An efficient D2D-based strategies for Machine Type Communications in 5G mobile systems

Miloud Bagaa[‖]*, Adlen Ksentini[§], Tarik Taleb[‖], Riku Jantti[‖], Ali Chelli* and Ilangko Balasingham*
[‖] Communications and Networking Department, Aalto University, Finland. Email: talebtarik@ieee.org and riku.jantti@aalto.fi
* Dep. of Ele. and Tel., NTNU, 7491 Trondheim, Norway. Emails:{miloud.bagaa,ali.chelli,ilangkob}@iet.ntnu.no
[§] IRISA, University of Rennes 1, France. Email: adlen.ksentini@irisa.fr

***Abstract*—Recent studies foresee that there would be roughly 50 billion of machine type communication (MTC) devices by 2020. Coping with the massive signaling overhead expected from these devices in $5G$ network is an important hurdle to tackle. In this paper, we have proposed two optimal solutions that use Device-to-Device (D2D) communications to lightweight the overhead of MTC devices on $5G$ network. Each scheme has a specific objective, and aims to manage the communications between devices and eNodeBs to achieve its objective. The proposed solutions nominate the devices that should communicate through D2D communication fashion and those that should directly communicate with eNodeBs. The first solution aims to reduce the energy consumption, whereas the second one aims to reduce the data transfer delay at the eNodeBs. The performance of the proposed schemes is evaluated via simulations and the obtained results demonstrate their feasibility and ability in achieving their design goals.**

***Index Terms*—LTE-A, MTC, D2D communication.**

I. Introduction

Driven by the success of M2M-based applications (such as Intelligent Transportation Services), recent studies foresee that there would be roughly 50 billion of MTC devices by 2020 [1]. The diversity of MTC applications, which need a high number of deployed devices, will put very high pressure on the cellular mobile network. Indeed, deploying the expected number of MTC devices in the cellular mobile network would face many challenges, such as the inadequacy of the current networks for the MTC traffic as there is more Uplink traffic than Downlink. Moreover, it may cause congestion and system overload in the whole network, i.e. in the Radio Access Network (RAN) part and the Core Network (CN) part, due to the huge amount of data/control traffic generated by MTC devices. In order to alleviate system overload and congestion at both parts of the network, 3GPP organization has established different studies groups [2]. So far, only system requirements without concrete solutions have been proposed. Meanwhile, several research works have been conducted, wherein existing signaling congestion avoidance and overload control solutions for MTC could be classified into two classes:

$i)$ Proactive class, where two approaches exist. The first one consists in separating the MTC traffic from the non-MTC traffic by using techniques like grouping MTC devices into groups or clusters [3]. After that, the access to RAN is differentiated for each group. For instance, by choosing different *RAN* access parameters in order to give more priority to the non-MTC traffic or improving group paging [4]. The second approach tries to reduce the amount of signaling traffic by aggregating the MTC requests either at the eNodeBs (Bulk signaling) or by creating profile ID that replaces common Information Element (IE) used by a group of MTC devices, which allow to reduce the size of signaling traffic. This latter solution is working like the Robust Header Compression (ROHC) principle.

$ii)$ Reactive class, which reacts to the congestion, caused by massive MTC attach/connect requests, by rejecting (mainly through admission control [5], [6]) or delaying these requests at the eNodeB.

Network controlled D2D communications was originally proposed in [7] where power control was proposed to be used to mitigate the impact of D2D links on cellular users sharing the same band. Since then the radio resource control and mode selection of D2D links have attracted a lot of attention in the research community. A recent review on the related radio resource control schemes can be found in [8]. Most of the works so far have focused on single operator case, but the EU project METIS also addressed the possibility for setting up D2D link between devices associated with different operators [9]. One of the first real implementations of D2D communications within LTE frame allowing interference cancellation between cellular UEs and D2D links has been demonstrated in [10]. The main idea is to use D2D communication between MTCs rather than the direct communication with the eNodeB. Thus, one MTC device is in charge of receiving the traffic from other devices, and then it aggregates (or not) and forwards the traffic toward the eNodeB. However, most of these works assume that the MTC are preconfigured to use only D2D communication, which is not optimal most of the time. In this paper, we propose new solutions that permit for the operator to select the subset of MTC along with the MTC head that use D2D, and the subset of MTC that use direct communication with the eNodeB. We propose two different solutions aiming to reduce separately the global energy consumption and the

This research work is partially supported by the TAKE 5 project funded by the Finnish Funding Agency for Technology and Innovation (TEKES), a part of the Finnish Ministry of Employment and the Economy.

This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme.

communication delay at the eNodeB.

The rest of the paper is organized as follows. In Section II, we formally define the problem of communication between devices and eNodeBs. The proposed solutions are presented in Section III. The simulation results are presented in Section IV. Finally, we conclude the paper in Section V.

