs is considered. This model is the simplified PL model, which avoids modeling any walls and considers only the distance R, expressed in meters, between the transmitter and receiver.

$$\text{PL (dB)} = 127 + 30 \cdot \log_{10}(R/1000). \tag{5}$$

2.2.2 Shadowing

Shadowing is caused by obstacles in the paths between the UEs and the BS. Shadowing is usually modeled by a log-normal distribution with a mean of 0 dB and standard deviation of 4 dB for the link between the HeNB and UE [16]. For all other links, including interference links, the standard deviation is 10 dB. The concept of a correlated shadow fading map proposed by Claussen for eight neighbors [18] is used, and the inter-site correlation value of 0.5 is considered [17].

2.2.3 Fast fading

Fast fading is the result of the constructive and destructive combination of randomly delayed, reflected, scattered, and diffracted signal components. This type of fading is relatively fast and is therefore responsible for the short-term signal variations that can occur when UEs or reflectors in an environment move short distances. In this paper, the fast-fading component of the signal is generated according to the WINNER II channel model [19]. The WINNER II channel model is suitable for indoor and outdoor users and is an evolution of the 3GPP spatial channel model extended (SCME). WINNER II covers more scenarios than 3GPP SCME and introduces additional features such as support for arbitrary 3D antenna patterns, multidimensional channel modeling for carrier frequencies between 2 and 6 GHz, and BWs up to 100 MHz. WINNER II is a more reliable and practical model for system-level simulations than the SCME model.

2.2.4 Antenna patterns

The 3D antenna patterns given by horizontal and vertical cuts are used for eNB as shown in Figure 3. The horizontal and vertical antenna patterns are modeled as

$$\begin{cases} A_H(\varphi_H) = -\min\left[12\left(\dfrac{\varphi_H}{\varphi_{3\text{dB}}}\right)^2, \ A_m\right] \\ A_V(\theta_V) = -\min\left[12\left(\dfrac{\theta_V - \theta_{\text{etilt}}}{\theta_{3\text{dB}}}\right)^2, \ \text{SLA}_v\right] \end{cases}, \tag{6}$$

where A_H and A_V are the horizontal and vertical antenna patterns, respectively [17]. φ_H and θ_V are the azimuth and vertical orientation, respectively. $\varphi_{3\text{dB}}$, $\theta_{3\text{dB}}$ are the 3-dB horizontal and vertical beam widths, respectively, and A_m and SLV_v are the backward attenuation and side lobe vertical attenuation, respectively. θ_{etilt} describes the downtilt of the electrical antenna.

Typical values for these parameters are as follows: $-180 \leq \varphi_H \leq 180$, $-180 \leq \theta_V \leq 180$, $\varphi_{3\text{dB}} = 70°$, $\theta_{3\text{dB}} = 10°$, $A_m = 25$ dB, $\text{SLA}_V = 20$ dB, and $\theta_{\text{etilt}} = (0°, 15°)$. The maximum value of
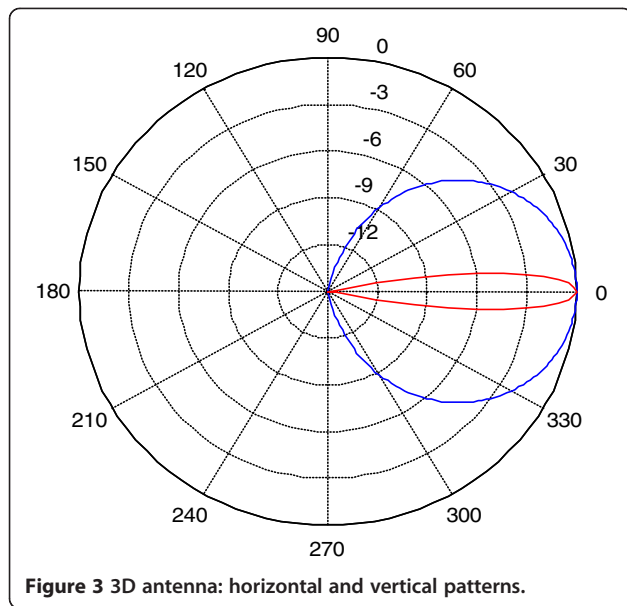
Figure 3 3D antenna: horizontal and vertical patterns.

the antenna gain is 15 dBi [17], which is suitable for an urban environment. Then, the 3D antenna patterns for eNB are represented as

$$A(\theta, \varphi) = - \min \left[A_V(\theta)_v + A_H(\varphi)_H, A_m \right]. \qquad (7)$$

The antenna patterns for UEs and HeNBs are assumed to be omnidirectional.

2.3 PHY layer

We abstract the PHY layer by considering the single-input single-output (SISO) antenna configurations, additive white Gaussian noise (AWGN) channel, and turbo code of rate 1/3. In LTE-A, the channel quality indicator (CQI) index is defined in terms of a channel coding rate and modulation schemes such as quadrature phase shift keying (QPSK), 16 quadrature amplitude modulation (QAM), and 64QAM [20].

The higher layers are modeled by considering the multiple-input multiple-output (MIMO) antenna configurations and the parameters mentioned in Table 1. In system-level simulations, the PHY layer, including all signal processing, cannot be explicitly simulated due to prohibitive complexity. To avoid actual signal processing, in system-level simulations, a link-to-system interface composed of a suitable PHY layer abstraction is typically used. Because the parameters of the PHY abstraction models such as frequency, symbol duration, and error correcting codes do not depend on the channel models or the multi-antenna techniques employed, the SISO-abstracted PHY layer can be used for MIMO configurations in the system-level simulations [21]. The goal of this PHY layer abstraction is essentially to obtain a block error rate (BLER) for a transport block (TB) without

Table 1 Simulation parameters

Parameters	Values
Link-level parameters	
Frequency	2.19 GHz
Bandwidth (B_{TOT})	20 MHz
Channel model	AWGN
AMC formats	QPSK: 1/3, 1/2, 2/3, 4/5
	16QAM: 1/2, 2/3, 4/5
	64QAM: 2/3, 4/5
Channel coding	Turbo code (1/3)
Number of antennas	SISO (1×1)
System-level parameters	
Number of eNBs	57 (two rings, 19 sites)
	(only 21 inner cells are in the region of interest)
Inter-eNB distance	500 m
Initial FR zone radius (R_{FR})	0.65
Number of HeNBs	126 (6/grid)
HeNB deployment	5×5 grid
UEs per eNB and HeNB	10, 2
eNB frequency reuse	3
Minimum coupling loss	70 dB (eNB)
	40 dB (HeNB)
Fading channel model	WINNER II
Path loss model	Urban (eNB)
	Grid Model (HeNB)
Shadowing	Claussen Fading Model
UE speed	3 km/h
Noise spectral density	−175 dBm/Hz
Number of antennas	4×4
MIMO modes	CLSM
Antenna pattern	TS 36.942 3D (eNB)
	Omnidirectional (HeNB)
UE receiver	Zero Forcing (ZF)
Traffic models	VoIP, video, gaming, HTTP, and FTP

having to carry out real signal processing. To assess the BLER of the received TBs, a set of AWGN link-level performance curves are employed. The SINR-to-BLER mapping then requires an effective SINR value, which is obtained from mapping the set of subcarrier SINRs assigned to the UE TB to an AWGN-equivalent SINR. The mutual information-based effective SINR mapping (MIESM) [22] is the method currently used to obtain a TB-effective SINR that can be used to map to the BLER obtained from the AWGN link-level simulation. The details regarding the PHY layer abstraction are provided in Section 5.2.

3 QoS provisioning in LTE-A

In an LTE-A network, users have diverse QoS requirements. The QoS differentiation enables communication service providers (CSPs) to manage the performance of mixed traffic flows. Hence, it is vital for CSPs to treat users based on their QoS requirements and service priorities. The bearer models and schedulers are used to meet the QoS requirements of the users. The bearer models discriminate the mixed traffic flows, while the schedulers allocate frequency bands to users. The frequency bands are allocated among users by considering metrics such as throughput, head-of-line (HOL) packet delays, priorities, and fairness.

3.1 Traffic models

To simulate typical traffic and network loading conditions, the realistic traffic models in [23] are used to assess the performance of the proposed QoS-DFFR scheme. The users are differentiated in terms of applications they use, e.g., VoIP, video, gaming, HTTP, and file transfer protocol (FTP). In this paper, to fully utilize the traffic models at a user level, the complete LTE protocol stack and bearer models are considered [24].

3.1.1 Voice over Internet Protocol traffic model

For voice over Internet Protocol (VoIP) applications, we use 'on' and 'off' Markov models, as shown in Figure 4 [23]. In a VoIP model, the probability of transitioning from active state 1 to non-active state 0 is 'a', while the probability of transitioning from non-active state 0 to active state 1 is 'c'. At a voice source rate of 12.2 kbps, a voice frame consisting of 244 bits is generated every 20 ms. A silence insertion descriptor (SID) packet consists of 15 bytes, and a SID packet is transmitted every 160 ms during silent periods.

