QoS-Aware Splitting and Radio Resource Allocation for Machine Type Communications

David Martin Amitu[†‡], Roseline Nyongarwizi Akol[†] and Peter Nakeba[‡]
[†]Department of Electrical & Computer Engineering, Makerere University
[‡]Avytel Informatica
Emails: †‡damitu@cedat.mak.ac.ug; †rnakol@cedat.mak.ac.ug and ‡nakebap@avytel.com;

***Abstract*—Machine Type Communications (MTC) networks need to resolve the key issues of massive access requirements, small data transmissions and diversity in Quality of service (QoS) requirements. The prominent approaches to address these challenges involve the use of MTC Gateways (MTCGs) as access points to the network, the use of MTC Devices (MTCDs) as relays and QoS clustering at the eNodeB. The MTCG-based approaches generally envision communication between MTCDs and the eNodeB through the MTCG only, for all MTCDs within the range of an MTCG. In this paper, we propose to divide MTCDs into delay-tolerant and delay-intolerant types, a process we coin "MTCD splitting". The proposed approach involves cluster formation, MTCD splitting and QoS-aware radio resource allocation. Our simulation results show better performance in terms of average throughput satisfaction and average QoS violation probability when MTCD splitting is employed as compared to existing techniques without MTCD splitting.**

***Keywords*—Quality of Service; Machine Type Communications; Radio Resource Allocation; MTCD Splitting; Cluster formation**

I. Introduction

Massive MTC's integration into cellular networks presents a number of challenges, for example support for a large number of devices per cell, small data transmissions from the large number of devices on the uplink, and the Quality of service (QoS) guarantees for MTC devices (MTCDs) with diverse requirements. The large number of MTCDs attempting eNodeB access concurrently leads to congestion at the control channels due to capacity limitation in the cellular networks. The small data transmissions waste a precious bandwidth resource in the networks and MTC QoS requirements from a large number of MTCDs are challenging to satisfy due to their diversity. Hence, techniques to handle massive MTC in order to alleviate some of these problems then become a necessity. Most existing literature have proposed approaches to handle a large number of MTCDs, for example, the use of an MTC Gateway (MTCG) [1], [2], [3] to support the massive access requirements, also the use of MTC Devices (MTCDs) as relays between the eNodeB and other MTCDs [4], [5]. Although these schemes reduce direct connections to the eNodeB, a significant delay is encountered which could hamper delay sensitive applications [6] and hence violating their QoS requirements. Dealing with MTC massive access while satisfying Quality of Service (QoS) requirements is crucial for the next generation networks [7], [8], [9]. In the next generation networks like 5G, the number of MTCDs and direct connections to the eNodeB from MTC Devices (MTCDs) are expected to grow [2]. Direct connections are efficient for delay sensitive MTC applications but as the number of MTCDs grow, the provision of QoS guarantees for delay sensitive applications become infeasible. In [6] MTC devices are grouped into clusters based on their QoS requirement at the eNodeB. All MTCDs access the eNodeB directly hence only direct connections are allowed in that case although the authors suggested that the same approach can be implemented on the MTCG which still becomes indirect access only.

The proposed approaches above can be classified into two main categories for dealing with massive MTC, (i) use of indirect connections to eNodeB (ii) allow direct connections to eNodeB and perform QoS clustering at the eNodeB. As pointed out earlier, allowing indirect connections only could impact on the MTCD delay requirements as their number grows and on the other hand direct connections are good for delay sensitive applications but due to the limitation in the control channel capacity, allowing access to all MTCDs becomes a challenge as the number of MTCDs grow. Consequently satisfying the throughput requirements for all direct connections to the eNodeB becomes infeasible as the MTCDs increase. Having discussed the problems presented by each of those approaches, we therefore propose a hybrid access scheme for massive MTC. In this scheme, MTCDs form clusters and the selected Cluster Head (CH) is responsible for splitting the MTCDs into direct access and indirect Access MTCDs based on the QoS requirements of the application running on each MTCD. This scheme allows for various applications with different QoS requirements to run in one MTCD. The direct access MTCDs are high priority and delay sensitive MTCDs that connect directly to the eNodeB for radio resource allocation whereas the indirect access MTCDs are low priority MTCDs that connect to the eNodeB through the Cluster Head (CH) for radio resource allocation. This paper is organized as follows, the next section presents the model and architecture used, section III presents the access splitting and resource allocation scheme, performance analysis is then presented in section IV and the paper is concluded in section V.