II. System model and problem definition

A. Problem definition

In Long Term Evolution Advanced (LTE-A), two types of communication are suggested to attach the MTC devices to eNodeBs: $i)$ MTC devices communicate directly with eNodeB; $ii)$ MTC devices communicate using a D2D fashion (i.e., form clusters) and then the cluster head forwards the aggregated data to the eNodeB. The direct communication between devices and eNodeBs has a positive impact on data transfer delay, whereas the communication through D2D fashion leads to reduce the energy consumption at the devices. The direct transmission of data to the eNodeBs overcomes the delay at the intermediate devices that can be involved when D2D communication is used. Meanwhile, the communication power for D2D mode is much smaller than the power required to communicate with a far eNodeB. Therefore, the devices which operate in D2D fashion experience a reduced energy consumption. Moreover, this operation mode reduces the contention at the eNodeBs as only few MTC devices would be concurrent to communicate with the eNodeBs. For such a network model, in this paper, we tackle the problem of communication between devices and eNodeBs to reduce the energy consumption and the data transfer delay, while maintaining the network contention below a maximum threshold. The notations used throughout this paper are summarized in Table I.

TABLE I: Notations used in the paper.

Notation	Description
$\mathcal{X}$	The set of devices in the network.
$\mathcal{Y}$	The set of eNodeBs in the network.
$X_{x,y}$	A boolean decision variable that shows if a device x selects y as next hop.
$\eta(x)$	A function that return the neighbors of node x in graph G.
$\mathcal{E}_{x,y}$	The energy consumption between devices x and y.
$\mathcal{E}$	A variable that represents the maximum amount of energy consumed in the network.
$\mathcal{E}_{MAX}$	The maximum amount of energy consumption tolerated in the network.
$\mathcal{D}_{x,y}^{[\lambda]}$	The delay of data transfer between devices x and y using the data generation rate λ.
$\mathcal{D}$	A variable that represents the maximum transfer delay in the network.
$\mathcal{D}_{MAX}$	The maximum amount of data transfer delay tolerated in the network.
$\mathcal{C}_y$	The contention at eNodeB y. Formally, $\mathcal{C}_y = \sum_{x\in\mathcal{X}} X_{x,y}$.
$\mathcal{C}_{MAX}$	The maximum contention tolerated in the network. We need to satisfy the condition $\max_{y\in\mathcal{Y}} \mathcal{C}_y \leq \mathcal{C}_{MAX}$.
$F_{x,y}$	An integer variable that mimics packet flow generated from the devices x to receiver y.

B. Communication modeling

In this section, we model the communications between the different MTC nodes. In order to improve the communication reliability, an automatic repeat request (ARQ) scheme is used for forwarding the information. The ARQ scheme allows resending a packet until successful reception at the destination or a maximum number of retransmissions M is reached. The number of ARQ retransmissions required for successful packet reception varies randomly according to the fading channel conditions.

Let x denotes the transmitting device, whereas y refer to the receiving node (device or eNodeB). The channel gain between these two nodes is referred to as $\alpha_{x,y}$. A Rayleigh block-fading channel is considered, where the channel gain $\alpha_{x,y}$ remains constant over one block[1] but changes independently from one block to another. The interferences generated by different nodes on the receiving node y are taken into account in our model. The fading coefficient from node t to node y is referred to as $\alpha_{t,y}$, which also follows a Rayleigh distribution. The received signal r_y at a destination node y can be expressed as

$$r_y = \alpha_{x,y}\sqrt{P_x}u_x + \sum_{t\in\mathcal{X}\cup\mathcal{Y}}^{t\neq x,y} \alpha_{t,y}\sqrt{P_t}u_t + n_y, \tag{1}$$

where P_x and P_t denote transmission powers of nodes x and t, respectively. The symbols transmitted by node x and node t are referred to as u_x and u_t, respectively. n_y is a zero-mean additive white Gaussian noise with variance N_0. The term $\gamma_{x,y}$ denotes the instantaneous received signal-to-noise ratio at node y given by $\gamma_{x,y} = P_x\alpha_{x,y}^2/N_0$ [11]. The mean value of $\gamma_{x,y}$ is denoted as $\bar{\gamma}_{x,y}$ which can be expressed as

$$\bar{\gamma}_{x,y} = \frac{P_x\mathbb{E}[\alpha_{x,y}^2]}{N_0}, \tag{2}$$

where $\mathbb{E}[\alpha_{x,y}^2]$ represents the channel variance, and $\mathbb{E}[\cdot]$ denotes the expectation operator. Using a distance dependent path loss model, the channel variance can be determined as [12]

$$\mathbb{E}[\alpha_{x,y}^2] = \left(\frac{d_0}{d_{x,y}}\right)^{\beta}, \tag{3}$$

where $d_{x,y}$ refers to the distance between nodes x and y, d_0 denotes a reference distance typically set to 1 m, and β denotes the path loss exponent. In our physical model, we take into account the effects of both path loss and fast fading, while the impact of shadowing is neglected. The instantaneous received signal-to-interference-plus-noise ratio (SINR) at node y is defined as [11]

$$\mathsf{SINR}_{x,y} = \frac{P_x\alpha_{x,y}^2}{N_0 + \sum_{t\in\mathcal{X}\cup\mathcal{Y}}^{t\neq x,y} P_t\alpha_{t,y}^2} = \frac{\gamma_{x,y}}{1+\sum_{t\in\mathcal{X}\cup\mathcal{Y}}^{t\neq x,y}\gamma_{t,y}}. \tag{4}$$

