EURASIP Journal on
Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

QoS-aware resource management for LTE-Advanced relay-enhanced network

Jacek Góra[1,2]

Abstract

Relaying is one of the major innovative concepts proposed in the recent years for cellular radio communication systems. It is a perfect solution for dealing with the issue of high variability of performance in cellular networks. By coordinated deployment at cell-edge or in shadowed areas, relay nodes can extend network coverage and increase the low end-user performance. Considering the advantages, relaying is recently being included in the standards of the fourth generation systems such as the LTE-Advanced and the WiMAX. However, one major problem of relaying is still to be resolved. Specifically, there are no concrete concepts for quality-of-service provisioning for relayed transmissions. This paper investigates the case of packet delivery times over multi-hop links in relay-enhanced networks. The discussion is specifically based on relaying implementation in the LTE-Advanced system. The quality-of-service satisfaction and its fairness for the base station and relay-node-connected users are analyzed in the framework of the utility theory. For the purpose of this analysis, utility functions are proposed for real-time traffic with minimum data rate and/or maximum packet delivery time requirements. Furthermore, several optimization concepts are proposed for managing multi-hop transmissions in a quality-of-service aware manner. The included analysis based on the LTE-Advanced system level simulations shows that the proposed optimizations have the potential to improve the overall quality-of-service satisfaction in a relay-enhanced system.

Keywords: LTE-Advanced; Relaying; Radio resource management; Utility theory; Quality-of-service; Delay

1 Introduction

In the year 2008, the International Telecommunication Union, Radiocommunication Sector (ITU-R) issued the M.2134 report [1] specifying requirements for the next generation of radio communication systems, the so-called International Mobile Telecommunication-Advanced (IMT-A). Specification of those requirements started a still ongoing process of developing new solutions extending capabilities of the existing radio communication systems. The research and development process leads to the definition of two major fourth generation (4G) systems, i.e., the Long Term Evolution-Advanced (LTE-A) [2,3] and the Worldwide Interoperability for Microwave Access (WiMAX) Release 2 [4,5].

Both the LTE-A and the WiMAX include a similar set of techniques to meet the IMT-A requirements [5-7]. The most significant building blocks considered are the following:

- Advanced multi-antenna techniques (multiple-input multiple-output, MIMO) [8],
- Bandwidth extension in the form of, e.g., carrier aggregation (CA) [9],
- Heterogeneous networks (HetNets) [10], and
- Improved interference mitigation techniques including interference avoidance [11] and coordinated multi-point (CoMP) transmission schemes [12].

One of the novel techniques proposed for the two 4G systems is relaying (in the above listing classified as part of the HetNet concept). The baseline of this technology is the introduction of a new type of access points, relay nodes (RNs), capable of dynamic setting-up its own backhaul (BH) link connection over a common radio interface, i.e., the same that is used for serving users (e.g.,

Correspondence: jacek.gora@nsn.com
[1]Technology and Innovation Department, Nokia Networks, Pl. Gen. J. Bema 2, 50-265 Wroclaw, Poland
[2]Faculty of Electronics and Communication, Poznan University of Technology, Ul. Piotrowo 3A, 60-965 Poznań, Poland

the LTE-A). Donor of the BH link connectivity is a standalone access point (donor node (DN)), or in other more advanced configurations, already operating, RN (i.e., the donor RN (DRN)). The main benefits of relaying envisioned for cellular systems are the following:

- Accuracy in coverage provisioning - RNs, as low power nodes, can be deployed exactly at the location where network coverage is required (including indoor or strongly shadowed areas) [13,14]. RNs specifically can be also deployed in locations where wireline BH provision is not possible (e.g., in public transport vehicles [15]).
- Low cost - RNs are commonly envisioned as low power and simple devices, in addition to not requiring fixed wireline BH connection, this enables for network operator savings in both capacity and operational expenditures [16].
- Flexibility of use - possibility to provide rapid and/or short-term deployments of network infrastructure (e.g., for mass events or for disaster/network malfunction recovery) without earlier planning or investments [17].

The basic application scenario for relaying (as specified by the 3rd Generation Partnership Project (3GPP) forum [18]) is coverage extension. The coverage extension scenario assumes that the main purpose of RNs in a cellular network is to provide additional system coverage and to enhance connection quality for the macro cell-edge users or the users located in macro coverage holes. Illustration of the relaying coverage extension scenario is presented in Figure 1.

The simplest application of relaying for coverage extension is deployment of RNs at the edge of a macro-cell coverage [19,20]. In such a case, the RNs initiate their own cells providing improved connection conditions for the nearby users (access, AC, link connectivity) and establish BH link connection to the overlaying macro base station (BS). This is the so-called two-hop relaying topology [21] (see Figure 2).

In a more advanced implementation, RNs may also establish BH link connection to other RNs. The topology is then called the multi-hop relaying [22] (see Figure 2). In general case, the multi-hop relaying topology may have the structure of a tree or a mesh. From an implementation perspective, however, the tree topology is generally preferred [23].

On the current development stage, the relaying technique is considered to provide coverage extension on the basic accessibility level. The LTE-A system specification does not provide yet any dedicated mechanisms for explicit capacity enhancement or quality-of-service (QoS) management. This is often pointed out as the main shortcoming of the existing relaying solutions [21,24].

The most relevant problem of the relaying technique regarding QoS provisioning is the introduction of additional delays to the packet delivery time in the radio interface. The additional delays are related to the multihop transmission and the RN signal processing times at each hop. Furthermore, the basic LTE-A relaying implementation assumes time domain multiplexing (TDM) of the RN BH and AC links (i.e., the in-band operation scheme [3,22,25]), which even further increases the end-to-end transmission time.

This paper investigates the QoS provisioning problem for the RN-connected users. The main focus is put on the packet delivery time issue, but the requirement of minimum data rate is also considered. Firstly, in Section 2, the QoS-provisioning problem for relayed transmissions is formulated. In Section 3, the weight of the problem is analyzed in the context of the LTE-A standard specification of relaying. This includes analysis of the available relay node configurations with respect to the impact they have on the end-user perceived packet delivery times. Next, in Section 4, resource management schemes based on the utility theory are proposed. The purpose of the proposed schemes is to optimize the resource allocation in a relay-enhanced network (REN) so to improve the general QoS satisfaction for all users. Efficiency of the proposed solutions is verified via LTE-A system-level simulations described in Section 5 of this paper. Finally, the work is concluded in Section 6.

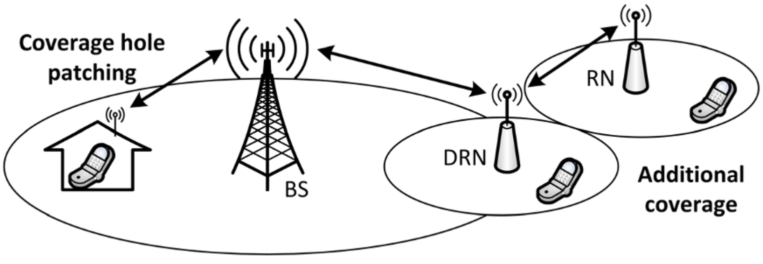

Figure 1 Relaying coverage extension scenario.

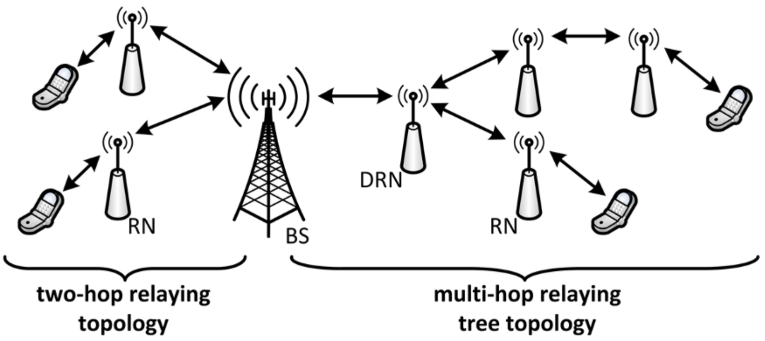

Figure 2 Two-hop and multi-hop relaying topologies.

2 Problem formulation

QoS provisioning for real-time traffic involves two elements [26]:

- Satisfaction of the minimum required data rate (i.e., the guaranteed bit-rate (GBR)), and
- Satisfaction of the maximum packet delivery time (i.e., the packet delay budget (PDB)).

Listing of the 3GPP standardized QoS classes with indication of the corresponding GBR and PDB requirements is depicted in Table 1.

To provide QoS satisfaction for a real-time traffic, both the GBR and PDB requirements need to be satisfied. If the GBR requirement is not met, it is not possible to guarantee the packet delivery times for all data packets. In such a case, the scheduling process in unstable [27], i.e., new data packets are created by the source node(s) faster than they are delivered to the target node(s).

On the other hand, even if the GBR requirement is satisfied, the PDB requirement cannot be assumed to be automatically satisfied [27]. It can be generally considered that the packet scheduling algorithm, if non-optimally implemented, may make some data packets wait in queue longer than allowed by the PDB requirement while instead scheduling for transmission packets with longer available time to drop.

Table 1 Standardized QoS classes [26]

QoS class	Priority	Bit rate requirement	PDB[a] (ms)	Packet error rate	Service example
1	2	GBR	100	10^{-2}	Live voice streaming
2	4		150	10^{-3}	Live video streaming
3	3		50	10^{-3}	Real-time gaming
4	5		300	10^{-6}	Buffered video streaming
5	1	Non-GBR	100	10^{-6}	IMS signalling
6	6		300	10^{-6}	Web traffic for privileged users
7	7		100	10^{-3}	Interactive gaming
8	8		300	10^{-6}	Web traffic for standard users
9	9				Elastic traffic

[a] Including on average 20 ms of delay in the core network.

In the case of relayed connections, the same rules apply. The task of the QoS satisfaction is, however, additionally complicated by the multi-hop nature of the transmission process.

