Abrignani *et al. EURASIP Journal on Wireless Communications and Networking* (2018) 2018:193
https://doi.org/10.1186/s13638-018-1206-2

EURASIP Journal on Wireless Communications and Networking

REVIEW Open Access

Scheduling M2M traffic over LTE uplink of a dense small cell network

Melchiorre Danilo Abrignani[1*], Lorenza Giupponi[2], Andrea Lodi[3] and Roberto Verdone[1]

Abstract

We present an approach to schedule Long Term Evolution (LTE) uplink (UL) Machine-to-Machine (M2M) traffic in a densely deployed heterogeneous network, over the street lights of a big boulevard for smart city applications. The small cells operate with frequency reuse 1, and inter-cell interference (ICI) is a critical issue to manage. We consider a 3rd Generation Partnership Project (3GPP) compliant scenario, where single-carrier frequency-division multiple access (SC-FDMA) is selected as the multiple access scheme, which requires that all resource blocks (RBs) allocated to a single user have to be contiguous in the frequency within each time slot. This adjacency constraint limits the flexibility of the frequency-domain packet scheduling (FDPS) and inter-cell interference coordination (ICIC), when trying to maximize the scheduling objectives, and this makes the problem NP-hard. We aim to solve a multi-objective optimization problem, to maximize the overall throughput, maximize the radio resource usage and minimize the ICI. This can be modelled through a mixed-integer linear programming (MILP) and solved through a heuristic implementable in the standards. We propose two models. The first one allocates resources based on the three optimization criteria, while the second model is more compact and is demonstrated through numerical evaluation in CPLEX, to be equivalent in the complexity, while it performs better and executes faster. We present simulation results in a 3GPP compliant network simulator, implementing the overall protocol stack, which support the effectiveness of our algorithm, for different M2M applications, with respect to the state-of-the-art approaches.

Keywords: M2M traffic scheduling, LTE uplink, Dense small cells network, Heterogeneous network, MILP

1 Introduction

Recent studies proposed by CISCO [1] predict that from 2014 to 2019, the traffic in mobile networks will grow by a factor of 10. Machine-type communications (MTC) and Machine-to-Machine (M2M) applications are announced to be one of the factors generating this increment in demand. MTC are defined as a form of data communications which do not need human interaction. Mobile operators like Telnor, Vodafone and Telefonica, to name a few, have created dedicated units or even companies to focus on M2M business opportunities. Large information technology (IT) vendors like IBM or HP also have ambitious plans to connect and exploit information generated by trillions of sensors. The M2M application space is vast and includes security, health monitoring, remote management and control, intelligent transport systems, ambient assisted living, etc. Communication challenges in the field are related with collecting and distributing the data efficiently, often in real time and with desired quality of service (QoS) requirements, in terms of, e.g. latency. The communication network plays an important part of this ecosystem, and its ability to support M2M services and traffic requirements will be crucial.

Cellular networks are expected to provide ubiquitous coverage to these extremely heterogeneous kinds of services and at low deployment costs. This is why significant effort has been lately devoted in standardization, where activities are on going in the 3rd Generation Partnership Project (3GPP) [2], IEEE [3] and European Telecommunications Standards Institute (ETSI) [4].

The advent of MTC, together with the demanding quality of experience (QoE) requirements of data applications, will generate a need for capacity increase, which can only be satisfied by a fundamental rethink of the radio access network, where heterogeneous nodes like remote radio heads (RRH), femto-, pico-, micro-,

*Correspondence: danilo.abrignani@unibo.it
[1]Department of Electrical, Electronic and Information Engineering (DEI), University of Bologna, Bologna, Italy
Full list of author information is available at the end of the article

small cells (SCs) in general, and traditional macrocells coexist in the same area, with an extremely high equipment density [5]. In these densified scenarios for future 5G networks, neighbouring base stations (BSs) most likely operate on the same channel due to the scarcity of spectrum resources, which make radio resource management (RRM) decisions tremendously complex.

In this paper, we focus on an ultra-densely [6] deployed network, where neighbouring base stations operate on the same channel, providing service to a urban scenario in a near-future smart city. We focus on a big boulevard, equipped with an ultra-dense street light small cell deployment, able to support both Human-to-Human (H2H) and M2M traffic. This cost-efficient and self-organized solution has been recently proposed by multiple vendors [5] in order to increase dramatically the density of nodes and to address the morphologies from dense urban to suburban. M2M traffic generated by most services/applications is bi-directional, and the network must be designed to support great amounts of uplink traffic. Furthermore, different applications have different requirements in terms of throughput, maximum tolerable packet loss rate, maximum delay, etc., which requires the implementation of intelligent scheduling algorithms to meet the requirements of all applications.

In this challenging ultra-dense scenario, where multiple M2M applications require satisfaction of their heterogeneous QoS requirements, the Long Term Evolution (LTE) scheduling functionality, located at the BS within the LTE medium access control (MAC) layer, plays a crucial role. It manages the limited radio resources at the access level, in a way that optimizes system performance in terms of a variety of criteria, such as throughput and fairness. The bandwidth is organized onto groups of sub-carriers, denoted as resource blocks (RBs), which are the minimum scheduling resolution in the time-frequency domain. The scheduling functionality performs the RB-to-user equipment (UE) assignment in each transmission time interval (TTI), handling shared radio resources amongst neighbour BSs. Decisions are based on the scheduling policies, taking into account network conditions, wireless channel quality and the QoS experienced by users at the service level, etc. Considerable work has been devoted in the literature to scheduling downlink (DL) traffic in densely deployed heterogeneous networks, considering also inter-cell interference coordination (ICIC) approaches [7]. The study of the uplink (UL), which is expected to be much more loaded in 5G scenarios, with M2M communications, even if it has been approached in traditional macrocell scenarios [8], is much less explored in ultra-dense networks where the component of interference plays a disruptive role.

Scheduling LTE's UL requires making considerations, for example, in terms of UE limited power budget, satisfaction of QoS requirements and enhancement of throughput vs fairness trade-off. Differently from the DL, where LTE adopts orthogonal frequency-division multiple access (OFDMA), the LTE's UL uses a pre-coded version of orthogonal frequency-division modulation (OFDM), called single-carrier frequency-division multiple access (SCFDMA). It helps in solving the undesirable high peak-to-average power ratio (PAPR) of OFDM, which would increase the cost of the UE terminal and drain the battery faster. However, the advantage of low power requirements is largely realized when resource contiguity is enforced in the RB allocations made to a single UE in the UL. This contiguous constraint is sufficient to make the UL LTE problem NP-hard [9].

We provide in this paper a solution for the LTE uplink scheduling problem, taking into account the interference coordination issues and the constraint of adjacency of RBs allocated to a single user. The problem of scheduling resources can be naturally approached through linear optimization tools. As a result, we model our scheduling problem through mixed-integer linear programming (MILP), and we first create a three-step model taking into account the criteria such as the overall throughput, the radio resource usage and the inter-cell interference (ICI). Then, we propose a more compact model, referred in the following as "unified", which solves the multi-criteria scheduling optimization in just one round. Finally, we present a greedy algorithm that solves the model and performs UL scheduling. We solve these problems through the IBM ILOG CPLEX optimization software [10]. We present a comparative study between the proposed models and the greedy algorithm, in order to evaluate the difference between the heuristic and the optimal solutions. We pay particular attention to the feasibility of implementation in the standard and to the time required to achieve a solution, considering that the scheduler has to be executed every TTI (1 ms). Also, the heuristic approach is characterized by low computational requirements, and so, it can be easily implemented in devices with reduced computational capability.

The designed heuristic algorithm has been implemented in a 3GPP compliant, high-fidelity, network simulator, Network Simulator 3 (NS3) LTE-EPC Network Simulator (LENA) [11], supporting the full protocol stack. Simulation results carried out in NS3 show the promising performance of our scheme for different M2M applications and with respect to the state-of-the-art approaches.

The rest of the paper is organized as follows. Section 3 positions this work with respect to the related literature. Section 4 presents the system model. Section 5

formulates the problem, the approach and the meaningful models. Section 6 describes our reference scenario. Section 7 shows the most important numerical results obtained by evaluating the proposed model. Section 8 discusses meaningful simulation results obtained using a network simulator. Finally, Section 9 summarizes the main conclusions.

2 Proposed methodology

As discussed in the previous section, the aim of this study is to provide an approach to schedule Long Term Evolution LTE uplink (UL) Machine-to-Machine (M2M) traffic in a densely deployed heterogeneous network. Small cells are foreseen to be deployed over the street lights of a big boulevard for smart city applications. We use a three-step approach to solve the problem: first of all, we describe theoretically and formalize mathematically the scheduling problem; then, we use a MILP model and solver, in combination with a simulation campaign, to validate the mathematical model. Great emphasis is given to the practical feasibility of the proposed approach and of its execution times, which have to be compatible with LTE MAC processing times. Finally, we define an algorithm which follows the model, and we test it on the NS3 LTE model, which is a popular 3GPP standard compliant network simulator, implementing the full protocol stack and offering the opportunity to obtain end-to-end statistics. The proposed scheduler is compared against the state-of-the-art solutions like round-robin and maximum fairness allocations. Different smart city applications are considered for performance evaluation, like the video surveillance and the traffic monitoring. Statistics are analysed in terms of throughput and end-to-end latency. Details on the system design, simulation and scenario are given in the following sections.

3 Review

In Table 1, we provide a comparison of recent proposals for scheduling LTE UL, based on the following criteria:

- Scenario: It indicates whether the proposed algorithm has been applied in traditional LTE macrocell, single, or multi-cell scenarios, or if it has been designed for application in heterogeneous dense networks.
- ICI: It indicates whether ICI is realistically taken into account. This is important, because ultra-dense deployments can cause many RBs not to be available for allocation.
- Model: It denotes if the solution is based on a model and which one.
- Allocation metrics: It indicates the driving scheduling criteria.
- QoS: It defines the QoS supported by the algorithm.
- Algorithm: It indicates if the solution is optimal or heuristic.
- Contiguous RBs: It denotes whether the model considers the constraint imposed by the implementation of SC-FDMA over the adjacency of RBs assigned to the same user. This is important, because this condition assures consistency with 3GPP. Notice that some authors consider instead OFDMA in the UL, and consequently, they do not consider the condition on the adjacency of the RBs.
- Solving time estimation: It indicates that the contribution evaluates the time to solve the problem, providing an analysis of the same model. This is important to establish the performances of the algorithm and its practical implementability.
- Numerical evaluation: It indicates whether the proposed model has been evaluated as a function of, e.g. the number of variables and constraints, the memory occupancy and the complexity.
- Performance evaluation: It indicates if the system performances of the proposed algorithm have been evaluated on a standard compliant network simulator implementing the complete protocol stack.
- M2M support: It indicates if the proposed scheduling model and algorithm takes into account the peculiarity of M2M traffic.

In this section, we do not analyse the works related to the scheduling of the DL, which are the great majority in the literature, and we only focus on 3GPP compliant solutions.