Theorem 1. *For any MTC device $x\in\mathcal{X}$ fails to transmit its packet to $y\in\mathcal{X}\cup\mathcal{Y}$ iff $\mathsf{SINR}_{x,y}$ falls below a threshold γ_{th}. This event is known as an outage event and occurs with a probability $\mathcal{P}_{x,y}$ which can be expressed as*

$$\mathcal{P}_{x,y} = 1 - \sum_{t\in\mathcal{X}\cup\mathcal{Y}}^{t\neq x,y} C_{t,y}\frac{\bar{\gamma}_{x,y}}{\bar{\gamma}_{x,y}+\bar{\gamma}_{t,y}\gamma_{\text{th}}}\exp\left(-\frac{\gamma_{\text{th}}}{\bar{\gamma}_{x,y}}\right), \tag{5}$$

where $C_{t,y} = \prod_{z\in\mathcal{X}\cup\mathcal{Y}}^{z\neq x,y,t}\dfrac{\bar{\gamma}_{t,y}}{\bar{\gamma}_{t,y}-\bar{\gamma}_{z,y}}$.

Proof: The proof is omitted due to space limitation. ■

[1] A block corresponds to the time duration necessary to send one packet.

C. Delay modeling

This section is devoted to the delay analysis of MTC systems. The MTC devices in the network are equipped with a buffer to store the packets before their transmission. The use of buffers improves the control of packet flow and reduces the network congestion. As mentioned in Section II-A, the MTC devices can attach to the eNodeBs directly or through the cluster head. The packets are generated at the MTC devices according to a Poisson distribution with a rate λ. For the packets received at the cluster head, the arrival of these packets follows a Poisson distribution with a rate λ. This section focus on the analyses of the average sojourn time in the buffer denoted as $\mathcal{D}_{x,y}^{[\lambda]}$ and the average waiting time for a data packet referred to as $\mathcal{W}_{x,y}^{[\lambda]}$.

The average waiting time $\mathcal{W}_{x,y}^{[\lambda]}$ in the buffer of x for a data packet is the average time spent by a packet in the buffer of x which begins from the generation/arrival[2] of that packet at the buffer of x until the start of its transmission. Note that the successful reception of a packet at the destination y occurs after a random number of retransmissions. To quantify the delay associated with the retransmission events, we measure the average sojourn time $\mathcal{D}_{x,y}^{[\lambda]}$ of a packet in the buffer of x, which is defined as the average time elapsed from the generation/arrival of the packet until its successful reception at the next destination node y. The packet's sojourn time in the buffer can be evaluated using the Pollaczek-Khinchin equation as [13]

$$\mathcal{D}_{x,y}^{[\lambda]} = \mathcal{W}_{x,y}^{[\lambda]} + \mathbb{E}(T_{x,y})T_F, \tag{6}$$

where T_F is the time required for a single transmission of a given packet and $\mathbb{E}(T_{x,y})$ is the average number of retransmissions for the packets sent from x to y. For the ARQ scheme, the packet is retransmitted until successful reception at the receiver y or a maximum number of retransmissions M is reached. In case of reception failure after M retransmissions the packet is discarded. The number of retransmissions $T_{x,y}$ varies randomly according to the conditions of the fading channel between the nodes x and y. The average number of retransmissions $\mathbb{E}(T_{x,y})$ can be expressed as [14]

$$\begin{aligned}\mathbb{E}(T_{x,y}) &= 1+\sum_{m=1}^{M-1} P(F^1,...,F^m) = 1+\sum_{m=1}^{M-1} (\mathcal{P}_{x,y})^m \\ &= \sum_{m=0}^{M-1} (\mathcal{P}_{x,y})^m = \frac{1-(\mathcal{P}_{x,y})^M}{1-\mathcal{P}_{x,y}},\end{aligned} \tag{7}$$

where $P(F^1,...,F^m)$ is the probability of a reception failure at the $1,\ldots,m$th retransmissions. Since the channel realizations in each transmission are independent identically distributed, the event of reception failure at each step are independent and have equal probabilities, thus $P(F^1,...,F^m) = (\mathcal{P}_{x,y})^m$. Using (7), we can conclude that the average number of retransmissions $\mathbb{E}(T_{x,y})$ increases as the outage probability $\mathcal{P}_{x,y}$ and the value of M increase.

[2]The term generation is utilized if the MTC device x communicate directly with the eNodeB y in this case the packet is generated at x, whereas the term arrival is used in case the device x forwards the traffic of another MTC device to the destination eNodeB y.