II. Model and Architecture

The proposed MTC system architecture is shown in Fig.1. The proposed MTCD splitting and radio resource allocation scheme is composed of three main steps, namely; cluster formation, MTC Device (MTCD) splitting and QoS aware radio resource allocation at the eNodeB. Once clusters are formed, a Cluster Head (CH) with ability to support a high number of MTCDs is then selected from among the MTCDs in each cluster. An MTC Cluster Head (MTC CH) is then respon-

978-1-5386-4649-6/18/$31.00

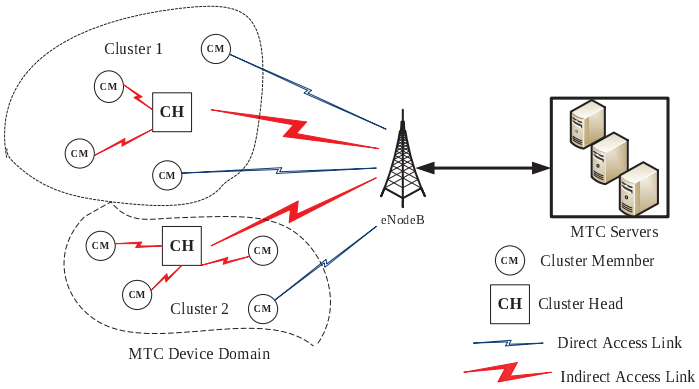

Fig. 1: MTC System architecture showing direct and indirect eNodeB access schemes.

sible for splitting the MTCDs in its cluster into direct access and indirect access MTCDs based on the QoS requirements. Each MTCD will then access the eNodeB either directly or indirectly through the MTC CH for radio resource allocation. The eNodeB has two objectives, first eNodeB has to maximize the number of admitted direct access MTCDs or the number of direct access MTCDs whose QoS requirements can be fully satisfied. Secondly, the eNodeB has to serve the indirect access MTCDs with best effort by minimizing the gap between the required radio resources and allocated resources.

III. MTCD Splitting and Resource Allocation

A. Cluster Formation

MTC Devices (MTCDs) self-organize into clusters based on their capability (buffer size and processing) or maximum number of MTCDs that can be supported when the MTCD acts as a Cluster Head (CH) and the node degree in case of a tie. In this step, the objective is to select a Cluster Head (CH) with the ability to support a high number of MTCDs, the selection is based on the assumption that the diversity of QoS requirements in MTC implies the diversity of capabilities of MTCDs. Each MTCD sends its capability metric and node degree to all its one-hop neighbors. If the node has the highest capability metric among all its one hop neighbors, then it can declare itself as an MTC CH by sending an MTC CH confirmation message to all its neighbors. If there are nodes in the neighborhood with high capability, then the MTCD selects the node with the highest capability metric as an MTC CH. If the nodes have the same capability metric (highest), then the nodes select the one with the largest node degree as an MTC CH. Once MTC CHs are selected, each MTCD associates with one of the MTC CHs and therefore, belongs to the specific cluster of the associated MTC CH. In this paper we mainly focus on MTCD access splitting and resource allocation, henceforth we assume capable MTC CHs have already been selected.

B. MTCD Splitting

The selected MTC CH splits the MTCDs in its cluster into direct access and indirect access MTCDs based on the QoS requirements. MTCDs with stringent QoS requirements are designated as direct access MTCDs and will connect directly to the eNodeB for resource allocation. MTCDs with less stringent QoS requirements are designated as indirect access MTCDs and will connect to the eNodeB through an MTC CH for resource allocation. In this step, the objective of the MTC CH is to maximize the number of indirect access MTCDs subject to MTC CHs capability (buffer size and processing speed) and the MTCD QoS requirements.

MTCDs in the cluster requesting for resource allocation forward their QoS requirements to the MTC CH for access splitting. Based on each MTCD's QoS requirements, an MTC CH designates it as direct access or indirect access MTCD. High priority (delay sensitive) MTCDs are designated as direct access MTCDs and will connect directly to the eNodeB for resource allocation. The current applications running in these MTCDs require stringent QoS requirements. On the other hand, low priority (delay tolerant) MTCDs are designated as indirect MTCDs and can therefore connect to the eNodeB through an MTC CH in its cluster which is responsible for access request procedures for the rest of indirect access MTCDs in the same cluster. If I is the set of indirect access MTCDs in the cluster, then the objective of an MTC CH in performing access splitting is to maximize $|I|$, the cardinality of the set of indirect access MTCDs.