The probability of being in active state 1 and non-active state 0 is calculated by (8) and is represented by P_1 and P_0, respectively:

$$\begin{cases} P_1 = \dfrac{c}{a + c} \\ P_0 = \dfrac{a}{a + c} \end{cases}. \tag{8}$$

3.1.2 Video traffic model

In video traffic model, a source rate of 64 kbps is assumed. Each frame of video data arrives at a regular interval T determined by the number of frames per second. Each frame is decomposed into a fixed number of slices; that is, the maximum eight slices per frame has been considered, and each is transmitted as a single packet. The size of these packets is modeled as a truncated Pareto distribution p_x with mean, maximum, location parameter (α), and shape parameter (k) values of 10 bytes, 250 bytes, 1.2 ms, and 20 bytes, respectively [23]:

$$p_x = \frac{\alpha_k^\alpha}{a + 1}, \ k \leq x < m, \tag{9}$$

$$p_x = \left(\frac{k}{m}\right)^\alpha, \ x = m, \tag{10}$$

The video encoder introduces a delay between the packets of a frame. These delay intervals are also modeled by a truncated Pareto distribution p_x, given in (9), with mean, maximum, location parameter (α), and shape parameter (k) values of 6 ms, 12.5 ms, 1.2 ms, and 2.5 ms, respectively.

3.1.3 Gaming traffic model

To simulate the random timing relationship between the arrival of a client traffic packet and the DL frame boundary, the starting time of a network gaming mobile is uniformly distributed within the specified range of [0, 40] ms. The initial packet arrival of the game is described by the uniform distribution u_x as

$$u_x = \frac{1}{b - a}, \quad \text{for } a \leq x \leq b \quad \text{where } a = 0, \ b = 40 \text{ ms}. \tag{11}$$
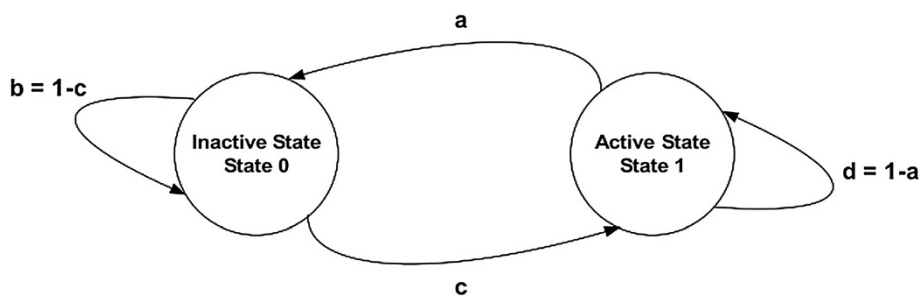

Figure 4 Two-state voice activity model.

The packet size and the packet arrival time are described by the largest extreme value distribution g_x, which is described as

$$g_x = \frac{1}{b} e^{-\frac{x-a}{b}} e^{-e^{-\frac{x-a}{b}}}, \tag{12}$$

where a = 55 ms and 120 bytes and b = 6 and 36 bytes for the packet arrival time and packet size, respectively.

3.1.4 Web-browsing HTTP traffic model

For web browsing, a web page consists of a main object and embedded objects. After receiving the main page, the web browser will parse for the embedded objects. The main parameters to characterize web browsing are the main size of an object S_M, the size of an embedded object S_E, the number of embedded objects N_D, reading time D, and parsing time for the main page T_P. The main object size S_M and the embedded object size S_E are generated by using truncated lognormal distribution l_x in (13) with the specific minimum, maximum, mean, and standard deviation values in [23]. The reading time D and parsing time T_P are calculated by using exponential distribution e_x in (14) with λ = 0.033 and λ = 7.69, respectively:

$$l_x = \frac{1}{\sqrt{2\pi}\sigma x} e^{\frac{-(\ln x - \mu)^2}{2\sigma^2}}, \ x > 0. \tag{13}$$

$$e_x = \lambda e^{-\lambda x}, \ x \geq 0. \tag{14}$$

3.1.5 Best effort (FTP) traffic model

An FTP is considered to be the best-effort traffic. An FTP session is a sequence of file transfers separated by reading time D. The two main FTP session parameters are size S of a file to be transferred and reading time D, i.e., the time interval between the end of download of the previous file and the user request for the next file. The files of size S are also generated by using (13), while reading time D is calculated by using (14) with the specific minimum, maximum, mean, λ, and standard deviation values given in [23].

3.2 Bearer model

In LTE-A specifications [24], the evolved packet system (EPS) bearer has been introduced to provide QoS differentiation. The EPS bearer maps a flow into a logical channel established between the UE and the gateway (GW). Moreover, a radio bearer is associated with each EPS bearer as a logical channel between UE and eNB. There are two types of bearers: the default bearer and the dedicated bearer. The default bearer is established when the UE is attached to the network, and it remains established during the lifetime of the connection, providing basic connectivity and exchange of control messages. By contrast, the dedicated bearer is established when there is a need to provide QoS to specific services such as VoIP, video, and web browsing. Orthogonally, bearers can be also classified as guaranteed bit rate (GBR) or non-guaranteed bit rate (non-GBR), according to the QoS requirements of the flow they map. In particular, the default bearer is a non-GBR; a dedicated bearer can be GBR or non-GBR. A set of QoS parameters is therefore associated with each bearer depending on the application data it carries, thus enabling differentiation among flows. Several classes of QoS have been identified through QoS class identifiers (QCIs), as given in Table 2 [25].

### 3.3 QoS-aware scheduling in LTE-A

To satisfy the users' QoS requirements during one transmission time interval (TTI) or subframe, which is the smallest scheduling interval in LTE-A, the frequency bands should be allocated efficiently among users. During scheduling, the frequency bands are usually allocated to users who maximize a metric in (15):

$$m_{j,k} = \max_{i}\{m_{i,k}\}. \tag{15}$$

This metric can be considered as the transmission priority of each user in a specific resource block. The value of the metric depends upon QCI values associated with each flow. The schedulers are used to differentiate the

Table 2 Standardized QCIs for LTE

QCI	Resource type	Priority	Packet delay budget (ms)	Packet error loss rate	Example services
1	GBR	2	100	10^{-2}	Conversational voice
2	GBR	4	150	10^{-3}	Conversational video (live streaming)
3	GBR	5	300	10^{-6}	Non-conversational video (buffered streaming)
4	GBR	3	50	10^{-3}	Real-time gaming
5	Non-GBR	1	100	10^{-6}	IMS signaling
6	Non-GBR	7	100	10^{-3}	Voice, video (live streaming), interactive gaming
7	Non-GBR	6	300	10^{-6}	Video (buffered streaming)
8	Non-GBR	8	300	10^{-6}	TCP-based (for example, www, e-mail), chat, FTP, p2p file sharing, progressive video, and others
9	Non-GBR	9			
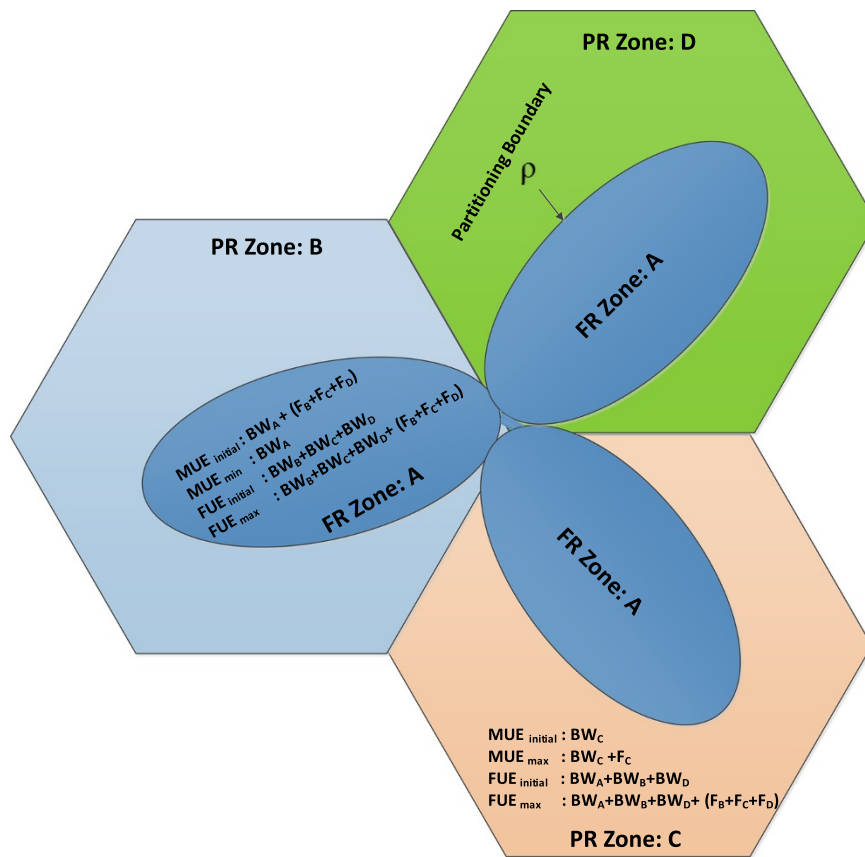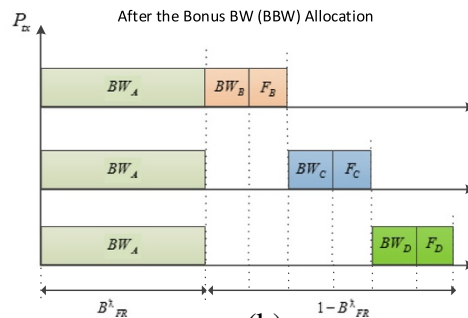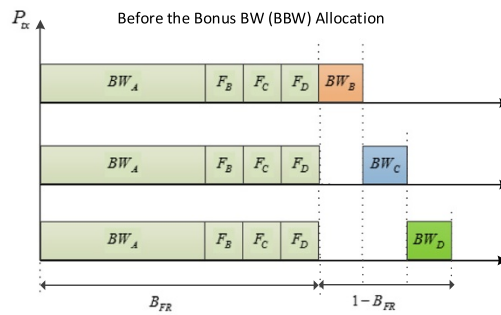
Figure 5 The proposed QoS-DFFR scheme. (a) QoS-DFFR bandwidth partitioning and **(b)** effect of bonus bandwidth allocation.