First of all, in case of an H-hop connection, each data packed is send over H component radio links. Thus, it is H times queued in buffers, processed for link adaptation, and transmitted over the radio interface. At each transmission, hop errors might be introduced, which require additional retransmissions and may cause further delays in the packet delivery. Considering an H-hop connection, the end-to-end delivery time (t_{e2e}) for a data packet can be generally estimated as the sum of the number of times required to perform the three aforementioned operations at each of the H transmission hops, i.e., as follows:

$$t_{e2e} = \sum_{h=1}^{H} \left(t_{p,h} + t_{q,h} + t_{t,h} \right) \tag{1}$$

where $t_{p,h}$ is the packet processing time, $t_{q,h}$ is the packet queuing time, and $t_{t,h}$ is the packet transmission time over a single hop (including retransmissions, if any). Subscript h indicates the number of a hop in the H-hop connection chain.

Based on Equation (1), the first estimation of the packet delivery time over an H-hop connection is that it is on average H times higher than the time expected for a single-hop connection. However, this simple estimation is not true as transmissions on the component links of a relayed connection are not independent.

The end-to-end packet transmission times over multihop links are also impacted by the RN buffer capacity and its fill level at each of the relaying hops. This might not be critical nor even noticeable if the system load is low (i.e., RN buffers are never fully loaded). However, as indicated in the work of Vitiello et al. [28], if the system load is high, RN buffers might get congested and become bottlenecks for multi-hop transmissions.

As shown in the author's earlier work [29], the impact of limited capacity of the RN buffers is especially noticeable if capacities of the RN BH and AC links are improperly balanced, e.g., as a result of sub-optimal resource allocation. If capacity of the RN BH link is lower than capacity of the RN AC link, data packets transmitted downlink to the user will get congested at the RN's donor node buffer. If the capacity of the RN BH link is higher than the capacity of the RN AC link, data packets transmitted downlink to the user will get congested at the RN buffer. For uplink transmissions, inverse process takes place.

Overall, it can be stated that the QoS-aware resource management for a multi-hop connection has to consider parameters of all its component links and involved RNs to secure the end-to-end QoS satisfaction. Such operation is beyond the existing LTE-A relaying specification that considers an RN as an access point with an autonomous resource management and packet scheduling functionalities (i.e., the layer-3 RN model [3,22]). This paper describes a multi-node resource management and scheduling scheme that can potentially solve this problem.

Before designing a resource management and scheduling procedure for relaying, it is curtail to get a full understanding of the RN configuration schemes supported by the LTE-A system. Therefore, in the next section of this paper, it is analyzed up to what extent the LTE-A relaying configurations are able to satisfy the QoS requirements of real-time services and what impact they have on the packet delay budget.

3 LTE-A relaying implementation

The basic relaying mode of operation considered in the LTE-A system standard and in most of the other implementations is the decode-and-forward (DF) approach [21]. In case of the DF relaying, a delay of at least one radio sub-frame is introduced at each RN in the multi-hop connection. The delay relates to the DF signal processing time, i.e., decoding of the signals received on the feeder link (i.e., RN BH in case of downlink transmissions) and encoding them again for transmission on the outgoing sink link (i.e., RN AC in case of downlink transmissions).

As the result of the DF processing, the transmissions taking place on the RN BH and AC links are not correlated. Therefore, the transmissions outgoing from a RN generate interference at the RN feeder link receiver [30]. To avoid the RN self-interference, two options are available [3]:

- Separation of the RN BH and AC transmissions by either allocation of orthogonal radio resources (e.g., frequency carriers as in the out-band relaying [3,22]) or by separation of the RN AC and BH antennas (e.g., by usage of directional antennas or antenna displacement).
- Time domain multiplexing of the RN BH and AC transmissions so that they are not active at the same time (i.e., the in-band relaying [3,22]).

Relaying implementations with the two self-interference avoidance options are analyzed next. The purpose of the analysis is to define the lower bounds of the end-to-end transmission delays ($t_{\mathrm{e2e}}^{\mathrm{LB}}$) for multi-hop connections involving RNs of either of the two configurations described above. The delay lower bounds are defined as the transmission times in an unloaded system, i.e., without the queuing times considered, i.e.,

$$t_{\mathrm{e2e}}^{\mathrm{LB}} = \inf(t_{e2e}) = \sum_{h=1}^{H}\left(t_{p,h} + t_{t,h}\right) \tag{2}$$

where inf(.) is the infinum function.

3.1 Full-duplex relaying

If sufficient separation is provided to the RN BH and AC links, the two links can be operated simultaneously (see Figure 3a), i.e., the RN can receive transmissions on the feeder link at the same time as it transmits on the sink link (full-duplex (FD) operation). In such case, the RN can forward data to the target node as soon as it receives and processes the transmission from the source node.

Based on the above characteristic, Equation (2) can be reformulated for the full-duplex relaying as follows:

$$t_{\mathrm{e2e}}^{\mathrm{LB-FD}} = \max_{h=1..H}\left(t_{t,h}\right) + \sum_{h=1}^{H} t_{p,h} \tag{3}$$

For a user data payload of size S, this is

$$t_{\mathrm{e2e}}^{\mathrm{LB-FD}} = \max_{h=1..H}\left(\left\lceil \frac{S}{y_h} \right\rceil\right) + \sum_{h=1}^{H} t_{p,h} \tag{4}$$

where y_h is the data rate achieved by the user's transmission on the component link h, and $\lceil . \rceil$ is the ceiling rounding function. Equation (4) can be next simplified as the following:

$$t_{\mathrm{e2e}}^{\mathrm{LB-FD}} = \left\lceil \frac{S}{\min\limits_{h=1..H}\left(y_h\right)} \right\rceil + \sum_{h=1}^{H} t_{p,h} \tag{5}$$

The minimum of the transmission data rates on the component links is, due to the 'bottleneck' mechanism,

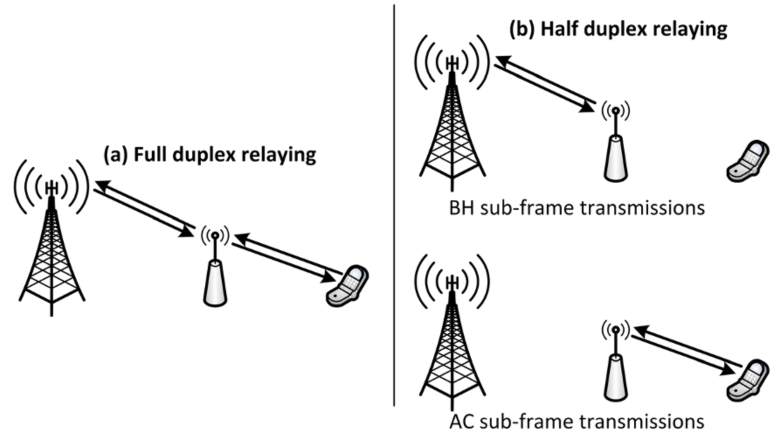

Figure 3 **Full-duplex (a) and half-duplex (b) relaying operation.**

the effective end-to-end data rate of the multi-hop connection (y_{e2e}), thus

$$t_{\mathrm{e2e}}^{\mathrm{LB-FD}} = \left\lceil \frac{S}{y_{e2e}} \right\rceil + \sum_{h=1}^{H} t_{p,h} \tag{6}$$

which can be finally reformulated as

$$t_{\mathrm{e2e}}^{\mathrm{LB-FD}} = \left\lceil \frac{S}{y_{\mathrm{e2e}}} \right\rceil + t_p^{\mathrm{BS}} + (H-1) t_p^{\mathrm{RS}} \tag{7}$$

where t_p^{RN} and t_p^{BS} are the packet processing times at RNs and BSs, respectively.

Compared to a single-hop transmission of the same end-to-end data rate (y_{e2e}), the H-hop FD relaying transmission time is longer by at least the (H −1) t_p^{RN} element. Considering that the RN processing delay t_p^{RN} should be in the range of 1 up to few milliseconds, the additional delay of a full-duplex multi-hop connection should not be critical considering the PDB requirements of most service types (see Table 1). However, in loaded systems, the queuing delay will be non-zero at some or all transmission hops. This delay, if not properly handled, can make the significant difference in the end-to-end packet delivery time. Therefore, a QoS-aware packet scheduling is so crucial for relaying. Proposal of such a scheduling algorithm is given in Section 4 of this paper.

3.2 Half-duplex relaying

In case sufficient separation of the RN BH and AC is not provided, the two links should be time domain multiplexed (TDM, see Figure 3b). The TDM-based resource partitioning implies that a RN operates in a half-duplex (HD) mode, i.e., the RN does not transmit and receive at the same time per transmission direction (downlink and uplink).

The TDM of RN BH and AC links provides protection from the RN self-interference without the need of either advanced hardware solutions or additional frequency resources as in case of the FD relaying. Considering the advantages, the TDM-based RN mode of operation is the main one currently considered in the LTE-A standardization [3,18].

The LTE-A system standard implements the RN TDM by reusing the LTE Release-8 multimedia broadcast over single frequency network (MBSFN) mechanism [31]. It defines that certain time sub-frames (duration of one LTE-A sub-frame is 1 ms) may be assigned for RN BH operation while the remaining sub-frames are used for the RN AC communication with users. Due to backward compatibility reasons, selection of the BH-enabled sub-frames is not fully flexible. The restrictions in the sub-frame configuration and the TDM itself generate additional delays on top of the delays already identified for the FD relaying (see Section 3.1).