Let L_{CH} be the buffer size of an MTC CH, L_i be the data (in bits) ready for transmission at the i^{th} MTCD, t be the average MTCD to MTC CH access delay, t_{CH} be the average MTC CH to eNodeB access delay, t_i^{req} be the i^{th} MTCD delay requirements, δ be the control parameter which ensures that MTCDs which satisfy access delays only can not connect through the cluster head and N is the number of MTCDs in the cluster. An MTC CH objective can be formulated as a simple optimization problem in problem 1;

$$\begin{aligned} &\text{Maximize } |I| \\ &\text{Subject to: } \sum_{i \in I} L_i - L_{CH} \leq 0 \\ &t + t_{CH} + \delta - t_i^{req} \leq 0, i = 1, 2, \cdots, N \\ &|I| - N \leq 0 \end{aligned} \tag{1}$$

The first constraint takes into account the data buffer size limitation in an MTC CH, the second constraint ensures that the delay requirements violation is minimized by only connecting MTCDs with delay requirements greater than the minimum average access delay through indirect access and the third constraint states that the indirect access MTCDs are part of the cluster.

1) Complexity and delay: The second constraint in problem 1 ensures that the complexity and delay of solving the problem does not grow with the number of MTCDs. The constraint allows an MTC CH to connect indirectly at maximum $\left\lfloor \frac{L_{CH}}{\sum_{i=1}^{|I|} L_i} \right\rfloor$ MTCDs. Therefore as N grows, $|I| < N$, hence the complexity and delay are proportional to L_{CH}, (complexity and delay $\propto L_{CH}$) which is limited in practice in MTCDs. With this illustration, an MTC CH is expected to perform access splitting with manageable complexity even as the number of MTCDs grows. For delay sensitive applications, the splitting is performed once, and the subsequent requests are handled directly by the eNodeB in order to minimize overhead delay, however if the applications requirements change, the MTCD has to connect to the MTC CH for splitting.

C. QoS-aware Radio Resource Allocation

Let D and I be the set of direct and indirect acess MTCDs respectively, due to limited network capacity, the guaranteed QoS requirements of all direct access MTCDs may not be fulfilled. We therefore define $D_s \subseteq D$ as the set of direct access MTCDs whose QoS requirements can be satisfied fully and $|D_s|$ is the cardinality of D_s. The eNodeB can only allow access to D_s whose QoS requirements can be guaranteed. Let β_D and β_I be the required radio resources for direct and indirect access MTCDs respectively. The relationship between the required resource blocks and the throughput can be given as;

$$\beta = \left\lceil \frac{\gamma}{\psi \times \epsilon} \right\rceil \tag{2}$$

Where β is the required number of resource blocks, $\psi = (S_c \times S_y)/T_s$, S_c is the number of subcarriers per resource block, S_y is the number of OFDM symbols per resource block, T_s is the timeslot duration, ϵ is the efficiency in bits/symbol depending on the modulation and coding scheme (MCS) used, γ is the required throughput. If $\alpha_D(j)$ and $\alpha_I(j)$ represent the binary allocation vector for direct and indirect access MTCDs respectively with 1 or 0 in the j^{th} position with allocated or unallocated resource respectively. Then the gap between the required resources and allocated resources for indirect access MTCDs, G_I is given by;

$$G_I = \beta_I - \sum_{j=1}^{M} \alpha_I(j) \tag{3}$$

Where M is the maximum number of radio resources (Resource Blocks) in the eNodeB (it depends on the available bandwidth). For the case of MTC communications with $1.4MHz$ bandwidth, the total number of resource blocks is 6, hence $M = 6$. The eNodeB has to solve the following multi-objective optimization problem;

$$\begin{gathered} \text{Maximize } |D_s| \\ \text{Minimize } \max_{l \in I} G_I \\ \text{Subject to: } \sum_{j=1}^{M} \alpha_D(j) = \beta_D \\ \sum_{j=1}^{M} \alpha_I(j) \leq \beta_I \\ \alpha_D(j) + \alpha_I(j) \leq 1 \\ \alpha_D(j), \alpha_I(j) \in \{0,1\}, \forall j = 1, 2, \cdots, M \end{gathered} \tag{4}$$

The first constraint in problem 4 represents guaranteed QoS requirements for direct access MTCDs. Constraint two represents the best effort service required by the indirect access MTCDs. The third constraint demonstrates that the direct and indirect access MTCDs cannot use the same radio resource at the same time and the last part of the constraints shows the binary nature of the resource allocation vectors.

Problem 4 is a multi-objective optimization problem and it has been proved to be NP-hard [10]. We propose splitting the problem into two parts and solve it sequentially as in [11]. As pointed out earlier, guaranteeing QoS requirements for all direct access MTCDs leads to the infeasibility of the problem, so we need to employ admission control to select the subset of direct access MTCDs whose QoS requirements can be fully satisfied, D_s. The above multi-objective problem is solved by first maximizing $|D_s|$ and then minimizing $\max_{l \in I} G_I$.

