Cellular-based machine-to-machine : congestion control and power management

Osama Arouk

ANNÉE 2016

UNIVERSITÉ DE RENNES 1

UNIVERSITE BRETAGNE LOIRE

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

Ecole doctorale Matisse

Osama Arouk

préparée à l'unité de recherche UMR 6074 (IRISA)
Institut de Recherche en Informatique et Système Aléatoires

Cellular-Based Machine -to-Machine: Congestion Control and Power Management

Communication Machine à Machine: Contrôle de Congestion et Gestion de l'Energie

Thèse rapportée par:

Lynda MOKDAD
Professeur, Université de Paris-Est, Créteil, France

Gabriel-Miro MUNTEAN
Maître de Conférences, Université de Dublin City, Irlande

et soutenue à Rennes
le 25 Mars 2016

devant le jury composé de :

Gabriel-Miro MUNTEAN
Maître de Conférences, Université de Dublin City, Irlande / Rapporteur

Hossam AFIFI
Professeur, Telecom Sudparis, France / Président

Adlen KSENTINI
Maître de Conférences - HDR, Université de Rennes 1, France / Co-directeur de thèse

César VIHO
Professeur, Université de Rennes 1, France / Directeur de thèse

I would like to dedicate this thesis to my loving mother, brother and his family who left the life before seeing this completed work ...

Acknowledgements

I would like, firstly, to express my gratitude to Dr. Adlen Ksentini and Prof. César Viho for their continues support during the whole phD period. Also, I would like to thank Dr. Yassine Hadjadj-Aoul for his support and help, especially during the first year of PhD. This PhD would not be finished without them. They gave me the best guidance, help and support to successfully go ahead in the PhD.

Also, all the thanks to the jury members for their presence, participation, and the time that they devoted to read this thesis. Their valuable comments and questions helped me to improve the quality of my dissertation.

My grateful is also to Prof. Tarik Taleb, who provided me the opportunity to join his team as a visiting researcher, in addition to his support during the period of PhD.

Without forgetting anyone, I would like to thank all my colleagues and friends for their support.

I would like to thank all the members of my team, Dionysos, at IRISA/INRIA for providing the perfect work environment. Special thanks and all my grateful are for the assistance of Dionysos team, Mme Fabienne Cuyollaa, for her precious support during the whole period of the PhD.

Finally, all my thanks to all the members of my family for their support and encouragement. The thanks will not be completed without mention my parents. I will never arrive to this stage without their help, support, and encouragement.

Abstract

The current and next generation wireless cellular networks (5G) have to deal with not only communications between people, known as Human-to-Human (H2H), but also with a massive deployment of Machine-Type-Communication (MTC). MTC, or alternatively Machine-to-Machine (M2M), can be viewed as devices connected among them without any human intervention. M2M can be considered as the cornerstone of Internet-of-Things (IoT) vision. It attracts a lot of attention, since it can be considered as a new opportunity and business market. Nowadays, there is a vast number of MTC applications, covering a large number of fields. Some of these applications are Healthcare, Intelligent Transport System (ITS), smart metering and smart grids, Public Safety (PS), forming the so-called smart city.

Deploying this type of applications in the current cellular mobile networks, especially Long Term Evolution (LTE) and LTE-Advanced (LTE-A), cannot be achieved before overcoming the accompanied challenges. Indeed, caused by the existence of a myriad of MTC devices, Radio Access Network (RAN) and Core Network (CN) congestion and system overload is one of these challenging issues. As the MTC devices are battery-equipped, power consumption is also a challenge.

In this thesis, we study the congestion and power consumption problems in the context of LTE and LTE-A networks featuring M2M communications. Regarding the congestion and system overload, the focus will be on the RAN part since it can be considered as the first defense line on the network.

The contributions of the thesis are organized on the following axis:

- Propose a general algorithm to predict the incoming traffic, so that the congestion in the network can be easily remedied.
- Study and propose a general analytical model of the Random Access Channel (RACH) procedure. The model can help to evaluate the congestion control methods targeting the RAN part.
- Depth study and propose methods improving the performance of Group Paging (GP) method, one of the methods approved by 3GPP to control the congestion.

Résumé

Les réseaux actuels et la prochaine génération des réseaux sans fil cellulaires (5G) doivent garantir, non seulement, les communications entre les gens (aussi connu sous le nom d'humain à humain - H2H), mais aussi à un déploiement massif de communication de type machine (MTC). MTC, ou encore Machine à Machine (M2M), peut être considérée comme des appareils qui peuvent établir des communications avec d'autres appareils sans aucune intervention humaine. M2M est aussi vue comme la pierre angulaire de la vision des objets connectés (IoT). Elle attire beaucoup d'attention, car elle peut être considérée comme une nouvelle opportunité pour les opérateurs de réseau et service IoT. Il existe aujourd'hui plusieurs types d'applications se basant sur MTC couvrant plusieurs domaines. On peut citer comme examples les applications suivantes: la santé, les systèmes de transport intelligents (ITS), les compteurs intelligents et les réseaux intelligents, et la sécurité publique (PS).

Le déploiement de ce type d'applications dans les réseaux mobiles cellulaires actuels, particulièrement Long Term Evolution (LTE) et LTE-Advanced (LTE-A) , ne peut être effectif sans surmonter les challenges posés par le déploiement d'un grand nombre d'équipement MTC dans la même cellule. En effet, le déploiement d'une myriade d'appareils MTC causera une congestion et une surcharge du système des réseaux d'accès radio (RAN) et du cœur de réseau (CN). Comme les appareils MTC sont équipés d'une batterie non rechargeable, la consommation d'énergie est aussi un défi.

Dans cette thèse, nous allons étudier les problèmes de congestion et de consommation d'énergie dans le contexte des réseaux LTE et LTE-A en présence des appareils M2M. En ce qui concerne la congestion et la surcharge de système, nous nous concentrons sur la partie RAN, puisqu'elle peut être considérée comme la première ligne de défense pour le réseau céllulaire.

Les contributions de cette thèse sont organisées sous les axes suivants:

- Proposition d'un algorithme générique pour prédire le trafic entrant, de sorte que la congestion dans le réseau peut être facilement résolue.
- Etude et proposition d'un modèle analytique générique de la procédure d'accès aléatoire au canal (RACH). Le modèle a pour but l'évaluation des méthodes de contrôle de congestion ciblant la partie RAN.

- Approfondissement et proposition des méthodes permettant d'améliorer la méthode Pagination de Groupe (GP) approuvée par le 3GPP pour contrôler la congestion.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

La mise en place de la ville intelligente nécessitera, sûrement, l'utilisation des technologies de l'information, pour les réseaux électriques, les pipelines de pétrole et de gaz, les systèmes d'eau, des bâtiments, des ponts, et même d'autres objets dans notre vie (par exemple, brosse à dents intelligente [7]), initiant le concept de l'Internet des objets (IoT). Une ville intelligente basée sur l'IoT peut être réalisé par les nouvelles technologies, comme le cloud computing [8] et la communication machine à machine (M2M).