The average waiting time $\mathcal{W}_{x,y}^{[\lambda]}$ for a data packet can be obtained as [13]

$$\mathcal{W}_{x,y}^{[\lambda]} = \frac{\lambda\mathbb{E}(T_{x,y}^2)T_F^2}{2(1-\rho)} + \frac{T_F}{2}, \tag{8}$$

where ρ is a parameter which should satisfy the following stability condition

$$\rho = \lambda\mathbb{E}(T_{x,y})T_F < 1. \tag{9}$$

The term $\mathbb{E}(T_{x,y}^2)$ represents the second-order moment of the number of retransmissions $T_{x,y}$ and can be expressed as [14]

$$\mathbb{E}(T_{x,y}^2) = 1+\sum_{m=1}^{M-1}(2m+1)P(F^1,...,F^m). \tag{10}$$

Using [15, Eq. (0.113)] and the equality $P(F^1,...,F^m) = (\mathcal{P}_{x,y})^m$, we can further simplify the expression of $\mathbb{E}(T_r^2)$ as

$$\begin{aligned}\mathbb{E}(T_{x,y}^2) &= 1+\sum_{m=1}^{M-1}(2m+1)\left(\mathcal{P}_{u,v}\right)^m = \sum_{m=0}^{M-1}(2m+1)\left(\mathcal{P}_{u,v}\right)^m \\ &= \frac{1-(2M-1)\left(\mathcal{P}_{u,v}\right)^M}{1-\mathcal{P}_{u,v}} + \frac{2\mathcal{P}_{u,v}\left(1-(\mathcal{P}_{u,v})^{M-1}\right)}{(1-\mathcal{P}_{u,v})^2}.\end{aligned} \tag{11}$$

From (11), it can be concluded that the second-moment $\mathbb{E}(T_{x,y}^2)$ of the number of retransmissions increases as the outage probability and M increase. Using (8), it can be clearly seen that the average waiting time $\mathcal{W}_{x,y}^{[\lambda]}$ is proportional to the second-moment $\mathbb{E}(T_{x,y}^2)$. Similarly, it can be deduced from (6) that the sojourn time $\mathcal{D}_{x,y}^{[\lambda]}$ is proportional to both the first-moment, $\mathbb{E}(T_{x,y})$, and the second-moment, $\mathbb{E}(T_r^2)$, of the number of retransmissions. Since both $\mathbb{E}(T_{x,y})$ and $\mathbb{E}(T_{x,y}^2)$ increase proportionally with the outage probability and M, consequently a larger average sojourn time $\mathcal{D}_{x,y}^{[\lambda]}$ and a longer average waiting time $\mathcal{W}_{x,y}^{[\lambda]}$ are experienced by the packets as the outage probability $\mathcal{P}_{x,y}$ and the value of M increase.

D. Energy consumption modeling

In this section, we study the energy consumption of MTC devices. To improve the communication reliability between any two devices x and y, an ARQ scheme is utilized for data transmission. The fact that the number of retransmissions varies depending on the channel conditions makes the consumed power a random variable. We will first study in this section the average consumed power and deduce from that the average consumed energy. The average consumed power $\bar{P}$ for the ARQ scheme can be determined as

$$\begin{aligned}\bar{P} &= P_x\cdot P(S^1) + 2P_x\cdot P(F^1,S^2)+... \\ &\quad +(M-1)P_x\cdot P(F^1,...,S^{M-1}) + MP_x\cdot P(F^1,...,F^{M-1}) \\ &= P_x\cdot\left(1+\sum_{m=1}^{M-1}P(F^1,...,F^m)\right) = P_x\cdot\left(1+\sum_{m=1}^{M-1}(\mathcal{P}_{x,y})^m\right) \\ &= P_x\cdot\mathbb{E}(T_{x,y}) = P_x\cdot\frac{1-(\mathcal{P}_{x,y})^M}{1-\mathcal{P}_{x,y}},\end{aligned} \tag{12}$$

where the term P_x stands for the power per retransmission[3] at a given node x. We denote by $P(S^1)$ the probability of successful reception at node y of the first transmission, while $P(F^1,...,S^{M-1})$ refers to the probability of a reception failure in the 1st, 2nd, ..., $(M-2)$th retransmissions and a successful reception at the $(M-1)$th retransmission. If the packet is

[3]We assume that the power is constant for all retransmissions.

successfully received after the first transmission (this event occurs with a probability $P(S^1)$), the amount of consumed power would be equal to P_x. If the packet is received correctly after two retransmissions (the probability of this event is $P(F^1, S^2)$), the amount of consumed power would be equal to $2P_x$. The consumed power would be equal to MP_x if the 1st, …, $(M-1)$th retransmissions fails (the probability of this event is $P(F^1, ..., F^{M-1})$). The average consumed power is obtained by summing up all the possible values of consumed power weighted by their respective probability of occurrence. The result in (12) shows that we can express the average consumed power as the product of two terms: the power per retransmission P_x and the average number of retransmissions $\mathbb{E}(T_{x,y})$.

The average consumed energy $\mathcal{E}_{x,y}$ can be obtained as

$$\begin{aligned}\mathcal{E}_{x,y} &= P_x \cdot T_F \cdot P(S^1) + 2P_x \cdot T_F \cdot P(F^1, S^2) + ... + (M-1)P_x \\ &\cdot T_F \cdot P(F^1, ..., S^{M-1}) + MP_x \cdot T_F \cdot P(F^1, ..., F^{M-1}) \\ &= P_x \cdot T_F \cdot \left(1 + \sum_{m=1}^{M-1} (\mathcal{P}_{x,y})^m\right) = P_x \cdot T_F \cdot \mathbb{E}(T_{x,y}) = \bar{P} \cdot T_F.\end{aligned} \tag{13}$$

Using (13), it can be noticed that there is a linear relation between the average consumed energy $\mathcal{E}_{x,y}$ and the average consumed power $\bar{P}$. On the other hand, it can be deduced from (12) that the average consumed power $\bar{P}$ increases proportionally with the average number of transmissions $\mathbb{E}(T_{x,y})$. We recall that as the outage probability $\mathcal{P}_{x,y}$ and the value of M increase, a larger value of $\mathbb{E}(T_{x,y})$ is witnessed which yields a higher average consumed power and energy $\bar{P}$ and $\mathcal{E}_{x,y}$, respectively. Note that a higher consumed power and energy lead to a shorter network lifetime.