1) Direct access MTCDs resource allocation: The direct access resource allocation problem involves finding and maximizing a subset of direct access MTCDs, $|D_s|$, whose QoS requirements can be fully satisfied. In other words, the eNodeB will only admit direct access MTCDs whose requirements can be guaranteed. Such a problem is equivalent to an Irreducible Infeasible Set (IIS) problem where a feasible subset of constraints is contained in a large infeasible set. IIS problems can be solved using elastic programming [12] with the help of elastic variables.

With elastic programming, each direct access MTCD k uses an elastic variable, e_k to compensate for unfulfilled QoS requirements. In order to identify the constraints that do not satisfy the requirements, the combination of sum of elastic variable minimization and filtering [13] can be performed whereby the set of inconsistent constraints is identified with positive elastic variables. We then go ahead and reformulate the first optimization problem as;

$$\begin{gathered} \text{Minimize } \sum_{k \in D} c_k * e_k \\ \text{Subject to: } e_k \geq 0, \forall k \in D \\ \sum_{j=1}^{M} \alpha^k_{D(j)} + e_k \geq \beta^k_D, \forall k \in D \\ \alpha^k_{D(j)} + \alpha^l_{I(j)} \leq 1, \forall k \in D, \forall j, \forall l \in I, \\ \alpha_D(j), \alpha_I(j) \in \{0,1\} \end{gathered} \tag{5}$$

Where c_k coefficients in the objective of problem 5 further define priority levels for different direct access MTCDs. For equal priority direct access MTCDs, $c_k = 1, \forall k \in D$. In our case, we do not differentiate these priorities for direct access MTCDs. The first constraint in problem 5, illustrates the non-negativity of elastic variables, the second constraint shows the use of elastic variables to compensate for the unfulfilled QoS requirements and therefore the allocated resources plus the elastic variables either meet or exceed the demand because the elastic variables are non-negative. The third constraint ensures that the same resource block can not be assigned to the direct and indirect access MTCDs at the same time and we can further expand it as follows;

Let $\widehat{\alpha^k_D}, \widehat{\alpha^l_I}, \mathbf{1}_M \in \mathcal{R}^M, k = 1, 2, \cdots |D|, l = 1, 2, \cdots |I|$, where $\mathbf{1}_M$ is a vector of all ones, then;

$$\begin{gathered} \widehat{\alpha^1_D} + \widehat{\alpha^1_I} \leq \mathbf{1}_M \\ \vdots \\ \widehat{\alpha^{|D|}_D} + \widehat{\alpha^{|I|}_I} \leq \mathbf{1}_M \end{gathered} \tag{6}$$

We then reorganize the above constraint into a matrix as;

$$\mathbf{A_D} = \begin{bmatrix} \widehat{\alpha_D^1} & \widehat{\alpha_D^1} \dots \widehat{\alpha_D^1} \\ \widehat{\alpha_D^2} & \widehat{\alpha_D^2} \dots \widehat{\alpha_D^2} \\ \vdots & \\ \widehat{\alpha_D^{|D|}} & \widehat{\alpha_D^{|D|}} \dots \widehat{\alpha_D^{|D|}} \end{bmatrix} \tag{7}$$

Where the selected column is the $|I|^{th}$ column vector of $\alpha_{\mathbf{D}}$ and therefore $\alpha_{\mathbf{D}} \in \mathcal{R}^{(|D|*M)\times|I|}$.
Similarly for the second component of the constraint, we form a matrix as;

$$\mathbf{A_I} = \begin{bmatrix} \widehat{\alpha_I^1} & \widehat{\alpha_I^2} \dots \widehat{\alpha_I^{|I|}} \\ \widehat{\alpha_I^1} & \widehat{\alpha_I^2} \dots \widehat{\alpha_I^{|I|}} \\ \vdots & \\ \widehat{\alpha_I^1} & \widehat{\alpha_I^2} \dots \widehat{\alpha_I^{|I|}} \end{bmatrix} \tag{8}$$

Where the selected row is the $|D|^{th}$ row vector of $\alpha_{\mathbf{I}}$ interms of $\widehat{\alpha_I} \in \mathcal{R}^M$ and therefore $\alpha_{\mathbf{I}} \in \mathcal{R}^{(|D|*M)\times|I|}$.
And for the last component of the constraint, Let $a_{ij} \in R^M$ where $a_{ij} = 1, \forall ij$ and we form a matrix;

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \dots a_{1|I|} \\ a_{21} & a_{22} \dots a_{2|I|} \\ \vdots & \\ a_{|D|1} & a_{|D|2} \dots a_{|D||I|} \end{bmatrix} \tag{9}$$