M2M, également connu sous le nom Machine-Type-Communication (MTC), peut être défini comme un type émergent de communication, permettant à des machines (appareils) de communiquer avec d'autres machines sans ou avec un minima d'intervention humaine. Historiquement, la communication M2M est considérée comme une forme de système de contrôle et d'acquisition de données (SCADA). Les appareils MTC permettent le déploiement d'une large variété d'applications (par exemple, Télésanté, surveillance et sécurité, systèmes de transport intelligent - ITS, automatisation de ville, etc.), dans un large éventail de domaines ; influançant différents marchés et environnements [9, 10].

Les communications M2M vont concerner la connexion d'un très grande nombre d'appareils MTC, approximativement 60 milliards de connexions M2M sont prévues d'ici l'aube de l'année 2020 [11]. De plus, d'autres predictions estiment que chaque personne, en moyenne, aura 1000 objets connectés à Internet en 2040 [12], ce qui ouvrira beaucoup de revenus et d'opportunités pour les opératuers de réseau et service IoT. Comme indiqué dans [13], environ 45% des connexions seront générées par les équipements M2M, tandis que le reste sera d'origine Machine à Humain (M2H), Humain

à Machine (H2M), et Humain à Humain (H2H). Par conséquent, les communications M2M sont devenues une technologie prometteuse attirant l'attention de nombreux opérateurs et fournisseurs.

Comme les communications H2H, les communications M2M doivent être fiables, sécurisées, évolutives et gérables [14]. Cependant, l'activation d'un grand nombre d'appareils MTC dans les réseaux mobiles cellulaires actuels peut générer une très grande quantité de trafic de signalisation et/ou de données. La gestion de cette énorme quantité de trafic dans les réseaux mobiles cellulaires ne peut être possible sans provoquer la surcharge et la congestion du Réseau d'accès radio (RAN) et / ou du réseau de coeur mobile (CN), composant l'architecture des réseaux LTE. En parallèle à la congestion, la gestion de la consommation d'énergie des appareils MTC représente un autre challenge [15, 16]. En effet, une fois installées, les batteries des appareils MTC ne seront plus remplacées, au moins pour de nombreuses années.

1.2 Contributions de la Thèse

La thématique générale de la thèse porte sur le contrôle de la congestion et la surcharge du réseau lorsque un grand nombre d'équipements MTC est déployé dans une cellule LTE. Plus précisement, nos contributions concernent la procédure d'accès aléatoire au canal (RACH), qui est la première procédure que le terminal doit effectuer pour obtenir l'accès au réseau. En effet, deux sujets principaux ont été abordés: l'optimisation des ressources et l'efficacité de la consommation énergétique.

Les contributions de cette thèse peuvent être classifiées en trois axes:

1. La prédiction du traffic: L'objectif de cette contribution [17] est de prédire le trafic entrant, de sorte que la congestion peut être fortement contrôlée. L'algorithme proposé dans cet ouvrage, à savoir Multi Channel Slotted Aloha - Optimal Estimation (MCSA - OE), se compose de deux parties. La première partie consiste à estimer et à adapter la probabilité d'inactivité du réseau en se fondant sur le nombre de canaux libres, c'est-à-dire les préambules inutilisés dans la procédure de RACH. Afin de limiter les grandes fluctuations dans les estimations, la méthode de prédiction proposée a été améliorée par un autre algorithme pour adapter le nombre d'arrivées (traffic) estimé. Il est à noter que ces deux algorithmes sont utilisés pour prédire le trafic MTC. Ils peuvent être employés par des méthodes de contrôle de congestion, comme la methode Access Class Barring (ACB), où le nombre d'arrivée estimé à chaque fois est utilisé pour mettre à jour les paramètres d'ACB. L'avantage des algorithmes proposés réside

dans leur efficacité pour estimer l'arrivée des traffics extrêmement denses, tels que ceux basés sur le modèle du trafic Beta. En outre, ils peuvent être utilisés pour obtenir l'efficacité quasi-optimale du réseau en terme de propobabilité de succès et l'utilisation du canal.

2. Modèle analytique de la procédure de RACH: Deux contributions ont été proposées [18, 19], avec l'objectif de fournir un modèle analytique pour évaluer les méthodes et les algorithmes traitant le problème de congestion du RAN et se concentrant sur la procédure de RACH. Le modèle proposé, à savoir General Recursive Estimation (GRE), est générique et adapté à tout type de trafic, même s'il est évalué avec le modèle du trafic Beta. L'avantage du modèle analytique proposé est qu'il est compatible avec la norme LTE / MTC, et il ne nécessite pas la modification des hypothèses prises par le 3GPP sur le trafic MTC.
3. Amélioration du processus Group Paging (GP) et l'optimisation de la consommation d'énergie: L'axe final de cette thèse comprend trois contributions [20–22]. L'objectif principal de ces travaux est d'améliorer la performance de la méthode GP, approuvé par le 3GPP, pour résoudre le problème de la congestion lors de l'interrogation ou pagination d'un grand nombre d'appareils MTC. Ces contributions sont constitués de deux parties:
   - La première partie représente l'amélioration de la méthode GP pour un cas particulier: tous les appareils MTC concernés sont connectés au réseau (en mode Radio Resource Control - RRC connecté), mais ils ont perdu la synchronisation avec la liaison remontante (ils ne sont pas synchronisés avec l'eNB) [20]. La méthode de contrôle de congestion proposée repose sur les identifiants (IDs) des apareils MTC concernés afin de distribuer les ressources disponibles, à savoir Controlled Ditribution of Resources (CDR). L'avantage de cette méthode est de rendre la procédure RACH sans contention, ce qui signifie que la probabilité de succès est toujours de 100 % et la probabilité de collision est toujours de zéro.
   - Afin d'améliorer la proposition mentionnée ci-dessus, une méthode plus générale a été proposée, visant à couvrir tous les états de la machine, quelque soit le mode RRC connecté ou RRC non-connecté [21, 22]. L'idée principale de ces travaux est de disperser les appareils MTC concernés sur l'intervalle disponible de pagination, au lieu de les laisser commencer la procédure de RACH en même temps, comme dans la méthode GP. La méthode proposée non-seulement améliore les performances de la méthode GP, mais également donne un moyen efficace pour choisir les paramètres de configuration du

réseau qui maximisent les performances. Outre l'utilisation efficace des ressources, la méthode proposée permet d'obtenir une réduction élevée de la consommation d'énergie.

1.3 Organisation du Manuscrit

Cette thèse est organisée comme suit:

Dans le Chapitre 3, on commence par l'introduction des informations générales sur la M2M et les efforts qui ont été faits jusqu'à présent pour supporter M2M dans les réseaux sans fils actuels. Par la suite, une attention particulière sur M2M dans le paysage de 3GPP est faite. Le chapitre se termine par l'introduction d'une classification générale des méthodes proposées dans la littérature sur le contrôle de congestion dans LTE. Cette classification est largement basée sur la partie du réseau (CN ou RAN) où la solution est déployée

Chapitre 4 se concentre sur la prédiction du trafic et l'optimisation du réseau. Après un état de l'art sur les modèles existants pour la prédiction du traffic dans le cadre de M2M, un nouveau mécanisme, à savoir Multi-canal Slotted ALOHA - Optimale Estimation (MCSA - OE), est introduit. Ce mécanisme se compose de deux algorithmes ; (*i*) estimation et adaptation de la probabilité d'inactivité (en fonction du nombre de canaux libres), (*ii*) adaptation du nombre d'arrivés estimé pour éviter les fluctuations de l'estimation. Les valeurs estimées peuvent être utilisées pour ajuster les paramètres de la méthode ACB afin de contrôler la congestion.

