Oliver-Balsalobre *et al. EURASIP Journal on Wireless Communications and Networking* (2016) 2016:7
DOI 10.1186/s13638-015-0508-x

EURASIP Journal on
Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Self-tuning of scheduling parameters for balancing the quality of experience among services in LTE

Pablo Oliver-Balsalobre[1]*, Matías Toril[1], Salvador Luna-Ramírez[1] and José María Ruiz Avilés[2]

Abstract

Improving the Quality of Experience (QoE) offered to subscribers has become a major issue for mobile network operators. In this paper, a self-tuning algorithm for adjusting parameters in a multi-service packet scheduler of a radio base station based on network statistics is proposed to balance QoE across services in Long Term Evolution (LTE). The aim of the algorithm is to ensure that all users achieve the same average QoE regardless of the type of service. For this purpose, the proposed heuristic algorithm iteratively changes service priority parameters to re-prioritize services so that those services with the lowest QoE increase their priority. Unlike previous approaches, the proposed algorithm takes QoE (and not Quality of Service) into account. Method assessment is carried out in a dynamic system-level LTE simulator. Simulation results in a typical scenario show that tuning service priority parameters can significantly increase the QoE of the worst service without affecting the overall network QoE.

Keywords: Long term evolution, Quality of experience, Self-organizing networks, Optimization, Balance, Re-prioritization

1 Introduction

With the success of smartphones and tablets, traffic in mobile broadband networks has dramatically changed due to the introduction of new services. Although recent radio access technologies, such as Worldwide Interoperability for Microwave Access (WiMAX) and Long Term Evolution (LTE), are prepared to offer a wide range of services, the launch of new services poses new challenges for network operators [1]. Likewise, continuous advances in terminals and, most importantly, in user expectations force operators to update the way they manage their networks. To provide the best end user experience, mobile operators are changing their management processes, currently focused on the network performance, to a more modern approach focused on user opinion. As a result, Customer Experience Management (CEM) has now become a key factor to differentiate among operators offering similar networks and services [2]. In such a user-centric approach, traditional objective Quality-of-Service (QoS) metrics are substituted by subjective Quality-of-Experience (QoE) metrics.

In parallel, the explosive growth of the size and complexity of mobile networks makes it very difficult for operators to manage their network. Such a need for increasing operational efficiency has stimulated intense research and standardization activity in the field of Self-Organizing Networks (SON) [3–5]. Most SON use cases in the literature only deal with basic radio aspects, such as radio network coverage, connection quality or capacity and power consumption [6, 7]. Although multi-layer, multi-vendor, and multi-technology issues have been addressed recently [8], less attention has been paid to the problems originated by the co-existence of multiple services in the same network and how these problems can be solved by SON.

Traffic and service management in current mobile networks is done by dynamic packet scheduling (PS) algorithms [9–11]. PS algorithms dynamically assign radio resources (i.e., frequency, time slot, and power) to user data requests based on their QoS constraints [9, 12]. Basic schedulers only deal with multiple users of the same

*Correspondence: pob@ic.uma.es
[1] Ingeniería de Comunicaciones, Universidad de Málaga, Campus de Teatinos S/N, 29071 Malaga, Spain
Full list of author information is available at the end of the article

service [13]. More sophisticated schedulers allocate more resources to users experiencing worse QoS to satisfy some fairness constraint [14, 15]. Such a QoS balance between users is evaluated from a theoretical perspective in several studies (e.g., [16–19]). However, these studies do not specify how the balance situation is accomplished.

To deal with the different service requirements, the 3rd Generation Partnership Project (3GPP) has defined several QoS Class Identifiers (QCI) to differentiate among service classes [20]. Based on QCIs, schedulers can prioritize among services. Some scheduling algorithms are proposed to provide differentiated services, QoS and fairness by assigning appropriate weights to each user queue (e.g., weighted and deficit round robin and weighted fair queuing). However, these schemes do not exploit multi-user diversity gain and hence do not achieve optimal system performance. More advanced scheduling algorithms combine both multi-service and multi-user diversity gain capabilities [19, 21, 22]. In [22], a scheduling algorithm is proposed to deal with real-time and non-real-time traffic in a proportional fair manner. More recent works [23–28] propose QoE-aware schedulers whose aim is to optimize the overall QoE while ensuring a minimum QoE for all users. All of them decide the exact resources assigned to every single user in real time, which makes them suitable for minimum QoS/QoE assurance. However, the aim of most schedulers is to ensure a minimum QoS/QoE for the worst users, rather than equalizing the average QoS/QoE per service. Thus, QoE balance between users or services is not guaranteed. Moreover, implementing these advanced schedulers would require upgrading network equipment, which is not desired by network operators that have already made an important investment to upgrade to the latest radio access technology.

Alternatively, tuning parameters of existing schedulers can be done to optimize the overall QoE. In [29], a self-tuning algorithm for the contention window parameter is proposed that does not differentiate between services. Closer to this work, an adaptive proportional and integrative controller is used in [30] to adjust application priorities in order to ensure end-to-end delay requirements. Similarly, an adaptive controller is proposed in [31] for adjusting flow priorities to ensure a certain QoS level for multimedia services in terms of delay. In that proposal, each service has its own controller, whose decisions only depend on the QoS of that flow. This might cause instabilities when each flow tries to increase its priority individually in real time. More importantly, the aim of the controller is not to balance QoS among flows but to ensure that all flows reach their QoS target. Likewise, in congestion situations, when no flow fulfills its required QoS and all priority values reach their limits, it is not ensured that all services have the same QoS. Thus, its aim is not to equalize QoS among services, but to increase the overall system throughput. To the authors' knowledge, no method has been proposed to adjust service priority parameters in a multi-service multi-user scheduler of a radio base station with the aim of balancing the overall QoE per service under different traffic load conditions.

In this paper, a self-tuning algorithm for adjusting parameters in a multi-service packet scheduler of a radio base station based on network statistics is proposed to balance QoE across services in LTE. The aim of the algorithm is to ensure that all users achieve the same average QoE regardless of the type of service. For this purpose, the proposed heuristic algorithm iteratively changes service priority parameters to re-prioritize services so that those with the lowest QoE increase their priority. Unlike previous approaches, the proposed self-tuning algorithm takes QoE (and not QoS) into account. Method assessment is carried out in a dynamic system-level LTE simulator implementing a regular macrocellular scenario. The main contributions of this work are as follows: (a) a self-tuning algorithm for scheduler parameters to equalize QoE among services in LTE with any network load conditions and (b) simulation results that quantify the impact of equalizing QoE among services in a realistic multi-service LTE scenario. The rest of the paper is organized as follows. Section 2 describes the LTE system model, including the considered scheduling algorithm. Then, Section 3 presents the proposed self-tuning algorithm for scheduler parameters. Section 4 describes the simulation tool used to assess the algorithm, and Section 5 presents the results of the simulations. Finally, Section 6 presents the conclusions of the study.

2 System model

In this section, a system model for a multi-service LTE system is presented. First, a multi-service scheduling algorithm is outlined, identifying its key parameters. Then, traffic models for the services included in the traffic mix of current mobile networks are presented. Finally, user QoE models relating QoS performance indicators to end user experience are explained for each service.