Therefore $\mathbf{A} \in \mathcal{R}^{(|D|*M)\times|I|}$.
The second constraint can then be written compactly as;

$$\begin{gathered} \mathbf{A_D} + \mathbf{A_I} - \mathbf{A} \preceq 0 \\ \mathbf{A_D}, \mathbf{A_I}, \mathbf{A} \in \mathcal{R}^{(|D|*M)\times|I|} \end{gathered} \tag{10}$$

where "$\preceq$" means that all elements of the resulting matrix in 10 are non positive. Let $\widetilde{\alpha_D^k} = 1^T \widehat{\alpha_D^k}$ and since $k = 1, \cdots, |D|$ we have $\widetilde{\alpha_D} \in \mathcal{R}^{|D|}$, similarly $\widetilde{\alpha_I^l} = 1^T \widehat{\alpha_D^l}$ and since $l = 1, \cdots, |I|$, we have $\widetilde{\alpha_I} \in \mathcal{R}^{|I|}$. We can now go ahead and rewrite problem 5 as;

$$\begin{gathered} \text{Minimize } c^T e \\ \text{Subject to: } e \succeq 0 \\ \beta_D - \widetilde{\alpha_D} - e \preceq 0 \\ \mathbf{A_D} + \mathbf{A_I} - \mathbf{A} \preceq 0 \\ \beta_D, \widetilde{\alpha_D}, e, c, \in \mathcal{R}^{|D|}, \mathbf{A_D}, \mathbf{A_I}, \mathbf{A} \in \mathcal{R}^{(|D|*M)\times|I|} \end{gathered} \tag{11}$$

The optimal value of e_k is zero, therefore we determine $D_s = \{\forall k \in D | e_k = 0\}$. Once D_s and the resource allocations matrix R_{D_s} of dimension $|D_s| \times M$ are determined, the remaining resources are allocated to the indirect access MTCDs. Then the allocated resources for direct access, AL_D can be computed as;

$$AL_D = \sum_{i=1}^{|D_s|} \sum_{j=1}^{M} b_{ij} \tag{12}$$

Where b_{ij} is an element of R_{D_s} matrix and it represents the j^{th} resource block allocated to MTCD i.

2) Indirect access MTCDs resource allocation: The available resources for indirect access MTCDs depend on AL_D. Let the total available resources in the eNodeB be R_M, if AL_D has already been allocated for direct access MTCDs, then the available resources for indirect access MTCDs, $R_I = R_M - AL_D$. The indirect access problem can be formulated as;

$$\begin{gathered} \text{Minimize } \max_{l \in I} G_I \\ \text{Subject to: } \sum_{j=1}^{M} \alpha_I^l(j) \leq \beta_I^l, \forall l \in I \\ \alpha_D^k(j) + \alpha_I^l(j) \leq 1, \forall k \in D, \forall j, \forall l \in I \\ \sum_{j=1}^{M} \alpha_I^l(j) \leq R_I, \forall l \in I \end{gathered} \tag{13}$$

Applying the same procedure as in direct access, we reformulate problem 13 as;

$$\begin{gathered} \text{Minimize } \max_{l \in I} \beta_I - \widetilde{\alpha_I} \\ \widetilde{\alpha_I} - \beta_I \preceq 0 \\ \mathbf{A_D} + \mathbf{A_I} - \mathbf{A} \preceq 0 \\ \widetilde{\alpha_I} - \widetilde{R_I} \preceq 0 \\ \beta_I, \widetilde{\alpha_I}, \widetilde{R_I} \in \mathcal{R}^{|I|}, \mathbf{A_D}, \mathbf{A_I}, \mathbf{A} \in \mathcal{R}^{(|D|*M)\times|I|} \end{gathered} \tag{14}$$

where $\widetilde{R_I} = \mathbf{1} \times R_I, \mathbf{1} \in \mathcal{R}^{|I|}$.

D. QoS Violation Probability

In order to provide statistical delay guarantees, the effective bandwidth and the effective capacity can be applied to model the statistical traffic behavior based on large deviation theory [14], [15]. The large deviation theory states that for stationary arrival and service processes under sufficient conditions, the probability that the buffer length B exceeds a certain threshold B' decays exponentially fast as the threshold B' increases;

$$P(B > B') \approx e^{-\theta B'} \tag{15}$$

where θ is a positive constant called the QoS exponent.
Since delay is our main QoS metric of interest, an equivalent expression to equation 15 is;

$$P(delay > d') \approx e^{-\theta \Delta d'} \tag{16}$$

where d' is the delay bound and Δ is jointly determined by the arrival process and the service process.
On the other hand the effective bandwidth, $EB(\theta)$, specifies the maximum constant service rate needed to serve the given arrival process subject to a given θ and the effective capacity, $EC(\theta)$, is the duality of $EB(\theta)$ representing the maximum constant arrival rate that can be supported by the system subject to a given θ [16].
Let θ^* be the solution of $EB(\theta^*) = EC(\theta^*)$, then we can find Δ by;