Chapitre 5 se concentre sur la présentation d'un modèle général pour la procédure de RACH. Ce chapitre commence par présenter un état de l'art sur les méthodes de modélisation pour la procédure de RACH. Après cela, un nouveau modèle général, à savoir le General Recursive Estimation (GRE), est introduit. Le but de ce modèle général est de fournir un outil pour evaluer des méthodes de contrôle de congestion pour le RAN. Le modèle analytique proposé est testé sur un traffic basé sur la distribution Beta, qui est considéré comme un cas extrême d'activation des capteurs M2M.

L'objectif principal du chapitre 6 est l'amélioration du mécanisme GP proposé par le 3GPP. Après la présentation d'un état de l'art sur les méthodes existantes améliorant la méthode GP, deux nouvelles améliorations sont introduites. La première méthode, à savoir Controlled Distribution of Resources (CDR), est proposée afin d'améliorer la méthode GP dans un cas particulier, où les appareils MTC ont des IDs de la cellule d'attache, mais ils ont perdu la synchronisation de la liaison remontante. L'idée de CDR est d'attribuer les ressources disponibles sur la base des identifiants des appareils MTC,

c'est-à-dire chaque appareil détermine les ressources nécessaires pour être utilisées en se basant sur son ID. Par conséquent, la procédure de RACH devient sans contention. Afin d'améliorer la méthode GP indépendamment de l'état de machine MTC, une deuxième méthode, nommée Further Improvement - Traffic Scattering For Group Paging (FI-TSFGP), est proposée. Cette méthode est général, c'est-à-dire elle peut être appliquée indépendamment de l'état de la machine, et peut être adaptée à toutes les tailles du groupe, ce qui n'est pas le cas pour la méthode GP. Cette méthode tente de disperser les membres du groupe concerné sur l'intervalle disponible de pagination, au lieu de les laisser tous commencer en même temps (comme dans la méthode GP). De plus, FI-TSFGP détermine le nombre de machines qui doivent être activées à chaque fois, pendant l'intervalle disponible de pagination, afin d'optimiser les performances.

Enfin, le chapitre 7 conclut la thèse avec différentes directions et perspectives pour les travaux futurs.

Chapter 2

Introduction

2.1 Motivations

Smart city concept will be, surely, achieved by applying the next generation information technology in our everyday life, such as power grids, oil and gas pipelines, water systems, buildings, bridges, and even other objects in our life (e.g., smart toothbrush [7]), forming the so-called Internet-of-Things (IoT). The IoT for smart city can be realized by emerging technologies, such as cloud computing [8] and Machine-to-Machine (M2M) communication.

M2M, also known as Machine-Type-Communication (MTC), can be defined as an emerging type of communication, enabling machines (devices) to communicate to each other without or with a little human intervention. Historically, M2M communication is considered as a developed form of the industrial Supervisory Control And Data Acquisition (SCADA) system. MTC devices support a broad variety of applications (e.g., eHealth, surveillance and security, Intelligent Transport System (ITS), city automation, etc), in a wide range of domains impacting different markets and environments [9, 10]. This type of communications is expected to connect an enormous number of MTC devices, as 60 billion of M2M connections are forecasted by 2020 [11]. However, others expect that every person, on average, will have 1000 Internet-connected devices by 2040 [12], which will certainly offer many revenue and opportunities. As mentioned in [13], about 45% of the connections will come from M2M, while the rest will be from Machine-to-Human (M2H), Human-to-Machine (H2M), and H2H. Thus, M2M is becoming a promising technology, attracting the attention of many operators and vendors.

Like H2H communication, M2M communication needs to be reliable, secure, scalable, and manageable [14]. However, enabling this a huge number of MTC devices in the

current cellular mobile networks will generate a very large amount of signaling/data traffic. Managing this huge amount of traffic within the current cellular mobile networks may not be possible without causing the overload and congestion for Radio Access Network (RAN) part and/or Core Network (CN) part, constituting the LTE architecture. Besides, enabling low-cost and low power consumption M2M devices is also a big challenge [15, 16]. Regarding the power consumption, once installed, the MTC devices' batteries will not be replaced, at least for many years. Therefore, enabling low-cost and low power consumption M2M devices is the key enabler, besides the access network, to allow them to be ubiquitous in our live.

2.2 Contributions of the Thesis

The thematic general of the thesis concerns the control of congestion and system overload. This theme is achieved by managing the Random Access Channel (RACH) procedure, which is the first procedure that the terminal should do to get access to the network. Under this theme, two principal subjects were also targeted: resource optimization and power efficiency.

The contributions of the thesis can be divided into three main axis:

1. Traffic prediction: under this axis, there is one contribution [17]. The goal of this contribution is to predict the incoming traffic, so that the congestion can be highly controlled. The algorithm proposed in this work consists of two parts. The first part consists in estimating and fitting the idle probability by relying on the number of idle channels, i.e. the unused preambles in the RACH procedure. In order to limit the large fluctuations in the estimations, the proposed prediction method is improved by another algorithm to fit the estimated number of arrivals. Note that these two proposed algorithms are used to predict MTC traffic. They may be employed by congestion control methods, such as Access Class Barring (ACB) method, where the estimated number of arrivals at each time is used to update the ACB parameters. The advantage of the proposed algorithms is that they can well work under heavy traffic, such as Beta traffic model. Furthermore, they can be used to obtain the near optimal network performance regarding the success probability and resource utilization.
2. Analytical model of RACH procedure: there are two contributions under this axis [18, 19], with the objective to provide an analytical model to evaluate the methods and algorithms dealing with the RAN issue focusing on the RACH procedure. The proposed model is generic, and adapted to any type of traffic,

even it is evaluated with Beta traffic model. The advantage of the proposed analytical model is that it is aligned with the Long Term Evolution (LTE)/MTC standard, and it does not require any changes to the assumption already taken by the 3GPP on MTC traffic.

3. Group Paging (GP) improvement and power efficiency: the final axis comprises three contributions [20–22]. The main objective of these works is to improve the performance of GP method, approved by 3GPP, to remedy the congestion problem when paging a large number of MTC devices. These works fall in two parts:
   (a) The first part represents the improvement of GP method for a special case; all the concerned MTC devices are connected to the network (i.e. they are in Radio Radio Resource Control (RRC) connected mode), but they lost the uplink synchronization (i.e. they are out of synchronization) [20]. The proposed congestion control method relies on the identifiers (IDs) of the concerned MTC devices to distribute the available resources, i.e. controlled distribution of resources. The advantage of this method is that the contention RACH procedure becomes like a contention-free procedure, meaning that the success probability is always 100% and the collision probability is always zero.
   (b) In order to improve the above mentioned proposition, a more general method was proposed, aiming at covering all the machine states, i.e. whether it is in RRC connected mode or in RRC idle mode [21, 22]. The main idea in these works is to scatter the concerned MTC devices over the available paging interval, instead of leaving them to start the RACH procedure all at once, as in GP method. The proposed method not only improves the performance of GP method, but also gives an efficient way to choose the network parameters that maximize the network's performances. In addition to the efficient utilization of resources, the proposed method achieves high power reduction.