III. Machine-Type communications optimal schemes

In this section, we present two solutions to forward the data from MTC devices to eNodeBs which differ mainly on their objectives to deal with the problem. The first solution, named *EA-MTC* (energy aware scheme for machine type communication), that aims to reduce the energy consumption when forwarding data from devices to the eNodeBs. The second solution, named *DA-MTC* (delay aware scheme for machine type communication), that aims to reduce the data transfer delay when forwarding data from devices to the eNodeBs. In this paper, we aim to find optimal solutions to deal with the problem of data forwarding in MTC network. For this reason, the proposed solutions are formalized through linear programming. To lightweight the linear programing complexity of each solution, we aim to reduce the search space of the receivers for each MTC device x to a feasible set $\eta(x)$. It is intuitive that if two nodes are not able to communicate in an interference-free environment, they will definitely fail to communicate in presence of interference. We define the transmission range ρ_x for each MTC device x according to its transmission power and the receivers sensitivity γ_{th} then the set of potential receivers of this device is restricted to the receivers that are within a distance ρ_x from device x. Formally, $\eta(x)$ is constituted with all the receivers that belong to a circle centered at x with radius ρ_x. In fact, any receiver $y \in \mathcal{X} \cup \mathcal{Y}$ is a candidate parent for x (i.e., $x \in \eta(x)$) iff the euclidean distance between x and y does not exceed ρ_x (i.e., $d(x, y) \leq \rho_y$). The following lemma specifies how to determine the transmission range, ρ_x:

Lemma 1. *MTC device x can forward its data to a receiver y iff y is within a distance less than ρ_x from node x, where ρ_x is defined as follows:*

$$\rho_x = d_0 \sqrt[\beta]{\frac{P_x}{N_0 \gamma_{\text{th}}}}, \tag{14}$$

P_x is the transmitter power of device x, γ_{th} is the receiver sensitivity, and β is the path loss exponent.

Proof: Using (2) and (3) we can express the average received signal-to-noise ratio at node y as

$$\bar{\gamma}_{x,y} = \frac{P_x \mathbb{E}[\alpha_{x,y}^2]}{N_0} = \frac{P_x}{N_0}\left(\frac{d_0}{d_{x,y}}\right)^\beta. \tag{15}$$

In average, the transmission from x to y succeeds only if the average received signal-to-noise ratio at y exceeds the sensitivity of the receiver y denoted as γ_{th}, i.e., $\bar{\gamma}_{x,y} \leq \gamma_{\text{th}}$. It follows that

$$\gamma_{\text{th}} = \frac{P_x}{N_0}\left(\frac{d_0}{\rho}\right)^\beta, \quad \text{thus} \quad \rho_x = d_0 \sqrt[\beta]{\frac{P_x}{N_0 \gamma_{\text{th}}}}.$$

■

A. EA-MTC: energy aware scheme for machine type communication

In this subsection, we present *EA-MTC* that aims to forward the data from MTC devices to eNodeBs while minimizing the energy consumption. *EA-MTC* models the problem as an integer program. As mentioned in Table I, we define three variables: (i) $X_{x,y}$ is a boolean decision variable that is set to 1 if device $x \in \mathcal{X}$ selects $y \in \mathcal{X} \cup \mathcal{Y}$ as receiver; (ii) $\mathcal{E}$ is the maximum amount of energy can be consumed in the network; (iii) F is a matrix of integers (with elements $F_{x,y}$) that introduces a notion of flow in order to force the optimization task to connect all devices, with at most two hops, to eNodeBs $\mathcal{Y}$ without creating cycles. To guarantee that, we have to ensure that the number of entering flows at $\mathcal{Y}$ equals to the number of devices $\mid \mathcal{X} \mid$. The *EA-MTC* formulation is as follows:

$$\mathbf{min}\, \mathcal{E} \tag{16}$$

S.t,

$$\forall x \in \mathcal{X} : \sum_{y \in \eta(x)} X_{x,y} = 1. \tag{17}$$

$$\forall x \in \mathcal{X} : (\sum_{y \in \mathcal{X} \cap \eta(x)} X_{y,x}) \times (\sum_{t \in \mathcal{X} \cap \eta(x)} X_{t,x}) = 0 \tag{18}$$