As it can be observed, the great majority of the works investigate the traditional macrocell scenario, where only single or multiple cells are considered. Only one work refers to a macro-femto scenario, but heterogeneous ultra-dense network deployments have not been considered in the literature, for the specific problem tackled in this paper. Currently, these ultra-dense scenarios, e.g. stadium [12] or street light small cells, are of great interest for industry and standardization bodies, and consequently, innovative solutions have to be studied in these contexts. The ICI issue has only been marginally considered in the literature related with the UL schedulers. It has been studied only in the macrocell scenarios, where consequently the interference problem is less critical than in ultra-dense deployments. As for the model, the allocation metrics and QoS parameters involved in the optimization procedures, the literature offers many interesting readings. Many of them consider SC-FDMA and provide algorithms fulfilling the RB adjacency constraint imposed by this access scheme, but others do not, either because they neglect this issue or because they actually focus on OFDMA as access scheme for the LTE UL. The condition of contiguous RBs makes the problem NP-hard, so that only

Table 1 Related work comparison

Reference	Scenario	ICI	Model	Allocation metric	QoS metric	Algorithm	Contiguous RBs	Solving time estimation	3GPP compliant sys. level simulator	M2M support
[28]	M/Mu	Y	MIP	Channel-aware	Fairness	H	Y	N	N	N
[29]	M/S	N	N	Channel-aware	QoS class	H	Y	N	N	N
[9]	M/F	Y	Markov chain	Max. throughput	Fairness	H	Y	N	N	N
[30]	M/Mu	Y	N	Multi-cell channel-aware	Fairness	H	N	N	N	N
[31]	M/S	N	N	Channel-aware	Many	H	Y	N	N	N
[13]	M/Mu	N	Search tree	Fairness	Max. profit	O	Y	N	N	N
[32]	M/Mu	N	MIP	Channel-aware	Maximization profit	H	Y	N	Y	N
[33]	M/S/Mu	N	Y	Channel-aware	QoS class	H	N	N	N	N
[34]	M/S	N	Game theory	Max. throughput	Max. throughput	P	N	N	N	N
[35]	S	N	N	Max. throughput	Max. throughput	H	Y	N	N	N
[14]	Mu	N	N	Group-based	Delay	H	NS	N	N	Y
[36]	S	N	N	Channel-aware	Delay	H	N	N	N	Y
[16]	S	N	N	Semi-static	Max. throughput	H	Y	N	Y	Y
[17]	S	N	N	Aware bit-rate	QoS class	H	NS	N	N	Y
[20]	S	N	N	App specific parameters	Quality of video (QoV)	H	NS	N	N	Y
[15]	S	N	N	Channel-aware M2M/H2H	QoS	H	Y	N	Y	Y
[18]	S	N	MIP	Channel-aware M2M/H2H	QoS	H	Y	N	N	Y
[19]	S	N	MIP	Channel-aware M2M/H2H	QoS	H	Y	N	N	Y

M macrocell, *F* femtocell, *S* single cell, *Mu* multi-cell, *H* heuristic, *P* polynomial, *O* optimal, *NS* not specified

heuristic solutions can be provided. Calabrese et al. [13] provide a solution based on search tree model applied to groups of RBs. This solution is optimal, but the algorithm requires fulfilling constraints on the tree. Regarding the M2M support, [14, 15] support MTC, and only a part of them also supports H2H traffic, e.g. [16, 17] or [18, 19], that in addition consider the energy efficiency problem for M2M devices. The remaining contributions are specific solutions for M2M scheduling, like [20], which presents a radio resource assignment (RRA) and method for video surveillance systems. Amongst these works, only few of them have evaluated the performance of the proposed approaches in a 3GPP-oriented simulator. Finally, to the best of the authors' knowledge, the works presented in the literature only focus on system performance analysis without first providing a numerical evaluation of the performance of the proposed models. As a result, it is impossible to evaluate whether the solutions proposed in the literature are, for example, characterized by a reasonable solution

time, which makes them actually implementable in the standard.

Taking into account the above observations, this paper introduces the following novelties with respect to the state of the art:

1. It proposes a scheduling solution for an ultra-dense scenario.
2. It provides support to M2M traffic.
3. It proposes two MILP models for scheduling LTE UL of an ultra-dense heterogeneous network, characterized by high frequency reuse and high ICI.
4. The two proposed models aim at maximizing the overall throughput, optimizing the radio resource usage and minimizing the ICI.
5. The SC-FDMA is considered and its implementation constraints.
6. The proposed models are analysed and solved, and the performance is compared to a greedy solutions. The solving time is evaluated, in order to deduce the real feasibility of the proposed approach.
7. The greedy solution, i.e. the proposed heuristic algorithm, has been implemented in NS3 and evaluated against the state-of-the-art algorithms.
8. Two variations of the greedy algorithm are presented and implemented in NS3. Those algorithms are compared with maximum fairness (MF) and round-robin (RR) algorithms.

4 System model

We consider a heterogeneous ultra-dense cellular network composed of a set of $\mathcal{M}$ nodes, ranging from traditional macro to SCs. The $M{=}|\mathcal{M}|$ cells provide coverage over a highly capacity-demanding 5G network. All the cells operate in the same frequency band, which allows to increase the spectral efficiency per area through spatial frequency reuse. A SC-FDMA 3GPP LTE UL is considered, where the system bandwidth B is divided into m RBs, with $B = m \cdot B_{\mathrm{RB}}$. A RB represents one basic time-frequency unit that occupies the bandwidth B_{RB} over a TTI, equal to 1 ms. The RB is the smallest resource that can be assigned to a UE. Associated with each BS is n UEs, which at every TTI have to be scheduled onto the set of available RBs.

We aim at designing the multi-user resource assignment that distributes the m RBs amongst the n users, focusing on a frequency-domain packet scheduling (FDPS) model. We do not consider that MTC devices are enabled with dynamic power control. MTC devices are designed to be low-cost and low-complexity devices. Hence, it is reasonable that some features are missing. However, to get a more realistic scenario, in the simulations, different transmit power levels have been set according to different applications. For instance, devices that are transmitting a video surveillance streaming are more likely to be above the ground level; hence, those can be set with a lower transmit power level with respect to the devices that are more likely to be deployed under the ground level, such as smart meters or traffic sensors. A generic user j generates a profit p_j, and we aim at satisfying it by maximizing the overall profit. We assume the coherence time of the channel to be larger than a TTI, so that channel conditions are constant over a TTI. We impose the condition of contiguous RB allocation to the same user. The number of assigned RBs per user is flexible and spans between 0 and m.

The scheduling is carried out taking into account the information transmitted by the user in the UL over the Signalling Radio Bearers (SRB): scheduling requests (SR), to distinguish active users with data in buffers from idle users; Buffer Status Reports (BSR), to inform the BS about the amount of data needed to be transmitted; power headroom reports (PHR), to inform the BS about the available power at the user for the scheduling; sounding reference signal (SRS), used to provide information on the UL channel quality; and channel quality indicator (CQI), to measure the channel quality between UE and BS.

In addition to this, and in order to take into account the high level of interference that exists in an ultra-dense network with high frequency reuse, we propose:

- ICI phase: First, based on the measurements carried out by the same BS and on those of the users, the BS evaluates the blocks of contiguous available RBs, according to the measured interference. Other information exchanged over the X2 interface, such as the high interference indicator (HII) or the overload indicator (OI) [21, 22], can also be used to extract this information.
- Scheduling phase: Based on the availability of contiguous RBs, on the quality of the channel and on the QoS requirements of the traffic to be scheduled, RBs are properly allocated to the users.

5 MILP model for the scheduler

In this section, we describe our proposed approach to schedule LTE UL traffic in an ultra-dense heterogeneous network. We first carry out a three-step optimization process, driven by multiple objectives: the maximization of the overall throughput, the minimization of the radio resource usage and the minimization of the ICI. Successively, we generate a one-step model (the so called "unified" model), which considers all the previous objectives. These optimizations can be stated as MILP problems, as they contain integer and continuous variables, linear constraints and a linear objective function. The problem was already demonstrated to be NP-hard [9], and although each MILP is not extremely hard in practice, solving many of them might not be compatible with

real-time packet scheduling. Consequently, we propose in the second part of this section a greedy algorithm, which solves the optimization problem in computing times that are compatible with the application at hand. This algorithm has been designed paying special attention to the feasibility of implementation in the LTE and Long Term Evolution-Advanced (LTE-A) standard and to its solving time, considering that it has to be executed every TTI. In addition, the computational cost is low, which assures that it can be implemented also in devices with reduced computational capability. Instead, the MILP approach is used as a reference to evaluate the effectiveness of the greedy algorithm, see Section 7.1.

Before describing in details our MILP models and approaches, in Section 5.1, we introduce a set of definitions common to both the proposed MILP models.

In the algorithms' pseudocode (Algorithms 1 and 2), we use capital letters to reference matrix and arrays, e.g. F is a matrix, and F_i is the array identified by the ith row of the same matrix; scalar variables are identified by lower case letters, e.g. $f_{i,j}$ represents the element identified by i and j in matrix F .

5.1 Definitions

We introduce a binary variable $x_{i,j}$ to define the allocation of RB i to user j, namely,

$$x_{i,j} = \begin{cases} 1 & \text{if RB } i \text{ is the first assigned to UE } j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

In addition, we construct the bi-dimensional matrix F whose (constant) entry $f_{i,j}$ gives the minimum number of contiguous RBs (h) needed by user j to satisfy its traffic, under the hypothesis that RB i is assigned to user j as a first RB, i.e. if $x_{i,j} = 1$. Otherwise, for example, if it is not possible to use RB i as a first RB for user j, then $f_{i,j} = -1$.

$$f_{i,j} = \begin{cases} h : \sharp\text{of contiguous RBs assigned to UE j} & \text{if } x_{i,j} = 1 \\ -1 & \text{otherwise} \end{cases} \quad (2)$$

The value of $f_{i,j}$ depends both on the channel conditions and user demand. The procedure used to compute $f_{i,j}$ is described in Algorithm 1. This algorithm requires knowledge about (1) the number of contiguous RBs available starting from RB i, which is summarized by the vector $\mathrm{Av_{RB}}$, and which depends on the ICI conditions, and (2) the maximum modulation and coding scheme (MCS) allowed in each RB, which is contained in vector $\mathrm{MCS_{RB}}$. The ICI phase allows to derive the signal-to-interference-plus-noise ratio (SINR) associated with each RB, and assuming a target block error rate (BLER) of 10%, the $\mathrm{MCS_{RB}}$ can be easily calculated [11, 23]. The function $g(mcs, h)$ is used to determine the capacity c_i of the contiguous RBs starting in i, i.e. transport block (TB) size. This is defined by 3GPP through a lookup table [24, 25]. Finally, we define

$$b^j_{i,k} = \begin{cases} 1 & \text{if UE } j \text{ uses RB } i \text{ and } x_{k,j} = 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

by calculating $b^j_{i,k}$, we obtain for each user j a two-dimensional matrix where the kth column has $f_{k,j} = h$ values set to 1 and all the other values set to 0. In other words, if $x_{k,j}$ is equal to 1, all the values in the range $\left[b^j_{k,k}, b^j_{k+h-1,i}\right]$ must be set to 1. Notice that the entries $f_{i,j}$ (2) and $b^j_{i,k}$ (3) are constant, and they are calculated every TTI before solving the model.