2.3 Organization of the Manuscript

Chapter 3 starts by exploring a general information on M2M and efforts that have been doing to support M2M in the current wireless networks. After that, a special attention on M2M in the landscape of 3GPP is made, where a necessary background is presented. The chapter ends by introducing a general classification of the overload and

congestion control methods, which is broadly based on where the control is applied, i.e. in the RAN part or in the CN part.

Chapter 4 concentrates on the traffic prediction and network's optimization. After a state of the art on the existing models for traffic prediction in the context of M2M, a novel mechanism, namely Multi-Channel Slotted ALOHA-Optimal Estimation (MCSA-OE), is introduced. This mechanism consists of two algorithms; (*i*) estimate and fit the idle probability (based on the number of idle channels), (*ii*) fit the estimated number of arrivals to avoid the fluctuations in the estimation. The estimated values are further used to adjust the parameters of ACB method so as to control the congestion.

Chapter 5 focuses on presenting a general model for RACH procedure. This chapter starts by presenting a state of the art on the modeling methods for the RACH procedure. After that, a new and general model, namely General Recursive Estimation (GRE), is introduced. The aim of this general model is to help evaluating the congestion control methods in the RAN part. The proposed analytical model is tested on Beta traffic, since it is considered as the worst case of M2M traffic modeling.

The main focus of chapter 6 is on Group Paging (GP) improvement. After presenting a state of the art on the existing methods targeting GP method, two GP improvements are introduced. The first method, namely Controlled Distribution of Resource (CDR), is proposed to improve the GP method in a special case, where the MTC devices have IDs from the cell within its coverage they are, but they lost the uplink synchronization. The idea of CDR is to attribute the available resources based on the IDs of the devices, i.e. each device determines the resources to be used by its ID. Therefore, the RACH procedure becomes like contention-free one. In order to improve the GP method regardless of the state of the MTC devices, a second method, namely Further Improvement-TSFGP (FI-TSFGP), is proposed. This method is a general one, i.e. it can be applied regardless of the state of the machine, and it can be adapted for any group size, which is not the case for GP method. This method tries to scatter the members of the concerned group during the available paging interval, instead of leaving them to start all at the same time like GP method. Moreover, FI-TSFGP determines the number of devices that should be activated at each time, during the available paging interval, in order to maximize the performance.

Finally, concluding remarks are introduced in chapter 7, where different future directions and perspectives are presented.

Chapter 3

M2M in the landscape of 3GPP: Congestion Control and Power Management

One of the 5G requirements is to ensure the connection of massive numbers of wireless devices to wireless networks, including not only User Equipment (UE) but also objects like sensors and actuators that constitute the concept of Internet-of-Things (IoT). Enabling the automatic communication of sensors and actuators with remote servers and systems is also known as Machine-to-Machine (M2M) communication, or Machine-Type-Communication (MTC) in 3^{rd} Generation Partnership Project (3GPP) terminology.

However, deploying a huge number of MTC devices (its main characteristic) in the current wireless networks, more specifically in 3GPP mobile networks, will generate a very large amount of signaling and data traffic. Although the current cellular mobile networks are well dimensioned for regular traffic (mainly web traffic known to consume more downlink bandwidth than uplink bandwidth), they were not designed to support M2M traffic, which is known to be short in time and involves small and frequent uplink data transmissions [23]. Accordingly, managing such a huge amount of traffic (each cell is supposed to host tens of thousands of MTC devices) within the current cellular mobile networks may not be possible without causing congestion and system overload. In addition to the aforementioned challenging issue, another important challenge for M2M is the power consumption. Indeed, once installed, MTC devices' batteries would not be replaced, at least for many years.

In this chapter, after introducing the power consumption at the steps of the Random Access Channel (RACH) procedure, a survey on the efforts made by the 3GPP group

as well as the research community to support massive deployment of M2M devices in the future wireless networks is presented. More specifically, the focus will be on the congestion and system overload issue (which is the main target of the thesis) in cellular mobile networks. This chapter starts by introducing some of use cases to better understand the diversity of M2M applications. Then, relevant background on Physical and Media access control (MAC) layers (specifically Long Term Evolution (LTE) radio frame and RACH procedure) and 3GPP architecture for M2M are outlined. Finally, a comprehensive classification of Overload and congestion control mechanisms, for both Radio Access Network (RAN) part and Core Network (CN) part, are then presented.

3.1 M2M Use Cases

In order to show all the potential of M2M communication, this section will detail some of M2M use cases and its applications. It is important to note that it is difficult to recognize all the possible use cases because of the enormous diversity of M2M applications. For more use cases, the reader may refer to [24–26].

1. Smart Grid and Smart Metering: Smart meters give the consumers the ability to track consumption data, e.g. the consumption of gas, water, or electricity, and thus saving money and resources [27, 28]. This technique is beneficial for both the costumer and the supplier. M2M technology-equipped devices send the ongoing consumption of electricity, for example, via short-rang radio technologies, home area network, or even 3G/4G/5G networks, periodically or on-demand to the smart grid central server. The central server can monitor the ongoing balance of power demand versus energy available, and then it updates the cost of the energy based on peak and low power consumption periods. Therefore, the costumer can save money by, for example, using the electricity in low power consumption periods. More advanced mode can be applied, wherein a smart grid central server can trigger non-time-critical home machines, such as dishwashers and washing machines, and thus achieve load balancing [29] , while helping costumers to save money.
2. Healthcare: This nascent application aims to improve the patient care by monitoring/tracking the patients. The M2M medical-related services allow the patients with advanced age or chronic disease, as an example, to live independently [30]. Thanks to more accurate and fast reporting of changes in physical conditions by M2M devices, patients care can be improved. For example, a patient can wear bio-sensors that record fitness indicators such as

blood pressure, body temperature, heart rate, etc. The collected data will be then forwarded periodically or on-demand by the sensors to a M2M device acting as an aggregator, which in turn transmits the data via the network to a M2M server that stores and may react to the collected data. Chronic diseases such as heart disease, hypertension, sleep apnea and other recurring illness top the list of health threats. M2M services can help by monitoring the patient's health status and may react in case of emergency, no matter if the patient is at home, in the bus, or anywhere else.

3. Intelligent Transport System (ITS): ITS refers to the use of information and communication technologies to develop and improve transportation systems. ITS includes all types of communications, e.g. vehicle to vehicle or vehicle to infrastructure [31]. ITS makes the transport easier and safer. For example, when there is an accident, the vehicles in the surrounded area can be informed by the transport information center about the presence of a lane closure. Therefore, the vehicles can avoid that lane. One of the important ITS services is emergency Call (eCall), which helps to save lives. In case of an accident, the eCall system installed in the vehicle automatically sends a message containing the current location of the device or establishes a voice connection with the nearest Public-Safety Answering Point (PSAP) [32], where an appropriate decision has to be made based on the received message (in case of sending message). It should be noted that ITS is not restricted to road transport, but it includes rail, water, and air transports.
4. Tracking and tracing: Tracking and tracing of cargo and vehicles, for example, by providing related-vehicle information (e.g., location) in real time allows to improve transportation efficiency and to safeguard the cargo against theft when distributing the goods nationally or internationally. It also enables better forecast about the arrival and delivery of the goods. However, there are many examples on this use case, such as tracing and tracking animals, persons with Alzheimer disease, for example, fleet vehicles, stolen vehicles tracking [31], etc. In these use cases, the persons, or the objects, are equipped with M2M devices, which in turn send information periodically or on-demand to the M2M server.