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y} \cap \eta(x)} F_{x,y} = |\mathcal{X}|. \tag{19}$$

$$\forall x \in \mathcal{X} : \sum_{y \in (\mathcal{X} \cup \mathcal{Y}) \cap \eta(x)} F_{x,y} - \sum_{y \in \mathcal{X} \cap \eta(x)} F_{y,x} = 1. \tag{20}$$

$$\forall x \in \mathcal{X}, \forall y \in \eta(x) : 0 \leq F_{x,y} \leq |\mathcal{X}| X_{x,y}. \tag{21}$$

$$\sum_{x \in \mathcal{X}} \sum_{y \in \eta(x)} X_{x,y} \mathcal{E}_{x,y} \leq \mathcal{E}. \tag{22}$$

$$\forall x \in \mathcal{X}: \sum_{\forall y \in \mathcal{Y} \cap \eta(x)} X_{x,y} \mathcal{D}_{x,y}^{[\lambda_x]} \leq \mathcal{D}_{MAX}. \tag{23}$$

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{X} \cap \eta(x) : if \ X_{x,y} = 1 \ then$$

$$\mathcal{D}_{x,y}^{[\lambda_x]} + \sum_{\forall z \in \mathcal{Y} \cap \eta(y)} X_{y,z} (\mathcal{D}_{y,z}^{[\lambda_x]} + \mathcal{D}_{y,z}^{[\lambda_y]}) \leq \mathcal{D}_{MAX}. \tag{24}$$

$$\forall y \in \mathcal{Y}: \sum_{x \in \mathcal{X} \cap \eta(x)} X_{x,y} \leq \mathcal{C}_{MAX}. \tag{25}$$

In *EA-MTC*, we aim to minimize the energy consumption as much as possible while the the data transfer delay does not exceed a threshold $\mathcal{D}_{MAX}$ and the contention at eNodeBs does not exceed a threshold $\mathcal{C}_{MAX}$. Meanwhile, the constraints are used to ensure the following conditions: constraint (17) ensures that each MTC device has only one receiver; constraint (18) ensures the data forwarding path between each device in $\mathcal{X}$ and an eNodeBs in $\mathcal{Y}$ should not exceed two hops. If a device $x \in \mathcal{X}$ is selected as receiver by another device $y \in \mathcal{X}$, the former should not select another MTC device t as receiver. Formally, $\sum_{y \in \mathcal{X} \cap \eta(x)} X_{y,x} = 1 \Rightarrow \sum_{t \in \mathcal{X} \cap \eta(x)} X_{x,t} = 0$. While constraints (19), (20) and (21) are used for modeling the network connectivity and ensuring that all the devices can transmit their data to eNodeBs $\mathcal{Y}$. To make sure that the formed topology connects all the devices to the eNodeBs $\mathcal{Y}$, a packet flow is mimicked, which is generated and routed from the devices to the eNodeBs $\mathcal{Y}$. Each device in the network has to generate only one packet. Every relay node forwards the received packets plus its own packet to the eNodeBs $\mathcal{Y}$. The mimicked packet flow should be routed within the constructed tree. Constraint (19) captures the fact that the number of packets received by $\mathcal{Y}$ equals to the number of devices (i.e., $\mid \mathcal{X} \mid$). Constraint (20) ensures that each node generates only one packet. Constraint (21) forces the generated flow to be routed only within the constructed topology, from each device x to its receiver y ($X_{x,y} = 0 \Rightarrow F_{x,y} = 0$).

Constraint (22) ensures that sum of energy consumed by the devices $\mathcal{X}$ should do not exceed the maximum amount of energy variable $\mathcal{E}$ that we aim to minimize. Based on (17), each device x should select only one receiver, and hence $\sum_{y \in \eta(x)} X_{x,y} \mathcal{E}_{x,y}$ represents the energy consumption between device x and its receiver. Constraint (23) ensures that delay between a device and its receiver form eNodeBs $\mathcal{Y}$ should do not exceed $\mathcal{D}_{MAX}$. While constraint (24) ensures that if a device x selects another device y as receiver, the time required to forward packets from x to y and also the time to forward the packets received and generated at y to eNodeBs $\mathcal{Y}$ should not exceed $\mathcal{D}_{MAX}$. By constraints (23) and (24), we ensure that the maximum data transfer delay between devices and eNodeBs should not exceed the maximum tolerated delay in the network $\mathcal{D}_{MAX}$. Constraint (24) ensures that the maximum contention at each eNodeB $\mathcal{Y}$ should not exceed the maximum tolerated contention $\mathcal{C}_{MAX}$. In fact, we aim to ensure that each eNodeB $y \in \mathcal{Y}$ should not be selected as receiver by more than $\mathcal{C}_{MAX}$ devices.