Algorithm 1 algorithm to create $f_{i,j}$

```
{Initialization}
AV_RB ← array of available RB
MCS_RB ← array of maximum available MCS per RB
Define D ← array of the demand
Initialize f_{i,j} = −1 ← for all f_{i,j}
for j = 1 to n do
  for i = 1 to m do
    mcs = MCS_RB[ i]
    h = 1
    while h ≤ AV_RB[ i] do
      mcs = min{mcs, MCS_RB[ i + h − 1] }
      c_i = g(mcs, h)
      if d_j ≤ c_i then
        f_{i,j} = h
        BREAK
      else
        h ← h + 1
        CONTINUE
      end if
    end while
  end for
end for
```

5.2 MILP models

The three-step optimization process based on the first MILP model is described as follows:

1. Throughput maximization: The first objective is set by Eq. 4, where we aim at maximizing the overall served traffic, i.e. the amount of bytes transmitted during each TTI. The optimization is then characterized by three scheduling constraints: (i) *exclusivity*, a single RB cannot be used by more than one user, i.e. each RB can be allocated at most to one user j. This is captured in constraint (5) and in (7); (ii) *interference avoidance*, a user j cannot be allocated to an RB where an unacceptable level of interference

has been detected. This is reflected in constraint (6); and (iii) *adjacency*, all the RBs allocated to the user j have to be contiguous. This is described by constraint (7). The formulation of the first optimization step is then given by:

$$\max \sum_{j=1}^{n} p_j \sum_{i=1}^{m} x_{i,j} \tag{4}$$

subject to:

$$\sum_{i}^{m} x_{i,j} \leq 1,\ j = 1, \ldots, n \tag{5}$$

$$f_{i,j} x_{i,j} \geq 0;\ j = 1, \ldots, n;\ i = 1, \ldots, m \tag{6}$$

$$\sum_{j=1}^{n} \sum_{k=i}^{m} b^{j}_{i,k} x_{k,j} \leq 1,\ i = 1, \ldots, m \tag{7}$$

2. Minimization of allocated RBs: Once the first optimization has been carried out, and the served throughput is maximized, we aim at minimizing the number of allocated RBs, and consequently the radio resource usage. This means that amongst all the allocations, which maximize the profit, we select the one that minimizes the number of allocated RBs. This optimization is described in Eq. (8). The optimization is characterized by four constraints, three of them are the same as for the previous optimization: (i) *exclusivity*, (ii) *interference avoidance*, (iii) *adjacency* and (iv) *satisfied profit*, i.e. the optimization has to satisfy at least the same profit P, as achieved by the first optimization. This is captured in constraint (9). The formulation of the second optimization step is then given by:

$$\min \sum_{j=1}^{n} \sum_{i=1}^{m} f_{i,j} x_{i,j} \tag{8}$$

subject to: (5)–(7) and

$$\sum_{j=1}^{n} p_j \sum_{i=1}^{m} x_{i,j} \geq P \tag{9}$$

3. ICI minimization: Amongst the possible allocations maximizing the throughput and minimizing the number of assigned RBs, the third step aims at finding the best possible configuration in terms of ICI through the minimization of the *utilization factor* $r_{i,j}$ defined as:

$$r_{i,j} = \frac{d_j}{\sum_{k=i}^{i+f_{i,j}-1} c_i} \tag{10}$$

This is defined as the ratio between the demand of user j, d_j, and the corresponding TB size (i.e. c_i). This assures that the demand of the scheduled users is transmitted through the best TB, in terms of MCS and/or number of RBs, so as to achieve a reduced power spectral density (PSD) per RB. The optimization is characterized by five constraints, four of them have been defined in the two previous steps: (i) *exclusivity*, (ii) *interference avoidance*, (iii) *adjacency*, (iv) *satisfied profit* and (v) *minimum number of RBs*, i.e. the profit P has to be served through the same amount of RBs R, as computed through the second optimization process.

$$\min \sum_{j=1}^{n} \sum_{i=1}^{m} r_{i,j} x_{i,j} \tag{11}$$

subject to: (5)–(7), (9) and

$$\sum_{j=1}^{n} \sum_{i=1}^{m} f_{i,j} x_{i,j} \leq R \tag{12}$$

Finally, we present a "unified" model that aims at solving the same multi-criteria optimization problem addressed by the three-step approach, but using only one optimization step. It uses a set of positive integer coefficients to weight the three components of the objective function. The optimization can still be stated as a MILP problem.

$$\max \sum_{j=1}^{n} \sum_{i=1}^{m} (\alpha p_j - \beta f_{i,j} - \gamma r_{i,j}) x_{i,j} \tag{13}$$

subject to: (5)–(7)

This unified MILP model has some computational advantages in terms of compactness with respect to the three-step MILP approach. Computational experiments show that it is not significantly more difficult than each of the three MILPs in isolation; thus, it is clearly preferable because only one solution step is needed. Nevertheless, it is worth noting that the two methods are equivalent if and only if parameters α, β and γ are carefully selected so as to determine an order for the three objective functions. The introduction of the above parameters offers a novel flexibility to the model, which allows also to give combined levels of priority to the different objectives. Moreover, by normalizing the user profit p_j and being the maximum theoretical value $f_{i,j}$ equal to m, it is easy to set α, β and γ in order to satisfy the desired priority between the objective function parameters. Note that this is always true regardless the real meaning of p_j since it is a normalized value. It is worth mentioning that the use of weight in multi-objective optimization problems is widely accepted and commonly used in many fields of knowledge, not only in engineering but also in medical and genetic studies, economics and finance, and in general for ranking and classification problems where each term of the utility function has a different priority. In this respect, the interested reader is referred to the interesting

survey [26]. To select α, β and γ, we have run extensive simulations that we could not reproduce due to the space constraints and to the risk of being out of scope. Empirically, we have found that the most satisfactory trade-off is the one represented by the values selected selected later in the results section (Section 7). All these aspects are evaluated through the use of the IBM-CPLEX MILP solver [10].

5.3 Proposed algorithm

As anticipated, solving MILPs might not be computationally feasible in the real-time scheduling application at hand. As a result, we propose a greedy solution to solve the optimization problem described in the first part of this section. The pseudocode is reported in Algorithm 2. The algorithm's inputs are the same as those defined by the model, i.e. the profit of user j, p_j and the matrix F, computed by using Algorithm 1. During each TTI, all the active users, i.e. users that have data packets to transmit, compete to be scheduled. Users are sorted by their profit. They are scheduled starting from the user with the higher profit. The proposed algorithm is composed by two nested loop. The outer one selects the user with higher profit and stops when either there are no more users to be scheduled or there are no more RBs to assign. The inner loop, given a user, i.e. *cu*, selects the smallest available set of RBs from F that is able to satisfy the user demand. This loop stops when the user is scheduled or no more RBs are available for the user. Finally, the function *isFeasible()* verifies the *exclusivity* constraint, i.e. whether a set of contiguous RBs can be assigned to a user.

For sake of clarity, the MILP models and the proposed algorithm remain valid regardless of the kind of traffic considered, e.g. M2M, H2H or combination of both. In fact, what the algorithm is maximizing is the profit function that could be designed to take into account the differences in term of QoS requirements. We present some results in this sense in Section 8.4.

Algorithm 2 Greedy algorithm

```
{Input:}
P ← array of the profit
F ← Matrix defined in Algorithm 1
{Variables:}
Define int rb_availables ← total available RBs
Define int rb_assigned = 0 ← assigned RBs
UA = ∅ ← Set of User scheduled
sort(p_j, descending)
while UA ≤ Max Active Users or rb_assigned ≠ rb_available
do
  cu := User with the higher profit
  while ∃ a set of RBs in F for User cu do
    cc := smallest available set of RBs for user cu in F
    if isFeasible(cu, cc) then
      rb_assigned+ = cc
      UA = UA ∪ {cu}
      BREAK
    else
      remove cc from F
    end if
  end while
end while
```

6 Network setup

In this section, we first describe the high-level scenario and reference system architecture. We then define the simulation setup and meaningful M2M applications and their simulation model.

6.1 Reference system architecture

We focus on a smart city scenario where M2M traffic is served by a 3GPP LTE street light small cell network, characterized by high density. The high-level scenario is depicted in Fig. 1.

The architecture that we are going to consider includes M2M devices connected directly or via M2M gateways to the Evolved UMTS Terrestrial Radio Access Network (E-UTRAN) architecture. The evolved Node Bs (eNBs) in the E-UTRAN are connected to the Evolved Packet Core (EPC) that provides connectivity to the IP backbone. The Evolved Packet System (EPS) including E-UTRAN and EPC forms the M2M and the cellular access network. Besides getting access to E-UTRAN through an eNB, the machines can also get access through small cells, such as a relay node (RN) or a Home evolved Node B (HeNB). The aggregated M2M and H2H traffic collected by the small cells can be routed to a LTE gateway then to an eNB and, finally, to the EPC. In the rest of the paper, we refer to the RNs or HeNBs generically as SCs.

6.2 Simulation scenario

We implement our algorithm in LENA, the NS3 LTE module, which is characterized by a high fidelity implementation of the complete LTE protocol stack. The simulation details are shown in Table 2. The street light SCs are located in correspondence of lamp posts or similar street furnitures. The street lights are located every 25 m, and a small cell is located every three street lights, i.e. every 75 m, as represented in Fig. 2. The yellow and red circles represent the street lights without and with installed SC, respectively. Each SC has to provide traffic and schedule 60 UEs over an access segment based on 50 RBs, which corresponds to a 10-MHz LTE UL implementation. Without loss of generality, we consider a 10-MHz LTE implementation, which ensures 25 Mbps of theoretical maximum throughput to a single UE at the physical

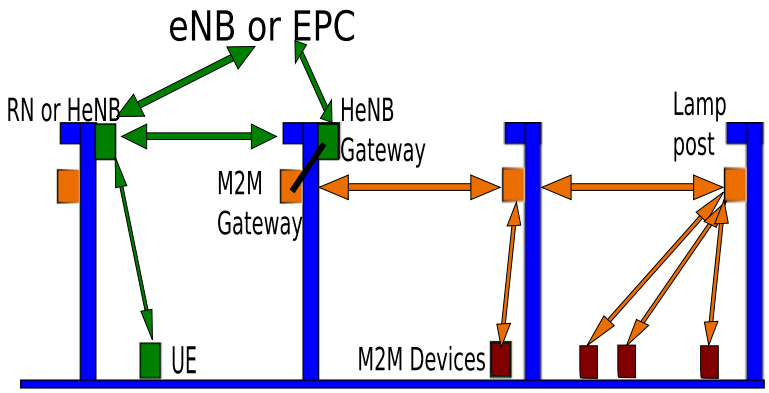

Fig. 1 High-level scenario

layer, so, enough for the kind of traffic that we are focusing on, and which allows multi-user scheduling per TTI.

6.3 Traffic models

In the scenario presented in Section 6.2, we foresee that multiple services coexist and need to be scheduled by different SCs in the same band. Because of that, in this section, we present the different applications and traffic models we focus on to take into account this aspect of a real scenario.

Figure 3 shows the state diagram of the traffic model of a M2M device, which is based on three states:

- OFF state: In this state, the devices are in a deep sleep mode and only a very low power clock is running. The devices move to the ON state when a timer expires.
- ON/monitoring mode: The devices in this state generate information on a time-driven fashion. In practice, the device monitors some physical variable and sends periodical information.
- ON/alarm mode: This model depends on the application and type of sensor the device is equipped with. In general, the device is triggered by a particular event, when there is the need to send more frequent information than in the monitoring mode. For instance, a temperature sensor in a building provides regular information, e.g. every 5–10 min, on the temperature in the building. However, if the temperature exceeds a certain threshold, this may be associated to a fire alarm, so that the device enters the alarm mode and sends information every 1–5 s.