Based on the aforementioned model usages, M2M applications can be classified into three types of applications (knowing that there is the same classification relative to Wireless Sensor Networks (WSN) [33]):

— Time Driven Applications: M2M applications establish connection to the server periodically every hour or half-hour, for example. This class can be found

in smart metering, where electricity meters send the consumption data every certain time.

— Event Driven Applications: M2M applications establish connection when an event occurs. The information sent by the M2M devices installed in the vehicle to PSAP in case of an accident represents a good example of this class. Another example is the detection of fire in the forest, where all the installed M2M devices will automatically wake up and send the measured data.

— Query Driven Applications: In this case, the establishment of connection to send information or do something, e.g. turn on the washing machine, is done by the service center. When receiving a query, the M2M device does the appropriate action.

3.2 M2M Standardization Efforts

The characteristics of M2M applications are widely different from H2H characteristics. M2M applications do not have the same characteristics and, therefore, not every system optimization scheme is suitable for every M2M application. Many organizations have been studying the M2M, such as International Telecommunication Union - Telecommunication sector (ITU-T) [34], European Telecommunications Standards Institute (ETSI) [35], and IEEE 802.16 [36]. Moreover, seven of major Information and Communications Technology (ICT) Standards Development Organizations (SDOs) have agreed to form a new global organization, named as oneM2M [37]. These seven SDOs are the Association of Radio Industries and Businesses (ARIB), the Telecommunication Technology Committee (TTC) of Japan, the Alliance for Telecommunications Industry Solutions (ATIS), the Telecommunications Industry Association (TIA) of the USA, the China Communications Standards Association (CCSA), the European Telecommunications Standards Institute (ETSI), and the Telecommunications Technology Association (TTA) of Korea. Many goals and benefits of oneMEM have been identified [38], but the main goal is to develop technical specifications and reports in order to ensure that M2M devices can successfully communicate on a global scale, i.e. common Service Layer [39].

3GPP has been also working on M2M. While the focus of oneMEM is on the common Service Layer that can be readily embedded within various hardware and software, the focus of 3GPP is on the underlying connectivity between the M2M Application Server and M2M devices through mobile networks [40]. From herein, the focus of this chapter is only on M2M within 3GPP.

3.3 M2M in the landscape of 3GPP

M2M Features

Because of the diversity of the characteristics of MTC applications, not every system optimization can be suitable for every MTC application. Therefore, 3GPP defined MTC features in order to provide structure for the different possibilities of system optimization [41]. These features insist specific requirements. Besides, there are common service requirements that need to be appropriately treated in order to apply M2M in 3GPP.

Common Service Requirements

Many common service requirements have been defined in 3GPP, such as device triggering, identifiers, and security [41]. Regarding the device triggering procedure, we can mention the following:

- The network should be able to trigger MTC devices in order to allow the establishment of communication with the remote MTC server, once a trigger demand from MTC server is received.
- If a trigger indication is received from a non-authorized MTC server, the network should be able to provide details, e.g. address, about the MTC server to the MTC user.

The identifiers should achieve many requirements. The efficient assignment of identifiers means that the system should be able to; *i*) uniquely identify the Mobile Equipment (ME) [42], *ii*) uniquely identify the MTC subscriber, *iii*) and provide mechanisms to efficiently manage MTC subscriber numbers and identifiers relative to MTC subscribers. Regarding the security issues, the security of Non-MTC devices should not be degraded by the above mentioned requirements.

Specific Service Requirements

The requirements are identified for specific features. So, if there will be, in the future, new features, new requirements should be satisfied. Some of the features and their requirements that are currently defined by 3GPP are:

1. Low/no mobility: MTC devices do not move (e.g., water metering), move infrequently but may move within small area (e.g., health monitoring at home) or wide area (e.g., mobile sales terminals) [43]. In this case, there should be

a reduction in resource usage, such as change of the frequency of mobility management procedure or location updates performed by the MTC devices.

2. Time controlled: This feature is intended for MTC devices that can tolerate data transmission/reception in defined time intervals. The objective is to restrict the access of MTC devices to the network and avoid unnecessary network load outside these predefined time periods. A requirement for this feature is that the network operator should be able to allocate (for a group of MTC devices) time periods during which the signaling/user plane traffic to/from the network is allowed (i.e., Grant Time Interval - GTI) or disallowed (i.e. Forbidden Time Interval - FTI).
3. Small data transmission: The MTC devices send/receive a small amount of data, e.g. in the order of $1K$ octets. The transmission of small amount of data should be possible with minimal network impact (e.g., signaling overhead and network resources).
4. MTC monitoring: It is intended for monitoring MTC device related events, e.g. change of the location, behavior not aligned with activated MTC feature(s). Therefore, mechanisms for detecting events, such as behavior not aligned with activated MTC feature and change of the location, should be provided.
5. Secure connection: It is the case when a secure connection is required between MTC device and MTC server. With this feature, a network security between MTC device and MTC server should be provided.
6. Group based MTC features: It is a feature that is applied to a MTC group. Therefore, the system should be optimized in order to handle MTC group and, also, a mechanism to associate a MTC device to a single group should be provided. Moreover, the MTC group should be uniquely identified within the 3GPP networks. Under this feature, there are two sub-features:
   - Group based policing: It is intended for the use with a MTC group when the network operator wants to enforce a combined Quality of Service (QoS) policy. Therefore, a maximum bit rate for the transmission/reception should be enforced.
   - Group based addressing: It is intended to be used when multiple MTC devices need to receive the same message.

Despite the wide diversity of M2M features, there are some characteristics that are common to all, or the majority, of the MTC devices, but different from that of H2H devices (see [23] for more information about traffic analysis for M2M):

1. The presence of a very large number of M2M devices in each cell (which is of major problem facing the deployement of M2M within cellular networks), e.g. about 36000 M2M devices in a cell of 2000m radius [10], or even more.
2. The proportion of the UpLink (UL) traffic to the DownLink(DL) traffic is large for M2M devices, while the case is the contrary for H2H devices [23].
3. The average session inter-arrival of the M2M devices is larger than that of H2H devices.

Deploying this huge amount of MTC devices (which will engender a very large amount of data/control traffic), with this broad diversity of characteristics and the differences between the M2M and H2H devices, will have a huge impact on the network and the services provisioned to the devices for which these networks are optimized, i.e. H2H devices. This impact will comprise the whole network: the Radio Access Network (RAN) part and the Core Network (CN) part. Thanks to the characteristics of the next generation of mobile cellular networks, e.g. LTE and LTE-A, high data traffic will not be an issue because theses networks support a high data rate. Thereby, the main problem of the deployment of MTC in the next generation of mobile cellular networks is the control traffic. The impact of deploying MTC devices comes at the form of network congestion or system overload (i.e., signaling), and it mainly appears because of one or more of the following reasons:

1. The synchronized behavior of an application, where a mass of MTC devices will transmit their all at the same time, e.g. generating data transmissions at precisely synchronous time intervals (e.g., exactly every hour or half an hour).
2. Malfunction/problem in the MTC server, e.g. MTC devices try/retry to connect to the MTC server which is down.
3. Problem in the serving network, e.g. MTC devices move at the same time to the local competing network once the serving network experienced a failure [40].
4. Malfunction of the MTC devices, e.g. the rejected MTC devices try to connect/attach the network all at the same time or to immediately reinitiate the same request.