2.1 Scheduling algorithm

The multi-service PS scheme considered in this work is a modified version of the classical exponential/proportional fair (EXP/PF) scheduler [21]. The original EXP/PF scheme is designed to support multimedia applications in a system with Adaptive Modulation and Coding (AMC) and Time Division Multiplexing (TDM). For this purpose, service requests are classified into real time (RT) or non-real time (NRT). Then, each request is given a priority value, itK, depending on its service type, with the following expressions:

$$K = \begin{cases} \exp\left(\dfrac{a_i W_i(t) - a\overline{W(t)}}{1+\sqrt{a\overline{W(t)}}}\right)\cdot PF_{\text{factor}} & i \in RT \\ \\ \dfrac{\omega(t)}{M(t)}\cdot PF_{\text{factor}} & i \in NRT \end{cases} \tag{1}$$

with

$$a\overline{W(t)} = \frac{1}{N_{RT}}\sum_{i\in RT} a_i W_i(t) \quad , \tag{2}$$

$$\omega(t) = \begin{cases} \omega(t-1) - \varepsilon & W_{\max} > \tau_{\max} \\ \omega(t-1) + \dfrac{\varepsilon}{\kappa} & W_{\max} < \tau_{\max} \end{cases} \tag{3}$$

$$a_i = -\frac{\log(\delta_i)}{\tau_{\max}} \quad . \tag{4}$$

In (1), $W_i(t)$ is the Head-Of-Line (HOL) packet delay of user i at time t, $a\overline{W(t)}$ represents the average delay of RT users, a_i is related to delay constraints and PF_{factor} is the fairness factor. $\omega(t)$ is a weight factor associated with NRT users and $M(t)$ is the average number of RT packets waiting at the eNodeB buffer at time t. In (2), N_{RT} is the number of RT users. In (3), $W_{\max}$ is the maximum HOL packet delay of all RT service users in the cell considered, $\tau_{\max}$ is the maximum delay constraint of RT services (in milliseconds), whereas ε and k are constants defining how $\omega(t)$ is updated depending on $W_{\max}$ and $\tau_{\max}$. Specifically, $\omega(t)$ is increased when $W_{\max} < \tau_{\max}$ (i.e., when delay constraints are being met by RT users), giving NRT users a higher priority by Eq. (1). Finally, in (4), δ_i is the maximum probability for HOL packet delay of user i to exceed its delay threshold (in this case, δ_i and $\tau_{\max}$ are shared by all RT users). The fairness factor, PF_{factor}, is computed as in the classical Proportional Fair (PF) algorithm [15],

$$PF_{\text{factor}} = \frac{r_i(t)}{R_i(t)} \quad , \tag{5}$$

$$R_i(t) = \left(1 - \frac{1}{t_c}\right)\cdot R_i(t-1) + \frac{1}{t_c}\cdot r_i(t-1) \quad , \tag{6}$$

where $r_i(t)$ is the achievable data rate of user i, $R_i(t)$ is its average data rate, and t_c is the averaging time constant, which is used to prioritize either throughput maximization or fairness. The reader is referred to [21] for a more detailed explanation of the behavior of the EXP/PF scheduler.

In the EXP/PF scheme, RT users always have a higher priority than NRT users when their HOL packet delays are approaching the maximum delay constraint, regardless of the experienced QoE. To change this behavior, and allow to re-prioritize services to some extent, the EXP/PF is modified here by adding a new parameter, referred to as Service Priority Index (SPI), as was already done in [32]. The new priority value, K', is computed from the previous value, K, as

$$K' = \min(\max(K,1),10)\cdot SPI_i \quad , \tag{7}$$

where SPI_i is a real value between 1 and 15 reflecting the service priority associated with user i, which can be used to gently re-prioritize services. For convenience, in (7), the value of K (i.e., the priority value computed by the classical EXP/PF) is limited between 1 and 10. Such limits ensure that a service with the highest SPI value (=15) always has a priority higher than one service with the lowest SPI value (=1), regardless of the value of K. Thus, the sensitivity of priority values to SPI changes is increased, which improves the ability of the self-tuning algorithm proposed later to equalize QoE among services. Without that limitation, K might be arbitrarily high, e.g., for users with an extremely high achievable data rate compared to their average data rate. For those users with a large K value, SPI reduction might not be enough to decrease their priority, so that SPI changes would not have an impact on the re-prioritization process.

2.2 Service models

In LTE, each service is associated with one QCI, which defines its performance objectives. In general, lower QCI values imply more restrictive services in terms of performance. This paper considers both RT and NRT services and thereby takes into account the whole QCI range [20]. Table 1 shows the main parameters of each service included in this work.

A first service is Voice over Internet Protocol (VoIP). As a conversational RT service, it is defined as a Guaranteed Bit Rate (GBR) service whose QCI value (=1) corresponds to the highest priority value for user data services [20]. In this work, the VoIP service is modeled as a data source generating packets of 20 bytes every 10 ms, with a bit rate of 16 kbps [33]. A call dropping model is also included, where a VoIP call is terminated when a user does not receive enough resources during one consecutive second.

A second service is buffered video streaming (hereafter, VIDEO for short). This service is defined as non-GBR with a less restrictive QCI value (=6). In this work, a simple model of the player's buffer at the client side is considered. The amount of video data in the buffer dynamically changes with the download bandwidth, video bit rate, and video playing rate. Initially, the buffer is filled with data. The larger the buffer size, the larger the initial delay. Later, if the buffer runs out (download bandwidth < video rate), the video stops (event known as *stalling*) and the player waits until the buffer is re-filled again. To avoid the use of an analytical traffic model, real video traces are used (http://trace.eas.asu.edu/tracemain.html) [34]. Such traces include frame arrival times and frame sizes of real video sequences obtained with an H.264/MPEG-4 AVC

Table 1 Service models parameters

Service	Service type	Main features
VoIP	GBR	Poisson arrivals
		Coding rate 16 kbps
		Exponential call duration (average 60 s)
		Call drop after 1 s without resources
VIDEO	non-GBR	Poisson arrivals
		H.264/MPEG-4 AVC
		Variable Bit Rate (VBR)
		352 × 288 pixel resolution, 25 frames per second
		Coded frame size based on traces
		Session drop based on session length
WEB	non-GBR	Poisson arrivals
		Several web pages per user (session)
		Exponential waiting time between pages (average 30 s)
FTP	non-GBR	One file per user
		Average file size 2 MB

codec. Video duration is randomly defined on a per-user basis with a uniform distribution to a maximum of 3 min. A video session drop model is also included, where a session is terminated if session time is more than twice the video duration.

The other two services are NRT services, namely web browsing with Hypertext Transfer Protocol (referred to as WEB) and file downloading with File Transfer Protocol (FTP). Both of them are best effort non-GBR services and are assigned to the lowest QCI values (=9). The WEB model in this work is inspired in [35]. For simplicity, a WEB session is modeled as the download of several web pages with inactivity time between them. FTP service is a typical file download service, where session time is determined by the time spent downloading the file at the maximum allowed data rate [35].

2.3 QoE models

A QoE model reflects the impact of QoS on end user experience. A common approach to build a QoE model is by means of utility functions. Utility functions are mathematical functions expressing some kind of preference relation. In the context of mobile networks, utility functions describe the relationship between the value of key QoS network performance statistics and the QoE perceived by users of a service. Since each service has different QoS performance targets, each service has a different utility function. In this work, the output of any QoE model is an estimate of the Mean Opinion Score (MOS) ranging from 1 (bad experience) to 5 (ideal experience) [36]. The following paragraphs describe the utility functions used for each service.

2.3.1 VoIP

The E-model [37] can be used to obtain an estimate of the voice quality, R ($\in [0, 100]$), from the average mouth-to-ear (i.e., one way) delay. In this work, only the delay in the downlink (DL) is considered to reduce the computational load of simulations. All other E-model parameters are set to their default values, described in [38]. Then, the voice quality R is translated into MOS with the formula:

$$\mathrm{MOS}_{\mathrm{VoIP}} = 1 + 0.035 \cdot R + R \cdot (R - 60) \cdot (100 - R) \cdot 7 \cdot 10^{-6}. \tag{8}$$

Note that the $\mathrm{MOS}_{\mathrm{VoIP}}$ is upper limited to 4.5 when the R parameter is rated to its maximum value, showing that, even in ideal test conditions, some individual may not rank the experience with the maximum.