Table 2 Simulation system parameters

Parameter	Value
Cellular layout	Circular cell
Inter-SC distance	75 m
SC radius	75 m
SC height	8 m
SC Ptx	10 dBm
Frequency	2.5 GHz
UL bandwidth	10 MHz
Simulation time	5 s
RBs assigned per SC	50
Users distribution	Uniform in SC radius
No. of users	60
Max Tx power of users	10 dBm
User antenna gain	0 dBi
Channel model	Friss channel model
Control plane	Ideal channel

Devices in ON state are supposed to be connected and synchronized with the LTE network. As for the simulation results, we focus on the devices in alarm mode, in order to consider the most demanding conditions to evaluate the scheduler.

For the purpose of the evaluation of the proposed algorithm, we consider a traffic based on the mix of three M2M applications:

1. Traffic monitoring: We consider a traffic monitoring application, in alarm mode, where there may be the need for exchange of several information to re-route human/vehicular traffic. The application is modelled by User Datagram Protocol (UDP) packets, with a periodicity of 10 ms.
2. Video surveillance-LQ (low quality): We consider in this case a continuous traffic, generated for example by a LQ streaming application or by devices that act as collectors of information from different sensors. The application is modelled by UDP packets, with a periodicity of 1 ms.
3. Video surveillance-HQ (high quality): We consider in this case a continuous traffic, generated for example

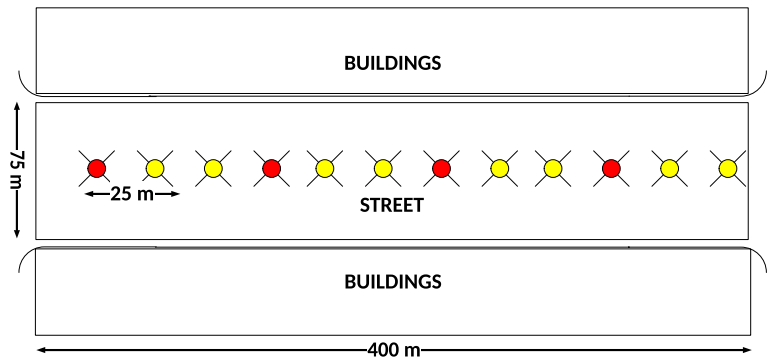

Fig. 2 Simulated scenario

by a HQ streaming application or by devices that act as collectors of information from different sensors. In both cases, there is the need to send a high amount of data. This could be modelled by a full buffer traffic generator.

7 MILP numerical results

In this section, we discuss the most important numerical results obtained by evaluating the proposed MILP models and the greedy algorithm. The approach that we follow is first to evaluate the proposed models, by solving the corresponding MILP problems through an optimization software, the IBM ILOG CPLEX Optimization Studio [10]. Then, we compare the results obtained through the solution of the MILP problems, to those obtained by the greedy algorithm, in order to evaluate the actual performance of the heuristic approach in relation to the optimal solution. We use Optimization Programming Language (OPL) to create a script that writes the mathematical models presented in Section 5.2 and test them on CPLEX. The first aim of this step is to show computationally the equivalence of the unified model (13), (5)–(7) with the three-step approach. As mentioned, such an equivalence depends on the selection of the values for the parameters α, β and γ, which have been set to 100, 10 and 1, respectively. Operatively, some preliminary tests have been run in order to prove that the value chosen for α, β and γ are appropriate. In other words, α should be large enough to ensure that the first term of the objective function has a higher priority than the others, i.e. no solution with a smaller value of the first term in (13) can be optimal. In the same way, β is chosen to be large enough to have a priority over the third term of the objective function.

The results summarized in Table 3 are averaged over 6000 channel realizations, where $n = 50$ RBs and $m = 60$ users. This high number of users has been selected to consider a 5G aligned scenario, where also M2M traffic is allowed.

Table 3 compares the two MILP approaches (namely, the unified model and the three MILPs of the three-step approach) in terms of the (1) number of instances solved by branching, (2) number of instances solved at the root node, (3) average number of needed branches and (4) average gap over the instances. The gap is defined as the difference between the upper bound of the problem (obtained through constraints relaxation) and the optimal feasible solution. A small gap indicates a good model, since the feasible solution is closer to the upper bound. We observe that the models are of very good quality in practice, i.e. for most of the instances, there is no need to branch, and the optimal solution is found at the root node.

For a deeper analysis and as an example, we report the results obtained for one specific instance in Table 4. Here, the results indicate (i) the value of the objective functions at the end of the execution, (ii) the number of branches, (iii) the reduced number of columns characterizing the

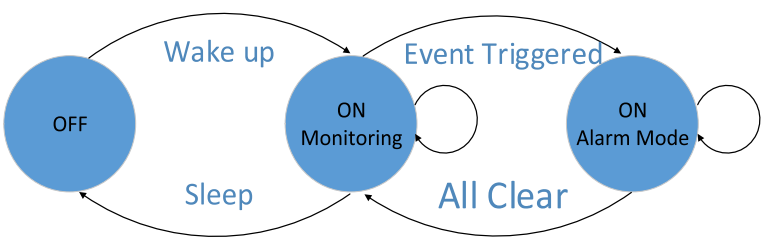

Fig. 3 M2M state diagram

Table 3 CPLEX Analysis results

Models	≥ 1 branches	Root solved	Avg. no. of branches	Avg. gap
Unified	49	6452	75.79	0.371
1st step	109	6692	7.07	0.391
2nd step	75	6726	65.07	2.537
3rd step	33	6768	9.34	0.939

model after the pre-solving phase, (iv) the pre-solve time and (v) the total time. We observe that the unified model behaves as the three-step approach (it obtains the same objective function value), so that it is possible to solve our scheduling problem in just one step. In particular, the dimensions of the reduced problems are exactly the same, but the unified approach allows to slightly reduce the execution time and to reduce the memory usage (although not shown in the table).

7.1 Algorithm comparison

In this section, we compare the results obtained by applying Algorithm 2 to those obtained by CPLEX. As a comparing metric, we use the gap, calculated over 150 instances. We redefine the gap as the difference between the solution provided by the greedy algorithm and the optimal feasible one provided by CPLEX. In this context, we focus on the gap's statistic distribution, in particular, on its probability mass function (PMF). Figure 4 shows the PMF of the gap when considering the unified model. It can be observed that in about 50% of the cases, the heuristic algorithm provides a solution equivalent to the optimal (0% GAP), and in 90% of the cases, the greedy solution lowers the performances with respect to the optimum, by less than 10%. As Fig. 4 suggests, the lower bound in terms of performance (which is very unlikely) is that the algorithm results in a GAP which is 18% lower than the optimal one. This last results has a significant impact on the algorithm evaluation, in fact, statistically, the presented algorithm results in a resource allocation scheme that is the best one regardless of the real profit function implemented. This will be more clear in Section 8 where the algorithm is tested on a standard compliant simulator and with different network traffic load.

Figure 5 shows the empirical cumulative distribution function (eCDF) of the solving time of the greedy

Table 4 CPLEX analysis results, details over an instance

Models	Unified	1st step	2nd step	3rd step
Obj value	863589.9274	8640	39	20.0726
No. of branches	19	10	13	18
Red. variables	1158	1158	1158	1158
Pre-solve time (s)	0.2	0.2	0.2	0.1
Total time (s)	0.06	0.19	0.09	0.08

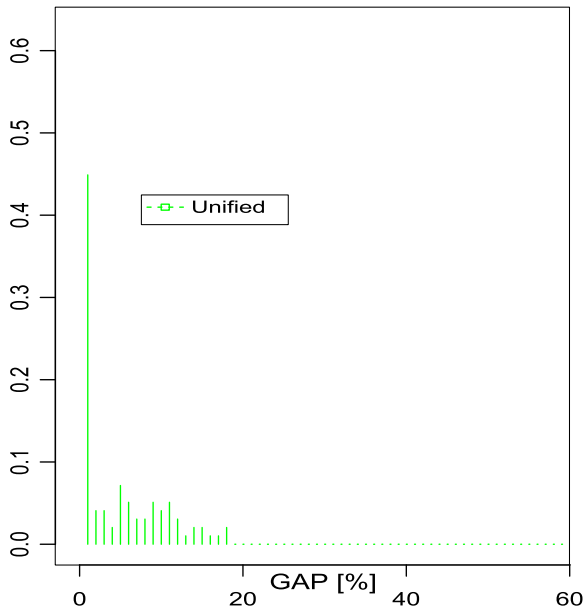

Fig. 4 PMF GAP wrt optimal solution

algorithm. It can be observed that the proposed algorithm achieves the solution in less than 0.75 ms in more than 80% of the instance and in less than 1 ms in 100% of the cases. These numbers have been obtained by running the algorithm 100 times for every instance. The code has been run on Ubuntu 13.10 operating system, CPU Intel Core i7 3.90 GHz and 8 GB of RAM. This makes the algorithm compliant with the standard's scheduling requirements. It is reasonable to foresee better results in terms of execution time on a dedicated hardware.

8 Simulations

In this section, we discuss the results obtained by applying the proposed scheduler to the three selected M2M traffic classes. These results are benchmarked to those provided by the state-of-the-art RR and MF schemes.

8.1 Benchmarks

8.1.1 Round-robin algorithm

The details on the RR implementation can be found in the LENA open documentation [11]. This algorithm first verifies how many users have sent a Buffer Status Report (BSR), i.e. the users that have something to transmit in the current TTI, then divides the total amount of available RBs by the number of active users, i.e. the users that have data to transmit. The minimum amount of RBs assigned for each user is three, this ensures a minimum of 3 bytes transmitted in a TTI, in case of worse channel quality condition, i.e. those associated to the lowest MCS index to ensure a transmission. The assignment starts from the first user that was not scheduled in the previous TTI and

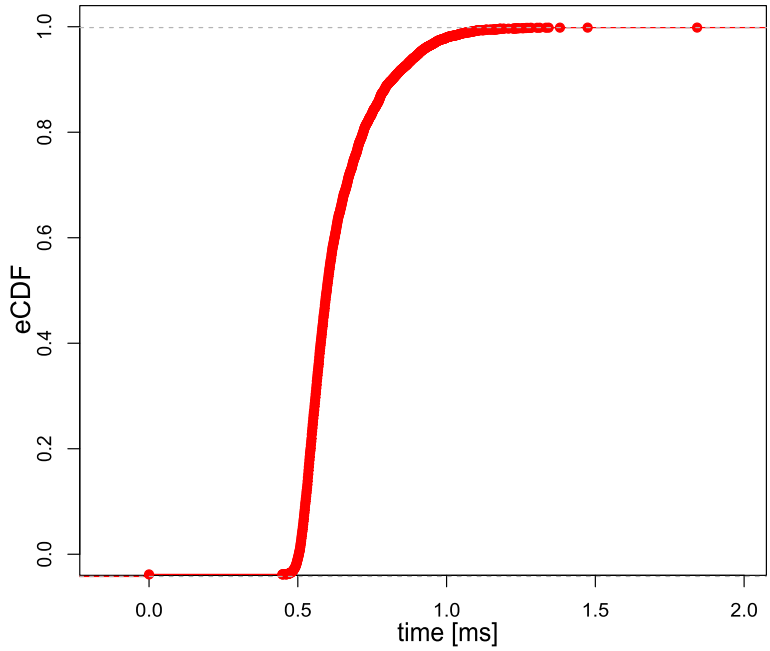

Fig. 5 ECDF of solving time

proceeds in a round-robin fashion. During the assignment phase, the algorithm chooses the lowest MCS between these available in the RB set assigned to the user.