This congestion may cause intolerable delays, packet loss or even service unavailability. From the aforementioned reasons, we can expect that the congestion can take place in one or more of the following places (as depicted in Fig. 3.1):

1. In the Radio Access Network (RAN), i.e. the eNB.
2. In one or more of the concerned nodes of CN, such as Mobility Management Entity (MME), Serving-GateWay (S-GW), PDN-GateWay (P-GW), etc.

3. On the link between the network and the Application Server (AS).

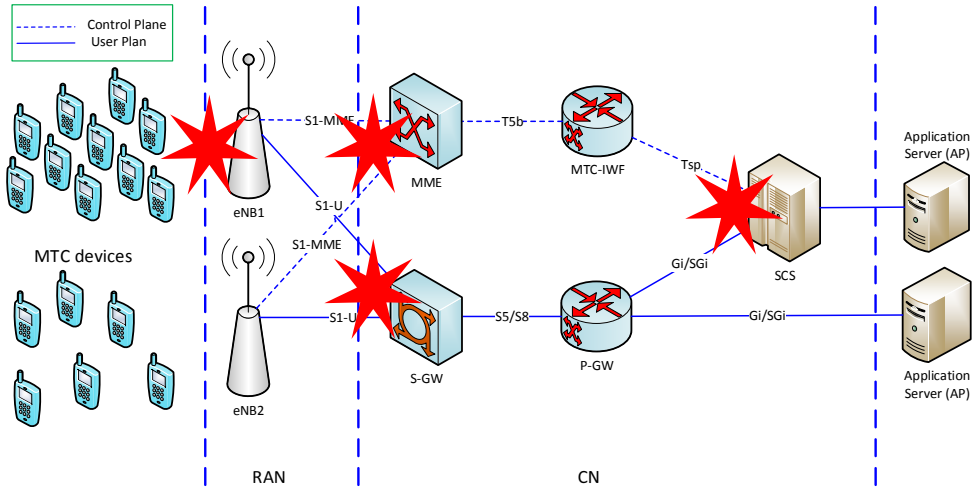

Figure 3.1 The congestion problem for MTC

Depending on the location where the congestion could happen, we have two types of congestion: RAN and CN congestions and system overloads.

Taking into account the fact that the RAN part can be considered as the first defense line of the network, the problem in question will thus be solved by an appropriate design of RAN congestion control methods. Therefore, the focus in the following of the thesis will be on the RAN part. Moreover, 3GPP has defined the RAN overload control as the first priority [44]. To show the need for efficient congestion and overload control methods, the authors in [1] have analyzed the LTE network's performance without applying any control method, Fig. 3.2, where Beta distribution model traffic is considered. From the figure, we remark that the peak number of preamble transmission (for both new arrivals and retransmission) is more than six times the total number of available preambles (which is equal to 54 in the considered study), engendering instantaneous collision probability exceeds 99%. Moreover, the access success probability is unacceptable, only 33.16% of the devices have access to network resources. Therefore, mechanisms to face such problems, guarantee network availability, and help the network to meet the performance requirements under such MTC load need to be investigated [45]. Many methods have been supported by the 3GPP SA2 group and proposed in the literature, targeting the goal (i.e., the congestion's problem). From the perspective of the way that the attach/connection request is initiated, the congestion control methods could be divided into push and pull based methods [46]. On one hand, in the push-based methods the MTC devices initiate the Random Access Channel (RACH) procedure. From the network's view point, these approaches are considered as a decentralized

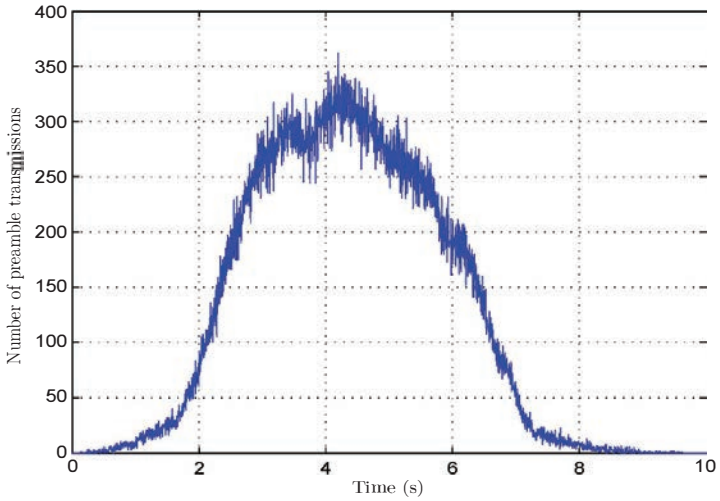

Figure 3.2 Number of simultaneous transmissions of preambles at each time without access overload control and with Beta distribution traffic model (number of preambles = 54) [1]

control model. On the another hand, in the pull-based methods the eNB initiates the RACH procedure. From the network's view point, these approaches are considered as a centralized control model [46]. However, in this manuscript we will introduce our perspective about the classification of the congestion control methods, which is based on the way of remedy the problem. The classification of RAN congestion and system overload control methods, based on our perspective, is illustrated in Fig. 3.11. Before elaborating the classification, some background on the physical layer and Media access control (MAC) layer will be introduced.

3.4 4G Networks Background

3.4.1 LTE Frame Structure

All time durations in LTE are defined in terms of basic time unit which is the sample period T_s, where $T_s = 1/30720000\ s$ and a sampling frequency $f_s = 1/T_s = 30.72\,M\,sample/s$ [2]. The downlink and uplink transmissions are organized into radio frames. In LTE and LTE-A, two radio frames are supported; Type 1 which is Frequency Division Duplex (FDD) and Type 2 that is Time Division Duplex (TDD) .

For frame structure Type 1, the radio frame has a length of $T_{frame} = 307200T_s = 10\,ms$ in the time domain and a length variable from 6 to 100 Resource Block (RB), which will be explained later, in the frequency domain. The $10\,ms$ radio frame is

divided into 10 equally sized subframes of length $1\,ms$, and each subframe is further divided into two equally sized slots of length $0.5\,ms$, as illustrated in Fig. 3.3.