The allowed BH MBSFN sub-frame patterns can be characterized with the following statistics that directly impact the performance of the HD relaying:

- Number of the BH sub-frames (K_{Bh}). This statistic determines the capacity of the RN BH link. It also controls the RN BH-AC capacity balancing for avoiding transmission bottlenecks [29]. The RN BH operation time share (σ) related to the MBSFN configuration is as follows:

$$\sigma = \frac{K_{\mathrm{Bh}}}{K} \tag{8}$$

where K = 40 sub-frames is the period of the MBSFN configuration. According to the current LTE-A specification, the BH operation share σ can be controlled in range 0% to 60% with a 7.5% resolution. With respect to the packet transmission time, the number of BH sub-frames also determines the expected waiting time between two consecutive sub-frames supporting BH transmission (t_{Bh2Bh}). This relation can be expressed with the following formula:

$$\mathrm{E}(t_{\mathrm{Bh2Bh}}) = \frac{1}{\sigma} \tag{9}$$

Similarly, the average waiting time between two consecutive sub-frames supporting AC transmission (t_{Ac2Ac}) can be expressed as follows:

$$\mathrm{E}(t_{\mathrm{Ac2Ac}}) = \frac{1}{1-\sigma} \tag{10}$$

- Concentration of the BH sub-frames. The less concentrated are the BH sub-frames (e.g., if they are evenly distributed in the 40 ms period), the lower is the expected time until the first available BH transmission event (t_{1Bh}) occurs. For example, the minimal delay that is applied to a data packet that becomes available for BH transmission in sub-frame k is as follows:

$$t_{\mathrm{1Bh}}(k) = \underset{\{k_{\mathrm{Bh}}\}}{\mathrm{MIN}}((k_{\mathrm{Bh}}-k) \bmod K) \tag{11}$$

where $\{k_{\mathrm{Bh}}\}$ is the set of indexes of the BH-assigned sub-frames, and mod is the modulo operation. The expected value of the t_{1Bh} delay is in such case

$$\mathrm{E}(t_{\mathrm{1Bh}}) = \frac{1}{K}\sum_{k=1}^{K} \underset{\{k_{\mathrm{Bh}}\}}{\mathrm{MIN}}((k_{\mathrm{Bh}}-k) \bmod K) \tag{12}$$

By analogy, the data packet received by an RN from BH link in sub-frame k_{Bh} experiences delay (t_{Bh2Ac}) of waiting until AC sub-frame of at least

$$t_{\mathrm{Bh2Ac}}(k_{\mathrm{Bh}}) = \underset{k\notin\{k_{\mathrm{Bh}}\}}{\mathrm{MIN}}((k-k_{\mathrm{Bh}}) \bmod K) \tag{13}$$

and the expected value of this delay is

$$\mathrm{E}(t_{\mathrm{Bh2Ac}}) = \frac{1}{\sigma K}\sum_{k_{\mathrm{Bh}}} \underset{k\notin\{k_{\mathrm{Bh}}\}}{\mathrm{MIN}}((k-k_{\mathrm{Bh}}) \bmod K) \tag{14}$$

All the delays defined with the above Equations (9) to (14) are expressed in terms of radio interface sub-frames. In the case of the LTE-A system, this is equivalent to milliseconds.

Considering the characteristic times of the MBSFN sub-frame configurations, it is possible to estimate the additional transmission delay related to the TDM of HD RNs. The additional delay results from the temporary unavailability of a specific RN link type at desired transmission time. The expected value of the TDM delay for two-hop relaying is as follows:

$$\begin{aligned}\mathrm{E}(t_{\mathrm{TDM}}) &= \mathrm{E}(t_{\mathrm{1Bh}}) + \mathrm{E}(t_{\mathrm{Bh2Ac}}) + \\ &+ \mathrm{MAX}\left((L-1)\mathrm{E}(t_{\mathrm{Bh2Bh}}), \left(L\tfrac{1-\sigma}{\sigma}-1\right)\mathrm{E}(t_{\mathrm{Ac2Ac}})\right)\end{aligned} \tag{15}$$

where L is the expected number of sub-frames required to transmit the data packet over the RN BH link. L can be considered as the 1-ms-resolution ceiling rounding of the time it takes to transmit the user data payload S with the BH link data rate y_{Bh}, i.e.,

$$L = \lceil S/y_{\mathrm{Bh}} \rceil \tag{16}$$

To calculate the lower bound of the HD relaying, the TDM transmission delay formulated in Equation (15) should be added on top of the FD-relaying transmission time formulated in Equation (8), i.e.,

$$\mathrm{E}\left(t_{\mathrm{e2e}}^{\mathrm{LB-HD}}\right) = \left\lceil \frac{S}{y_{\mathrm{e2e}}} \right\rceil + t_p^{\mathrm{BS}} + (H-1)t_p^{\mathrm{RS}} + \mathrm{E}(t_{\mathrm{TDM}}) \tag{17}$$

Comparison of the TDM delay overheads for various MBSFN configurations and L values is depicted in Figure 4. The TDM delay overhead is especially high in case of low σ configurations. This relates to high waiting times between the packet generation event and the first available BH sub-frame. Significance of this overhead is, however, decreasing with the expected number of active transmission sub-frames L (i.e., with bigger data packets and/or lower data rates). This is because the t_{1Bh} and t_{Bh2Ac} times are only applicable at the initial stage of a packet transmission.

In case of more than two-hop connection, the estimation of the TDM-related delay is less straightforward. This is because relations between the MBSFN configurations applied for two consecutive RN BH links need to be considered. The summary of expected values for the TDM-related delay for a tree-hop relaying is depicted in Figure 5. The depicted distributions are based on a statistical analysis of MBSFN sub-frame sequences with the assumption that all possible MBSFN configurations are equal probable.

Let us define an HD-relaying overhead as the ratio between the transmission time via TDM relays and the transmission time of the same data payload over a single-hop link. The collected data indicates that in the case of a three-hop HD relaying the TDM delay overhead is typically (median) at the level of 10× and in extreme cases can reach up to 75×. Again, the highest TDM delay overheads correspond to the low σ MBSFN configurations. This means that the delay in transmission over a three-hop HD-relaying connection is typically ten times higher than the transmission time over a single-hop link, but it can be also many times higher.

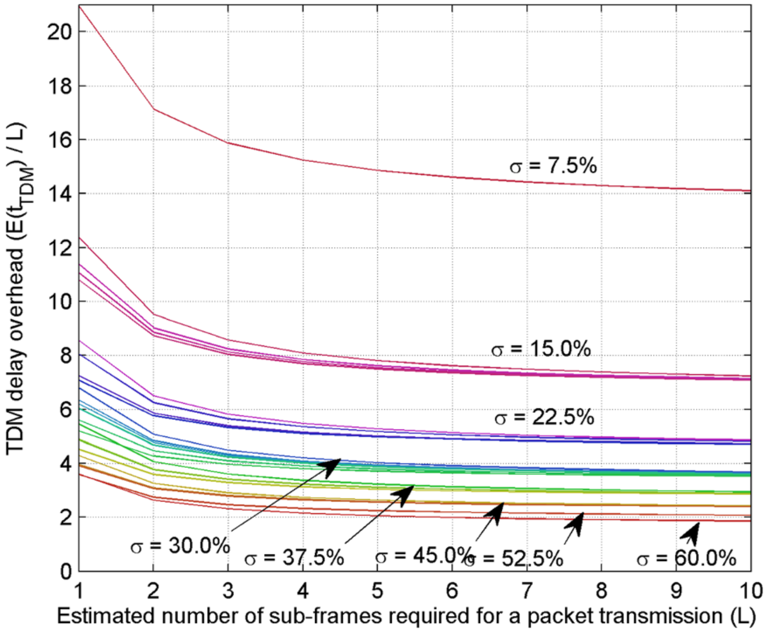

Figure 4 Expected TDM delay overhead for a two-hop relayed link for various MBSFN configurations.

Considering the typical values for the PDB requirement (see Table 1), it should be stated that the MBSFN configurations with low *σ* are unable to satisfy the maximum delay requirements. High *σ*, configurations provide significantly higher probability of the PDB requirement satisfaction, however, still do not guarantee it. Basing on the presented estimations, it is recommended that the HD relaying should be used only in two-hop topologies, while the multi-hop topologies should be based on the FD relaying only (e.g., out-band [3,22]).

The lower bound delay values given above for both the FD and HD relay configurations related just to the nature of multi-hop transmission and the available RN configurations. In a realistic case of a network handling simultaneous transmissions of multiple users, the queuing delays will need to be added on top of the delays estimated in this section. Those delays, however, depend on the packet scheduling algorithm implemented in the BS and in the RNs. The aspect of proper design of this algorithm is treated in the following section.

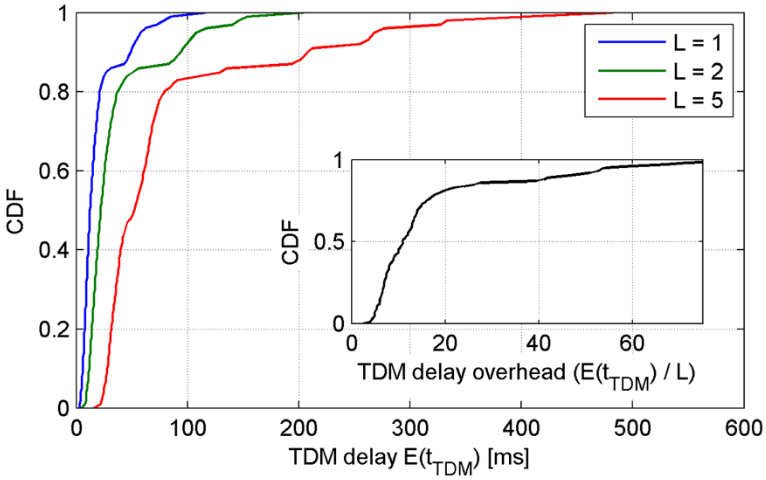

Figure 5 Expected TDM delay distribution over three-hop relayed link in relation to estimated number of effective transmission sub-frames (L).

4 QoS-aware resource management

In modern telecommunication networks, various service types may be used. This is a challenge for resource management functionalities as the services may be characterized with diverse QoS requirements. A feature of a good resource management algorithm should be to adapt its behavior the QoS requirement mixture present in the network at a given time. Examples of such algorithms for relay-less networks are described in [27,32,33].