8.1.2 Maximum fairness algorithm

The MF algorithm goal is to obtain the maximum possible fairness for each user. In order to achieve this goal, the active users are sorted with respect to their average bit rate, evaluated over a temporal window of 100 ms. At each TTI, the user with the lowest average bit rate is granted the entire bandwidth, e.g. if the system has 10 MHz of uplink bandwidth, the selected user is granted all available RBs in the current TTI.

8.2 Proposed algorithms

We implement in the NS3 LENA simulator two different instantiations of the greedy Algorithm 2, presented in Section 5, by considering two different profit functions.

We first implement a CDA, channel- and demand-aware version of the Algorithm 2, where the profit function p_j is the user demand, i.e. d_j, the amount of bits that the user j has to transmit. The main goal of this algorithm is to maximize the overall throughput.

A second implementation in turn considers that the profit function p_j is the delay δ_j, where we define δ_j as the delay in number of TTIs since the last time a user was scheduled. As a result, when the user j is scheduled, the algorithm resets δ_j to zero, and each time an active user cannot be scheduled, δ_j is incremented by one.

During the scheduling procedure, the users are sorted by δ_j and the users with larger value are scheduled first. We refer in the following to this algorithm as CADELTA, channel-aware delta algorithm. The main goal of CADELTA is to reduce the delay in the user resource assignment.

Both our solutions are channel-aware algorithms, where we select the most appropriate MCS per user and RB, so that the bit error rate of the physical channel is ensured to be lower than 10%.

8.3 Key performance indicators

We consider the following key performance indicators (KPIs) to compare CDA and CADELTA algorithms, with respect to RR and MF:

- Throughput: We consider the cumulative throughput (at radio link control (RLC) layer) of the RR simulation results and compare our algorithm in terms of variation of throughput. In other words, if an algorithm has an increment of throughput in the order of 30%, it means that the cumulative traffic served is 130% with respect to the RR one.
- Fairness: As for the fairness index, we use the well-known Jain index (J-index) as it is defined in [27] and in (14)

$$J = \frac{\left(\sum_{j=1}^{n} z_j\right)^2}{n \sum_{j=1}^{n} z_j^2} \tag{14}$$

where z_j is defined as $z_j = T_j/T_j^{opt}$, T_j and T_j^{opt} are the throughput and the optimal fair throughput of user j, respectively.

- Delay: We measure the delay at Packet Data Converge Protocol (PDCP) layer. This delay includes also the components associated to RLC layer, where different transmissions have to be received in order to aggregate a packet, before sending it to the PDCP layer.

8.4 Simulation results

We consider 60 UEs uniformly distributed in each SC coverage area. In this contribution, we show the results over 15 different realizations of our scenario. In the figures, the different realizations are addressed as simulation rounds. We present the simulation results for the different traffic models presented in Section 6.3. Each UE runs only one application, and the UEs are uniformly distributed amongst available applications.

Figure 6 represents the throughput variation with respect to RR performances, as a function of the fairness. In this first set of results, we only consider simulations with individual classes of traffic, without mixing multiple classes. In particular, Fig. 6 shows the results in terms of fairness vs throughput for traffic monitoring and video surveillance (LQ), respectively. The former is characterized by a less demanding traffic compared to the latter. With traffic monitoring, all the algorithms behave similarly, in particular, the throughput variation is in the order of 7% while the fairness is almost 1. When considering video surveillance (LQ), simulation results show that both CDA and CADELTA outperform the RR in terms of throughput, by more than 100 and 80%, approximately, respectively. On the other hand, the proposed algorithms provide a reduction in fairness between 10 and 15% with respect to the MF algorithm, which, as it was expected, provides the best fairness results. Figures 7 and 8 show the performance of the algorithms in terms of delay, defined at PDCP layer. We observe that, when considering video surveillance (LQ), our proposed solutions outperform RR, while MF achieves approximately the same performances. On the other hand, in case of traffic monitoring, our algorithms perform better, but in absolute terms, the delay reduction is negligible.

We observe that channel- and demand-aware (CDA) performs worst than MF in terms of delay. This was an expected results; as in Fig. 6, CDA results in a 20% lower fairness with respect to MF. This suggests that some users in the network are less likely to be scheduled, which results in a higher delay. Similarly, the results presented in Fig. 8 should be read by observing also Fig. 6. In this case, the fairness is always one. This results in a slightly better performance of CDA with respect to MF, in terms of delays.

We present now a second simulation campaign, where we mix different traffic types and we stress the network with more demand considering both low quality (LQ) and high quality (HQ) video surveillance applications. Simulation results are shown in Figs. 9 and 10. The intense demand has a direct impact on the MF approach, which seriously deteriorates the throughput performance, obtaining 20% less throughput than RR. On the other

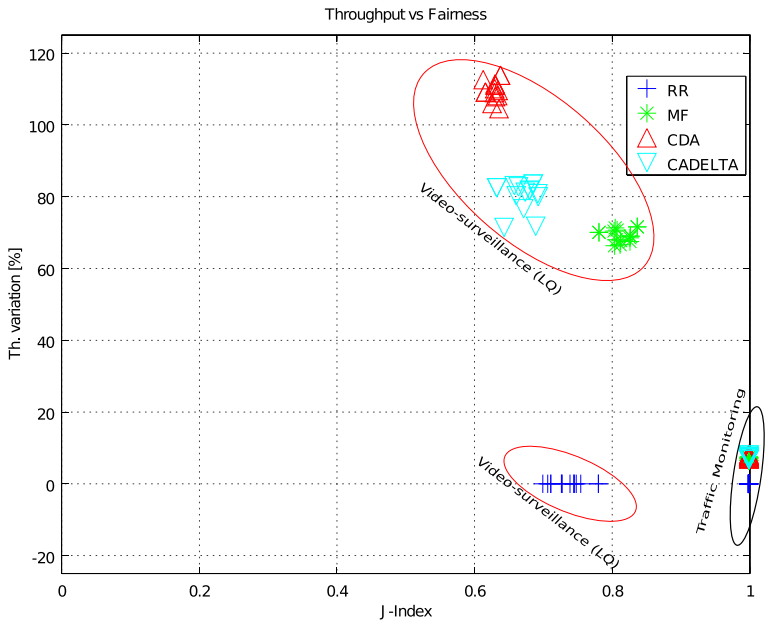

Fig. 6 Throughput vs fairness—video surveillance (LQ), traffic monitoring

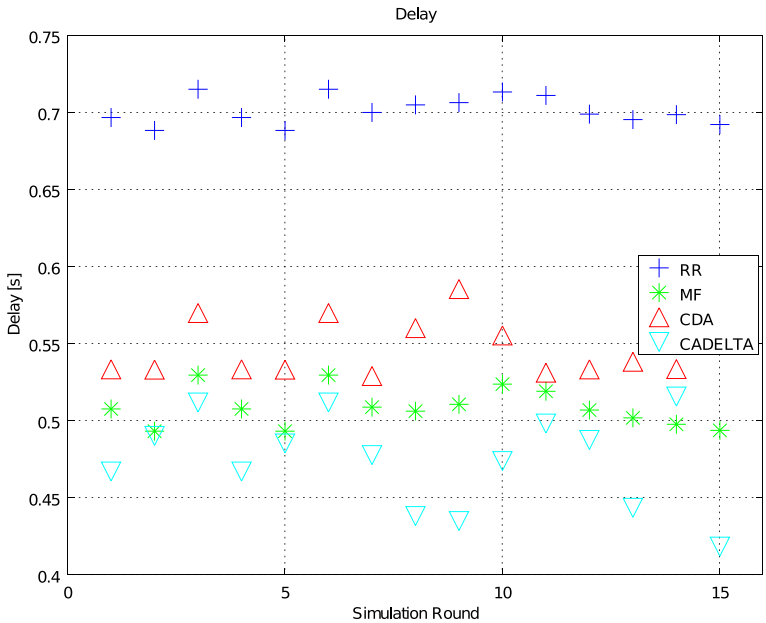

Fig. 7 Delay—video surveillance (LQ)

hand, both our solutions are more robust to the traffic change and perform well, with an increment of throughput with respect to RR in the order of 70–80%. Comparing Figs. 6 and 9, it is possible to observe that in the former, CDA results on average in a lower fairness than channel- and delay-aware (CADELTA), while in the latter is exactly the opposite. This suggested that CDA is less sensible to the traffic changes; in fact, in all the results presented in this work, the average fairness is never lower than 0.6. Meanwhile, CADELTA is more sensible; in fact, in different network traffic loads, it is possible to observe that the average fairness span between 0.5 (more intensive traffic in the network) and 1 (less intensive traffic in the network). However, regardless the network's traffic, CADELTA is the best performing in terms of delays and CDA is the best performing in terms of throughput.

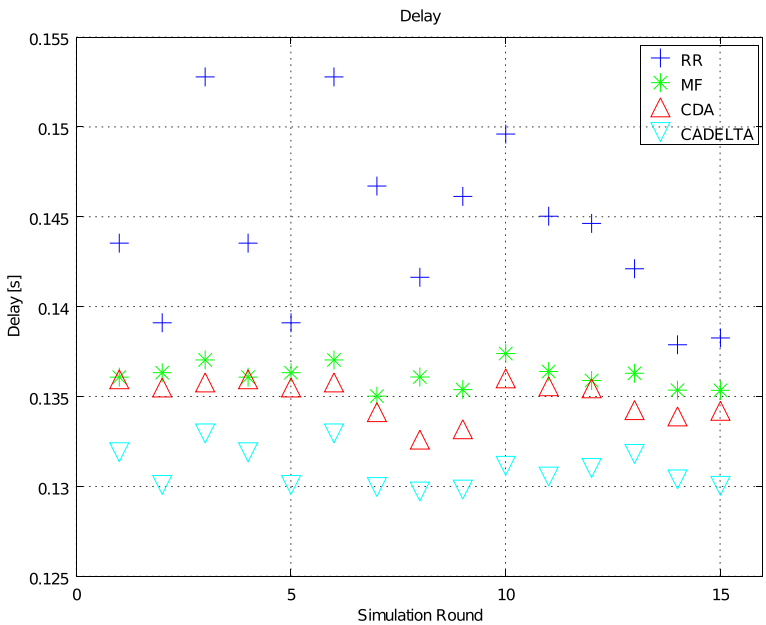

Fig. 8 Delay—traffic monitoring

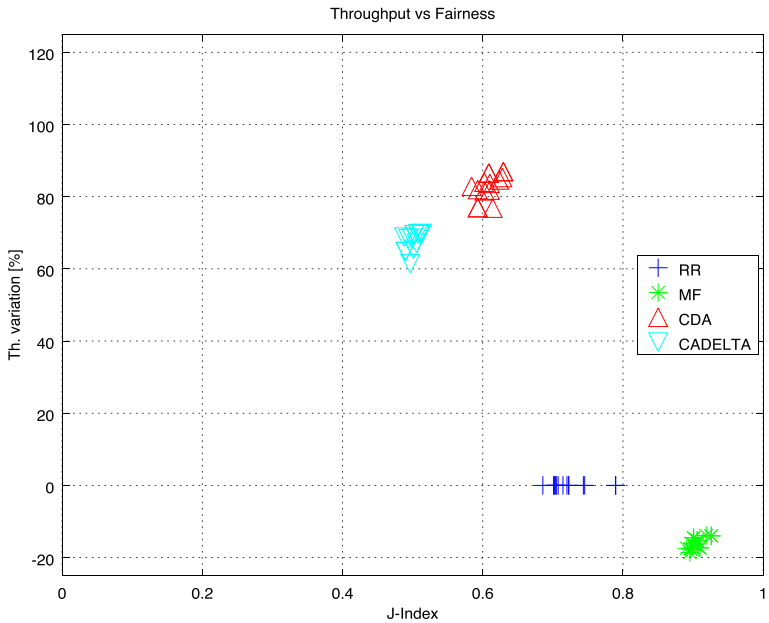

Fig. 9 Throughput vs fairness—video surveillance LQ and video surveillance HQ

Same trends are observed in Fig. 10, where both CDA and CADELTA have an average delay in the order of 0.4 s against the 0.65 s obtained on average by RR and MF algorithms, providing a performance improvement in the order of 50%.