specific flows and to schedule the users based on their QoS requirements and service priorities. In this paper, we consider the exponential proportional fair (EXP/PF) scheduler [26] to separate the real-time (RT), non-real-time (NRT), and best-effort (BE) flows.

The EXP/PF scheduler gives priority to the real-time packets with respect to non-real-time packets when the packets' HOL delay is very close to the delay threshold. For real-time flows, the metric is given as

$$
\begin{aligned}
m_{i,j} &= \exp\left(\frac{\beta_i D_{\mathrm{HOL},i} - \chi}{1 + \sqrt{\chi}}\right)\frac{T_{i,j}}{R_i}, \\
\text{where} \quad \chi &= \frac{1}{N_{\mathrm{rt}}}\sum_{i=1}^{N_{\mathrm{rt}}}\beta_i D_{\mathrm{HOL},i}, \\
\beta &= -\frac{\log\delta_i}{\tau_i},
\end{aligned}
\tag{16}
$$

where $T_{i,j}$ is the instantaneous available data rate for the ith flow and the jth sub-channel, R_i is the available past data rate, $D_{\mathrm{HOL},i}$ is the HOL packet delay, δ_i is the maximum probability that the packet delay can increase $D_{\mathrm{HOL},i}$, and N_{rt} is the number of active DL real-time flows.

For non-real-time and BE flows, the metric will be based on a simple PF scheduler [27]. The PF scheduler metric is defined as the ratio between the instantaneous data rate and the past average data rate:

$$
m_{i,j} = \frac{T_{i,j}}{R_i}.
\tag{17}
$$

4 Proposed QoS-based dynamic FFR scheme for load balancing and interference management in HetNets

The conventional and proposed QoS-DFFR scheme is based on dividing the cell area into two zones, i.e., center and edge zones. In the center zone, which is named an FR zone, the interference is lower and reuse-1 is employed. In edge zones, which are named as PR zones, the users have high co-channel interference, and therefore, reuse-3 is employed. The FR zone is represented as A, while the PR zones are labeled as B, C, and D in Figure 5a.

4.1 Conventional FFR scheme
In conventional FFR, the total BW is divided into center and cell-edge frequency bands:

$$
B_{\mathrm{TOT}} = B_{\mathrm{FR}} + 3B_{\mathrm{PR}},
\tag{18}
$$

where B_{TOT} is the total available BW, and B_{FR} and B_{PR} denote the FR and PR BW, respectively. The normalized FR and PR BW are represented as β_{FR} and α_{PR}, respectively:

$$
\beta_{\mathrm{FR}} = B_{\mathrm{FR}}/B_{\mathrm{TOT}}, \ \text{with}\ \beta_{\mathrm{FR}}\in[0,1].
\tag{19}
$$

$$
\alpha_{\mathrm{PR}} = \frac{1}{3}\left(1 - \beta_{\mathrm{FR}}\right).
\tag{20}
$$

The FR and PR zones are separated by a partitioning boundary ρ, as shown in Figure 5a. The frequency bands allocated to the FR and PR zones depend on the size of each zone. The users are uniformly distributed in the sector; hence, by varying the partitioning boundary ρ, as described in [15], the number of frequency bands allocated to each sector also varies. Taking this into account, a normalized BW density (Hertz per unit of surface) is defined as

$$
N_{\mathrm{BW\_density}} = \begin{cases} \dfrac{\beta_{\mathrm{FR}}}{A_{\mathrm{FR}}(\rho)} \\[2ex] \dfrac{\frac{1}{3}\left[1-\beta_{\mathrm{FR}}\right]}{A_{\mathrm{PR}}(\rho)} \end{cases}
\tag{21}
$$

where $A_{\mathrm{FR}}(\rho)$ and $A_{\mathrm{PR}}(\rho)$ denote the area of the FR and PR zones, respectively. In a conventional FFR scheme, frequency bands are equally divided in the PR zones, but this division is valid only for constant network loading conditions. For mixed traffic flows, the QoS requirement of users per sector is different; i.e., some sectors need more BW, while others do not request the same amount of BW during the same TTI. Therefore, under the conventional FFR scheme, user capacity per sector is optimized without considering network loading conditions. The capacity for FR and PR zones in the conventional FFR scheme is described as

$$
C_{\mathrm{FR},\rho,\beta_{\mathrm{FR}}}(x,y) = \frac{\beta_{\mathrm{FR}}}{A_{\mathrm{FR}}(\rho)}\log_2[1 + \mathrm{SINR}_{\mathrm{FR}}(x,y)].
\tag{22}
$$

$$
C_{\mathrm{PR},\rho,\alpha_{\mathrm{PR}}}(x,y) = \frac{\alpha_{\mathrm{PR}}}{A_{\mathrm{PR}}(\rho)}\log_2[1 + \mathrm{SINR}_{\mathrm{PR}}(x,y)].
\tag{23}
$$

Hence, in the conventional FFR scheme, a significant amount of BW is wasted because the scheme does not consider the dynamic network loading conditions. This BW wastage results in a less spectrally efficient system.

4.2 Proposed QoS-based dynamic FFR scheme
The conventional FFR scheme works well only under constant network loading conditions. In this paper, we propose a QoS-DFFR scheme that considers the dynamic network loading conditions while efficiently allocating frequency bands to users. In the proposed QoS-DFFR scheme, the concept of BBW is employed. The BBW is the BW that can be dynamically allocated to users based on their QoS requirements and service priorities. The main reason to use the concept of BBW with the FFR scheme is to dynamically allocate the extra bands to more demanding users. If
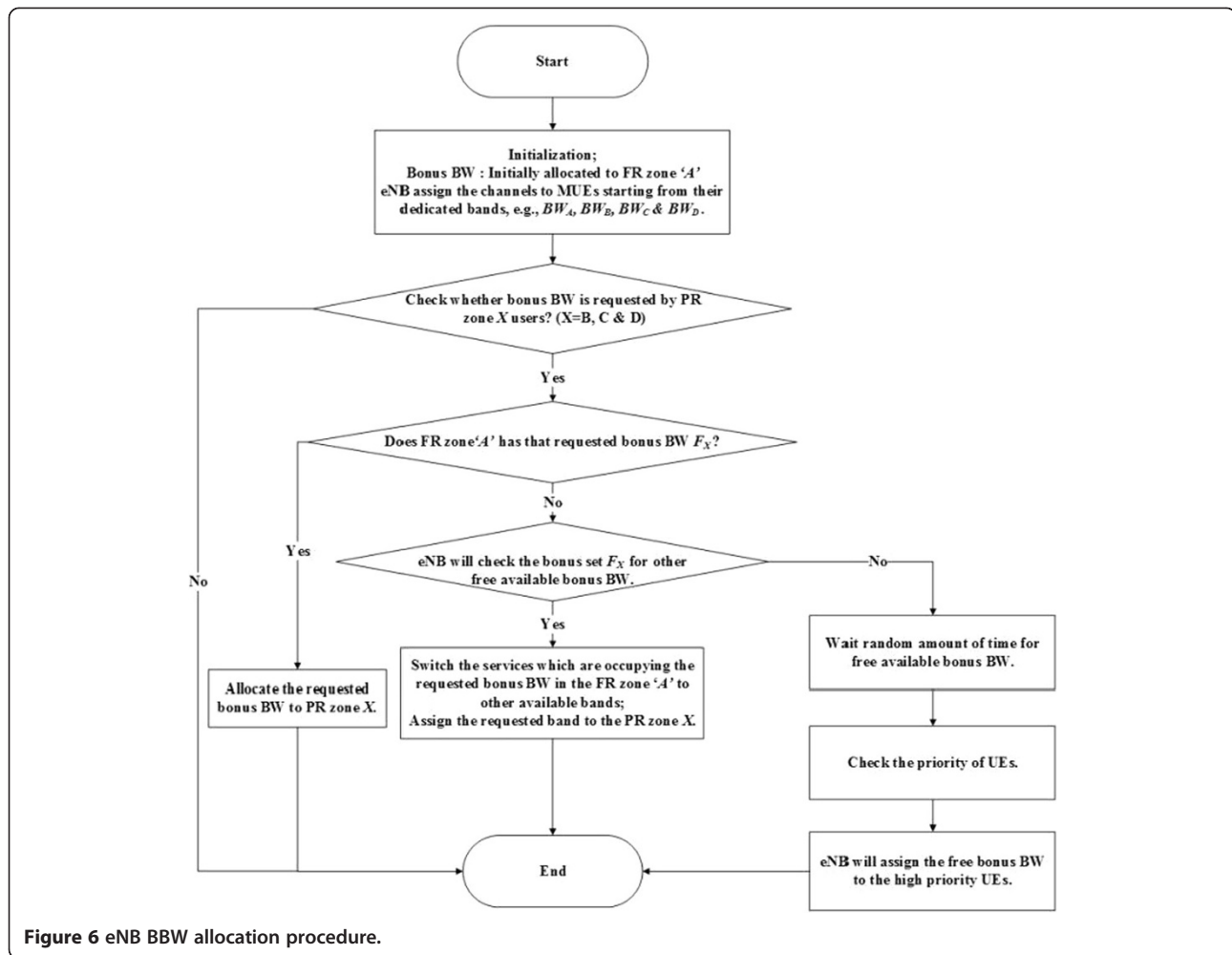
Figure 6 eNB BBW allocation procedure.