One approach to the QoS-aware resource management problem is defined on the basis of the utility theory. The utility theory is a theory of economy that deals with wealth redistribution and trade. In the context of resource management, it is often considered as a method for optimizing resource allocation with respect to the user-perceived performance fairness. As such, it has been studied *inter alia* by Fishburn [34,35], Kelly [36,37], and Lan [38-40].

In this section, firstly, the resource management framework based on the utility theory is described with the generic equations derived describing the procedure (Section 4.1). Secondly, the framework is adapted for the specific case of multi-hop transmissions (Section 4.2). Finally, proposals of generic utility functions for various QoS-bounded traffic types are proposed (Section 4.3), and the overall resource management and scheduling metrics are derived for the traffic types (Section 4.4). The formulas describing the proposed resource management and scheduling procedure are defined here in a generic manner. Depending on specific QoS requirements (GBR and delay) of a traffic type, one of the utility functions proposed in Section 4.3 can be used to calculate appropriate resource allocation and scheduling metrics for the transmission.

4.1 Principles of utility theory

The system optimization with accordance to the utility theory is based on case-specific utility functions. The utility functions are a quantitative description of one's preferences and satisfaction. In the field of telecommunications, the preferences may reflect the QoS requirements of a service. In such case, the utility functions represent the objective or subjective level of the user's experience of using the network.

With respect to the above definition, each user j active in a network can be characterized with a certain utility u_j related to its achieved performance. The achieved user performance is the outcome of the applied resource allocation scheme x. In such case, the utility of the system is as follows:

$$U_{\mathrm{Sys}}(x) = \sum_{j\in J} u_j(x) \tag{18}$$

where J is the set of active users, and U_{Sys} is the utility of the whole system.

Target of the resource management optimization with respect to the utility theory is to find the resource allocation scheme $\dot{x}$ that maximizes the cumulated system utility, i.e.,

$$\dot{x} = \underset{x\in X}{\arg\max}\left(U_{\mathrm{Sys}}(x)\right) \tag{19}$$

where X is the set of all possible resource allocation schemes.

The optimization problem (19) can be solved, e.g., by means of the Lagrange multipliers method. The problem (19) is bounded with respect to the maximum resource availability

$$x = \left\{x_{j,r}\right\} \Leftrightarrow \underset{r\in R}{\forall} \sum_{j\in J} x_{j,r} \leq 1, \tag{20}$$

i.e., where the resource allocation scheme x can be defined as a set of factors x_j,r, each denoting allocation of resource element r (r $\in R$) to user j. When considering an instantaneous resource allocation, $x_{j,r}$ takes {0,1} values for all r and j.

The resource allocation $x = \{x_{j,r}\}$ has a direct impact on the transmission data rates of users by granting them access to channel capacities per resource element

$$y_j = \sum_{r\in R} C_{j,r} x_{j,r} \tag{21}$$

where $C_{j,r}$ is capacity of the radio link of the user j on the resource element r.

Considering Equations (20) and (21), the Lagrange function for the problem (19) is as follows:

$$\begin{aligned} &\mathscr{L}(y,x,z,\lambda,\mu) = \\ &= \sum_{j\in J}\left(u_j\left(y_j\right) - \lambda_j\left(y_j - \sum_{r\in R} C_{j,r}x_{j,r}\right)\right) + \\ &+ \sum_{r\in R}\left(\mu_r\left(1 - z_r - \sum_{j\in J} x_{j,r}\right)\right) \end{aligned} \tag{22}$$

where λ_j and μ_r are the Lagrange multipliers, and z_r is a factor balancing inequality in Equation (20).

Solution to the problem (19) is a stationary point of the Lagrange function (22). To find it, partial derivatives of the Lagrange function (22) are calculated as follows:

$$\frac{\partial\mathscr{L}}{\partial y_j} = \frac{\partial u_j\left(y_j\right)}{\partial y_j} - \lambda_j \tag{23}$$

$$\frac{\partial\mathscr{L}}{\partial x_{j,r}} = \lambda_j C_{j,r} - \mu_r \tag{24}$$

$$\frac{\partial \mathscr{L}}{\partial z_r} = -\mu_r \tag{25}$$

$$\frac{\partial \mathscr{L}}{\partial \lambda_j} = y_j - \sum_{r \in R} C_{j,r} x_{j,r} \tag{26}$$

$$\frac{\partial \mathscr{L}}{\partial \mu_r} = z_r - 1 + \sum_{j \in J} x_{j,r} \tag{27}$$

and they are compared to zero, thus

$$\begin{cases} \lambda_j = \dfrac{\partial u_j\left(y_j\right)}{\partial y_j} & \text{if } y_j > 0 \\ \lambda_j \geq \dfrac{\partial u_j\left(y_j\right)}{\partial y_j} & \text{if } y_j = 0 \end{cases} \tag{28}$$

$$\begin{cases} \mu_r = \lambda_j C_{j,r} & \text{if } x_{j,r} > 0 \\ \mu_r \geq \lambda_j C_{j,r} & \text{if } x_{j,r} = 0 \end{cases} \tag{29}$$

$$\begin{cases} \mu_r = 0 & \text{if } \mathrm{z}_r > 0 \\ \mu_r \geq 0 & \text{if } \mathrm{z}_r = 0 \end{cases} \tag{30}$$

Based on the above equations, the following interpretation of the factors λ_j, μ_r, and z_r can be stated:

- λ_j is a marginal cost of utility, i.e., the price of changing value of the user's *j* utility function.
- μ_r is the systems cost of using resource element *r*.
- z_r indicates if the resource element *r* is available for assignment (i.e., is not assigned to any of the users).

The above set of equations cannot be solved directly without using concrete utility functions for the users. A generic solution is to use an iterative resource allocation with priority metric M_{jr} derived by combining Equations (27) and (28)

$$M_{j,r} = C_{j,r}\lambda_j = C_{jr}\frac{\partial u_j\left(y_j\right)}{\partial y_j} \tag{31}$$

The priority metric indicates what increase of the user's (and system's) utility (aka the marginal utility [41]) can be expected when allocating resource element *r* to user *j*.

Of course, the iterative implementation of the problem is sub-optimal. It is convergent to the optimal state, however, the resource allocation decisions are done in each iteration based on the past state of the system (λ_j) and estimates of the future state of the radio links ($C_{j,r}$). The optimality of the iterative implementation is as good as the knowledge of the conditions that will occur when the scheduled transmission will be executed. On the other hand, advantage of the iterative solution is the possibility of its direct implementation in a real system. The system can calculate the priority metric for each user and resource and use it to assign resources for the next transmission time interval.

The above derivation corresponds to the optimization approach focused on the maximization of the system utility, i.e., the best effort (BE) approach. The approach allocates resources always to the user that can provide the highest marginal utility for the system. As a result, users in poor radio conditions (i.e., with low C_{jr}, e.g., at cell-edge) may never be granted resources. From an individual user perspective, such variation in the achievable performance indicates low reliability of the transmission quality and is typically inacceptable.

When performance fairness in the network is more desired than the total system performance, the so-called α-fairness utility can be used. The α-fairness utility is a utility recalculation function that corresponds to a certain fairness parameter α. The α-fairness function is defined as [42] follows:

$$u_j^{\alpha} = \begin{cases} \frac{\left(u_j\right)^{\alpha-1}}{\alpha-1} & \text{for } \alpha \neq 1 \\ \ln\left(u_j\right) & \text{for } \alpha = 1 \end{cases} \tag{32}$$

With α =1, the α-fairness optimal solution corresponds to the traditional proportional fair (PF) resource allocation satisfying the Nash's definition of fairness [43].

With respect to the α-fairness utility, the resource allocation priority metric can be redefined as the following:

$$M_{jr}^{\alpha} = C_{jr}\lambda_j^{\alpha} = C_{jr}\left(u_j\right)^{\alpha-2}\frac{\partial u_j\left(y_j\right)}{\partial y_j} \tag{33}$$

where λ_j^{α} is the α-fair marginal cost of utility for the user *j*.

Further in this paper, only the PF approach (α =1) is considered as the one providing significantly higher ubiquity of performance compared to the BE approach.

4.2 Utility-based resource management for relaying

Systems enhanced with RNs can use a resource management approach similar to the one described in the previous section. The utility-theory-based description of the optimization process can be extended over multi-hop connections, but additional constraints need to be introduced. The additional constraints correspond to the interdependencies of the consecutive relay links in a multi-hop connection. Specifically, the additional constraints are [29] the following:

- The data rate on the RN BH should be equal to the cumulated data rates on the RN AC links to users and to subordinate RNs

$$y_n^{\mathrm{BH}} = \left(\sum_{j \in J_n} y_{j,n}^{\mathrm{AC}}\right) + \left(\sum_{k \in N_n} y_{k,n}^{\mathrm{BH}}\right) \tag{34}$$

where $y_n{}^{\mathrm{BH}}$ is the BH data rate of the RN n, $y_{j,n}{}^{\mathrm{AC}}$ is the data rate of user j on the AC of the RN n, $y_{k,n}{}^{\mathrm{BH}}$ is the BH data rate of a subordinate RN k connected to the RN n, and J_n and N_n are the sets of users and RNs connected to the RN n, respectively.

- In case of HD RNs and FD RNs with frequency domain resource partitioning, the same resources cannot be assigned to the BH and AC of an RN

$$x_{n,r}^{\mathrm{BH}}x_{n,r}^{\mathrm{AC}} = 0, \forall r \in R, n \in N \tag{35}$$

where $x_{n,r}{}^{\mathrm{BH}}$ and $x_{n,r}{}^{\mathrm{AC}}$ are the resource element r allocation factors to BH and AC of the RN n, respectively, and R is the full set of system resources.

In the LTE-A standardization, the DF RNs are equipped with similar resource management capabilities as traditional BSs. This enables implementation of the distributed resource management schemes (see Figure 6). In this approach, each access point (BS or RN) decides about the configuration of only the links it directly supports. Specifically, a BS controls resource allocation only to the direct links to users and first-hop RN BH. By analogy, a RN controls AC communication with the users connected to it and BH links of the next-hop RNs, if any. The distributed management scheme is based on local system status information, thus, in some cases may make sub-optimal decisions. It is, however, also significantly less burdened with the control information provisioning overheads.