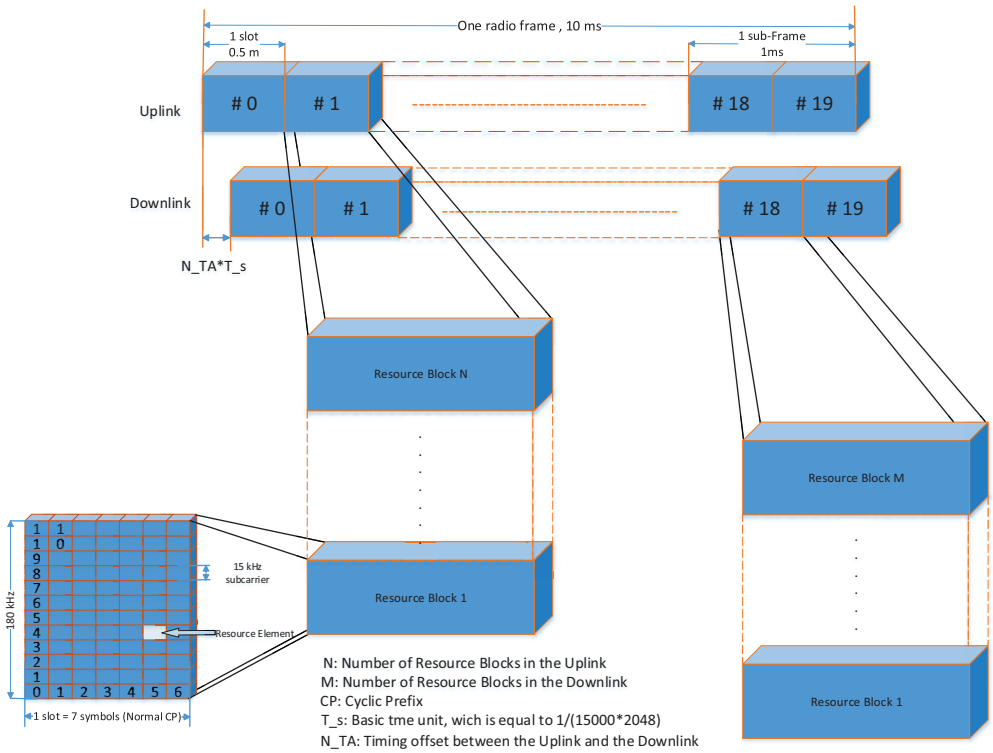

Figure 3.3 LTE radio frame structure for FDD

Regarding frame structure of Type 2 (Fig. 3.4), it is applicable to TDD. Each radio frame of Type 2 is also of length $T_{frame} = 307200T_s = 10\,ms$, dividing into two half-frames of length $5\,ms$. The half-frame is further sub-divided into five subframes of length $T_s = 1\,ms$, while each subframe consists of two slots of length $T_{slot} = 0.5\,ms$. However, for frame structure of Type 2, there are many uplink/downlink configurations, where the supported ones are listed in Table 3.1. Generally, there are three types of subframes; downlink subframe "D" for downlink transmission, uplink subframe "U" reserved for uplink transmission, and special subframe, denoted as "S". The special subframe "S" consists of three fields; Downlink Pilot Time Slot "DwPTS", Uplink Pilot Times Slot "UwPTS", and Guard Period "GP" (you may refer to [2] for more details). For the release 12 of 3GPP, the subframes 0 and 5 and "DwPTS" are always reserved for downlink transmission, while the field "UwPTS" and the subframe that immediately follows are reserved for uplink transmission.

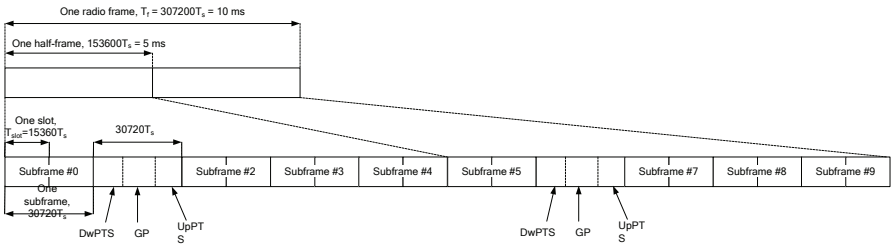

Figure 3.4 LTE radio frame structure for TDD [2]

Uplink-downlink configuration	Downlink-to-Uplink Switch-point periodicity	Subframe number									
		0	1	2	3	4	5	6	7	8	9
0	5 ms	D	S	U	U	U	D	S	U	U	U
1	5 ms	D	S	U	U	D	D	S	U	U	D
2	5 ms	D	S	U	D	D	D	S	U	D	D
3	10 ms	D	S	U	U	U	D	D	D	D	D
4	10 ms	D	S	U	U	D	D	D	D	D	D
5	10 ms	D	S	U	D	D	D	D	D	D	D
6	5 ms	D	S	U	U	U	D	S	U	U	D

Table 3.1 Uplink-downlink configurations for frame structure Type 2

3.4.2 Slot Structure and Physical Resources

To go further inside, the slot itself corresponds to $N_{symb}^{UL}/N_{symb}^{DL}$ symbols for Uplink/Downlink in the time domain, which are seven or six symbols, for normal and extended Cyclic Prefix (CP), respectively [2]. This symbol is an Orthogonal Frequency-Division Multiplexing (OFDM) symbol in the case of downlink, and Single Carrier-Frequency Division Multiple Access (SC-FDMA) in the case of uplink. In the frequency domain, the slot consists of $N_{RB}^{UL} \times N_{sc}^{RB}$ sub-carriers in the Uplink and $N_{RB}^{DL} \times N_{sc}^{RB}$ sub-carriers in the Downlink. Note that N_{RB}^{UL}/N_{RB}^{DL} is the number of Resource Block (RB) in the Uplink/Downlink and N_{sc}^{RB} is the number of sub-carriers in one RB. Resource Element (RE), the smallest modulation structure in LTE, consists of exactly one symbol in the time domain and one sub-carrier in the frequency domain. Many REs are grouped together to form one RB, which generally consists of $N_{symb}^{UL}/N_{symb}^{DL}$ symbols in the time domain, and N_{sc}^{RB} sub-carriers in the frequency domain. For normal CP, one RB consists of 12 sub-carriers, which is equal to $180kHz$, and seven OFDM/SC-FDMA symbols in the time domain. More details can be found in Table 3.2.

3.4.3 UE State Machine

A LTE terminal can be in one of the two Radio Resource Control (RRC) states: RRC_CONNECTED and RRC_IDLE [47, 48, 20], as illustrated in Fig. 3.5. When

Configuration	Sub-carrier spacing Δf (kHz)		N_{sc}^{RB}		N_{symb}^{UL}	N_{symb}^{DL}
	Uplink	Downlink	Uplink	Downlink		
Normal CP	15	15	12	12	7	7
Extended CP	15	15	12	12	6	6
		7.5		24		3

Table 3.2 Parameters of Physical resource blocks

a terminal turns on, it will be in RRC_IDLE state, in which it does not belong to a specific cell and thus no RRC context is established [48]. By consequence, the terminal can not neither receive nor transmit specific data. However, it can receive broadcast information (by e.g. monitoring paging channel), like cell system information required to communicate with the network. In order to move to the RRC_CONNECTED state, the terminal has to perform the RACH procedure, which will be detailed in the next section. In the another state, i.e. RRC_CONNECTED, the terminal belongs to a

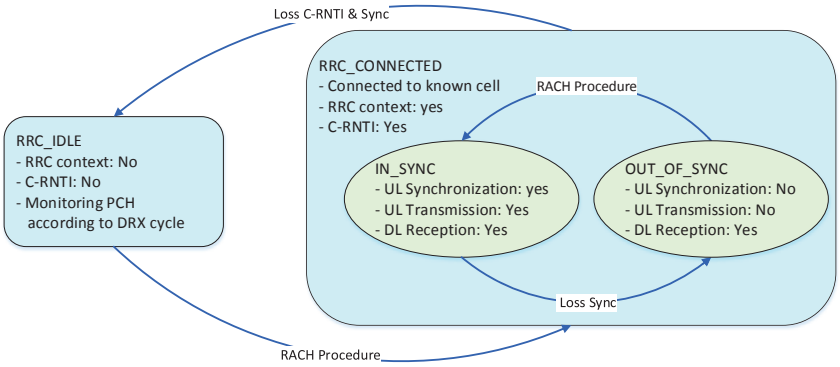

Figure 3.5 UE State Machine in LTE

specific cell. Moreover, it has RRC context, allowing the terminal to transmit/receive unicast data to/from the network. In this state, the terminal has, among other things, a temporal identity, named as Cell-Radio Network Temporary Identifier (C-RNTI), assigned by the cell to which the terminal is attached. Depending on whether there is uplink synchronization, the RRC_CONNECTED state can be sub-divided into two substates: IN_SYNC and OUT_OF_SYNC. As long as the uplink is synchronized, the uplink transmission is possible. Otherwise, the terminal has to perform the RACH procedure in order to restore the uplink synchronization.