we allocate all of the available BW to the BBW, then the main purpose of dividing the area into PR and FR zones, to avoid interference among users by utilizing the orthogonal frequency bands, would not be fulfilled. Hence, the system will behave more like the reuse-1 system than FFR with reuse-3. This increases the peak throughput at the cost of mean and edge throughput. Thus, in the proposed QoS-DFFR scheme, eNB initially partitions the BW into seven parts as

$$B_{\text{TOT}} = BW_{\text{A}} + BW_{\text{B}} + BW_{\text{C}} + BW_{\text{D}} \\ + (F_{\text{B}} + F_{\text{C}} + F_{\text{D}}), \qquad (24)$$

Of the above seven BWs, four parts of the frequency bands are initially allocated to the FR zones. These parts include the BBW, i.e., $BW_{\text{A}} + F_X$, where F_X is the BBW with $X = \{B, C, \text{and } D\}$, while the other three parts are equally divided in the PR zones according to (27). The number of BBW bands in the proposed scheme depends upon the number of PR zones. By considering BBW allocation, the tessellation property of FFR-3 will remain valid

because it always has one FR zone and three PR zones in the scenario.

The amount of dedicated BW and BBW is proportional to the size of the area. To optimize users' per-sector capacity according to the varying network loading conditions, the partitioning boundary ρ is varied in every subframe, i.e., the BW allocated to each zone will be different per subframe. By considering the uniform distribution of the users over each cell, varying the partitioning boundary ρ effectively alters the amount of physical resources that can be allocated to each user.

The exact number of bands to be allocated to each FR zone depends upon the radius of that zone [14]. The radius of the FR zone will be adjusted according to network loading conditions and the users' service priorities in every subframe. That is, if more users with high service priority are located in the PR zone, then the radius of the FR zone would be decreased and vice versa. Hence, these bands are named as the BBW is taken out of the FR zone and will be allocated to the most demanding PR zone. For static FFR, it has been proven in [14] that in order to achieve
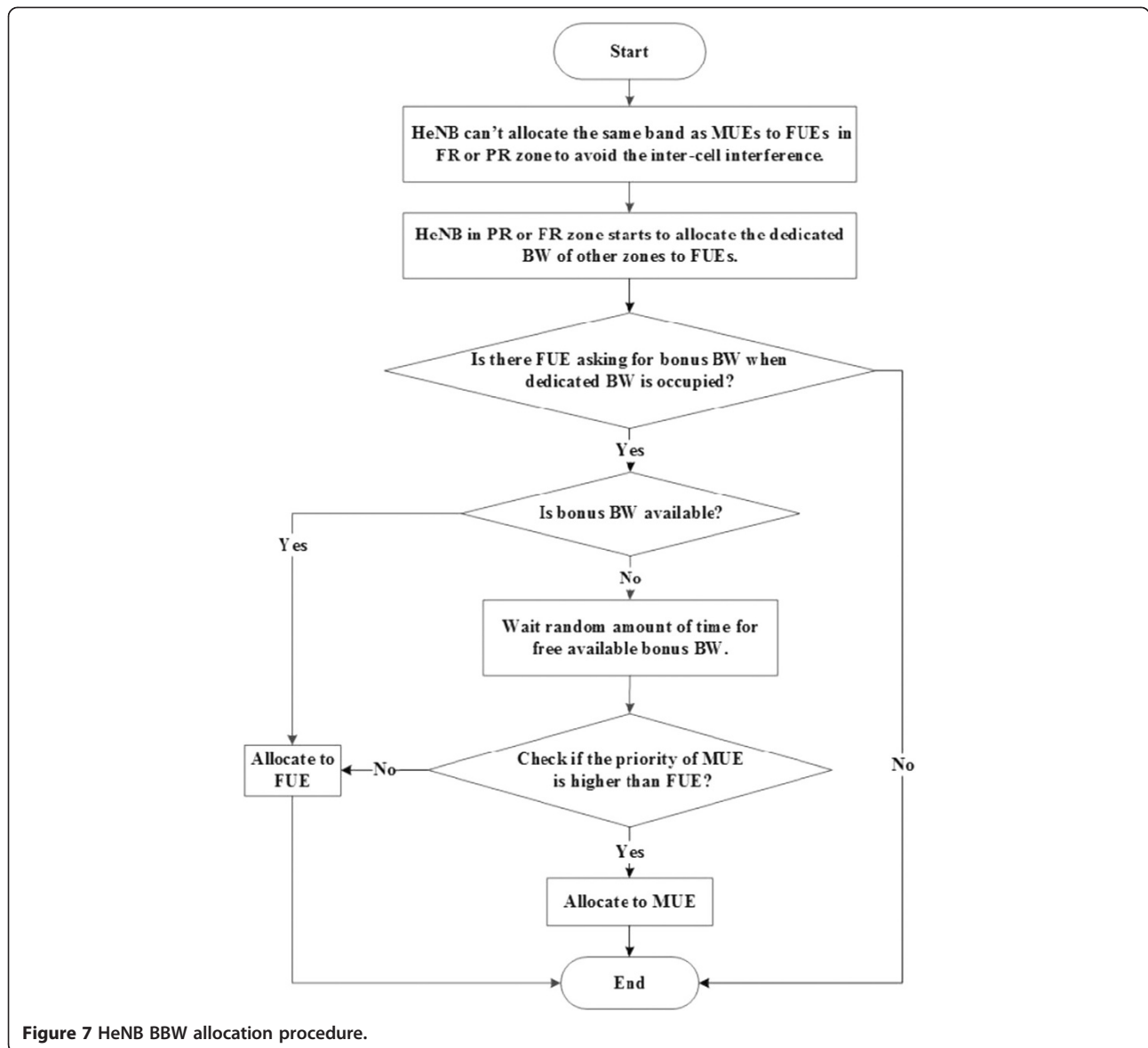
Figure 7 HeNB BBW allocation procedure.

optimum average throughput in the FR zone, the FR zone radius (R_{FR}) should be 0.65 of the normalized radius (R) of the cell. Hence, we start our simulation by considering $R_{FR} = 0.65$ and then vary it in steps of $\Delta R_{FR} = 0.05$ as the demand in the PR zone varies due to dynamic FFR conditions. Thus, the exact values of the FR zone, PR zone, and BBW in each subframe duration can be calculated as follows:

$$B_{FR}(i) = B_{TOT}\left(\frac{R_{FR}(i)}{R}\right)^2$$
$$= BW_A(i) + F_X(i),$$

where $R_{FR}(i)$: radius of FR zone in sub-frame(TTI)i

R : radius of the cell
$$(25)$$

$$BW_A > BW_X.$$
$$(26)$$

$$B_{PR}(i) = \frac{B_{TOT} - B_{FR}(i)}{3}$$
$$= BW_X.$$
$$(27)$$

$$F_X(i) = B_{Tot}\left(\frac{R_{FR}(i-1)}{R}\right) - B_{Tot}\left(\frac{R_{FR}(i)}{R}\right),$$

where $R_{FR}(i)$: the radius of FR zone in current sub-frame (TTI) i,

$R_{FR}(i-1)$: the radius of FR zone in previous sub-frame (TTI) i
$$(28)$$

The constraint in (26) is reasonable because all users in the macrocell are uniformly distributed, and a larger area contains a higher number of users.