With respect to the utility theory, the distributed resource management process considers RNs as typical users. Based on Equation (34), the RN utility function can be defined as a combination of the utilities of the users connected to the RN, however, related to the RN BH link capacity, not their respective AC link capacities

$$u_n(x) = \left(\sum_{j\in J_n} u_j(x)\right) + \left(\sum_{k\in N_n} u_k(x)\right) \tag{36}$$

The Lagrange function for the relay-enhanced network is thus

$$\begin{aligned}&\mathscr{L}(y,x,z,\lambda,\mu) = \\ &= \sum_{j\in J}\left(u_j\left(y_j\right) - \lambda_j\left(y_j - \sum_{r\in R} C_{j,r}x_{j,r}\right)\right) + \\ &+ \sum_{n\in N}\left(u_n\left(y_n^{\mathrm{BH}}\right) - \lambda_n\left(y_n^{\mathrm{BH}} - \sum_{r\in R} C_{n,r}x_{n,r}\right)\right) + \\ &+ \sum_{r\in R}\left(\mu_r\left(1 - z_r - \sum_{j\in J} x_{j,r} - \sum_{n\in N} x_{n,r}\right)\right) + \\ &+ \sum_{n\in N}\sum_{r\in R}\left(\phi_{n,r}x_{n,r}\left(\sum_{j\in J_n} x_{j,r} + \sum_{k\in N_n} x_{k,r}\right)\right)\end{aligned} \tag{37}$$

Again, solution to the optimization problem (19) can be found by calculating partial derivatives of the Lagrange function (37) and comparing them to zeros (the derivation is to big extent the same as for relay-less network, thus, not repeated here). From the derivation, the PF marginal cost of utility for a RN is as follows:

$$\lambda_n^{\mathrm{PF}} = \left[\left(\sum_{j\in J_n} \frac{1}{u_j\left(y_j\right)} \frac{\partial u_j\left(y_j\right)}{\partial y_j}\right) + \left(\sum_{k\in N_n} \lambda_k^{\mathrm{PF}}\right)\right] \tag{38}$$

and as in (31), the resource allocation priority metric for a RN ($M_{n,r}{}^{\mathrm{PF}}$) is

$$M_{n,r}^{\mathrm{PF}} = C_{n,r}\lambda_n^{\mathrm{PF}} \tag{39}$$

Further in this paper, a distributed iterative resource allocation procedure is considered. The procedure is based on the above definition of marginal cost of utility for RNs. The procedure is described in details in Algorithm 1.

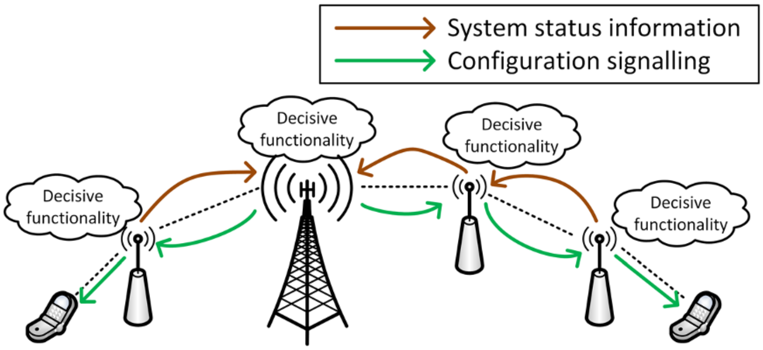

Figure 6 Distributed resource management scheme for multi-hop relay-enhanced networks.

Algorithm 1 Iterative distributed resource allocation procedure for multi-hop relay-enhanced networks

```
1. Calculate marginal cost of utility for every user basing on the historical performance statistics.

2. Propagate 'bottom-up' information of the marginal cost of utility, i.e., calculate marginal cost of utility
for every RN based on the marginal costs of utility of users connected to this RN and its direct
subordinate RNs.

3. For every served node, (user or subordinate RN) calculate the corresponding resource allocation
priorities basing on the marginal cost of utility.

4. At every BS and RN, allocate each resource element to the served node (user or subordinate RN) with
the highest resource priority metric.

5. Repeat the procedure in the next TTI.
```

4.3 Proposal of utility function for real-time traffic

In the state-of-the-art literature, various proposals of utility functions can be found. The described utility functions are, however, not always realistic. For example, it is often forced to make the utility functions concave to guarantee that there always is a unique solution to the resource assignment optimization problem. In this section, the utility functions for real-time services are proposed with consideration of the standardized QoS requirements defined in [26] (see also Table 1).

As explained in Section 2, the two main QoS requirements for real-time traffic are the GBR and PDB. A utility function for a real-time traffic should, therefore, reflect the two requirements and, thus, it is proposed here to be formulated as

$$u_j(x) = w_j u_j^{\mathrm{GBR}}\left(y_j\right)\, u_j^{\mathrm{PDB}}\left(t_j\right) \tag{40}$$

where y_j is the data rate of the user j, t_j is its packet delivery time, u_j^{GBR} is the utility related to the GBR requirement satisfaction, u_j^{PDB} is the utility related to the PDB requirement satisfaction, and w_j is priority weight of the service type related to the GBR of the service.

With respect to the resource allocation scheme x, the formula (40) is subject to Equation (21) and

$$\mathrm{E}\left(t_j\right) = S/y_j \tag{41}$$

In case of a GBR traffic as, e.g., audio and/or video live streaming, data is consumed by the receiving application at a specific rate matching the rate in which the data is generated by the source application. If the available transmission data rate is above the data generation rate, the excessive link capacity will be unused and the utility of the transmission will not increase above a certain maximum level. On the other hand, if the available transmission data rate is below the data generation rate at the source application, the target application will show loss of data, e.g., breaks in audio/video streaming. Therefore, the GBR utility function should have a form of a step function, with the step steepness depending on the acceptable level of packet loss. In line with this characteristic, it is proposed here to define the GBR utility function as the following modified logistic function:

$$u_j^{\mathrm{GBR}}\left(y_j\right) = \frac{1}{1 + \exp\left(\frac{a_j^{\mathrm{GBR}} - y_j}{b^{\mathrm{GBR}}}\right)} \tag{42}$$

where a_j^{GBR} is a parameter related to the GBR value of service used by the user j, and b^{GBR} is a parameter controlling the steepness of the GBR utility function in proximity of $y_j = a_j^{\mathrm{GBR}}$. The shape of the GBR utility function and its PF modification for various traffic types are depicted in Figure 7.

The PBD utility, on the other hand, should have the form of a step-down function, i.e., with the highest utility value available for delays shorter than the maximum packet delivery time and zero utility for higher delays. This is because in a real-time traffic, data packets are utilized by the receiving application in a certain sequence. If a data packet is not delivered on time before the receiving application expects it, the packet is of no use even if fully delivered (e.g., audio frame is of no use, if its emission event is passed). Likewise, if the data packet is delivered ahead of the expected time, it will

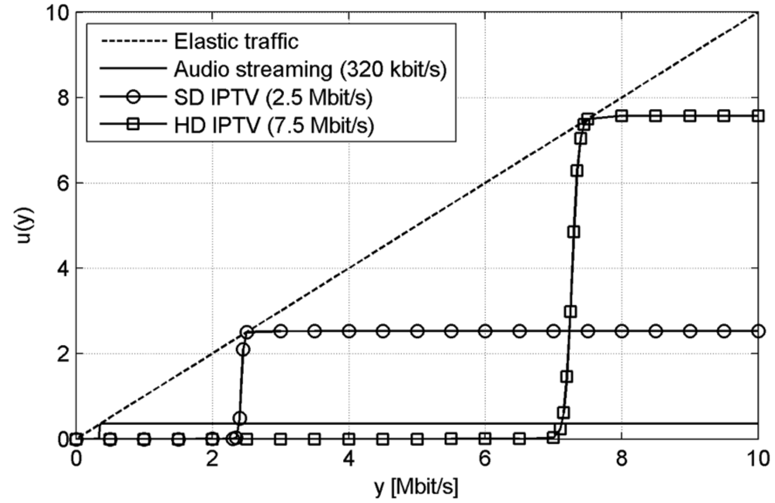

Figure 7 GBR utility function for common traffic types.

wait in the application buffer and not increase the QoS. A function meeting this description is as follows:

$$u_j^{\mathrm{PDB}}(t_j) = \frac{1 - \exp\left(-\frac{t_j^{\mathrm{Max}} - t_j}{b^{\mathrm{PDB}}}\right)}{1 + \exp\left(-\frac{t_j^{\mathrm{Max}} - t_j}{b^{\mathrm{PDB}}}\right)} \tag{43}$$

where t_j^{Max} is the maximum packet delivery time for the user j, and b^{PDB} is a parameter controlling the steepness of the PDB utility function. Impact of the PDB utility on the overall utility function for real-time traffic is illustrated in Figure 8.

4.4 Packet scheduling algorithms

Based on the above proposed utility functions for real-time traffic, resource management can be done in a cellular system in a QoS-aware manner. Next, several options for the QoS-aware resource allocation are defined. The approaches differ in terms of the criteria they aim at optimizing.