2.3.2 Video streaming

In a buffered video-streaming service, such as YouTube or Netflix, the key indicators defining the QoE of a user are the initial delay and the number and duration of stallings. In this work, the utility function for video streaming is [39]:

$$\mathrm{MOS}_{\mathrm{VIDEO}} = 4.23 - 0.0672 L_{ti} - 0.742 L_{fr} - 0.106 L_{tr} \quad , \tag{9}$$

where L_{ti} denotes the initial buffering time (in seconds), L_{fr} is the average frequency of stallings (in seconds^{-1}), and L_{tr} is the average stalling duration (in seconds) [39]. As in the previous case, the model is upper limited to 4.23.

2.3.3 FTP

The QoE associated with FTP service depends on the average user throughput during the connection. The formula used to obtain the MOS value is [40]:

$$\mathrm{MOS}_{\mathrm{FTP}} = \max(1, \min(5, 6.5 \cdot T - 0.54)) \quad , \tag{10}$$

where T is the average user throughput of a user (in Mbps).

2.3.4 Web browsing

Similar to the FTP case, the level of satisfaction in web browsing is measured on the basis of user throughput. In this case, the MOS is calculated as [40]

$$\mathrm{MOS}_{\mathrm{WEB}} = 5 - \frac{578}{1 + \left(\frac{T + 541.1}{45.98}\right)^2} \quad , \tag{11}$$

where T is the average user throughput of a user (in kbps). The shape of this utility function defines the web browsing service not as restrictive as FTP. Thus, a web browsing user needs fewer resources than an FTP user to receive a satisfactory service.

3 Balancing algorithm

SON optimization algorithms aim to solve network problems by modifying Radio Access Network (RAN) parameters [4, 5]. In this work, a self-tuning algorithm is proposed to balance the QoE among services by adjusting the service priority parameters in the scheduler of the eNodeB. The aim of tuning is to re-prioritize services so that the average QoE per service is the same in the long term. For this purpose, SPIs are modified based on statistical QoS measurements collected per service in the network management system. As a result, users of services with worse QoE increase their priority and receive more radio resources.

The algorithm is conceived as an iterative process that decides the new SPI values based on an estimate of the QoE per service in the previous iteration (hereafter referred to as optimization loop). To avoid abrupt parameter changes, the controller is designed with an incremental structure, where SPI parameters are modified progressively. Thus, the output of the decision-making process is the positive or negative step added to the current SPI values.

For simplicity, it is assumed here that all eNodeBs in the network share the same set of SPI values (i.e., tuning is done on a network basis). Thus, the algorithm is divided into a set of controllers (one per service) in charge of modifying the corresponding SPI parameter. The inputs to each controller are the QoE for the optimized service and that of the other services measured across the whole network. The output of each controller is the new value of the SPI parameter for the optimized service that is used in all schedulers in the network.

In the following paragraphs, two variants of the algorithm are described, differing in the drivers used to guide the tuning process.

3.1 Unweighted strategy

In a first option, referred to as *unweighted strategy*, the aim is to equalize the average QoE of all services, i.e., the arithmetic mean of the QoE experienced by connections of a service,

$$\overline{\mathrm{QoE_j}} = \frac{1}{N_j}\sum_{i=1}^{N_j}\mathrm{QoE_i} \quad , \tag{12}$$

where j is the evaluated service, N_j is the number of users of service j, and QoE_i is the QoE perceived by user i, estimated from QoS statistics. The sum considers that all users of the same service have equal target QoE. With this aim, the input to each controller is the average QoE of that service and the average of the average QoE of the other services, computed as

$$\overline{\mathrm{QoE}}_{k\neq j} = \frac{1}{N_s - 1}\sum_{k\neq j}\overline{\mathrm{QoE_k}} \quad , \tag{13}$$

where N_s stands for the number of active services on the network. Then, the QoE difference is calculated as

$$\Delta\overline{\mathrm{QoE_j}} = \overline{\mathrm{QoE_j}} - \overline{\mathrm{QoE}}_{k\neq j} \quad , \tag{14}$$

where $\overline{\mathrm{QoE}}_{k\neq j}$ is the average QoE of those services different from j. Such a difference is used as a measure of the distance and direction from the balance situation.

A classical proportional controller is used to modify SPIs. The response of the controller is represented in Fig. 1. It can be observed that SPI changes are inversely proportional to $\Delta\overline{\mathrm{QoE}}_j$. Thus, if a service experiences a QoE larger than the other services, its SPI is decreased. The change is more aggressive if $\Delta\overline{\mathrm{QoE}}_j$ is higher than 1.

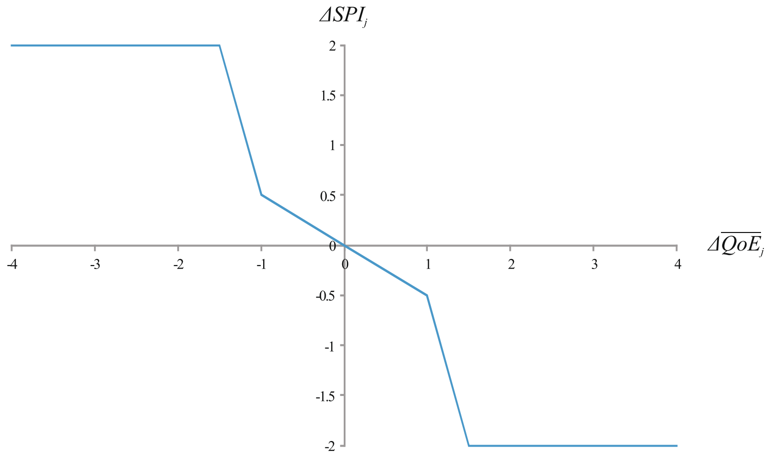

Fig. 1 Controller in the unweighted strategy

The two slopes provide a gain scheduling mechanism to achieve an adequate trade-off between speed of response and system stability. The lower slope for low QoE differences aims to reduce system sensitivity to ensure stability at the end of the balancing process. The higher slope for larger QoE differences ensures fast convergence to equilibrium and fine granularity to reduce QoE differences. Finally, upper and lower limits ensure that the largest variation of the SPI between consecutive loops is 2.

3.2 Weighted strategy

For network operators, the total number of satisfied users is a key driver. In this case, the percentage of users of each service becomes a very important parameter, since services with more users should be prioritized against those with fewer users. To favor services with more users, the indicator used to measure the QoE of a service is modified to include a weight dependent on the number of users of the service, as

$$\overline{\mathrm{QoE}}_j^W = \frac{\overline{N}}{N_j} \cdot \overline{\mathrm{QoE}}_j \quad , \tag{15}$$

where the superscript ^{W} denotes *weighted*, N_j is the number of users of that service j in the network, and the weight factor $\overline{N}$ represents the average number of users per service. Note that the larger the value of N_j, the lower the value of $\overline{\mathrm{QoE}}_j^W$, reflecting a worse value of the weighted QoE indicator for that service. The lower QoE of more populated services is compensated for the controller by increasing their priority so that those services receive more resources. Similarly to the unweighted strategy, there is no distinction of target QoE among users within the same service.