Table 3 User percentages for different traffic categories

Application	Traffic category	Percentage of users
FTP	Best Effort	10%
Web browsing/HTTP	Non-real time	20%
Video	Real time	20%
VoIP	Real time	30%
Gaming	Real time	20%

To allocate the dedicated FR bands, PR bands, and the BBW, the main steps are as follows:

Initially, the eNB will allocate the dedicated bands and BBW in the FR zone. As soon as the dedicated bands are assigned, available BBW is dynamically assigned to more demanding PR zones. Next, HeNBs (while allocating the dedicated BWs to FUEs) will start to allocate orthogonal frequency bands alternatively to MUEs in order to avoid co-channel interference. To avoid interference between the MUEs and FUEs, the orthogonal frequency bands are allocated between the eNB and HeNBs by using the orthogonal frequency band allocation algorithm proposed in [28].

The main steps involved in that algorithm to allocate the orthogonal resources for HeNBs are as follows: First, the set of usable frequency sub-bands J_U for the HeNB is initialized to the set of all available frequency sub-bands J in a macrocell. When HeNB is turned on, it senses the neighboring macrocell signals and estimates the received signal strength indication (RSSI) value (R_j) for each frequency sub-band. Let T denote the set of RSSI values for all available frequency sub-bands in the macrocell, and let R_j^* denote the highest RSSI value. If the RSSI value of sub-band A matches the highest value, then the HeNB is located in the FR zone. Otherwise, the HeNB is located in one of the PR zones. In this case, the HeNB will avoid using sub-band A and can use the other sub-bands B, C, and D of the PR zones. Finally, the same procedure is adopted by the HeNBs to avoid using the same frequency bands as the eNBs in the PR zones.

In every subframe to allocate the BBW, the eNB and the HeNB will check on the demand raised for BBW in PR zones X by MUEs and FUEs, respectively. If MUEs or FUEs in PR zone X demand the BBW, then the eNB or HeNB will check whether the requested BBW has been occupied by the UEs in the FR zone. If BBW is available, the eNB or HeNB will allocate it to PR zone X; otherwise, the eNB or HeNB will switch the FR zone users to other available frequency bands and will allocate the requested frequency bands to PR zone X. If no free frequency band is available in FR zone and there is still demand from PR zone users, the users will wait a random amount of time for free available BBW. If BBW is available, it will be allocated to users by comparing priorities and QoS requirements of users in PR zones by using Table 2. The flowcharts of the proposed eNB and HeNB BBW allocation procedures are shown in Figures 6 and 7, respectively.

The QoS-DFFR scheme is implemented independently in each cell to dynamically allocate the BBW based on the service priorities of the users. Therefore, the BBW allocation will not break the orthogonality of FFR-3 as it considers the orthogonal frequency-band allocation algorithm [28]. Similarly, BBW allocation in the center cell will not interfere with the cells in the first and second tiers because the cell is partitioned into FR and PR zones, as shown in Figure 5. This partitioning helps to avoid inter-cell interference. After BBW is allocated to one of the PR zones, the remaining BW in the FR zone is described as

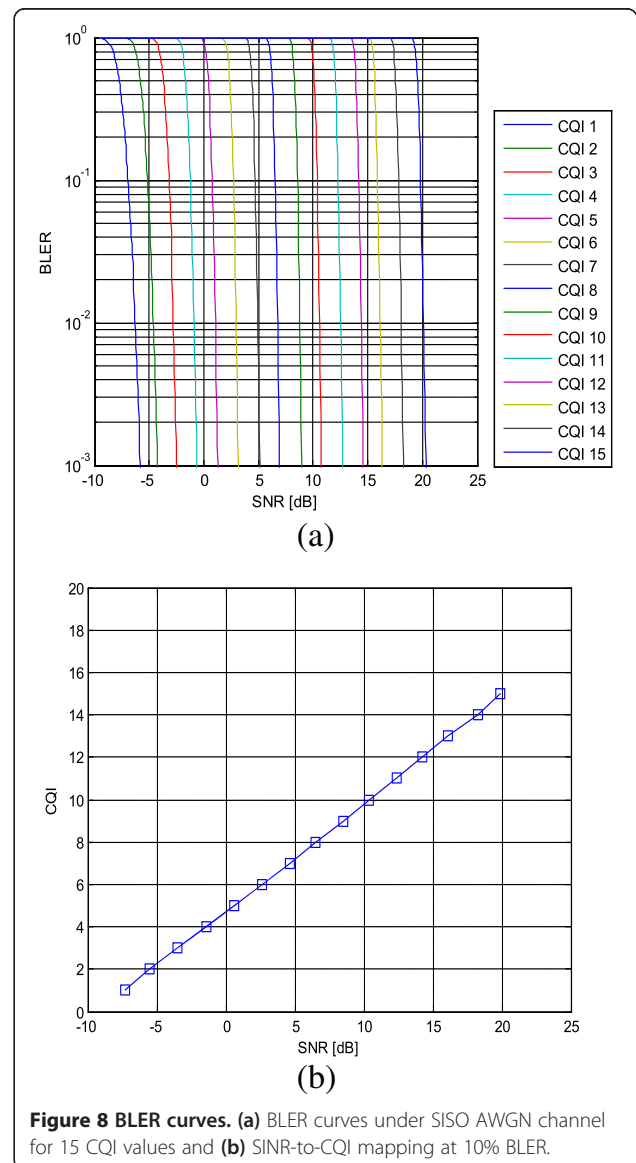$$B^X_{FR} = BW_A - F_X. \qquad (29)$$

Figure 8 BLER curves. (a) BLER curves under SISO AWGN channel for 15 CQI values and **(b)** SINR-to-CQI mapping at 10% BLER.

Therefore, after BBW allocation, the ratio of BW allocated to each of the FR and PR zones can be given as

$$\beta^{\lambda}_{\text{FR}}(\rho) = B^{\lambda}_{\text{FR}}(\rho)/B_{\text{TOT}}, \text{with} \quad \beta^{\lambda}_{\text{FR}}(\rho) \in [0,1], \quad (30)$$

$$\alpha_X{}^{\lambda}_{\text{PR}}(\rho) = \frac{1}{3}\left[1-\beta^{\lambda}_{\text{FR}}(\rho)\right] + \frac{F_X}{B_{\text{TOT}}}, \quad (31)$$
$$\text{with} \quad \alpha_B{}^{\lambda}_{\text{PR}}(\rho) \neq \alpha_C{}^{\lambda}_{\text{PR}}(\rho) = \alpha_D{}^{\lambda}_{\text{PR}}(\rho),$$

where $\beta^{\lambda}_{\text{FR}}(\rho)$ is the ratio of the BW allocated to the FR zone and $\alpha_X{}^{\lambda}_{\text{PR}}(\rho)$ is the ratio of the BW allocated to each of the PR zones X after BBW allocation. In (31), $\frac{F_X}{B_{\text{TOT}}}$ is the normalized BBW allocated to one of the PR zones X. As a zone changes its size, the resources that can be allocated to users also changes. Thus, to consider this effect, the normalized BW density is defined in (32) as a function of partitioning boundary ρ:
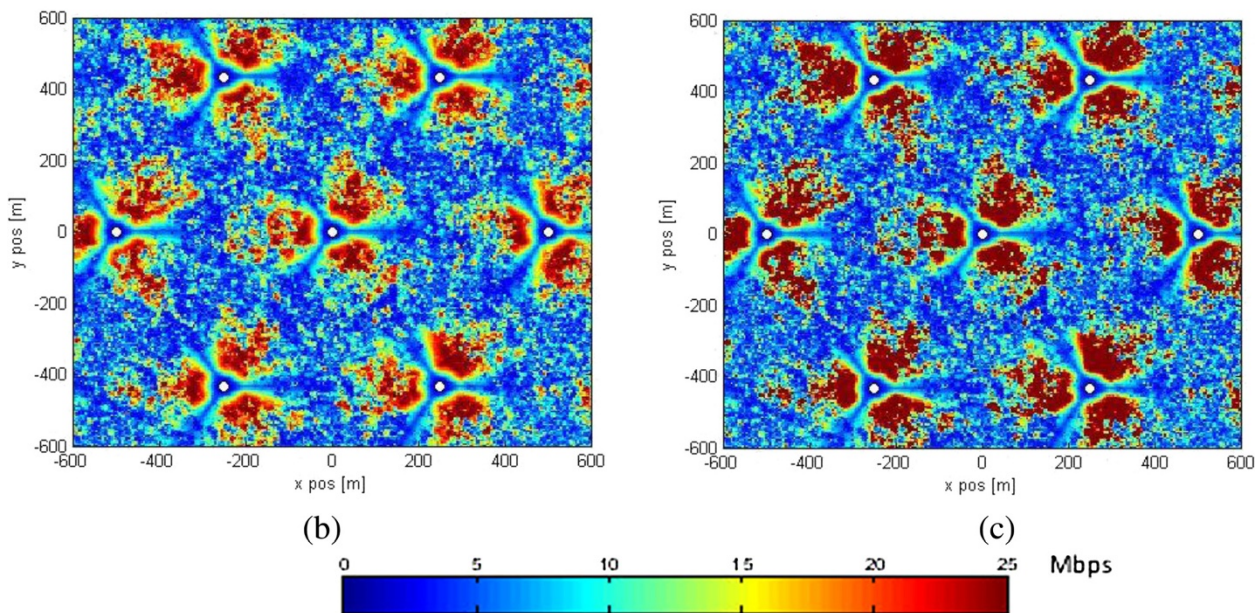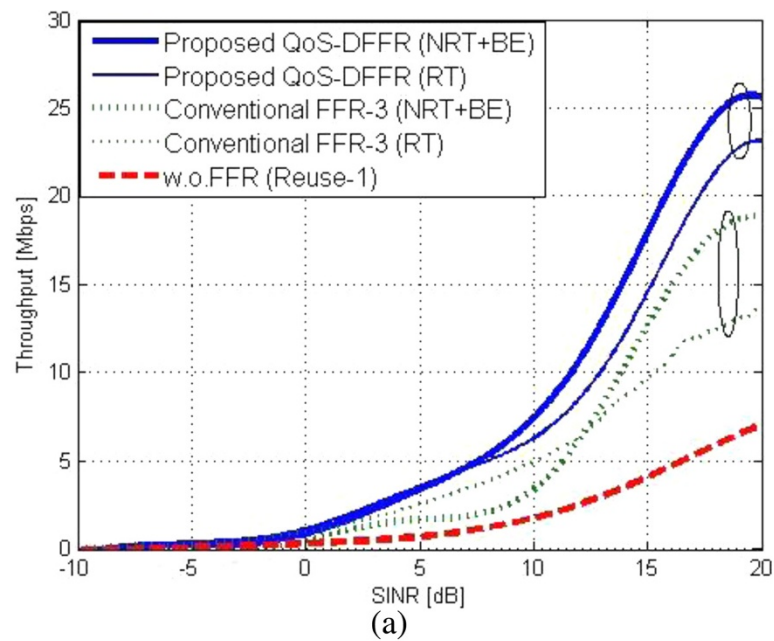