$$EB(\theta^*) = EC(\theta^*) = \Delta \tag{17}$$

The effective bandwidth function of a given arrival process $\{A(t), t > 0\}$ is given as;

$$E_b(u) = \frac{\Gamma(u)}{u} = \lim_{t\to\infty} \frac{1}{ut} \log\left(\mathbf{E}\left[e^{uA(t)}\right]\right) \quad (18)$$

where $\Gamma(u)$ is the log-moment generating function of the process and $\mathbf{E}[.]$ is the expectation.
Let ϵ be the acceptable probability of d' violation. The QoS requirements dictate that $e^{-\theta\Delta d'} \leq \epsilon$. If we assume the MTCD request arrival process to be poisson, then we can determine Δ as;

$$\Delta = \frac{\lambda\left(e^{\theta^*} - 1\right)}{\theta^*} \quad (19)$$

where λ is the arrival rate of the poisson process.
Therefore for certain QoS requirements we can determine θ^* as;

$$\theta^*\Delta = -\frac{log\epsilon}{d'} = \lambda\left(e^{\theta^*} - 1\right) \quad (20)$$

$$\theta^* = \log\left(1 - \frac{\log\epsilon}{\lambda d'}\right) \quad (21)$$

And the minimum constant service rate, η, can be computed as;

$$\eta = -\frac{\log\epsilon}{\theta^* d'} \quad (22)$$

Therefore at the service rate, η, we compute QoS violation probability as the ratio of the number of requests whose QoS requirements are violated to the total number of requests.

IV. Performance Evaluation

In this section, the metrics used and the simulation results are presented.

A. Metrics

1) Average throughput satisfaction: Average throughput satisfaction which is the ratio of the allocated resource blocks (satisfied throughput requirements) to the requested resource blocks (requested throughput).

For direct access MTCDs, the average throughput satisfaction, S_D^{av} is computed as;

$$S_D^{av} = \frac{1}{N_D}\sum_{k=1}^{N_D}\left(\frac{\sum_{i=1}^{M}\alpha_D^k(i)}{\beta_D^k}\right) \quad (23)$$

Where β_D^k corresponds to the requested resources by a direct access MTCD k, α_D^k are the allocated resources to a direct access MTCD k, M is the total number of resource positions with 1 or 0 in position i with or without a resource respectively and N_D is the number of direct access MTCDs.

Similarly for indirect access MTCDs, average throughput satisfaction, S_I^{av} is computed as;

$$S_I^{av} = \frac{1}{N_I}\sum_{k=1}^{N_I}\left(\frac{\sum_{i=1}^{M}\alpha_I^k(i)}{\beta_I^k}\right) \quad (24)$$

We compare throughput satisfaction performance when splitting is employed by the MTC CH and the case without splitting such that either all MTCDs are direct access ($N_D \to N$) or indirect access ($N_I \to N$).

2) Average QoS violation probability: As with throughput satisfaction, the average QoS violation probability is also compared for the case when splitting is employed and the case without splitting.

B. Simulation Results

The simulation parameters used are shown in table I.

Parameter	Value
B	1.4 MHz
N	10-100
S_c	12
S_y	7
T_s	0.5 ms
ϵ	0.2
d'	100 ms
M	6
L_{CH}	10240 bits

TABLE I: Simulation parameters.

1) MTCD Splitting: The results in Fig. 2 (a) show that the average number of indirect access MTCDs is higher than direct access MTCDs for up to 90 MTCDs and then drops as the number of MTCDs increase, this is because an MTC CH maximizes the number of indirect access MTCDs in the cluster, $|I|$ but as the MTCDs, N increase, $N - |I| > |I| \to |D| > |I|$, hence the number of direct access MTCDs exceeds indirect access MTCDs because $|I|$ has limitations such as data buffer size in an MTC CH.

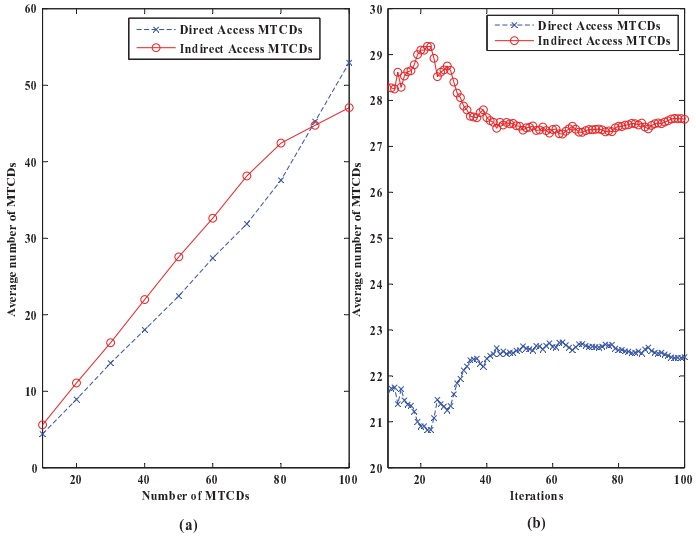

Fig. 2: (a) MTCD Splitting with varying number of MTCDs (b) MTCD Splitting with 50 MTCDs over time

However, if the number of MTCDs is fixed ($|I|$) for example to 50 MTCDs, Fig. 2 (b) shows that the number of indirect access MTCDs is always maximized.