However, the above optimization problem is not linear due to the constraints defined by (18) and (24). In order to simplify the solution, the following transformations are applied to (18) and (24) in order to convert the model to a linear program. Constraint (18) can be transformed as follow:

$$\forall x \in \mathcal{X}: \sum_{y \in \mathcal{X} \cap \eta(x), x \neq y} X_{y,x} \leq (1 - \sum_{y \in \mathcal{X}, x \neq y} X_{x,y}) |\mathcal{X}| \tag{26}$$

It is obvious that $\sum_{y \in \mathcal{X} \cap \eta(x), x \neq y} X_{y,x} \leq |\mathcal{X}|$. From (26), if device x selects a device y as receiver, then all the other devices are forbidden to select it as receiver. Otherwise, can be a receiver at most for all the devices $|\mathcal{X}|$. Meanwhile, constraints (24) is transformed to linear constraints as follows:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{X} \cap \eta(x):$$

$$\mathcal{D}_{x,y}^{[\lambda_x]} + \sum_{\forall z \in \mathcal{Y} \cap \eta(y)} X_{y,z} (\mathcal{D}_{y,z}^{[\lambda_x]} + \mathcal{D}_{y,z}^{[\lambda_y]}) \leq \mathcal{D}_{MAX} + A(1 - X_{x,y}), \tag{27}$$

where A is a large number ($A \to \infty$).

From (27), according to the values of $X_{x,y}$, the following could be noted: If $X_{x,y} = 0$, $\mathcal{D}_{x,y}^{[\lambda_x]} + \sum_{\forall z \in \mathcal{Y} \cap \eta(y)} X_{y,z} (\mathcal{D}_{y,z}^{[\lambda_x]} + \mathcal{D}_{y,z}^{[\lambda_y]})$ can take any values without any restriction as $A \to \infty$. Otherwise, i.e., $X_{x,y} = 1$, we have $\mathcal{D}_{x,y}^{[\lambda_x]} + \sum_{\forall z \in \mathcal{Y} \cap \eta(y)} X_{y,z} (\mathcal{D}_{y,z}^{[\lambda_x]} + \mathcal{D}_{y,z}^{[\lambda_y]}) \leq \mathcal{D}$, which is equivalent to (24).

Based on the above analysis, the optimization problem would be transformed to the following linear program:

$$\mathbf{min}\, \mathcal{E}$$

S.t, (17), (19), (20), (21), (22), (23), (25), (26) and (27).

B. DA-MTC approach

In this subsection, we present the *DA-MTC* approach, which aims to gather the data from different MTC devices as fast as possible. Using the same approach as for *EA-MTC*, *DA-MTC* is modeled as an integer program. The *DA-MTC* formulation is as follows:

$$\mathbf{min}\, \mathcal{D}$$

S.t,
(17), (19), (20), (21), (25), (26) and

$$\sum_{x \in \mathcal{X}} \sum_{y \in \eta(x)} X_{x,y} \mathcal{E}_{x,y} \leq \mathcal{E}_{MAX}. \tag{28}$$

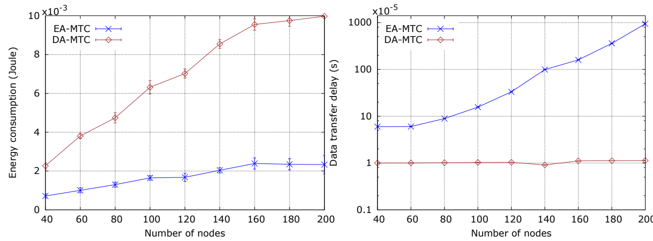
$$\forall x \in \mathcal{X}: \sum_{\forall y \in \mathcal{Y} \cap \eta(x)} X_{x,y} \mathcal{D}_{x,y}^{[\lambda_x]} \leq \mathcal{D}. \tag{29}$$

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{X} \cap \eta(x):$$

$$\mathcal{D}_{x,y}^{[\lambda_x]} + \sum_{\forall z \in \mathcal{Y} \cap \eta(y)} X_{y,z} \mathcal{D}_{y,z}^{[\lambda_x]} \leq \mathcal{D} + M \times (1 - X_{x,y}). \tag{30}$$

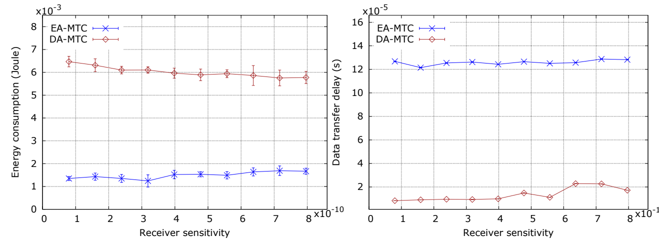
IV. SIMULATION RESULTS

In this section, we evaluate the proposed solutions, *EA-MTC* and *DA-MTC* via simulation. The solutions are evaluated through python, extended package for graph theory called networkx [16] and IBM ILOG CPLEX version 12.6.1. The algorithms are evaluated in terms of the following metrics: i) Energy consumption, which is defined as the amount of energy consumed by all devices. ii) End-to-end delay, which is defined as the maximum time required for each device to succeed to forward all its packets to the eNodeBs $\mathcal{Y}$. In the simulation results, each plotted point represents the average of



(a) Energy consumption (b) Delay of data transfer

Fig. 1: The impact of devices number on the network performances at each solution



(a) Energy consumption (b) Delay of data transfer

Fig. 2: The impact of receiver sensitivity on the network performances at each solution

10 executions. The plots are presented with 95% confidence interval. The evaluation of different algorithms is performed by varying the receiver sensitivity threshold γ_{th} and the number of devices $|\mathcal{X}|$. We conduct two set of experiments. Firstly, we vary the number of nodes $|\mathcal{X}|$ and fix γ_{th} to $10^{-9.4}$. Then, we vary γ_{th} while fixing $|\mathcal{X}|$ to 100.