Figure 9 UE throughput performance. (a) SINR-to-throughput mapping curves, **(b)** UE throughput color map without BBW, and **(c)** UE throughput color map with BBW.

$$N_{\text{BW\_density}}(\rho) = \begin{cases} \dfrac{\beta^{\breve{x}}_{\text{FR}}(\rho)}{A_{\text{FR}}(\rho)} \\ \dfrac{\frac{1}{3}\left[1-\beta^{\breve{x}}_{\text{FR}}(\rho)\right]}{A_{\text{PR}}(\rho)} \end{cases}, \quad (32)$$

where $\beta_{\text{FR}}(\rho)$ denotes the normalized BW in the FR zone, while the other parameters are as described in (21).

Thus, in the proposed QoS-DFFR scheme, users-per-sector capacity is optimized by considering network loading conditions. Hence, users-per-sector capacity after allocating the BBW can be calculated for the FR and PR zones as

$$C^{\breve{x}}_{A(\text{FR}),\rho,\beta^{\breve{x}}_{\text{FR}}(\rho)}(x,y) = \frac{\beta^{\breve{x}}_{\text{FR}}(\rho)}{A_{\text{FR}}(\rho)} \log_2[1 + \text{SINR}_{\text{FR}}(x,y)], \quad (33)$$

$$C^{\breve{x}}_{X,\rho,\alpha^{\breve{x}}_{\text{PR}}(\rho)}(x,y) = \frac{\alpha_X{}^{\breve{x}}_{\text{PR}}(\rho)}{A_{\text{PR}}(\rho)} \log_2[1 + \text{SINR}_{\text{PR}}(x,y)], \quad (34)$$

where $C^{\breve{x}}_{A(\text{FR}),\rho,\beta^{\breve{x}}_{\text{FR}}}(x,y)$ and $C^{\breve{x}}_{X,\rho,\alpha^{\breve{x}}_{\text{PR}}(\rho)}(x,y)$ are the capacities in the FR and PR zones after BBW allocation. Hence, due to the allocation of BBW to the most demanding and high-priority PR zone, users-per-sector capacity is optimized.

4.3 Inter-zone scheduling
To efficiently allocate the frequency bands and BBW between users in each sector, the EXP/PF rule scheduler is employed. (The details of the scheduler were discussed in Section 3). These schedulers distribute the frequency bands by maintaining fairness, HOL delays, and priorities between users. Hence, the scheduler optimizes per-user throughput and results in an increase in system throughput.

Thus, the proposed QoS-DFFR scheme reduces co-channel interference and optimizes the sector, per-user, and system throughput by satisfying network loading conditions.

5 Performance evaluation
5.1 Simulation environment
To evaluate the performance of the proposed QoS-DFFR scheme, system-level simulations [29] are performed under the network scenario described in the section that reviewed the system model. The performance is evaluated with a 4 × 4 MIMO antenna configuration, employing a closed-loop spatial multiplexing (CLSM) mode. In addition, instead of full-queue case simulations, the realistic traffic models are considered to assess performance under typical traffic and network loading conditions.

Users with different QoS requirements are randomly distributed with applications according to the percentage

outlined in Table 3 [23]. The simulations are performed while considering the practical simulation parameters listed in Table 1.

5.2 Link-level abstraction
In LTE-A, the user measures the DL received signal quality by using the reference signals and reports to eNB the preferred modulation and coding schemes (MCS). This report is signaled using a CQI index. For LTE, 15 different MCSs are defined to keep the BLER below 10%.

To evaluate the BLER of the received TBs, a set of AWGN link performance curves [30] for a 20-MHz BW are considered, as shown in Figure 8a. Then, according to the proper BLER-SINR curve, i.e., depending on the applied MCS, it will be decided whether the packet has been correctly received or not. If the packet is not received correctly, then the packet is considered erroneous and is discarded. The SINR-to-BLER mapping requires an effective SINR, i.e., SINR_{eff}. SINR_{eff} is obtained by mapping the set
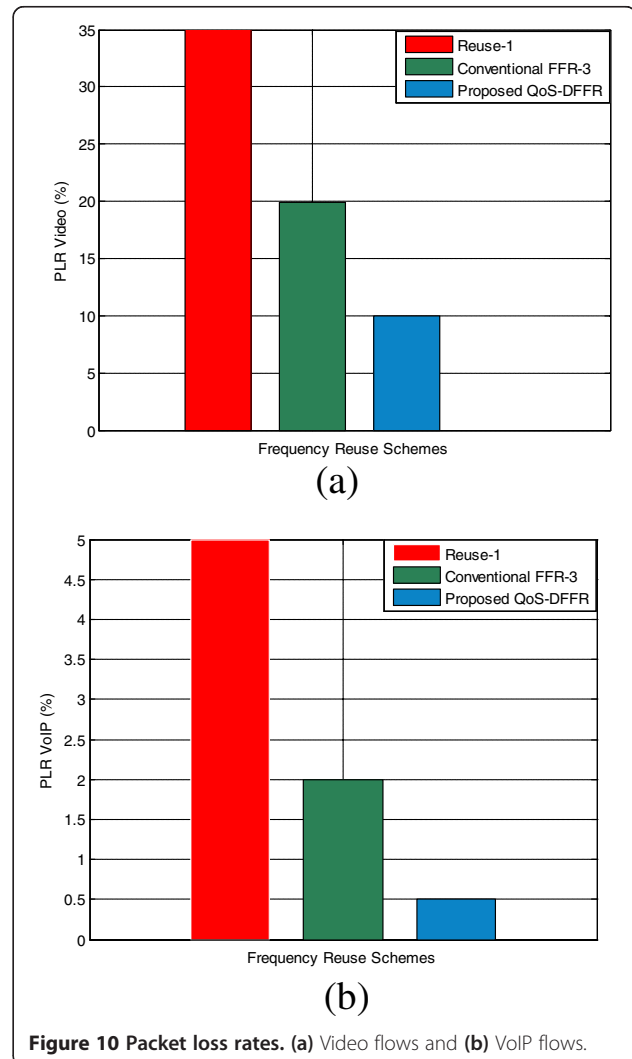

Figure 10 Packet loss rates. (a) Video flows and **(b)** VoIP flows.

of subcarriers SINR assigned to users TB to the AWGN-equivalent SINR. We use the MIESM [21] to do averaging, which is described as

$$\mathrm{SINR}_{\mathrm{eff}} = \alpha_1 \cdot I^{-1}\left(\frac{1}{N}\sum_{k=1}^{N} I\left(\frac{\mathrm{SINR}_k}{\alpha_2}\right)\right), \qquad (35)$$

where the parameters α_1 and α_2 adapt to different MCSs and are calibrated by means of link-level simulations to fit the compression function to the AWGN-BLER results. N is the number of samples and the function $I(.)$ is the mutual information.

5.3 Simulation results
We compare the performance of the conventional FFR schemes (i.e., reuse-1 and reuse-3) with the proposed QoS-DFFR scheme under the HetNet scenario in terms of the user's average throughput, packet loss rate, interference, and cell-edge and peak throughputs.