In the weighted strategy, the balancing process aims to reduce the difference between the weighted QoE indicator of each service and the mean value of the other services, computed as

$$\Delta\overline{\mathrm{QoE}}_j^W = \overline{\mathrm{QoE}}_j^W - \overline{\mathrm{QoE}}_{k\neq j}^W \quad , \tag{16}$$

$$\overline{\mathrm{QoE}}_{k\neq j}^W = \frac{1}{N_s - 1} \sum_{k\neq j} \overline{\mathrm{QoE}}_k^W \quad , \tag{17}$$

where $\Delta\overline{\mathrm{QoE}}_j^W$ is the difference of weighted QoE of a service against that of the other services, $\overline{\mathrm{QoE}}_{k\neq j}^W$ is the average weighted QoE of the other services, and N_s is the number of services in the network. As shown in Fig. 2, the shape of the controller is exactly the same as in the unweighted case.

It should be pointed that, in both strategies, equalizing QoE across services does not necessarily increase the overall QoE. However, it is expected that, in normal situations, increasing the priority of services with the lowest QoE should improve the overall system QoE. Such a belief is based on the shape of utility functions, shown in (8)–(11). With them, the QoE increase obtained by reassigning more resources to an under-prioritized service is often larger than the loss of QoE caused by taking those resources from over-prioritized services that are receiving more resources than the strictly needed.

4 Performance analysis

In the previous section, two balancing algorithms based on re-prioritizing services have been presented. Several tests are now carried out to assess their value. For clarity, the simulation setup is first introduced and results are presented later.

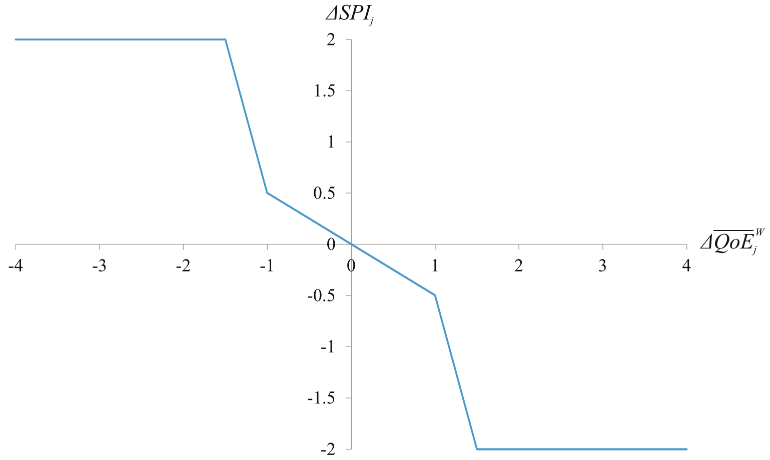

Fig. 2 Controller in the weighted strategy

4.1 Analysis setup

In the absence of an analytical model or a live LTE system, performance assessment is done in a dynamic system-level LTE simulator [33]. The considered macrocellular scenario, shown in Fig. 3, consists of 19 tri-sectorized sites with 57 cells evenly distributed in space. Table 2 shows the main parameters of the simulator. System bandwidth is configured to 6 Physical Resource Blocks (PRBs) to reduce the number of users needed to achieve a high network load, and thus reduce the computational load of simulations. Likewise, a hexagonal cellular layout and uniform spatial distribution have been selected to ease the analysis of results. The reader is referred to [33] for a more detailed explanation of the configuration parameters and the tool itself.

Table 3 shows the traffic mix used in the simulations, which is inspired in [35] and [3]. The average number of users per cell is large enough to ensure that the PRB utilization ratio is close to 100 %, so that services compete for radio resources and service priority has an impact on end user performance.

For repeatability, random variables are pre-generated to ensure that every optimization loop is carried out under exactly the same conditions. Thus, performance differences between loops are only due to changes in the SPI configuration.

Table 4 shows the values of the internal scheduler parameters in this work. The most important one of those parameters, presented in 2.1, is t_c, which has a direct influence on the PF_{factor} and therefore on the scheduling process. The value $t_c = 1.25$ in (6) means that the weight of past history (previous average data rate) is 0.2 and the weight of present (instantaneous achievable data rate) is 0.8.

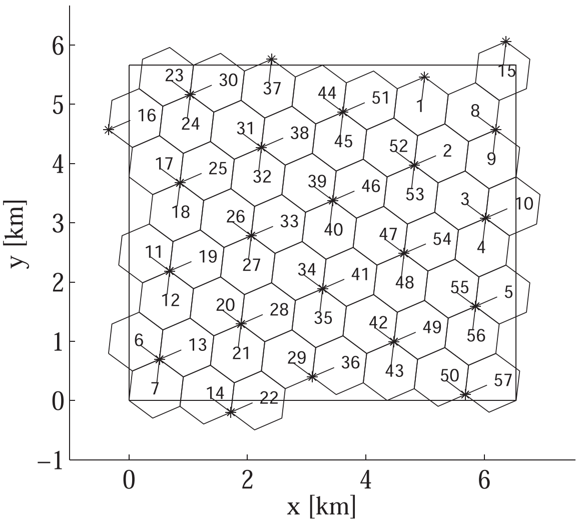

Fig. 3 Simulation scenario

Table 2 Simulation parameters

Cellular layout	Hexagonal grid
	57 cells (3 × 19 sites)
Transmission direction	Downlink
Carrier frequency	2.0 GHz
System bandwidth	1.4 MHz (6 PRBs)
Cell radius	0.5 km
Inter-site distance	1.5 km
Propagation model	Okumura-Hata with wrap around
	Log-normal slow fading, $\sigma = 8$ dB, correlation distance = 20 m
	Multipath fading, ETU model
Mobility model	Random direction, constant speed, 3 km/h
Service model	VoIP: GBR, Poisson arrivals, mean call duration: 60 s
	VIDEO: non-GBR, Poisson arrivals, buffered video, H.264, real traces (http://trace.eas.asu.edu/tracemain.html)
	FTP: non-GBR, Poisson arrivals, mean file size: 2 MB [35]
	WEB : non-GBR, Poisson arrivals, mean web page size/waiting time [35]
Base station model	Tri-sectorized antenna, MIMO 2 × 2, $\text{EIRP}_{\text{max}} = 43$ dBm
Scheduler	EXP/PF multi-service modified with SPI parameter
Power control	Equal transmit power per PRB
Link adaptation	CQI based
RRM features	Handover, call access control
HO parameter settings	TimetoTrigger = 100 ms
	HO margin = 3 dB
User distribution	Uniform spatial distribution
Dropped call model	Radiolinktimeout = 1 s
Time resolution	10 TTI (10 ms)
Simulated network time	2 h (24 optimization loops * 5 min per loop = 120 min)

Three experiments are carried out. The aim of the first experiment is to show how the basic balancing algorithm manages to equalize the QoE across services. For this purpose, the unweighted algorithm is tested with an initial SPI configuration where all services begin with the

Table 3 Traffic mix

Service	QCI (Traffic category)	Share of users (%)
VoIP	1 (Real time)	50
VIDEO	6 (Streaming)	20
WEB	8 (Interactive)	20
FTP	9 (Best effort)	10

Table 4 Scheduler internal parameters

Parameter	Value
t_c	1.25
ε	1
κ	10
$\tau_{\max}$	100
δ_i	0.01

same intermediate SPI value (=7). The aim of the second experiment is to check the impact of the initial SPI configuration. For this purpose, the unweighted algorithm is tested with an initial SPI configuration where the SPI is different for each service. Specifically, $\text{SPI}_{\text{VoIP}} = 2$, $\text{SPI}_{\text{VIDEO}} = 13$, $\text{SPI}_{\text{FTP}} = 5$, and $\text{SPI}_{\text{WEB}} = 3$. The last experiment aims to show how the weighted algorithm manages to prioritize the most populated services. Each experiment consists of 24 optimization loops (5 min per loop). Thus, 2 h of network time are simulated. It is checked a posteriori that, with the proposed controller, and for the initial QoE imbalance between services in the simulated scenario, such a number of loops is enough to ensure that the system reaches equilibrium.