Figures 11 and 12 depict the performance of the four algorithms with a second mix of traffic based on traffic monitoring and video surveillance HQ. This simulation campaign confirms the tendency of MF to poorly perform when a high-demand traffic profile is offered, obtaining a performance of 60 and 80% below those provided by CDA and CADELTA, respectively. In terms of delay, our solutions perform always better than RR, while the improvement with respect to MF is negligible.

Finally, Figs. 13 and 14 show a simulation campaign where all the defined traffic classes are considered. In

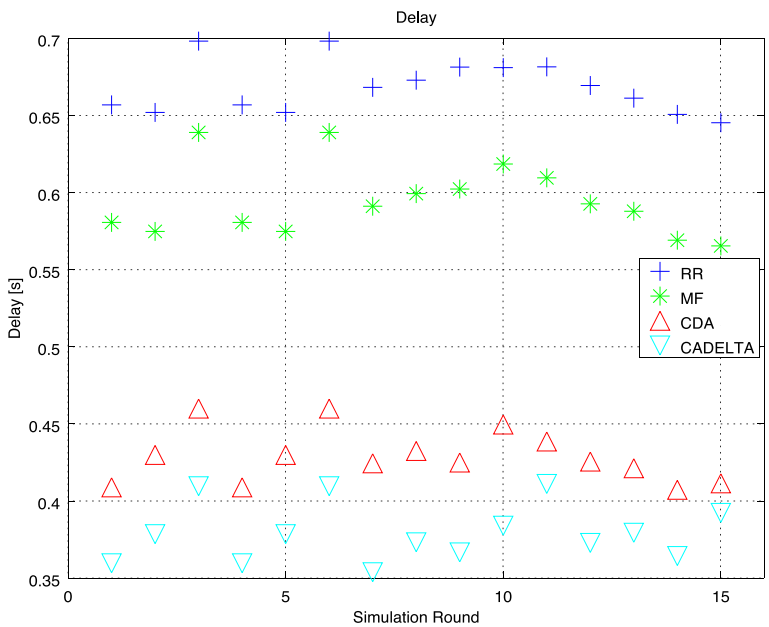

Fig. 10 Delay–video surveillance LQ and video surveillance HQ

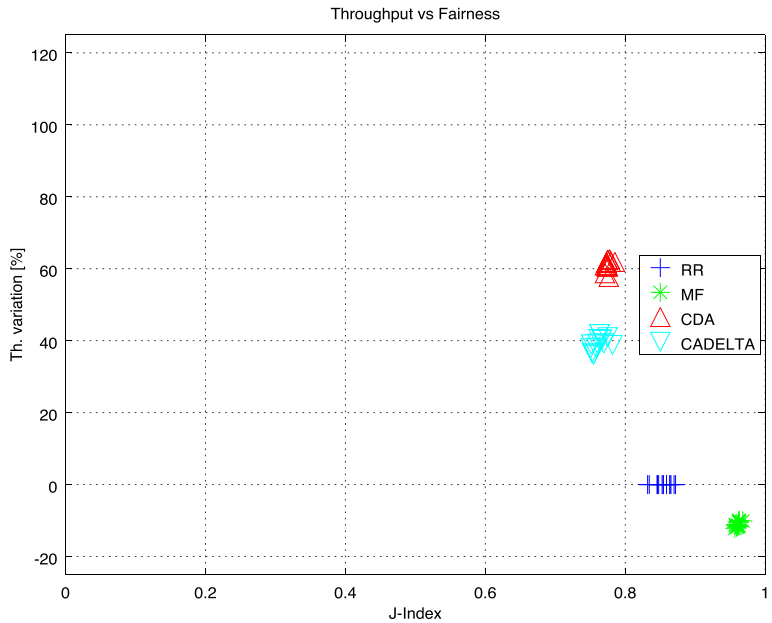

Fig. 11 Throughput vs fairness—traffic monitoring and video surveillance (HQ)

particular, the traffic mix consists of 25% of traffic monitoring, 25% of video surveillance (LQ) and 50% of video surveillance (HQ). We observe that, in this case, all the algorithms result in a reduced fairness, with respect to the previous combinations. In this case, CA and CADELTA perform very similarly to RR. On the other hand, in terms of throughput, CDA and CADELTA outperform by 100% and 80%, respectively, compared to RR. With respect to MF, the improvement is even larger, i.e. 120 and 100%, respectively. Finally, in terms of delay, we observe that CDA and CADELTA outperform both RR and MF.

These four simulation campaigns, based on considering multiple combinations of M2M traffic, have shown that CDA and CADELTA offer a high increment of throughput and reduction of delay, with respect to the

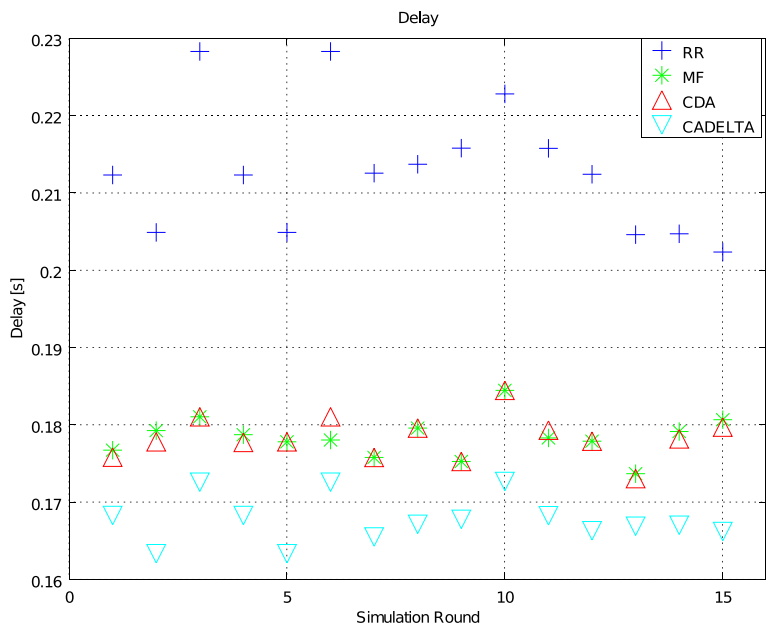

Fig. 12 Delay—traffic monitoring and video surveillance (HQ)

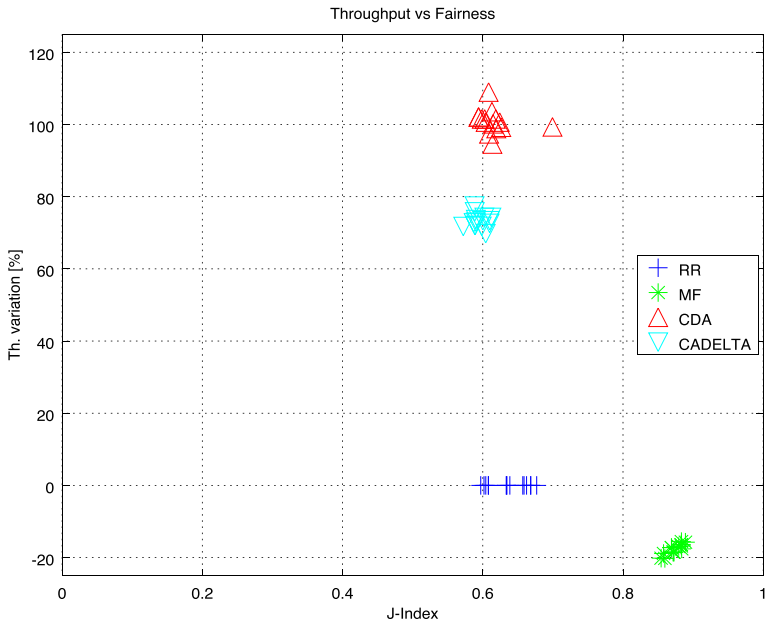

Fig. 13 Throughput vs fairness—traffic monitoring (25%), video surveillance (LQ) (25%) and video surveillance (HQ) (50%)

benchmarks. This is achieved at the expense of a reduced (in the order of 10–15%) fairness. This behaviour is robust to changes in terms of traffic. In addition, our scheme is parametric with the profit function, and different QoS parameters can be optimized through it. For example, comparing CDA and CADELTA, we observe that each algorithm gives higher priority to the QoS parameter that has been designed to optimize, i.e. the CDA always provides better performance in terms of throughput, while CADELTA performs better in terms of delays.