5.3.1 UE throughput enhancement by the QoS-based dynamic FFR scheme
Users with different QoS requirements are dropped in the simulation environment, and the comparison is carried out by considering three different frequency reuse schemes: reuse-1, reuse-3, and the proposed QoS-DFFR scheme. For all of the aforementioned frequency resource configurations, we compare the UE average throughput in Figure 9a at a SINR of 15 dB in which reuse-1 shows the worst performance owing to the full reuse of the entire

frequency band. This results in high co-channel interference between the users. The UE throughput increases for mixed traffic by using reuse-3 because the co-channel interference decreases by partitioning the frequency bands into FR and PR zones. Thus, the throughput reaches 9 Mbps for real-time traffic and 13 Mbps for non-real-time-plus-BE traffic. The real-time flows have low throughput peak values compared to non-real-time-plus-BE flows because real-time users are not always in an active state. However, using the proposed QoS-DFFR scheme, UE throughput increases and shows approximately 40% and 28% throughout improvement compared with the conventional FFR for real-time and non-real-time flows, respectively.

The reason for this improvement is the per-sector capacity optimization and user priority consideration for BBW allocation. The effect of allocating the BBW by the proposed scheme over the entire HetNet system can be clearly seen in the throughput color map in Figure 9. The map shows high-intensity red areas when the BBW is allocated in (c) as compared to (b). The increase in intensity of the red color indicates that the throughput of the users increased in that region.

5.3.2 Packet loss rate of real-time users
The packet loss rate (PLR) is calculated as the ratio of t packets discarded for not meeting the delay budget to the sum of packets arriving at the eNB station buffer. Thus, PLR is given as
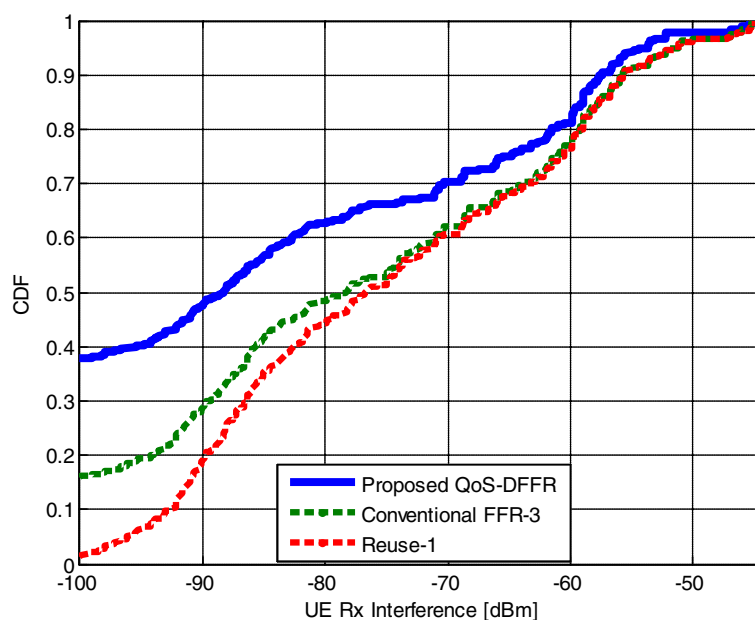

Figure 11 UE receive interference.

$$PLR = \frac{\sum\limits_{i=1}^{K}\sum\limits_{t=1}^{T}p_{\text{discard}_i}(t)}{\sum\limits_{i=1}^{K}\sum\limits_{t=1}^{T}p_i(t)}, \quad (36)$$

where p_{discard_i} is the discarded packets of user i and K is the total number of active users in the system. In particular, network performance of non-real-time and BE flows is evaluated by throughput, while for the real-time traffic (i.e., video and VoIP), the PLR is usually used as a metric. The PLR for video and VoIP flows is plotted using a bar chart for different frequency-reuse schemes, as shown in Figure 10. The PLR of video and VoIP flows shows that the proposed QoS-DFFR is able to provide the best QoS owing to less packet loss compared to the conventional schemes. There are two main reasons behind the reduction in PLR: the per-sector capacity optimization by dynamically assigning BBW, and the use of the EXP/PF rule scheduler, which gives priority to users with high HOL packet delay.

From Figure 10, it is clear that in the proposed QoS-DFFR scheme, the PLR for video and VoIP is reduced noticeably from the baseline configuration schemes by approximately 25% and 4.5%, respectively. The above discussion concludes that the proposed QoS-DFFR scheme is highly effective in interference-limited scenarios and in QoS-demanding conditions such as a LTE-A scenario.

5.3.3 DL interference reduction by the proposed QoS-DFFR scheme

In a HetNet environment, MUEs near HeNBs suffer from high interference because eNBs and HeNBs are operating on the same frequency bands. By using FFR schemes (i.e., frequency bands), partitioning the interference is reduced as compared with the reuse-1 scheme. The interference received by MUEs is compared at very low interference range (e.g., between −100 and −90 dBm) for the proposed QoS-DFFR and other conventional schemes.

The results in Figure 11 clearly demonstrate that approximately 50% of users using the proposed QoS-DFFR scheme lie in the specified range of low interference, while for standard schemes, only 20% to 30% of users are in this range. There are two main reasons behind the interference reduction using the proposed QoS-DFFR scheme: 1) FUEs use orthogonal frequency bands as compared to MUEs, and 2) the BBW allocation based on the service priority of users reduces contention in the frequency bands. Hence, by using the proposed QoS-DFFR scheme, the interference received by MUEs is reduced as compared with the conventional schemes.
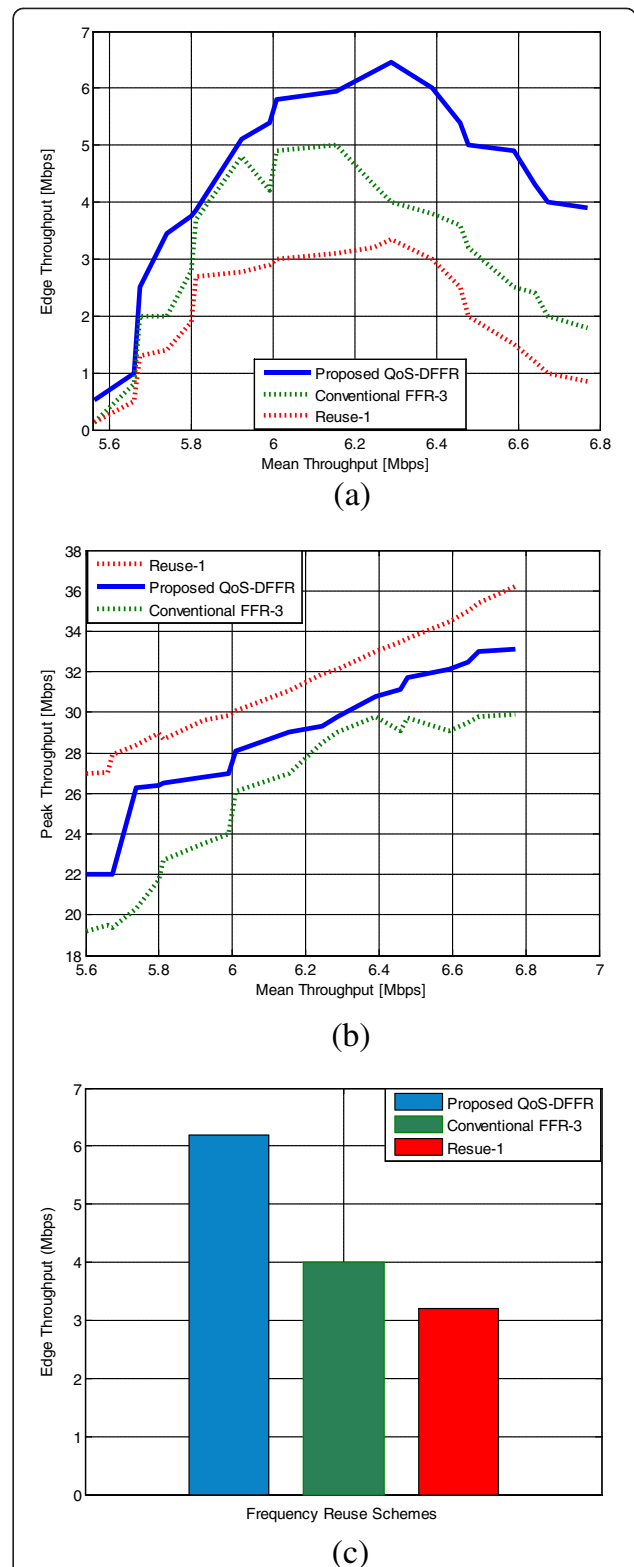

(a)

(b)

(c)

Figure 12 Throughput comparison. (a) Mean vs. edge throughput, **(b)** mean vs. peak throughput, and **(c)** edge throughput at the mean throughput of 6.3 Mbps.