To assess the value of a SPI configuration, several network performance indicators are used. The main figure of merit from the network perspective is the average QoE of the worst service, $\min\{\overline{\text{QoE}_j}\}$ or $\min\{\overline{\text{QoE}_j}^W\}$. Such a choice is consistent with the way operators monitor network QoE, where the worst users receive most of the attention to reduce churn rates. Note that the worst service in equilibrium (i.e., at the end of the optimization process) is not necessarily the same as in the first loop. The selection of the unweighted or weighted variant of the figure of merit depends on the aim of the balancing process. If the aim is to improve fairness among services, regardless of the number of users per service, the unweighted variant must be selected. In contrast, if the aim is to benefit the most populated service, the weighted variant is the proper choice. Other important network performance indicators are the global service QoE, defined in the unweighted version as the arithmetic mean of the average QoE of all services

$$\text{QoE}_{\text{global}} = \frac{1}{N_s}\sum_{j=1}^{N_s}\overline{\text{QoE}_j} \quad , \tag{18}$$

and the maximum QoE difference among services, defined as

$$\Delta\overline{\text{QoE}}_{\max} = \max\left\{\overline{\text{QoE}_j}\right\} - \min\left\{\overline{\text{QoE}_j}\right\} \quad . \tag{19}$$

For the weighted version, the overall QoE is obtained from the average user QoE, computed as

$$\text{QoE}^{\text{W}}_{\text{global}} = \frac{1}{N_T}\sum_{i=1}^{N_T}\text{QoE}_i \quad , \tag{20}$$

where QoE_i is the quality experienced by each user treated and N_T is the total number of connections, and the maximum service QoE difference is computed as

$$\Delta\overline{\text{QoE}}^{W}_{\max} = \max\left\{\overline{\text{QoE}_j}^{W}\right\} - \min\left\{\overline{\text{QoE}_j}^{W}\right\} \quad . \tag{21}$$

Although the above-described network performance indicators are computed in every optimization loop, the focus is on the values obtained at the end of the tuning process. To assess the algorithm from the control perspective, the whole trajectory is evaluated by checking convergence speed and stability. The former is given by the number of loops to reach equilibrium, while the latter is based on the absence of fluctuations in system parameters.

5 Simulation results

The results of the unweighted algorithm are first presented, since they are easier to analyze. The results of the weighted variant are discussed later.

5.1 Unweighted strategy/equal initial SPI

In the first experiment, the initial SPI value for all services is set to an intermediate level (i.e., 7). Figure 4 presents the evolution of the QoE and SPI of each service. In Fig. 4a,

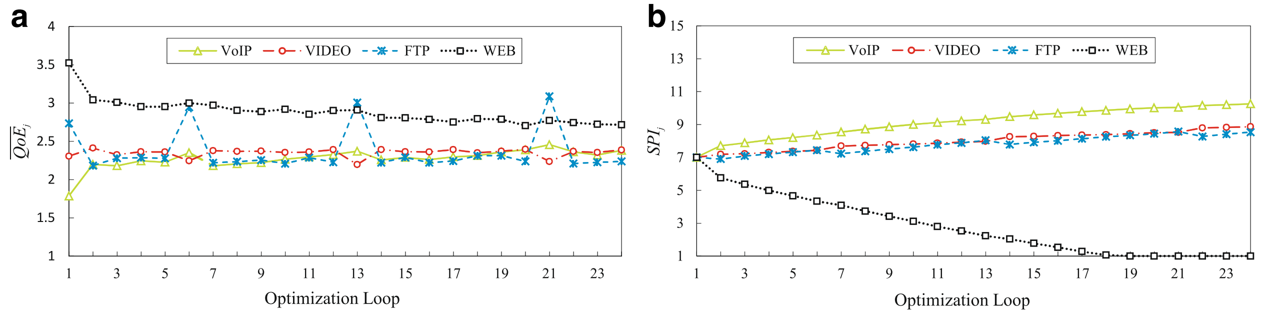

Fig. 4 Results in the unweighted strategy with equal initial SPI settings. **a** Average service QoE. **b** SPI parameter

it is observed that, with the initial SPI settings (i.e., loop 1), the RT service (VoIP) experiences the lowest QoE, whereas the WEB service has the largest QoE. It is inferred that, with the initial configuration, the scheduler benefits WEB by allocating enough resources to download web pages when needed. This is just a consequence of the low throughput threshold in the utility function of WEB, presented in (11), which make it the least restrictive service.

In Fig. 4b, it is observed that, in only three loops, the algorithm already manages to balance the QoE of VIDEO, VoIP, and FTP by reducing the SPI of services with QoE above the average (i.e., WEB) and increasing the SPI of those below the average (i.e., VoIP, VIDEO, and FTP). Thereafter, the algorithm tries to increase the QoE of VoIP, VIDEO, and FTP at the expense of WEB. Even if the SPI of WEB reaches its lower limit, the SPIs of all other services keep increasing.

Peaks in the $\mathrm{QoE}_{\mathrm{FTP}}$ curve are easily explained by observing the evolution of SPI parameters, shown in Fig. 4b. A comparison of both figures reveals that abrupt changes of QoE occur when $\mathrm{SPI}_{\mathrm{FTP}}$ has just become greater than $\mathrm{SPI}_{\mathrm{VIDEO}}$, i.e., when the priority of the video service becomes less than the FTP service. A closer analysis shows that this happens whenever $\mathrm{SPI}_{\mathrm{VIDEO}}$ falls below the SPI of another service. This is due to the fact that the video service occupies more than 50 % of PRBs in the network, which makes the system quite sensitive to changes in the priority of the video service.

Figure 5 compares the evolution of the global QoE against that of the minimum service QoE (primary axis) and maximum QoE difference (secondary axis). As expected, the balancing algorithm achieves nearly a fourfold reduction of the maximum QoE difference from 1.74 to 0.47. As a result, the QoE of the worst service (VoIP, VIDEO, or FTP, depending on the loop) is increased from 1.8 to 2.25. Such an improvement is obtained without changing $\mathrm{QoE}_{\mathrm{global}}$, except for the positive peaks observed in $\mathrm{QoE}_{\mathrm{FTP}}$.

5.2 Unweighted strategy/different initial SPI

To check the influence of initial parameter settings, the unweighted algorithm is initialized with an uneven SPI configuration selected at random, where VIDEO has larger SPI, WEB and FTP have lower SPI, and VoIP has the lowest SPI, close to the minimum.

Figure 6 shows the evolution of QoE and SPI for each service. Again, it is observed that the balancing algorithm manages to reduce QoE differences among services. Nonetheless, WEB remains the best service despite reaching the minimum SPI value (=1) at the beginning of the process. This is because WEB is the least demanding service. From the comparison of both figures, it is deduced that every time $\mathrm{SPI}_{\mathrm{VoIP}}$ crosses the SPI of the other services, the QoE of VoIP increases, especially when crossing the VIDEO service. This was expected since VIDEO is the service demanding the largest amount of resources.

Figure 7 shows the evolution of the three figures of merit. In the figure, it is observed that $\Delta\overline{\mathrm{QoE}}_{\mathrm{max}}$ on the secondary axis decreases by more than six times (from 1.93 to 0.31) after tuning SPIs. Likewise, $\min\{\overline{\mathrm{QoE}_j}\}$ on the primary axis improves 90 % (from 1.18 to 2.25). In addition, the global QoE on the primary axis improves as a result of the balancing process.