As a final simulation campaign, we modify our algorithm in order to be able to optimize not only one QoS parameter but a combination of them. We refer to this implementation of our proposal as CAA—channel- and application-aware—which combines the principles of

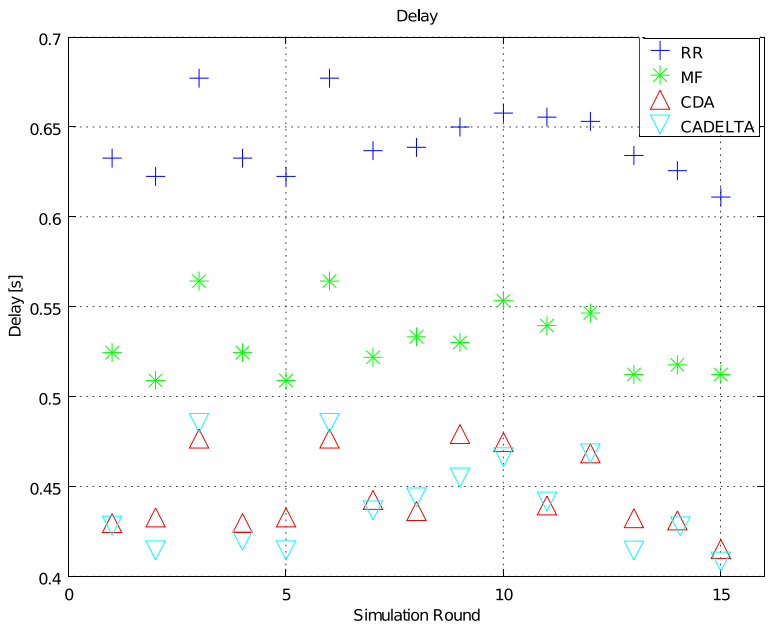

Fig. 14 Delay—traffic monitoring (25%), video surveillance (LQ) (25%) and video surveillance (HQ) (50%)

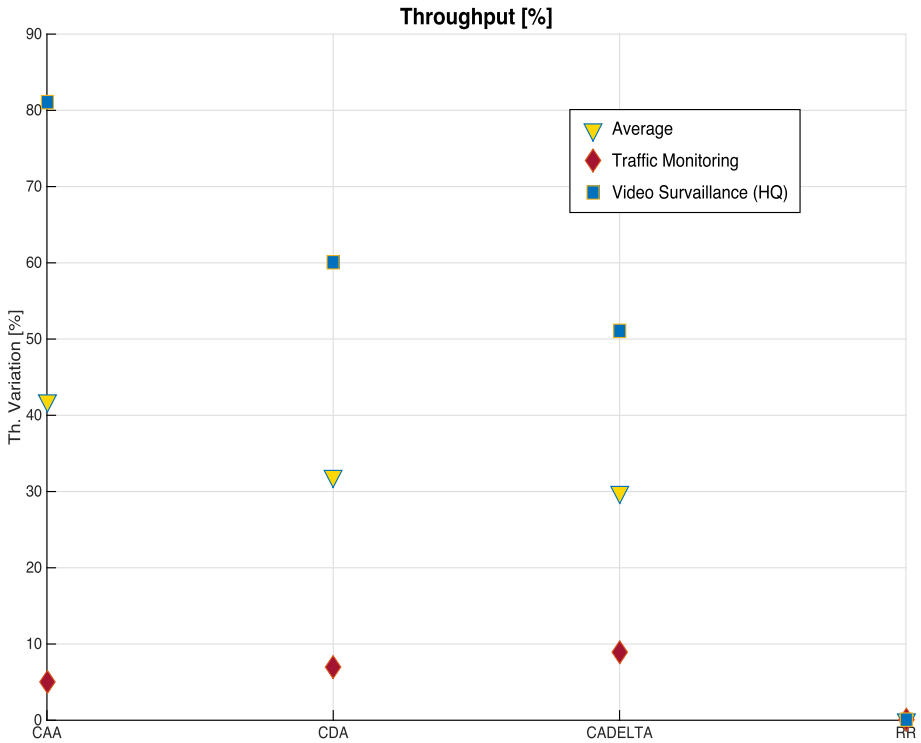

Fig. 15 Channel- and application-aware throughput—traffic monitoring and video surveillance (HQ)

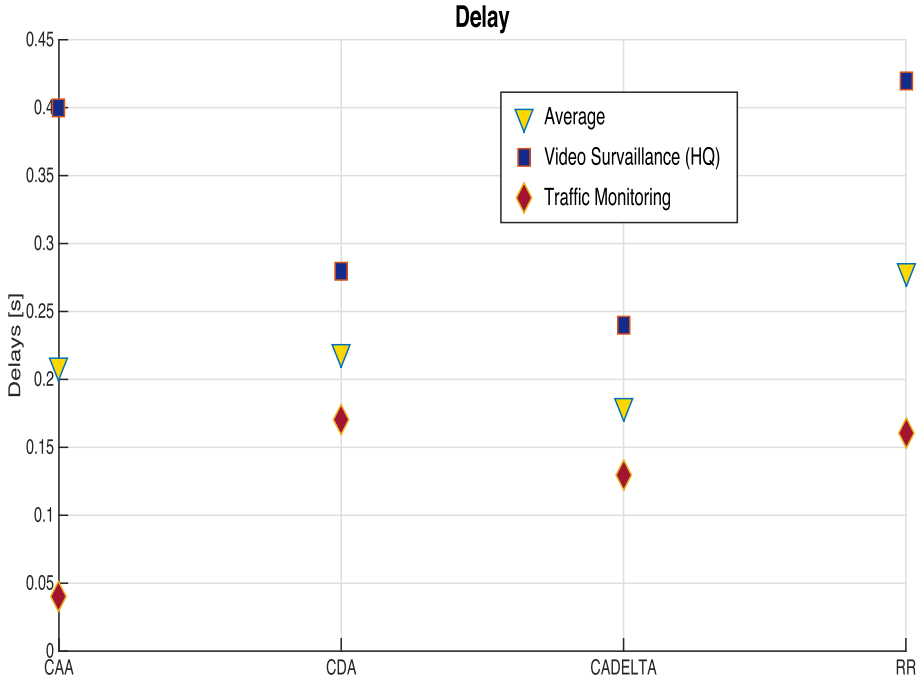

Fig. 16 Channel- and application-aware delay—traffic monitoring and video surveillance (HQ)

CADELTA and CDA. In particular, the profit function for CAA is defined as follows:

$$p_j = \phi \times d_j + \theta \times \delta_j \tag{15}$$

where ϕ and θ are coefficients that depend on the particular application and traffic class. For instance, a traffic class which is more sensitive to the delay than to the throughput will be characterized by a high ϕ and a lower θ. Figures 15 and 16 show the corresponding results. Simulation results are compared to RR, CDA and CADELTA, and the obtained values are averaged over all the simulations. We consider a mix of two classes of traffic: video surveillance (HQ), the most demanding class, and the traffic monitoring, the class most sensitive to delays. Simulation results show that CAA correctly works by discriminating between classes of applications and prioritizing the QoS parameter to be optimized. In particular, in terms of throughput, we observe an increment in the video surveillance (HQ) traffic, and a slight reduction (about 2%) for the traffic monitoring application. In terms of delay, we observe that the CAA offers extremely low latency values for the traffic monitoring class, while it increases the delay of the video surveillance application. On average, the behaviour is through very similar to that provided by CDA and CADELTA. To conclude, with this last implementation of our algorithm, we have shown that the proposed framework allows to target multiple QoS parameters depending on the specific kind of traffic to be served.

9 Conclusions

In this paper, we have presented a framework to address the scheduling of multiple and heterogeneous M2M applications over an ultra-dense small cell network, deployed over the street lights for smart city applications. We have focused on the UL scheduling problem, which, due to the constraint imposed by the standard, requires the allocation of contiguous RBs to the same user. This simple limitation makes the problem NP-hard. We have presented a multi-objective optimization to maximize the throughput of the network, to minimize the high ICI generated due to the intense spatial reuse in the small cell deployment and to maximize the radio resource usage. We have modelled this optimization through two different MILP models, which are formulated in order to allow for the application of multiple scheduling policies. The first model gradually optimizes the allocation of resources in order to meet the three targeted optimization objectives. The second model allows for a more compact representation of these objectives. We have proven that the second compact model is equivalent to the first one, based on a three-step optimization. We have found the optimal solution through a CPLEX analysis, which proves that the compact model performs better and executes faster. As a solution suitable for implementation in real-world networks, we have proposed a feasible and fast heuristic algorithm that solves the NP-hard problem. We have shown that this greedy solution lowers the performance by no more than 10% with respect to the theoretically optimal solution, in more than 90% of the cases, and its execution time runs on non-dedicated hardware in less than 1 ms, thus meeting the standard scheduling constraints.

We have implemented two different versions of the greedy algorithm on a standard-compliant network simulator, i.e. the LTE module of NS3, implementing with high fidelity, the full LTE protocol stack. We proved the superiority of our solutions, in terms of delay and throughput, with respect to the benchmarks like round-robin and maximum fairness approaches, considering multiple M2M applications and heterogeneous mixes of traffic.

Abbreviations

3GPP: 3rd generation partnership project; BLER: Block error rate; BS: Base station; BSR: Buffer status report; CA: Channel-aware; CAA: Channel- and application-aware; CADELTA: Channel- and delay-aware; CDA: Channel- and demand-aware; CSI: Channel state information; DL: Downlink; E-UTRAN: Evolved UMTS terrestrial radio access network; eCDF: Empirical cumulative distribution function; eNB: Evolved node B; EPC: Evolved packet core; EPS: Evolved packet system; ETSI: European telecommunications standards institute; FDPS: Frequency-domain packet scheduling; H2H: Human-to-human; HeNB: Home evolved node B; HII: High interference indicator; HQ: High quality; ICI: Inter-cell interference; ICIC: Inter-cell interference coordination; IoT: Internet of things; IT: Information technology; KPI: Key performance indicator; LENA: LTE-EPC network simulator; LQ: Low quality; LTE: Long term evolution; LTE-A: Long term evolution-advanced; M2M: Machine-to-machine; MAC: Medium access control; MCS: Modulation and coding scheme; MF: Maximum fairness; MILP: Mixed-integer linear programming; MIP: Mixed-integer programming; MTC: Machine-type communications; NS3: Network simulator 3; OFDM: Orthogonal frequency-division modulation; OFDMA: Orthogonal frequency-division multiple access; OI: Overload indicator; OPL: Optimization programming language; PAPR: Peak-to-average power ratio; PDCP: Packet data converge Protocol; PF: Proportional fair; PMF: Probability mass function; PSD: Power spectral density; QCI: QoS class indicator; QoE: Quality of experience; QoS: Quality of service; RB: Resource block; RLC: Radio link control; RN: Relay node; RR: Round-robin; RRA: Radio resource assignment; RRH: Radio remote head; RRM: Radio resource management; SC: Small cell; SC-FDMA: Single-carrier frequency-division multiple access; SINR: Signal-to-interference-plus-noise ratio; SRB: Signalling radio bearers; SRS: Sounding reference signal; TB: Transport block; TTI: Transmission time interval; UDP: User Datagram protocol; UE: User equipment; UL: Uplink

Acknowledgements

This work was partially funded by Spanish MINECO grant TEC2017-88373-R (5G-REFINE). The research leading to these results was also partially supported by Generalitat de Catalunya grant 2017 SGR 1195.

Funding

This work was partially funded by Spanish MINECO grant TEC2017-88373-R (5G-REFINE). The research leading to these results was also partially supported by Generalitat de Catalunya grant 2017 SGR 1195.

Availability of data and materials

All relevant figures have been added in the contribution. However, the authors can provide extended data upon request.

Authors' contributions

Hereafter, the individual contributions of the authors are as follows. MDA and LG made substantial contributions to all the phases from conception and design to acquisition of data and analysis and interpretation of data. Moreover, they participated in drafting the article and editing the revisions. AL made a

substantial contribution in the conception and design as well as the interpretation of data. He has been particularly active in the definition of the MILP model and algorithm. Moreover, he gives final approval of the version to be submitted and any revised version. RV made a substantial contribution in the conception and design as well as the interpretation of data. He has been particularly active in the scenario setup and interpretation of the simulation results. Moreover, he gives final approval of the version to be submitted and any revised version. All authors read and approved the final manuscript.