4.4.1 GBR-aware scheduling

The basic and the most common approach to resource management considers only optimization with respect to the required data rates. The GBR-aware scheduling

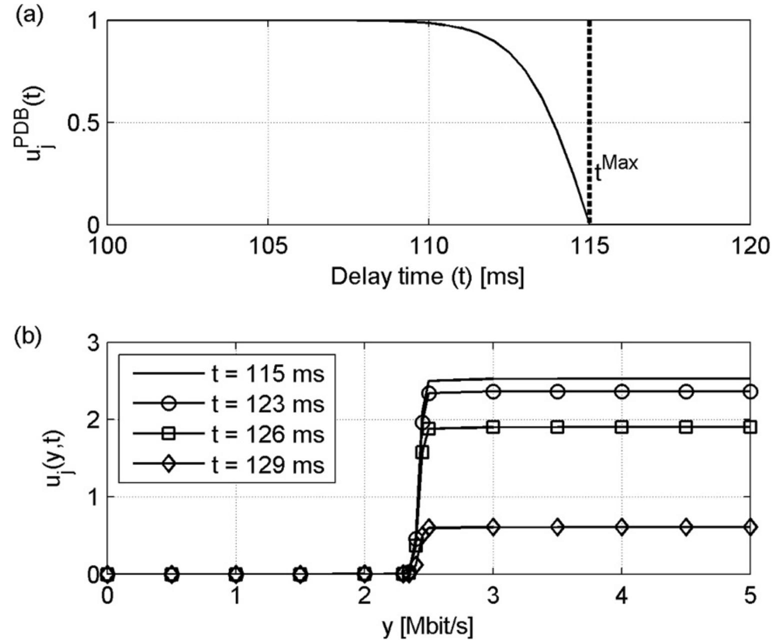

Figure 8 Impact of packet delay on utility of a 2.5 Mbit/s SD IPTV service (a-b).

algorithm tries to achieve a certain distribution of data rates to users with respect to their requirements. With respect to the utility theory framework described earlier, by inputting the GBR utility function to the general formula for the PF marginal cost of utility, the GBR-aware marginal cost of utility for the PF resource management is as follows:

$$\lambda_j^{\mathrm{GBR}} = \frac{1}{u_j^{\mathrm{GBR}}\left(y_j\right)} \frac{\partial u_j^{\mathrm{GBR}}\left(y_j\right)}{\partial y_j}, \tag{44}$$

and next by solving the derivative in (44) with respect to (42), there is

$$\lambda_j^{\mathrm{GBR}} = \frac{\exp\left(\frac{a_j^{\mathrm{GBR}} - y_j}{b^{\mathrm{GBR}}}\right)}{b^{\mathrm{GBR}}\left(1 + \exp\left(\frac{a_j^{\mathrm{GBR}} - y_j}{b^{\mathrm{GBR}}}\right)\right)} \tag{45}$$

4.4.2 PDB-aware scheduling

In case of traffic with specified requirements on the maximum packet delivery time, the PDB-aware scheduling algorithm can be used. The PDB-aware resource allocation prioritizes the delay optimization criteria; however, it also considers the minimum data rate requirement. This is because, as stated earlier, it is not possible to provide stable PDB satisfaction without satisfaction of the minimum data rate. By inputting the PDB utility function to the general formula for the PF marginal cost of utility, the PDB marginal cost of utility for the PF resource management is

$$\lambda_j^{\mathrm{PDB}} = \frac{1}{u_j^{\mathrm{PDB}}\left(t_j\right)} \frac{\partial u_j^{\mathrm{PDB}}\left(t_j\right)}{\partial y_j} \tag{46}$$

where considering the transmission and queuing times

$$t_j = \frac{S}{y_j} + t_{q,j} \tag{47}$$

thus, by solving the derivative in (46) with respect to (43), there is

$$\lambda_j^{\mathrm{PDB}} = \frac{2\exp\left(-\frac{t_j^{\mathrm{Max}} - t_j}{b^{\mathrm{PDB}}}\right)}{b^{\mathrm{PDB}}\left(1 - \exp\left(-2\frac{t_j^{\mathrm{Max}} - t_j}{b^{\mathrm{PDB}}}\right)\right)} \frac{S}{y_j^2} \tag{48}$$

4.4.3 Full QoS-aware scheduling

In the most complex approach, radio resources are assigned with respect to both the GBR and PDB requirements. In this case, the marginal cost of utility is calculated as the combination of the formulas (45) and (48)

$$\lambda_j^{\mathrm{QoS}} = \lambda_j^{\mathrm{GBR}} + \lambda_j^{\mathrm{PDB}} \tag{49}$$

5 Performance evaluation

This section covers simulation-based evaluations of the concepts described in the former sections of this paper. Firstly, simulation assumptions used for the analysis are described. Later, the generated results are presented and discussed.

5.1 Simulation assumptions

The data presented further in this paper is generated via computer simulations based on the widely accepted methodology for the LTE system evaluations specified by the 3GPP forum in [3]. For aspects not covered by the 3GPP recommendation (e.g., multi-hop relaying), supplementary assumptions are taken from [44].

The tool used to perform the evaluations presented further in this paper is a dynamic system level simulator created by the author on the MATLAB platform. It models operation of an LTE-A network including layers 1 to 3 of the protocol stack. All relevant network nodes (BSs, RNs, and UEs) and radio interfaces are modeled explicitly. The modeling is dynamic, which means that there is a timeline simulated that drives dynamic mechanisms such as user mobility, fast fading, resource allocation, traffic dynamics (data packets creation, transmission, reception and dropping), etc. Operation of the tool has be verified and calibrated vs. multiple other similar tools as part of the European Union founded project ARTIST4G [44].

The simulation assumptions that are clearly defined in [3] and [44] (e.g., radio propagation and network node models) are not repeated here as the two documents are publically available. With respect to the parameters that are ambiguous in [3] and [44], in this paper, it is specifically assumed that

- Only downlink transmission direction is simulated.
- Total system bandwidth is 20 MHz with full resource reuse at each cell.
- Macro BSs are deployed on a hexagon grid with 1,732 m inter-site distance (ISD).
- In relay-enhanced network scenarios, ten RNs are used per macro BS sector. The RNs are deployed in two tiers along the edges of macro sectors (for details, see [44]).
- RNs utilize the standardized LTE cell-selection procedure [45] (i.e., based on the higher received signal power) to select their respective donor cells. The selection is, however, restricted with respect to the capability of forming multi-hop topologies. The multi-hop topology support is a parameter of the simulations.
- MBSFN configuration for HD RNs is adapted to provide BH/AC capacity balancing [27] (typically high σ configurations are used), and the MBSFN

patterns with low BH sub-frame concentration are used where possible.
- Users are deployed uniformly in each macro BS sector, and they utilize the LTE cell selection procedure to attach to BSs or RNs.

The assumptions defined explicitly for this paper cover also traffic models. Specifically, two scenarios are assumed here: low and high load. In the low load scenario, there are on average ten users per macro BS sector, and in the high load scenario, the average number of users per sector is 20. In both load scenarios, it is assumed that all users use constant bit-rate real-time traffic services with the following distribution:

- Two thirds of all users request an audio streaming transmission with 320 kbit/s GBR, and
- One third of all users request a standard definition (SD) IP television (IPTV) transmission with 2.5 Mbit/s GBR.

For all considered traffic types, the additional assumptions are as follows:

- Data packet size is 1,374 bytes [46], and
- Maximum packet delivery time in radio interface is 130 ms (150 ms PDB with 20 ms delay assumed for core network [26]).

All the users are active all the time continuously generating data packets according to their respective service data rates and packet sizes.

The traffic assumptions correspond to the following average load conditions:

- Low load scenario:
  Relay-less system: 75% load
  Relay-enhanced system: 50% load

- High load scenario:
  Relay-less system: 95% to 100% load
  Relay-enhanced system: 80% load

With respect to the assumed traffic types, the system efficiency is assessed in terms of the following:

- GBR satisfaction, i.e., ratio between the data rate achieved by a user and its GBR setting,
- Packet dropping rate (PDR), i.e., percentage of packets that are not fully delivered to the target user within the 120 ms time limit, and
- Average and guaranteed (i.e., 95th percentile) packet delivery times.

5.2 Results and discussion

In Section 4.4, three approaches to resource management are proposed. Those management schemes are compared next on the basis of relay-less and relay-enhanced networks. Figure 9 depicts statistics of the GBR satisfaction and average packed drop rate (PDR) for macro BS connected users in a relay-less system. Figure 10, on the other hand, presents the distributions of the packet delivery times for the relay connected users in the presence of a three-hop HD relaying topology.

The PDB-aware scheduling focuses on satisfying the packet delivery time requirement for all users resulting in the lowest PDR at the initial stage of the system operation (see Figure 9b). The approach, however, does not consider directly the data rate requirement, thus it provides relatively low GBR satisfaction (on average 84% in the high load scenario). As result, many users do not achieve their required minimum data rate (even in the low load scenario). For such users, the PDR increases significantly over time (up to average 6% PDR in the low load scenario and 21% PDR in the high load scenario). As illustrated in Figure 10, the PDB-aware scheduling tries to deliver as many packets as possible by increasing scheduling priority for the packets with short times to drop (visible as increased probability for the highest packet delivery times). The actions are, however, insufficient considering the unsatisfied GBR requirement.

The GBR-aware approach, on the other hand, focuses on satisfying the data rate requirement and does not consider the packet delivery time requirement. This approach results in higher average GBR satisfaction level (88% in the high load scenario) and improved long-term PDR statistics (on average 5.5% PDR in the low load scenario and 17.5% PDR in the high load scenario). The approach, however, does not take any actions with respect to the packets with short times to drop (see right-hand part of Figure 10), thus some packets are dropped even though they could be delivered on time.

Finally, the full QoS-aware approach considers both the GBR and PDB requirements. Such dual approach results in the highest performance statistics. In terms of the GBR satisfaction, it provides similar average satisfaction as the GBR-aware approach (87% in the high load scenario) but improved performance for users in the worst traffic conditions (10% gain at 5th percentile GBR satisfaction level). It terms of the PDB requirement, it also results in the lowest PER levels (on average 4.5% PDR in the low load scenario and 14% PDR in the high load scenario).