2) Resource Allocation: The results in Fig. 3 show that a high average throughput satisfaction is achieved when MTCD splitting is employed as compared for the case without splitting. The reason for the better performance is that, by allowing

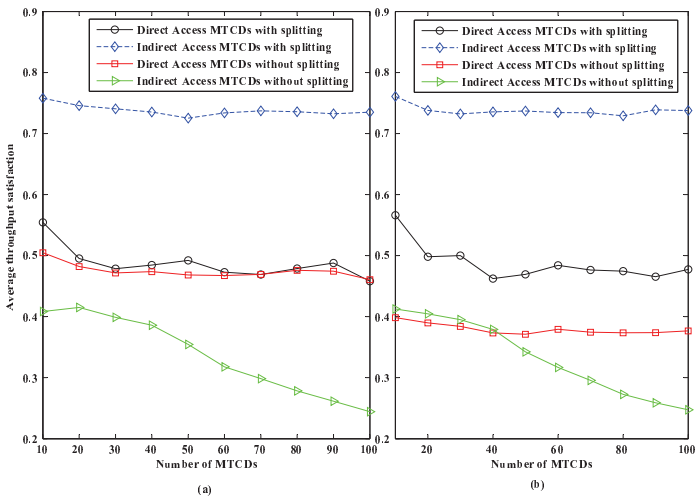

Fig. 3: Average throughput satisfaction for MTCDs with and without splitting (a) with 100% eNodeB support for direct access MTCDs without splitting (b) with 80% eNodeB support for direct access MTCDs without splitting

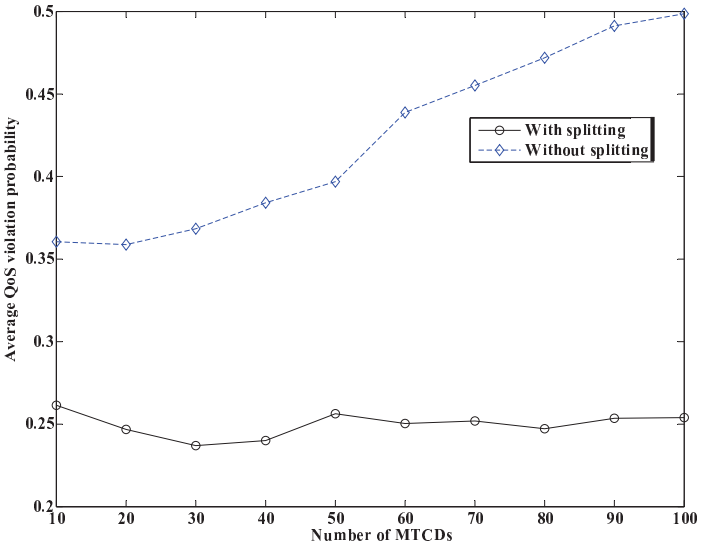

Fig. 4: Average QoS violation probability with and without splitting; without splitting means that all MTCDs access the eNodeB directly or all MTCDs access the eNodeB through the MTCG or a relay and with splitting means that our proposed splitting mechanism is employed for eNodeB access.

many MTCDs with less stringent requirements to connect indirectly, the MTC CH is responsible for access procedures of the cluster, which then minimizes the effect of congestion in the control channels and hence the throughput requirements of more indirect MTCDs are also satisfied. Also when splitting is employed, an eNodeB can maximize the number of fully satisfied direct access MTCDs and at the same time minimizing the gap between requested resources and allocated resources by the indirect access MTCDs. The average throughput satisfaction for direct access MTCDs with and without splitting is close in Fig. 3 (a) because the eNodeB is assumed to support 100% direct connections when no splitting is employed, which is not always the case and therefore that is the ideal scenario for direct access MTCDs without splitting. On the other hand, Fig. 3 (b) shows a drop in average satisfaction performance of direct access MTCDs without splitting when the eNodeB can only support 80% direct connections as expected because in the presence of a large number of MTCDs, all attempting direct access to the eNodeB concurrently, the eNodeB can only satisfy a small portion of throughput requirements. Indirect access MTCDs without splitting achieve the lowest average throughput satisfaction performance and greatly drops with the number of MTCDs because an indirect access relay like an MTCG or MTC CH can only connect a limited number of MTCDs to the eNodeB in the presence of a large number of MTCDs due to its inherent limitations.