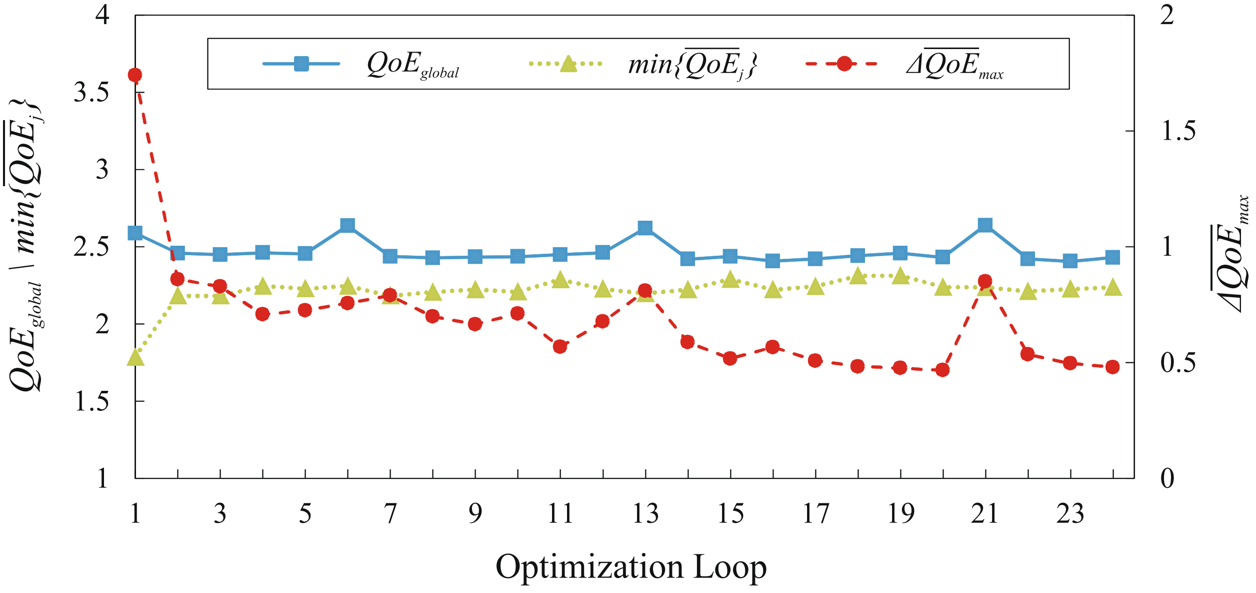

Fig. 5 Figures of merit in the unweighted strategy with equal initial SPI settings

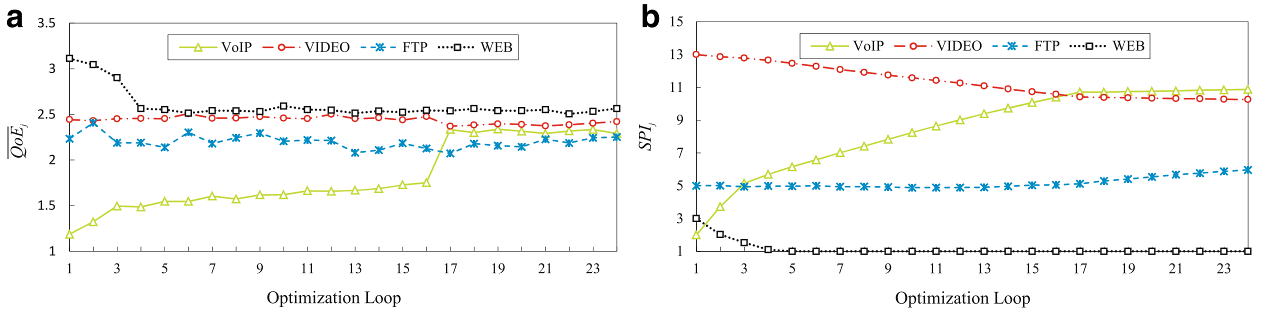

Fig. 6 Results in the unweighted strategy with uneven initial SPI settings. **a** Average service QoE. **b** SPI parameter

5.3 Weighted strategy/equal initial SPI

The last experiment shows how the weighted algorithm improves the average end user experience by prioritizing the most populated services. User ratios in Table 3 show that, in the considered case, VoIP is the service with more users (50 %) and FTP is the one with fewer users (10 %).

Figure 8a shows the evolution of the indicator balanced by the weighted algorithm (i.e., the average service QoE divided by the number of users per service), while Fig. 8b presents the SPI configuration trend. It is observed that the SPI of VoIP reaches the maximum value (=15) almost immediately, since VoIP is the most populated service. After that, $\overline{\text{QoE}}_{\text{VoIP}}^{\text{W}}$ barely changes. In fact, the situation is almost stable after loop number 13, since thereafter SPI_{FTP} and SPI_{WEB} stagnate at 1 (i.e., the minimum value) and only $\text{SPI}_{\text{VIDEO}}$ varies very slowly. It is observed that the balancing process reaches saturation in this cas e.

Figure 9 shows the evolution of the global service QoE on the primary axis, and the maximum weighted QoE difference, $\Delta\overline{\text{QoE}}_{\text{max}}^{\text{W}}$, and the minimum weighted QoE, $\min\{\overline{\text{QoE}_j}^{W}\}$, on the secondary axis. As expected, $\Delta\overline{\text{QoE}}_{\text{max}}^{\text{W}}$ is halved at the end of the balancing process. Likewise, the minimum weighted service QoE improves by 70 % as a result of increasing the priority of the most populated service (i.e., VoIP). A beneficial side effect is that $\text{QoE}_{\text{global}}{}^{\text{W}}$ improves by 15 %, especially in the first iterations.

The case of weighted strategy and uneven initial SPI settings (not shown here for brevity) gives the same conclusions.

6 Conclusions

In this paper, a self-tuning algorithm for adjusting parameters in a multi-service packet scheduler of a LTE base

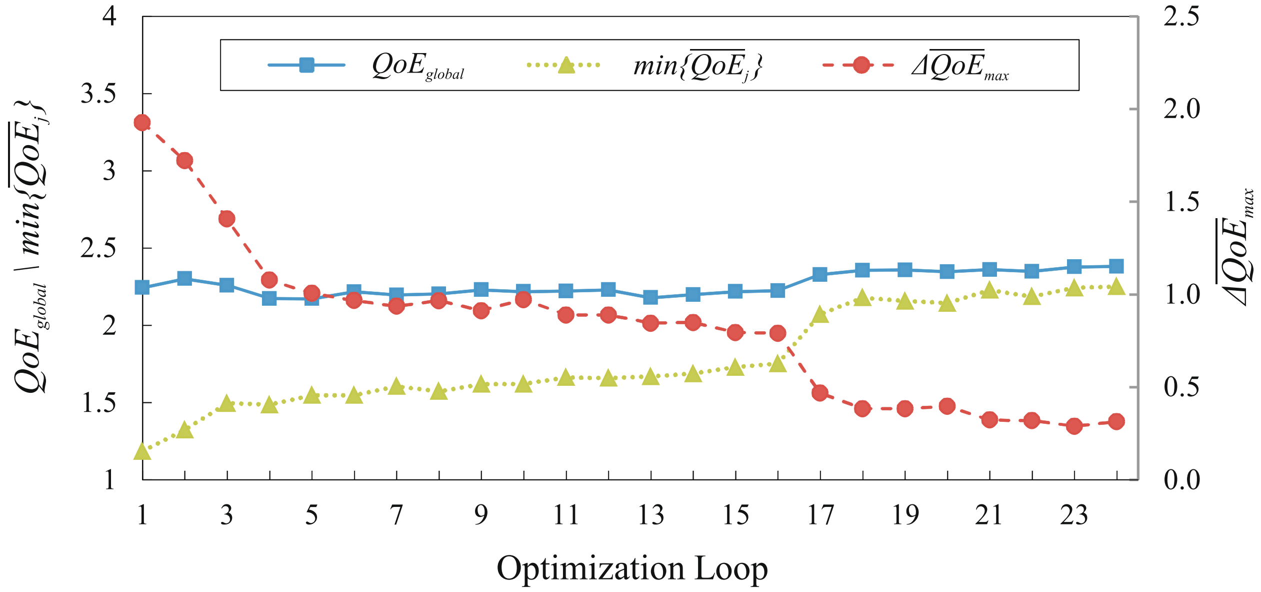

Fig. 7 Figures of merit in the unweighted strategy with uneven initial SPI settings