Figure 10 presents also results for a modified (multi-hop) QoS-aware scheduling. The modified QoS-aware approach considers both the GBR and PDB requirements, but in case of multi-hop connection, it assumes reduction of the maximum packet delivery time

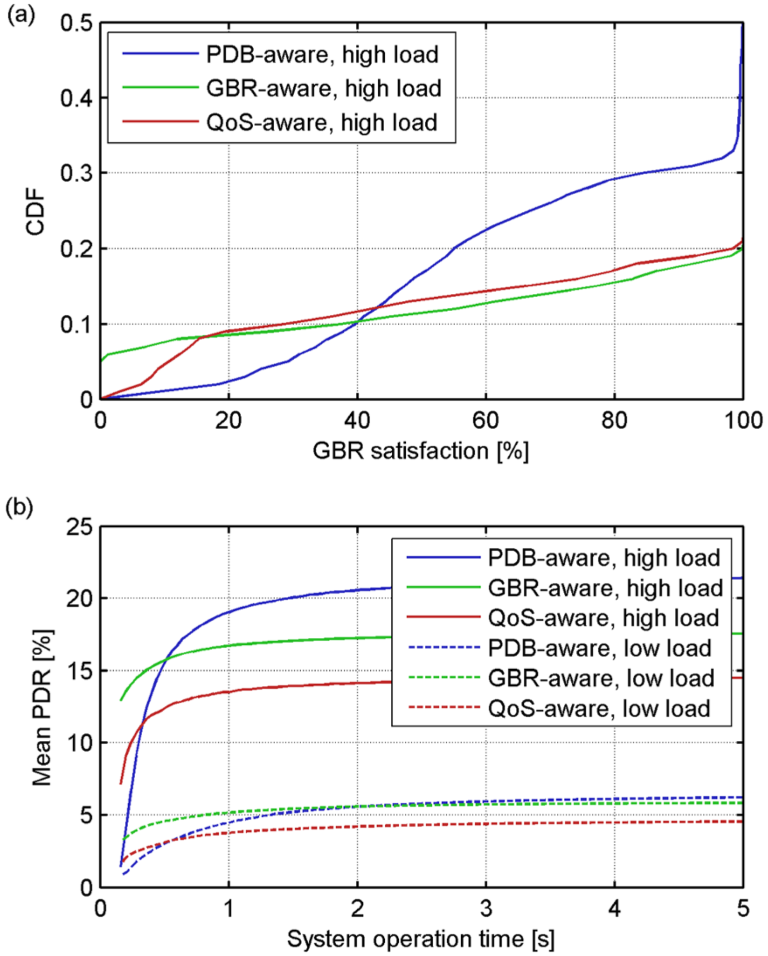

Figure 9 Comparison of resource management schemes in relay-less system (a-b).

proportionally to the number of end-to-end relaying hops. Specifically, in the conducted simulations, it is assumed that the maximum packet delivery time for a hop h in an H-hop connection is

$$t_{j,h}^{\mathrm{Max}} = t_{j}^{\mathrm{Max}} - \frac{H-h}{H} 30\ \mathrm{ms} \tag{50}$$

The purpose of the modification is to improve the on-time packet delivery for the relay-connected users. The effect of the modification is illustrated in Figure 10. Transmissions to users connected over three-hop and two-hop relaying connections have distinctive peaks in packet delivery times at 80 and 110 ms, respectively. The peaks correspond to the reduced maximum packet delivery times of 70 and 100 ms, respectively. As a result, the average PDR for the relay-connected users is reduced from 16% when using the basic QoS-aware approach to 14% PDR when using the multi-hop QoS-aware approach.

Basing on the data presented in Figures 9 and 10, it can be concluded that the joint GBR- and PDB-aware approach is the most effective in providing the overall satisfaction for users using real-time traffic. This method is used next, to analyze packet delivery times for various relaying topologies. Specifically, delays related to the FD and HD RN operations in two- and three-hop topologies are compared. In this analysis, only the low load scenario is considered as it gives a clear picture of the relaying-originated delays without the impact of macro BS overload. Figure 11 depicts distributions of packet delivery times for the considered scenarios and corresponding statistics are summarized in Table 2.

The packet delivery time statistics confirms the analysis conducted in Section 3. The FD relaying has a minor impact on the end-to-end delay. On average, the FD two-hop relaying generates 1.5 ms additional delay. The FD three-hop relaying introduces on average 2.3 ms additional delay, however, it also increases the maximum delay by 11 ms (compared to 7 ms for macro BS-connected users). The results are in line with the analysis included in Section 3.1 and especially with formula (4).

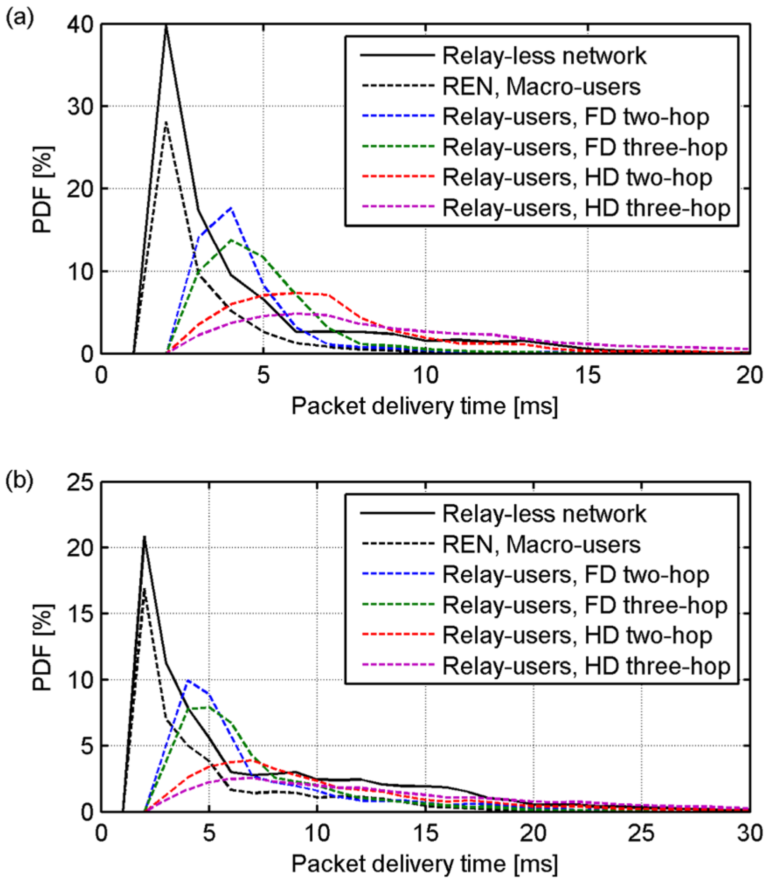

Figure 10 Comparison of resource management schemes in relay-enhanced system, high load scenario (a-b).

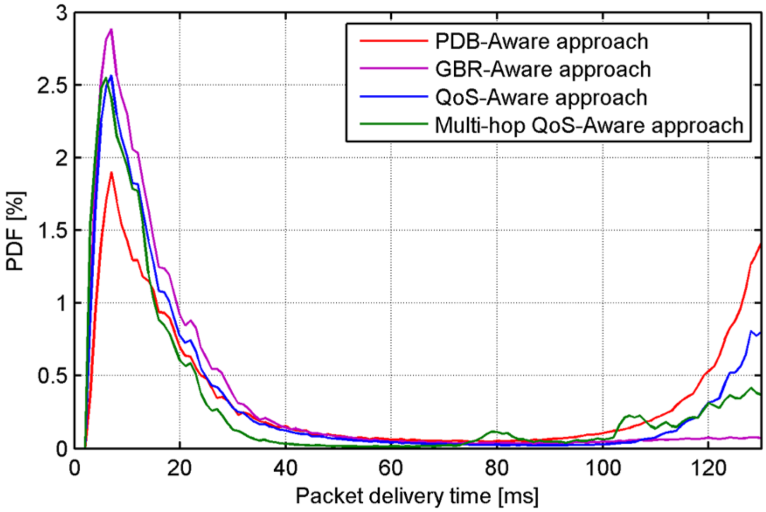

Figure 11 Packet delivery times for low load relaying scenarios.

Table 2 Summary of packet delivery statistics for macro BS and relay-connected users in the low load scenario

	Macro BS users	Relay users
Relay-less system	Average: 10.6 ms 95th percentile: 122 ms	-
FD two-hop relaying	Average: 1.1 ms 95th percentile: 7 ms	Average: 2.6 ms 95th percentile: 10 ms
FD three-hop relaying		Average: 3.4 ms 95th percentile: 18 ms
HD two-hop relaying		Average: 4.5 ms 95th percentile: 21 ms
HD three-hop relaying		Average: 7.2 ms 95th percentile: 85 ms

In case of HD relaying, the impact on packet delivery times is higher than in case of FD relaying. The average packet delivery time for two-hop relaying is 4.5 ms and for three-hop relaying, it is 7.2 ms. Again, the simulation results are in line with theoretical predictions presented in Figures 4 and 5 (see results for high σ and L =1). In case of HD relaying, the highest delays are significantly increased in comparison to the direct BS-user transmission (up to 85 ms in case of three-hop connection). This indicates that in case of bigger packet sizes and/or lower data rates available for the multi-hop connection, problems with satisfying the packet delivery time might occur on wide scale even for users with satisfied GBR requirement.

6 Conclusions

This paper deals with the topic of satisfying packet delivery time requirement for multi-hop relayed transmissions. The issue is a highly relevant topic in the context of recent development of cellular systems. This is because the relaying concept is included in the standards of fourth generation systems; however, the relevant standardization bodies do not provide dedicated mechanisms for QoS provisioning over relayed links.

The presented discussion of the topic focuses first on the analysis of the two available relaying configurations: full-duplex and half-duplex operation. The full-duplex relaying is characterized with a small additional delay overhead related to the RN processing time. The theoretical analysis and the conducted simulations show that the lower bound of the additional delay is in the range of few milliseconds and proportional to the number of transmission hops. The statistically highest delays are, however, more than doubled when comparing a three-hop FD topology with a direct BS-user transmission.