Authors' information

Dr. Melchiorre Danilo Abrignani (GSM13-M'16) received the M.S. Degree in Telecommunication Engineering in 2012 and his Ph.D. Degree in Electronics, Telecommunications, and Information Technologies in 2016 at the University of Bologna. His research interests include radio resource management for cellular networks and MAC, routing and IoT applications for wireless sensor networks. Dr. Abrignani has been involved in the FP7 Network of Excellence NEWCOM♯. He also contributed to COST Action IC1004.

Dr. Lorenza Giupponi (GSM'02-M'07) received the Telecommunications Engineering degree from the University of Rome La Sapienza in July 2002 and the Ph.D. from the Technical University of Catalonia (UPC) in 2007. She joined the Radio Communications Group of UPC in 2003 with a grant of the Spanish Ministry of Education. During 2006 and 2007, she was an assistant professor in UPC. In September 2007, she joined the CTTC where she is currently a Senior Researcher in the Mobile Networks Department of the Communication Networks Division. Since 2007, she is also a member of the Executive Committee of CTTC, where she acts as the Director of Institutional Relations. She is the co-recipient of the IEEE Consumer Communications and Networking Conference 2010 (IEEE CCNC 2010) and of the IEEE Third International Workshop on Indoor and Outdoor Femto Cells 2011 best paper awards. Since 2015, she is a member of the Executive Committee of NS3 consortium. She is an IEEE Senior Member.

Prof. Andrea Lodi (M'00) holds Canada's main chair in operations research. Prof. Lodi received the Ph.D. in System Engineering from the University of Bologna in 2000, and he has been a Herman Goldstine Fellow at the IBM Mathematical Sciences Department, NY, in 2005–2006. He is the author of more than 80 publications in the top journals of the field of Mathematical Optimization. He serves as an Associated Editor for several prestigious journals in the area.

Roberto Verdone (S'94-M'95) received the Laurea Degree in Electronics Engineering and Ph.D. degree from the University of Bologna, Bologna, Italy, in 1991 and 1995, respectively. Since 2001, he has been a Full Professor of Telecommunications with the University of Bologna. He has authored or coauthored more than 100 research papers, mostly in IEEE journals or conferences. His research interests include both infrastructure-based radio networks and infrastructureless radio networks. Over the last 10 years, he has investigated radio resource management for cellular systems and MAC, routing and topology aspects of the wireless sensor networks.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Department of Electrical, Electronic and Information Engineering (DEI), University of Bologna, Bologna, Italy. [2]Centre Technologic de Telecomunicacions de Catalunya, Castelledefels, Barcelona, Spain. [3]Polytechnique Montréal, Montreal, Canada.

Received: 31 July 2016 Accepted: 19 July 2018
Published online: 09 August 2018

References

1. online Cisco Visual Networking index: Global Mobile Data traffic Forecast update. 2016-2021 White paper online at https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html
2. 3GPP, Service requirements for machine type communications (2014). 3GPP TS 22.368 v13.0.0
3. IEEE, DRAFT Amendment to IEEE Standard for WirelessMAN-Advanced Air Interface for broadband Wireless Access Systems: Enhancements to Support Machine-to-Machine Applications (P802.16p-11/0033) (2011). Std. P802.16p-11/0033
4. ETSI, Machine-to-Machine communications (M2M); functional architecture, Std. v.2.1.1 (2013). ETSI TS 102 690
5. Small Cell Forum, Small cell market status. SCF White Paper (2012). Available online at: https://www.smallcellforum.org/
6. J Zander, Beyond the ultra-dense barrier: paradigm shifts on the road beyond 1000× wireless capacity. IEEE Wirel. Commun. **24**(3), 96–102 (2017). https://doi.org/10.1109/MWC.2017.1500377WC
7. AS Hamza, SS Khalifa, HS Hamza, K Elsayed, A survey on inter-cell interference coordination techniques in OFDMA-based cellular networks. IEEE Commun. Surv. Tutor. **15**(4), 1642–1670 (2013)
8. N Abu-Ali, A-EM Taha, M Salah, H Hassanein, Uplink scheduling in LTE and LTE-Advanced: tutorial, survey and evaluation framework. IEEE Commun. Surv. Tutor. **16**(3), 1239–1265 (2014)
9. X Xiang, C Lin, X Chen, XS Shen, Toward optimal admission control and resource allocation for LTE-A femtocell uplink. IEEE Trans. Veh. Technol. **PP**(99), 1–1 (2014). https://doi.org/10.1109/TVT.2014.2351837
10. IBM, *IBM ILOG CPLEX V12.1 - User's Manual For CPLEX, IBM Corp*, (2009). IBM Corp. The manual is available online at: https://www.ibm.com/support/knowledgecenter/SSSA5P_12.6.1/ilog.odms.cplex.help/CPLEX/homepages/usrmancplex.html
11. CTTC, *LTE Simulator Documentation, Release M6 edn. CTTC*, (2013). ns-3 LTE module documentation available online: https://www.nsnam.org/docs/models/html/lte.html
12. Fujitsu, 4G femtocell solutions for stadiu m environments (2011). Available online at https://www.fujitsu.com/us/Images/4G-Femtocell-for-Football-Stadiums.pdf
13. FD Calabrese, PH Michaelsen, C Rosa, M Anas, CU Castellanos, DL Villa, KI Pedersen, PE Mogensen, in *Vehicular Technology Conference, 2008. VTC Spring 2008*. Search-tree based uplink channel aware packet scheduling for UTRAN LTE (VTC Sping, Singapore, 2008), pp. 1949–1953. https://doi.org/10.1109/VETECS.2008.441
14. S-Y Lien, K-C Chen, Y Lin, Toward ubiquitous massive accesses in 3GPP Machine-to-machine communications. IEEE Commun. Mag. **49**(4), 66–74 (2011). https://doi.org/10.1109/MCOM.2011.5741148
15. AM Maia, D Vieira, MF de Castro, Y Ghamri-Doudane, in *Global Communications Conference (GLOBECOM), 2014 IEEE*. A mechanism for uplink packet scheduler in LTE network in the context of Machine-to-Machine communication (Globecom, Austin, 2014), pp. 2776–2782. https://doi.org/10.1109/GLOCOM.2014.7037228
16. S Zhenqi, Y Haifeng, C Xuefen, L Hongxia, in *Information and Communications Technologies (IETICT 2013), IET International Conference On*. Research on uplink scheduling algorithm of massive M2M and H2H services in LTE, (Beijing, 2013), pp. 365–369. https://doi.org/10.1049/cp.2013.0070
17. I Abdalla, S Venkatesan, in *Mobile and Wireless Networking (MoWNeT), 2013 International Conference on Selected Topics In*. A QoE preserving M2M-aware hybrid scheduler for LTE uplink, (Montreal, 2013), pp. 127–132. https://doi.org/10.1109/MoWNet.2013.6613808
18. A Aijaz, M Tshangini, MR Nakhai, X Chu, AH Aghvami, *Energy-efficient uplink resource allocation in LTE networks with M2M/H2H co-existence under statistical QoS guarantees*, vol. 62. (IEEE, NY, 2014), pp. 2353–2365. https://doi.org/10.1109/TCOMM.2014.2328338
19. A Azari, G Miao, in *2015 IEEE Wireless Communications and Networking Conference (WCNC)*. Lifetime-aware scheduling and power control for cellular-based M2M communications (WCNC, Istanbul, 2015), pp. 1171–1176. https://doi.org/10.1109/WCNC.2015.7127635
20. Y Liao, C Wang, D Yang, W Chen, *Uplink scheduling for LTE 4Gg video surveillance system*. (IEEE, New Orleans, 2014). https://doi.org/10.1109/WCNC.2015.7127670
21. 3GPP, X2 general aspects and principles (release 8) (2007). 3GPP TS 36.420 V8.0.0 (2007-12)
22. 3GPP, X2 Application Protocol (X2AP) (release 10) (2010). 3GPP TS 36.423 V10.1.0 (2010-03)
23. CTTC, The LTE-EPC Network Simulator (LENA) Project. http://networks.cttc.es/mobile-networks/software-tools/lena/. Accessed 1 Aug 2018
24. 3GPP, Conveying MCS and TB size via PDCCH (2008). 3GPP R1-081483
25. 3GPP, LTE - Evolved Universal Terrestrial Radio Access (E-UTRA) physical layer procedures (2009). 3GPP TS 36.213, ETSI TS 136 213 v8.8.0

26. RT Marler, JS Arora, Survey of multi-objective optimization methods for engineering. Struct. Multidiscip. Optim. **26**(6), 369–395 (2004). https://doi.org/10.1007/s00158-003-0368-6
27. R Jain, D Chiu, W Hawe, A quantitative measure of fairness and discrimination for resource allocation in shared computer system. DEC Res. Rep. TR-301 (1984). https://arxiv.org/abs/cs/9809099
28. K Yang, S Martin, TA Yahiya, in *Computers and Communication (ISCC), 2014 IEEE Symposium On*. Interference aware resource allocation for LTE uplink transmission (ISCC, Madeira, 2014), pp. 1–6. https://doi.org/10.1109/ISCC.2014.6912475
29. A Afifi, KMF Elsayed, A Khattab, in *Computers and Communications (ISCC), 2013 IEEE Symposium On*. Interference-aware radio resource management framework for the 3GPP LTE uplink with QoS constraints (ISCC, Split, 2013), pp. 000693–000698. https://doi.org/10.1109/ISCC.2013.6755029
30. P Frank, A Muller, H Droste, J Speidel, in *Personal Indoor and Mobile Radio Communications (PIMRC), 2010 IEEE 21st International Symposium On*. Cooperative interference-aware joint scheduling for the 3GPP LTE uplink (PIMRC, Istambul, 2010), pp. 2216–2221. https://doi.org/10.1109/PIMRC.2010.5671678
31. L Ruiz de Temino, G Berardinelli, S Frattasi, P Mogensen, in *Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium On*. Channel-aware scheduling algorithms for SC-FDMA in LTE uplink (IEEE, NY, 2008), pp. 1–6. https://doi.org/10.1109/PIMRC.2008.4699645
32. F Ren, Y Xu, H Yang, J Zhang, C Lin, Frequency domain packet scheduling with stability analysis for 3GPP LTE uplink. IEEE Trans. Commun. **12**(12), 2412–2426 (2013). https://doi.org/10.1109/TMC.2012.223
33. R Ruby, V Leung, in *Communication Networks and Services Research Conference (CNSR), 2011 Ninth Annual*. Towards QoS assurance with revenue maximization of LTE uplink scheduling (CNRS, Montreal, 2011), pp. 202–209. https://doi.org/10.1109/CNSR.2011.37
34. E Yaacoub, Z Dawy, Achieving the Nash bargaining solution in OFDMA uplink using distributed scheduling with limited feedback. AEU Int. J. Electron. Commun. **65**(4), 320–330 (2011). https://doi.org/10.1016/j.aeue.2010.03.007
35. S-B Lee, I Pefkianakis, A Meyerson, S Xu, S Lu, in *INFOCOM 2009, IEEE*. Proportional fair frequency-domain packet scheduling for 3GPP LTE uplink (InfoCom, Rio de janeiro, 2009), pp. 2611–2615
36. AS Lioumpas, A Alexiou, in *GLOBECOM Workshops (GC Wkshps), 2011 IEEE*. Uplink scheduling for Machine-to-Machine communications in LTE-based cellular systems, (2011), pp. 353–357. https://doi.org/10.1109/GLOCOMW.2011.6162470