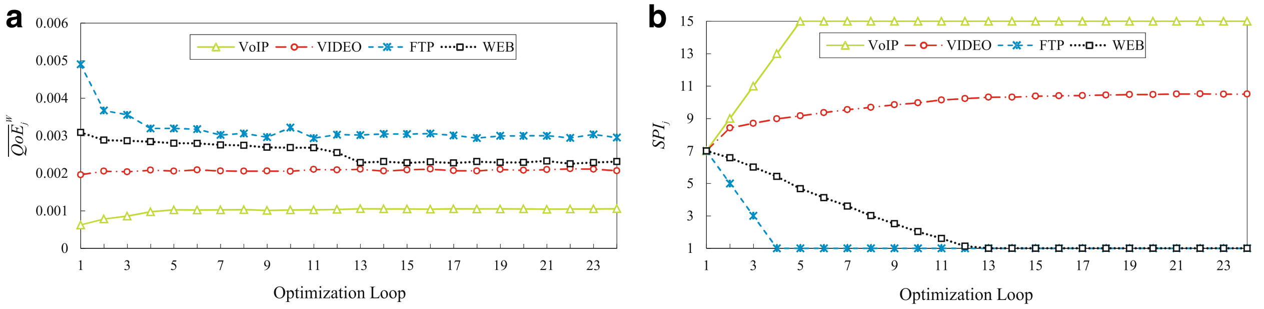

Fig. 8 Results in the weighted strategy. **a** Average service QoE. **b** SPI parameter

station based on network statistics has been proposed. The aim of the algorithm is to ensure that all users achieve the same average QoE regardless of the type of service. For this purpose, the proposed iterative algorithm changes service priority parameters to re-prioritize services so as to equalize the QoE among services. Controlling QoE instead of QoS makes it easier to compare services with very different QoS constraints. Two variants of the algorithm have been presented, depending on whether the aim is to improve the average service QoE (unweighted approach) or the average user QoE (weighted approach).

Method assessment has been carried out in a dynamic system-level LTE simulator implementing the downlink in a regular scenario. Results have shown that the unweighted version of the algorithm can equalize the QoE of services by changing the service priority parameter from different initial settings. Thus, the QoE of the worst service is doubled by re-prioritizing services properly. Likewise, the weighted version of the algorithm improves the average user QoE by 15 % by increasing the priority of the most populated services. The unweighted strategy is the preferred one if fairness among services is desired, regardless of the number of users per service. On the other hand, the weighted strategy should be selected when the aim is to favor the most populated service.

It should be pointed out that, if the considered utility functions were others, a different situation might be reached at the end of the SPI adjustment process. Generally, a more optimistic utility function for a service, showing a higher MOS with the same QoS, would lead to a decrease in the SPI of that service and, hence, a lower service priority. However, it is worth noting that the balancing algorithm would remain the same, regardless of the utility functions.

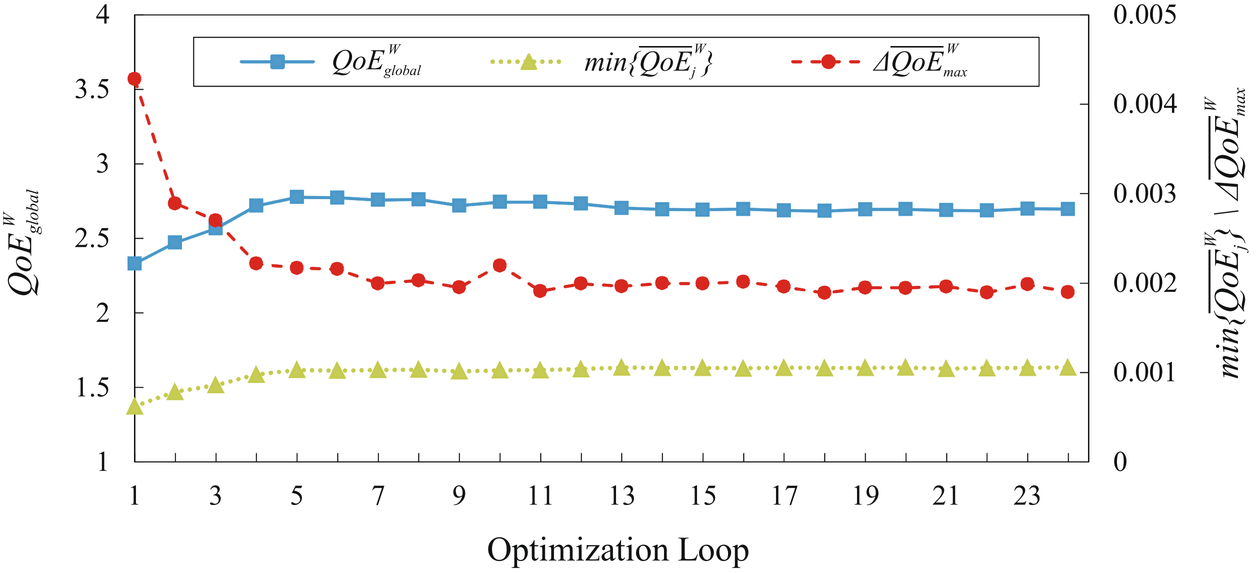

Fig. 9 Figures of merit in the weighted strategy

It is left for future work to design more sophisticated controllers that ensure optimal network performance by applying classical optimization techniques instead of simple balancing rules. It is also intended to analyze how the proposed controller influences QoS metrics.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work has been funded by the Spanish Ministry of Economy and Competitiveness (TIN2012-36455), and Optimi-Ericsson, Agencia IDEA (Consejería de Ciencia, Innovación y Empresa, Junta de Andalucía , ref. 59288) and FEDER.

Author details

[1]Ingeniería de Comunicaciones, Universidad de Málaga, Campus de Teatinos S/N, 29071 Malaga, Spain. [2]Ericsson, Severo Ochoa 51, 29590 Málaga, Spain.

Received: 29 July 2015 Accepted: 20 December 2015
Published online: 07 January 2016