The relaying delay overhead is a more significant issue in case of HD RNs. In this case, the delay depends strongly on the configuration of time-domain multiplexing of RN BH and AC sub-frames. In the worst case, the end-to-end transmission time over an FD two-hop connection can be even 20 times higher than for the direct connection. The situation can be, however, improved with proper sub-frame configuration.

In the second part of this paper, resource management schemes based on the utility theory are proposed for managing relay-enhanced networks. The proposed concepts take into consideration the minimum data rate and/or the maximum packet delivery times to provide QoS satisfaction for users. From the considered schemes, the one based on a complex data rate and transmission time shows to provide the best results. The scheme can be further optimize for operation over multi-hop relayed links.

The overall conclusions from the study are that the full-duplex relaying is characterized with significantly lower delay overheads than the half-duplex relaying, and thus the full-duplex relaying is generally recommended. As for the half-duplex relaying, it is recommended to be used with sub-frame configurations characterized with high percentage of time available for backhaul link operation and low concentration of the backhaul sub-frames. It is also recommended to use the half-duplex relaying only with low number of end-to-end hops (maximum of three).

Competing interests

The author declares that he has no competing interests.

Acknowledgement
The author is a scholar within Sub-measure 8.2.2 Regional Innovation Strategies, Measure 8.2 Transfer of knowledge, Priority VIII regional human resources for the economy, Human Capital Operational Programme co-financed by European Social Fund and state budget.

Received: 1 February 2013 Accepted: 13 October 2014
Published: 29 October 2014

References

1. ITU-R, *M.2134 Requirements Related to Technical Performance for IMT-Advanced Radio Interface(s)* (2008)
2. 3GPP, *TR 36.814 v9.0.0: Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects* (2010)
3. H Holma, A Toskala, *LTE for UMTS: Evolution to LTE-Advanced*, 2nd edn. (John Wiley & Sons, Chichester (UK), 2011)
4. WiMAX Forum, *WMF-T23-005-R020v01: Mobile Radio Specifications; Release 2.0* (2012)
5. WiMAX Forum, *WiMAX and the IEEE 802.16m Air Interface Standard* (2010)
6. S Parkvall, E Dahlman, A Furuskär, Y Jading, M Olsson, S Wänstedt, K Zangi, LTE-Advanced - Evolving LTE Towards IMT-Advanced, in *IEEE Vehicular Technology Conference (VTC). Fall* (2008), pp. 1–5
7. PE Mogensen, T Koivisto, KI Pedersen, IZ Kovács, B Raaf, K Pajukoski, MJ Rinne, LTE-Advanced: The Path Towards Gigabit/S in Wireless Mobile Communications, in *International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE)* (2009), pp. 147–151
8. Q Li, G Li, W Lee, M Lee, D Mazzarese, B Clerckx, Z Li, MIMO techniques in WiMAX and LTE: a feature overview. IEEE Commun Mag **48**(5), 86–92 (2010)
9. KI Pedersen, C Rosa, H Nguyen, LGU Garcia, Y Wang, Carrier aggregation for LTE-Advanced: functionality and performance aspects. IEEE Commun Mag **49**(6), 89–95 (2011)
10. K Safjan, S Strzyż, J Góra, Kontrola interferencji oraz poprawa wydajności heterogenicznych sieci LTE. Przegląd Telekomunikacyjny Wiadomości Telekomunikacyjne **2011**(6), 1–4 (2011)

11. N Gresset, H Halbauer, J Koppenborg, W Zirwas, H Khanfir, Interference avoidance techniques: improving ubiquitous user experience. IEEE Veh Technol Mag **7**(4), 37–45 (2012)
12. R Irmer, H Droste, P Marsch, M Griger, G Fettweis, S Brueck, HP Mayer, L Thiele, V Jungnickel, Coordinated multipoint: concepts, performance and field trial results. IEEE Commun Mag **49**(2), 102–111 (2011)
13. R Irmer, F Diehm, On Coverage and Capacity of Relaying in LTE-Advanced in Example Deployments, in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)* (2008), pp. 1–5
14. T Beniero, S Redana, J Hämäläinen, B Raaf, Effect of Relaying on Coverage in 3GPP LTE-Advanced, in *IEEE Vehicular Technology Conference (VTC)* (Spring, 2009), pp. 1–5
15. Y Sui, A Papadogiannis, T Svensson, The Potential of Moving Relays - a Performance Analysis, in *IEEE Vehicular Technology Conference (VTC)* (2012), pp. 1–5
16. E Lang, S Redana, B Raaf, Business Impact of Relay Deployment for Coverage Extension in 3GPP LTE-Advanced, in *IEEE International Conference on Communications (ICC) Workshops* (2009), pp. 1–5
17. A Bou Saleh, S Redana, B Raaf, J Hämäläinen, Comparison of Relay and Pico Enb Deployments in LTE-Advanced, in *IEEE Vehicular Technology Conference (VTC)* (2009), pp. 1–5
18. 3GPP, RP-110911, *Relays for LTE - Core Part (3GPP Work Item Description)* (2011)
19. A Bou Saleh, Ö Bulakci, J Hämäläinen, S Redana, B Raaf, Analysis of the impact of site planning on the performance of relay deployments. IEEE Trans Veh Technol **61**(7), 3139–3150 (2011)
20. O. Bulakci, S. Redana, B. Raaf, J. Hämäläinen, Performance Enhancement in LTE-Advanced Relay Networks Via Relay Site Planning, in *IEEE Vehicular Technology Conference (VTC)* (Spring, 2010), pp. 1–5
21. ARTIST4G WP3, D3.3, *Relay Networks Specific Resource Management Features* (2011)
22. ARTIST4G WP3, D3.4, *Relay Configurations* (2011)
23. 3GPP, TS 36.806 v9.0.0, *Evolved Universal Terrestrial Radio Access (E-UTRA); Relay Architectures for E-UTRA (LTE-Advanced)* (2010)
24. ARTIST4G WP3, D3.5a, *Enhancements to Type-1 Relay Implementation* (2012)
25. 3GPP, TS 36.216 v10.3.1, *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer for Relaying Operation* (2011)
26. 3GPP, TS 23.203 v11.7.0, *Policy and Charging Control Architecture* (2012)
27. B Sadiq, R Madan, A Sampath, Downlink Scheduling for Multiclass Traffic in LTE, in *EURASIP Journal on Wireless Communications and Networking* (2009), pp. 1–18
28. F Vitiello, T Riihonen, J Hämäläinen, S Redana, On Buffering at the Relay Node in LTE-Advanced, in *IEEE Vehicular Technology Conference (VTC)* (2011), pp. 1–5
29. J Góra, S Redana, Resource Management Issues for Multi-Carrier Relay-Enhanced Systems, in *EURASIP Journal on Wireless Communications and Networking* (2012), pp. 1–8
30. T Riihonen, S Werner, R Wichman, ZB Eduardo, On the Feasibility of Full-Duplex Relaying in the Presence of Loop Interference, in *IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)* (2009), pp. 275–279
31. H Holma, A Toskala, *LTE-Advanced 3GPP Solution for IMT-Advanced*, 1st edn. (John Wiley & Sons, Chichester (UK), 2012), pp. 1–223
32. J Huang, VG Subramanian, R Agrawal, R Berry, Downlink scheduling and resource allocation for OFDM systems, in *Proc. of the Conference on Information Sciences and Systems (CISS)* (2006)
33. R Agarwal, V Majjigi, R Vannithamby, JM Cioffi, Efficient Scheduling for Heterogeneous Services in OFDMA Downlink, in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM)* (2007), pp. 3235–3239
34. PC Fishburn, *Utility theory for decision making* (John Wiley & Sons, Chichester (UK), 1970), pp. 1–246
35. PC Fishburn, Utility theory, INFORMS Manage. Sci Theory Series **14**(5), 335–378 (2010)
36. F Kelly, Charging and rate control for elastic traffic. Eur Trans Telecommun **8**(1), 33–37 (1997)
37. FP Kelly, AK Maulloo, DKH Tan, Rate control for communication networks: shadow prices, proportional fairness and stability. J Oper Res Soc **49**(3), 237–252 (1998)
38. T Lan, D Kao, M Chiang, A Sabharwal, An Axiomatic Theory of Fairness In Network Resource Allocation, in *IEEE International Conference on Computer Communications (INFOCOM)* (2010), pp. 1–9
39. T Lan, M Chiang, An Axiomatic Theory of Fairness in Resource Allocation, in *IEEE International Conference on Computer Communications (INFOCOM)* (2010), pp. 1–9
40. C Joe Wong, S Sen, T Lan, M Chiang, Multi-Resource Allocation: Fairness-Efficiency Tradeoffs in a Unifying Framework, in *IEEE International Conference on Computer Communications (INFOCOM)* (2012), pp. 1206–1214
41. L Rittenberg, T Tregarthen, *Principles of Microeconomics* (Flat World Knowledge, Inc, Washington, D.C. (USA), 2009), pp. 1–432
42. M Uchida, J Kurose, An Information-Theoretic Characterization of Weighted Alpha-Proportional Fairness, in *IEEE International Conference on Computer Communications (INFOCOM)* (2009), pp. 1053–1061
43. J Nash, The bargaining problem. Econometrica **18**(2), 155–162 (1950)
44. ARTIST4G WP3, D3.5, *Performance Evaluations of Advanced Relay Concepts* (2012)
45. 3GPP, TS 36.300 v11.3.0, *Evolved Universal Terrestrial Radio access (E-UTRA) and Evolved Universal Terrestrial Radio access Network (E-UTRAN); Overall Description; stage 2* (2012)
46. G Thompson, *IPTV - What does it really mean and how does it work? SMPTE [Online* (2013). Available at http://www.smpte-profdev.org/resources/SMPTE_PDA_Now_01-17-2008_IPTV.pdf Accessed 13 Jan

doi:10.1186/1687-1499-2014-178
Cite this article as: Góra: **QoS-aware resource management for LTE-Advanced relay-enhanced network.** *EURASIP Journal on Wireless Communications and Networking* 2014 **2014**:178.