5.3.4 Edge and peak throughput gains

Here, we discuss three key performance indicators (KPIs), which are typically derived from the throughput empirical cumulative distribution function (ecdf) related to FFR. The KPIs consist of mean throughput, edge throughput, and peak throughput. The terms 'edge' and 'peak' refer to the 5% and 95% points of the users' throughput ecdf, respectively. These can be interpreted as the performance of a user at the cell edge and cell center, respectively. Figure 12a shows that to maintain the same mean throughput (i.e., around 6.3 Mbps for users), the cell-edge throughput of the proposed QoS-DFFR scheme is best because it reaches approximately 6.2 Mbps of cell-edge throughput, while the other conventional schemes reach between 3 and 5 Mbps. Thus, the cell-edge throughput shows twice the improvement as the conventional schemes. The improvement in cell-edge throughput performance is more visible in Figure 12c, where bar charts are plotted at the fixed mean throughput value of 6.3 Mbps compared to the proposed QoS-DFFR scheme performance with the other schemes. In the case of peak throughput, the results are quite different from the case of cell-edge throughput. The peak performance is best for reuse-1 and worst for the conventional FFR-3, while the proposed scheme lies between the two schemes. The proposed QoS-DFFR scheme's peak throughput performance is better than that of the reuse-3 scheme because the former manages the frequency bands dynamically and more efficiently by considering QoS requirements and service priorities of users.

Degradation of peak throughput performance occurs in the reuse-3 case because the frequency bands are partitioned into FR and PR zones. Hence, there is a trade-off between the peak throughput performance degradation and interference reduction when using the reuse-1 and reuse-3 schemes.

Interference is a more serious problem in the current LTE-A system, which in turn makes this trade-off as a prominent criterion to efficiently operate LTE systems. Therefore, the proposed QoS-DFFR scheme is well suited to improve the cell-edge user's throughput performance, at the cost of a very small peak throughput degradation of the system. On the other hand, this peak throughput degradation is balanced by interference reduction achieved by partitioning the frequency resources between the FR and PR zones in the HetNet scenario.

6 Conclusions

FFR is an effective and straightforward scheme for interference management under the HetNet scenario. In this paper, a novel QoS-DFFR scheme was proposed for a two-tier HetNet scenario in LTE-A systems. The proposed QoS-DFFR scheme efficiently allocates the non-occupied FR zone frequency bands, i.e., BBW, to cell-edge users by considering QoS requirements and service priorities of users. Furthermore, the proposed scheme optimizes the capacity of each PR zone according to dynamic traffic load conditions by allocating BBW to the most demanding and high-priority PR zone. The proposed scheme was validated by system-level simulations. The simulation results showed a remarkable reduction in packet loss rate, i.e., a reduction of approximately 25% and 5% for video and VoIP services, respectively. Moreover, the throughput of cell-edge users almost doubled when compared to that of conventional static frequency allocation schemes such as reuse-1 and reuse-3. Hence, the proposed QoS-DFFR scheme exploits limited spectral resources more efficiently and thus reduces the blocking probability of users. This results in a reduction in packet loss rate and an increase in average throughput, cell-edge users' throughput, and system efficiency of the users in the network.

References
1. Qualcomm Document Center, LTE Advanced: heterogeneous networks, 2011. https://www.qualcomm.com/media/documents/files/lte-heterogeneous-networks.pdf. Accessed 20 Dec 2013
2. R Bendlin, V Chandrasekhar, R Chen, A Ekpenyong, E Onggosanus, From homogeneous to heterogeneous networks: a 3GPP Long Term Evolution rel. 8/9 case study, in *Proceedings of the 45th Annual Conference on Information Sciences and Systems* (Baltimore, MD, 2011), pp. 1–5
3. Qualcomm Document Center, 1000× data challenge: more small cells taking HetNets to the next level, http://www.qualcomm.com/1000x. Accessed 01 Jan 2014
4. Small Cell Forum Ltd, http://www.smallcellforum.org. Accessed 05 Jan 2014
5. ME Sahin, I Guvenc, MR Jeong, H Arslan, Handling CCI and ICI in OFDMA femtocell networks through frequency scheduling. IEEE Trans. Consum. Electron. **55**(4), 1936–1944 (2009)
6. H Li, X Xu, D Hu, X Qu, X Tao, P Zhang, Graph method based clustering strategy for femtocell interference management and spectrum efficiency improvement, in *Proceedings of the IEEE 6th International Conference on Wireless Communications Networking and Mobile Computing* (Chengdu, China, 2010), pp. 1–5
7. Y Yang, L Ling, P Zhao, W Wang, Cognitive adaptive power ratio updating algorithm in soft frequency reuse scheme, in *Proceedings of the IEEE 78th Vehicular Technology Conference* (Las Vegas, USA, 2013), pp. 1–5
8. S Park, W Seo, Y Kim, S Lim, D Hong, Beam subset selection strategy for interference reduction in two-tier femtocell networks. IEEE Trans. Wireless Commun. **9**(11), 3440–3449 (2010)
9. 3GPP, R1-106052: per cluster based opportunistic power control, in *3GPP TSG RAN WG1 Meeting 63* (Jacksonville, FL, USA, 2010)
10. W Pramudito, E Alsusa, Enhancing a hybrid resource management technique for energy and QoS optimization in fractional frequency reuse based cellular networks. IEEE Trans. Commun. **61**(12), 4948–4960 (2013)
11. L Zhang, L Yang, T Yang, Cognitive interference management for LTE-A femtocells with distributed carrier selection, in *Proceedings of the IEEE 72nd Vehicular Technology Conference* (Ottawa, Canada, 2010), pp. 1–5

12. M Rahman, H Yanikomeroglu, Enhancing cell-edge performance: a downlink dynamic interference avoidance scheme with inter-cell coordination. IEEE Trans. Wireless Commun. **9**(4), 1414–1425 (2010)

13. M Assaad, Optimal fractional frequency reuse (FFR) in multicellular OFDMA system, in *Proceedings of the IEEE 68th Vehicular Technology Conference* (Calgary, Canada, 2008), pp. 1–5

14. N Saquib, E Hossain, DI Kim, Fractional frequency reuse for interference management in LTE-advanced HetNets. IEEE Commun. Mag. **20**(2), 113–122 (2013)

15. M Taranetz, JC Ikuno, M Rupp, Capacity density optimization by fractional frequency partitioning, in *Proceedings of the 45th Annual Asilomar Conference on Signals, Systems, and Computers* (Pacific Grove, California, 2011), pp. 1398–1402

16. 3GPP, R4-092042: simulation assumptions and parameters for FDD HeNB RF requirements, in *3GPP TSG RAN WG4 Meeting 51* (San Francisco, CA, USA, 2009)

17. ETSI 3rd Generation Partnership Project (3GPP), *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) System Scenarios (release 8), 3GPP TR 36.942* (ETSI, Sophia Antipolis, 2009)

18. H Claussen, Efficient modeling of channel maps with correlated shadow fading in mobile radio systems, in *Proceedings of the IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications* (Berlin, Germany, 2005), pp. 512–516

19. L Hentila, P Kyosti, M Kaske, M Narandzic, M Alatossava, *MATLAB implementation of the WINNER phase II channel model ver1.1*, 2007. http://www.ist-winner.org/phase_2_model.html. Accessed 10 Jan 2014

20. ETSI 3rd Generation Partnership Project (3GPP), *Evolved Universal Terrestrial Radio Access (E-UTRA), Physical Layer Procedures (Release 8), 3GPP TS 36.213* (ETSI, Sophia Antipolis, 2008)

21. F Kaltenberger, I Latif, R Knopp, On scalability, robustness and accuracy of physical layer abstraction for large-scale system-level evaluations of LTE networks, in *Proceedings of the 47th Annual Asilomar Conference on Signals, Systems, and Computers* (Pacific Grove, USA, 2013), pp. 1644–1648

22. X He, K Niu, Z He, J Lin, Link layer abstraction in MIMO-OFDM system, in *Proceedings of the International Workshop on Cross Layer Design* (Jinan, China, 2007), pp. 41–44

23. 3GPP, R1-070674, LTE physical layer framework for performance verification, in *3GPP TSG-RAN Meeting 48* (St Louis, MO, USA, 2007)

24. ETSI 3rd Generation Partnership Project (3GPP), *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN), 3GPP TS 36.300* (ETSI, Sophia Antipolis, 2013)

25. ETSI 3rd Generation Partnership Project (3GPP), *Group Services and System Aspects, Policy and Charging Control Architecture, TS 23.203 V8.9.0* (ETSI, Sophia Antipolis, 2008)

26. JH Rhee, JM Holtzman, DK Kim, Scheduling of real/non-real time services: adaptive EXP/PF algorithm, in *Proceedings of the 57th IEEE Semiannual Vehicular Technology Conference* (Jeju, Korea, 2003), pp. 462–466

27. JG Choi, S Bahk, Cell-throughput analysis of the proportional fair scheduler in the single-cell environment. IEEE Trans. Veh. Technol. **56**(2), 766–778 (2007)

28. TH Kim, TJ Lee, Throughput enhancement of macro and femto networks by frequency reuse and pilot sensing, in *Proceedings of the IEEE Performance, Computing and Communications Conference* (Austin, TX, USA, 2008), pp. 390–394

29. C Mehlfuhrer, JC Ikuno, M Simko, S Schwarz, M Wrulich, M Rupp, The Vienna LTE simulators – enabling reproducibility in wireless communications research. EURASIP J. Adv. Signal. Process. **2011**, 29 (2011)

30. M Wrulich, JC Ikuno, D Bosanska, M Rupp, Simulating the long term evolution physical layer, in *Proceedings of the European Signal Processing Conference (EUSIPCO)* (Glasgow, Scotland, 2009), pp. 1472–1478