References

1. D Soldani, SK Das, M Hassan, JA Hassan, GD Mandyam, Traffic management for mobile broadband networks. IEEE Commun. Mag. **49**(10), 98–100 (2011)
2. A Banerjee, Revolutionizing CEM with subscriber-centric network operations and QoE strategy. White paper, Heavy Reading (2014)
3. Next Generation Mobile Networks Recommendation on SON and O&M requirements. Technical report, NGMN (2008)
4. J Ramiro, K Hamied, *Self-Organizing Networks (SON): Self-Planning, Self-Optimization and Self-Healing for GSM, UMTS and LTE*. (Wiley, UK, 2011)
5. S Hämäläinen, H Sanneck, C Sartori, *LTE Self-Organizing Networks (SON): Network Management Automation for Operational Efficiency Hardcover*. (Wiley, UK, 2012)
6. Use Cases related to Self Organising Network, Overall Description. Technical report, NGMN (2007)
7. 3GPP TR 36.902, *LTE; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions*. (V9.2.0, ETSI, Sophia Antipolis Cedex, France, 2010)
8. 4G Americas, Self-optimizing networks in 3GPP Release 11: The benefits of SON in LTE. Technical report (2013)
9. H Holma, A Toskala, *LTE for UMTS-OFDMA and SC-FDMA Based Radio Access*. (Wiley, UK, 2009)
10. KI Pedersen, TE Kolding, F Frederiksen, IZ Kovács, D Laselva, PE Mogensen, An Overview of Downlink Radio Resource Management for UTRAN Long-Term Evolution. IEEE Commun. Mag. **47**(7), 86–93 (2009)
11. FRM Lima, TF Maciel, WC Freitas, FRP Cavalcanti, Resource Assignment for rate maximization with QoS guarantees in multiservice wireless systems. IEEE Trans. Veh. Technol. **61**(3), 1318–1332 (2012)
12. S Sesia, I Toufik, M Baker, *LTE, the UMTS Long Term Evolution: from Theory to Practice*. (Wiley, USA, 2009)
13. R Kwan, C Leung, J Zhang, Proportional fair multiuser scheduling in LTE. IEEE Signal Proc. Lett. **16**(6), 461–464 (2009)
14. R Kwan, C Leung, J Zhang, Multiuser scheduling on the downlink of an LTE cellular system. Res. Lett. Commun. **2008** (2008). doi:10.1155/2008/323048
15. RK Almatarneh, MH Ahmed, OA Dobre, in *IEEE 72nd Vehicular Technology Conference Fall (VTC 2010-Fall)*. Performance Analysis of Proportional Fair Scheduling in OFDMA Wireless Systems (IEEE, Ottawa, ON (Canada), 2010), pp. 1–5
16. L Wang, AH Aghvami, in *IEEE Global Telecommunications Conference (GLOBECOM '99)*. Optimal power allocation based on QoS balance for a multi-rate packet CDMA system with multimedia traffic, vol. 5 (IEEE, Rio de Janeiro, Brazil, 1999), pp. 2778–2782
17. N Bansal, KR Pruhs, Server scheduling to balance priorities, fairness, and average quality of service. SIAM J. Comput. **39**(7), 3311–3335 (2010)
18. H Ackermann, S Fischer, M Hoefer, M Schöngens, Distributed algorithms for QoS load balancing. Distrib. Comput. **23**(5–6), 321–330 (2011)
19. T Farkhondeh, YS Chan, JJ Lee, Scheduling with Quality of Service Support in Wireless System. Google Patents. EP Patent 2,277,329 (2014). http://www.google.com/patents/EP2277329B1?cl=en. Access date: October 2014
20. 3GPP TS 23.203, Technical Specification Group Services and System Aspects; Policy and charging control architecture. (V13.1.0, ETSI, Sophia Antipolis Valbonne, France, 2014)
21. J-H Rhee, JM Holtzman, D-K Kim, in *IEEE 57th Semiannual Vehicular Technology Conference (VTC 2003-Spring)*. Scheduling of Real/Non-real Time Services: Adaptive EXP/PF Algorithm, vol. 1 (IEEE, Jeju, Korea, 2003), pp. 462–466
22. X Li, Y Zaki, Y Dong, N Zahariev, C Goerg, in *IEEE 6th Joint IFIP Wireless and Mobile Networking Conference (WMNC)*. SON Potential for LTE Downlink MAC Scheduler (IEEE, Kyoto, Japan, 2013), pp. 1–7
23. S Khan, S Duhovnikov, E Steinbach, W Kellerer, MOS-Based Multiuser Multiapplication Cross-Layer Optimization for Mobile Multimedia Communication. Adv. Multimed. **2007**(1) (2007). doi:10.1155/2007/94918
24. P Ameigeiras, JJ Ramos-Munoz, J Navarro-Ortiz, P Mogensen, JM Lopez-Soler, QoE oriented cross-layer design of a resource allocation algorithm in beyond 3G systems. Comput. Commun. **33**(5), 571–582 (2010)
25. S Thakolsri, W Kellerer, E Steinbach, in *IEEE International Conference on Communications (ICC)*. QoE-based cross-layer optimization of wireless video with unperceivable temporal video quality fluctuation, (2011)
26. M Shehada, S Thakolsri, Z Despotovic, W Kellerer, in *IEEE 14th International Symposium on Wireless Personal Multimedia Communications (WPMC)*. QoE-based Cross-Layer Optimization for Video Delivery in Long Term Evolution Mobile Networks (IEEE, Brest, France, 2011), pp. 1–5
27. A El Essaili, L Zhou, D Schroeder, E Steinbach, W Kellerer, in *IEEE 13th International Workshop on Multimedia Signal Processing (MMSP)*. QoE-driven Live and On-demand LTE Uplink Video Transmission (IEEE, Hangzhou, China, 2011), pp. 1–6
28. F Wamser, D Staehle, J Prokopec, A Maeder, P Tran-Gia, in *Proceedings of the 24th International Teletraffic Congress (ITC), no. 15*. Utilizing Buffered Youtube Playtime for QoE-Oriented Scheduling in OFDMA Networks (International Teletraffic Congress (ITC), Kraków, Poland, 2012)
29. P Patras, A Banchs, P Serrano, A control theoretic approach for throughput optimization in IEEE 802.11e EDCA WLANs. Mob. Netw. Appl. **14**(6), 697–708 (2009)
30. W He, K Nahrstedt, X Liu, End-to-end delay control of multimedia applications over multihop wireless links. ACM Trans. Multimed. Comput. Commun. Appl. (TOMCCAP). **5**(2) (2008). doi:10.1145/1413862.1413869
31. H Luo, M-L Shyu, Quality of service provision in mobile multimedia-a survey. Human-centric Comput. Inf. Sci. **1**(1), 1–15 (2011)
32. D Soldani, HX Jun, B Luck, in *IEEE 73rd Vehicular Technology Conference (VTC 2011-Spring)*. Strategies for Mobile Broadband Growth: Traffic Segmentation for Better Customer Experience (IEEE, Budapest, Hungary, 2011), pp. 1–5
33. P Muñoz, I de la Bandera, F Ruiz, S Luna-Ramírez, R Barco, M Toril, P Lázaro, J Rodríguez, Computationally-efficient design of a dynamic system-level LTE simulator. Int J. Electron. Telecommun. **57**, 347–358 (2011)
34. P Seeling, M Reisslein, Video transport evaluation with H.264 video traces. IEEE Commun. Surv. Tutor. **14**(4), 1142–1165 (2012)
35. 3GPP TSG-RAN1#48, *LTE physical layer framework for performance verification*. (R1-070674, St. Louis, MI, USA, 2007)
36. G Gómez, J Lorca, R García, Q Pérez, Towards a QoE-Driven Resource Control in LTE and LTE-A Networks. J. Comput. Netw. Commun. (2013). doi:10.1155/2013/505910
37. International telephone connections and circuits – General Recommendations on the transmission quality for an entire international telephone connection; One-way transmission time. (ITU-T Recommendation G.114, Geneva, Switzerland, 2003)
38. International telephone connections and circuits – General definitions; *The E-model, a computational model for use in transmission planning*. (ITU-T Recommendation G.107, Geneva, Switzerland, 1998), p. 8
39. RK Mok, EW Chan, RK Chang, in *IFIP/IEEE International Symposium on Integrated Network Management (IM)*. Measuring the Quality of Experience of HTTP Video Streaming (IEEE, Dublin, Ireland, 2011), pp. 485–492
40. J Navarro-Ortiz, JM Lopez-Soler, G Steay, in *IEEE European Wireless Conference (EW)*. Quality of Experience Based Resource Sharing in IEEE 802.11e HCCA (IEEE, Lucca, Italy, 2010), pp. 454–461