Fig. 1 illustrates the performance of different protocols as a function of number of nodes $|\mathcal{X}|$. The first observation we can draw from this figure that the increase in the number of devices has a negative impact on the energy consumption and the end-to-end delay. Fig. $1(a)$ shows that *EA-MTC* outperforms *DA-MTC* in terms of energy consumption whatever the number of devices. Moreover, the gap between *EA-MTC* and *DA-MTC* increases proportionally with number of devices in the network. While Fig.$1(b)$ shows that *DA-MTC* outperforms *EA-MTC* in terms of end-to-end delay. We observe that the end-to-end delay in *DA-MTC* remain constant whatever the number of nodes while it increases proportionally with the number of nodes when using *EA-MTC*.

Fig. 2 illustrates the performance of different protocols as a function of receiver sensitivity γ_{th}. From Fig. $2(a)$, we can observe that *EA-MTC* outperforms *DA-MTC* in terms of energy consumption whatever γ_{th} values. Fig. $2(b)$ shows that *DA-MTC* outperforms *EA-MTC* in terms of end-to-end delay whatever the receiver sensitivity γ_{th} values. Results obtained from the conducted simulations demonstrate the efficiency of each proposed solution in achieving its key design goals.

V. Conclusion

One key vision of the upcoming $5G$ network is to interconnect a huge number of MTC devices to the internet. An important challenge is to cope with the amount of signaling to be generated by these devices. To deal with this problem, we have proposed two solutions, named *EA-MTC* and *DA-MTC*, that manage the communication between devices and eNodeBs to achieve their objectives. The first solution favors the reduction of energy consumption, whereas the second one favors the reduction of data transfer delay. Simulation results proved the efficiency of each solution in achieving its design goal.

References

[1] L. Changwei, "Telco development trends and operator strategies," *Win-Win Magazine*, no. 13, pp. 19 – 22, 2012.

[2] 3GPP TS 23.682 "architecture enhancements to facilitate communications with packet data networks and applications" version 11.0.0, 2012-03-12.

[3] A. Tsai, L. Wang, J. Huang, and T. Lin, "Overload control for machine type communications with femtocells," in *IEEE VTC'12*. IEEE, 2012, pp. 1–5. [Online]. Available: http://dx.doi.org/10.1109/VTCFall.2012.6399236

[4] O. Arouk, A. Ksentini, and T. Taleb, "Group paging optimization for machine-type-communications," in *Communications (ICC), 2015 IEEE International Conference on*, June 2015, pp. 6500–6505.

[5] A. Amokrane, A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, "Congestion control for machine type communications," in *Communications (ICC), 2012 IEEE International Conference on*, June 2012, pp. 778–782.

[6] A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, "Cellular-based machine-to-machine: overload control," *Network, IEEE*, vol. 26, no. 6, pp. 54–60, November 2012.

[7] P. Janis, C.-H. Yu, K. Doppler, C. Ribeiro, C. Wijting, K. Hugl, and O. Tirkkonen, "Device-to-device communication underlaying cellular communications systems," *Int'l J. of Communications, Network and System Sciences*, vol. 2, no. 3, 2009.

[8] Z. Uykan and R. Jantti, "Transmission-order optimization for bidirectional device-to-device (d2d) communications underlaying cellular tdd networks -a graph theoretic approach," *IEEE Journal on Selected Areas in Communications*, no. 99, 2015.

[9] B. Cho, K. Koufos, R. Jäntti, Z. Li, and M. A. Uusitalo, "Spectrum allocation for multi-operator device-to-device communication," in *2015 IEEE International Conference on Communications, ICC 2015, London, United Kingdom, June 8-12, 2015*. IEEE, 2015, pp. 5454–5459. [Online]. Available: http://dx.doi.org/10.1109/ICC.2015.7249191

[10] Z. Liang, K. Jussi, M. Nicolas, Y. Beyene, K. Ruttik, R. Jntti, and O. Tirkkonen, "Creating secondary spectrum usage opportunity for d2d communication with interference cancellation," in *IEEE DySPAN'15*, June 2015, pp. 6500–6505.

[11] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels*, 2nd ed. New Jersey, USA: John Wiley & Sons, 2005.

[12] A. F. Molisch, *Wireless Communications*. Wiley-IEEE Press, 2005.

[13] W. Chan, T.-C. Lu, and R.-J. Chen, "Pollaczek-Khinchin formula for the M/G/1 queue in discrete time with vacations," *IEE Proceedings - Computers and Digital Techniques*, vol. 144, no. 4, pp. 222–226, Jul. 1997.

[14] A. Chelli, E. Zedini, M.-S. Alouini, J. R. Barry, and M. Pätzold, "Performance and delay analysis of hybrid ARQ with incremental redundancy over double Rayleigh fading channels," *IEEE Trans. on Wireless Commun.*, vol. 13, no. 11, pp. 6245–6258, Nov. 2014.

[15] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products, Seventh Edition*. Academic Press, 2007.

[16] Networkx, "http://networkx.lanl.gov."