Fig. 4 shows that the average QoS violation probability when splitting is employed is lower than the case when it is not employed. This is because splitting ensures that delay sensitive MTCDs connect directly to the eNodeB for resource allocation and at the same time the number of direct connections is reduced by maximizing the number of indirect access MTCDs at the MTC CH.

V. Conclusion

We have proposed an MTCD splitting and resource allocation technique for Machine Type Communications in this paper by exploiting the diversity in MTCDs in order to select capable nodes to perform splitting and then prioritize the resource allocation for direct access MTCDs at the eNodeB while serving indirect access MTCDs with best effort. We then demonstrated the effectiveness of this approach through analysis and simulations and the results show that it can achieve better throughput satisfaction rates and low QoS violation probability on average as compared to the existing approaches where splitting is not employed.

References

[1] K. Zheng, F. Hu, W. Wang, W. Xiang, and M. Dohler, "Radio resource allocation in lte-advanced cellular networks with m2m communications," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 184 – 192, July 2012.

[2] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jntti, and A. Ghosh, "Machine-type communications: Current status and future perspectives toward 5g systems," *IEEE Communications Magazine*, vol. 53, pp. 10 – 17, September 2015.

[3] P. K. Verma, R. Verma, A. Prakash, A. Agrawal, K. Naik, R. Tripathi, M. Alsabaan, T. Khalifa, T. Abdelkader, and A. Abogharaf, "Machine-to-machine (m2m)communications: A survey," *Journal of Network and Computer Applications*, vol. 66, pp. 83 – 105, May 2016.

[4] C.-Y. Tu, C.-Y. Ho, and C.-Y. Huang, "Energy-efficient algorithms and evaluations for massive access management in cellular based machine to machine communications," *in Proceedings of IEEE Vehicular Technology Conference (VTC Fall)*, pp. 1 – 5, September 2011.

[5] C.-Y. Ho and C.-Y. Huang, "Energy-saving massive access control and resource allocation schemes for m2m communications in ofdma cellular networks," *IEEE Wireless Communication Letters*, vol. 1, no. 3, pp. 209 – 212, June 2012.

[6] P. Si, J. Yang, S. Chen, and H. Xi, "Adaptive massive access management for qos guarantees in m2m communications," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 7, pp. 3152 – 3166, July 2015.

[7] 3GPP, "Service requirements for machine-type communications (MTC)," Sophia Antipolis Cedex, France, TS 22.368, September 2012.

[8] M. Beale, "Future challenges in efficiently supporting m2m in the lte standards," *in Proc. IEEE Wireless Communications and Network Conference Workshops*, pp. 186 – 190, 2012.

[9] R. Liu, W. Wu, H. Zhu, and D. Yang, "M2m-oriented qos categorization in cellular network," *in Proceedings of 7th Wireless Communications, Networking and Mobile Computing (WiCOM)*, no. 12, pp. 1 – 5, September 2011.

[10] H. Aissi, C. Bazgan, and D. Vanderpooten, "Complexity of the min-max and min-max regret assignment problems," *Operations Research Letters*, vol. 33, no. 6, pp. 634 – 640, November 2005.

[11] A. Hatoum, R. Langar, N. Aitsaadi, R. Boutaba, and G. Pujolle, "Cluster-based resource management in ofdma femtocell networks with qos guarantees," *IEEE Transactions on Vehicular Techology*, vol. 63, no. 5, pp. 2378 – 2391, June 2014.

[12] G. Brown and G. Graves, "Elastic programming: A new approach to largescale mixed integer optimization," *ORSA/TIMS Conference, Las Vegas, NV, USA*, November 1975.

[13] J. W. Chinneck and E. W. Dravnieks, "Locating minimal infeasible constraint sets in linear programs," *ORSA Journal on Computing*, vol. 3, no. 2, pp. 157 – 168, 1991.

[14] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Transactions on Wireless Communication*, vol. 2, no. 4, pp. 630 – 643, July 2003.

[15] T. Jia and Z. Xi, "Cross-layer modeling for quality of service guarantees over wireless links," *IEEE Transactions on Wireless Communications*, vol. 6, no. 12, pp. 4504 – 4512, December 2007.

[16] S.-Y. Lien, Y.-Y. Lin, and K.-C. Chen, "Cognitive and game-theoretical radio resource management for autonomous femtocells with qos guarantees," *IEEE Transactions on Wireless Communications*, vol. 10, no. 7, pp. 2196 – 2206, June 2011.