3.4.4 Attach Procedure

Let us suppose that a terminal turned off. Once turned on, it will be in the RRC_IDLE state, meaning that it does not belong to a specific cell. Before transmitting/receiving specific data, the terminal should be connected to the network, more specifically LTE-based network. To do so, the terminal will first perform cell search. In the following, the main steps to move from idle to connected (i.e., to get connected) will be summarized:

1. Cell search and Synchronization procedures: after turning on, the terminal will find the appropriate cell, e.g. depending on the signal strength, and acquire time and frequency synchronizations [49].
2. Acquiring the cell system information: once the cell search and synchronization are finished, the UE receives and decodes information system, Master Information Block (MIB),broadcasted by the network, on the Physical Broadcast Channel (PBCH), which is needed to communicate to the network.
3. RACH Procedure: after acquiring the downlink synchronization and the cell system information, the terminal needs to acquire the uplink synchronization in order to connect to the network. This is done by the RACH procedure, explained below in details. We recall that the downlink synchronization allows the terminal to receive the information broadcasted by the network, not to receive specific information. In order to receive specific information, the terminal should require it from the network, and thus the terminal has to first do the RACH procedure to request specific data.

RACH Procedure

A UE trying to connect to the network has to perform RRC connection setup procedure (see Fig. 3.6) [49, 3]. The first four signaling steps concern the random access procedure, also known as Initial Ranging (IR) [50–53], and they are detailed below. Two forms of RACH procedure exist: contention-based and contention-free random access procedures. The first one is used, for example, when a UE moving form RRC_IDLE to RRC_CONNECTED or a UE trying to recover the uplink synchronization. The contention-free procedure can be used, for example, for handover or downlink data arrival [49].

Before starting the RACH procedure, many serving cell-related information is supposed to be available. Some of this information are (you may refer to [54, 48] for more information):

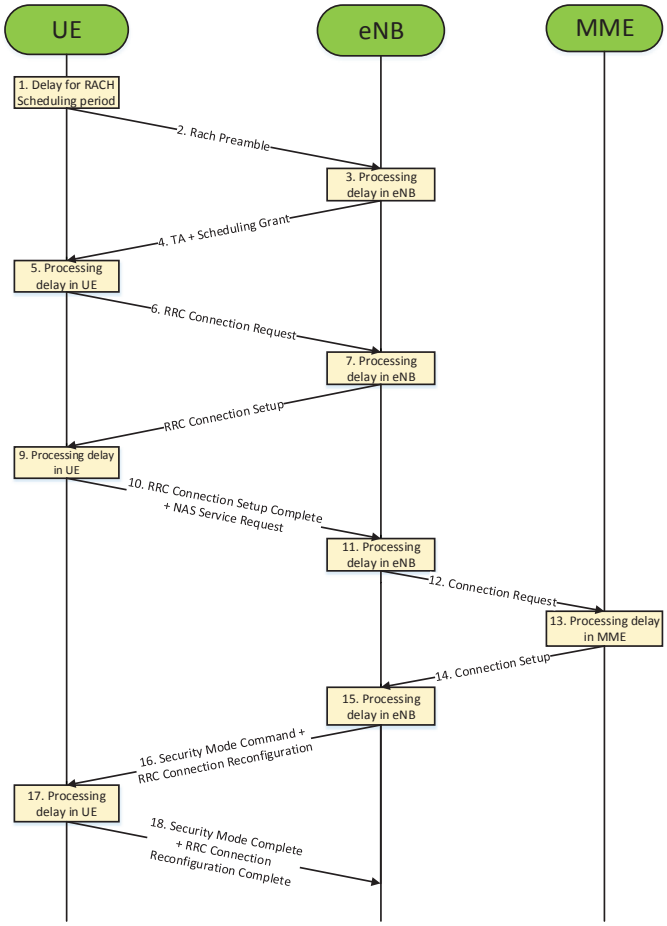

Figure 3.6 Control-Plane activation procedure [3]

1. $prach-ConfigIndex$: It represents the set of PRACH resources available for Random Access (RA) preamble transmission, i.e. when and where the preamble can be transmitted. It is given by higher layers. Table 3.3 lists the values of this parameter and also the subframe numbers within which the PRACH preamble transmission is possible.
2. $numberOfRA-Preambles$, $sizeOfRA-PreamblesGroupA$, and $sizeOfRA-PreamblesGroupB$: The parameter $numberOfRA-Preambles$ defines the number of available RA preambles, while $sizeOfRA-PreamblesGroupA$ and $sizeOfRA-PreamblesGroupA$ determine the number of RA preambles available in the groups A and B, respectively. The aim behind defining two RA preamble groups is to let the network, more specifically eNB, knowing the size of resources that the terminal needs. Therefore, the choice of a group is a two-digit information sent with the preamble.

PRACH Configuration Index	Preamble format	System frame number	Subframe number
0	0	Even	1
1	0	Even	4
2	0	Even	7
3	0	Any	1
4	0	Any	4
5	0	Any	7
6	0	Any	1, 6
⋮	⋮	⋮	⋮
58	3	Any	2, 5, 8
59	3	Any	3, 6, 9
60	N/A	N/A	N/A
61	N/A	N/A	N/A
62	N/A	N/A	N/A
63	3	Even	9

Table 3.3 PRACH configuration Index values for frame structure Type 1 [2]

3. $ra-ResponseWindowSize$: It is the size of Random Access Response (RAR) window, defined below, in a subframe unit. It can take the following values: 2, 3, 4, 5, 6, 7, 8, and 10 subframes.
4. $powerRampingStep$: It is the power ramping factor, explained in section 3.4.5, in a dB unit. This parameter can take the following values: 0, 2, 4, and 6 dB.
5. $preambleTransMax$: This parameter defines the maximum number of transmissions of the preamble when a failure takes place. The maximum number of preamble transmission can take one of the following values: 3, 4, 5, 6, 7, 8, 10, 20, 50, 100, or 200.
6. $preambleInitialReceivedTargetPower$: It is the initial preamble power, detailed in section 3.4.5, in a dBm unit. It takes the values between $-120\,dBm$ and $-90\,dBm$ with a step 2, i.e. -120, -118, -116, ..., -116, -92, $-90\,dBm$.
7. $maxHARQ-Msg3Tx$: It is the maximum number of Msg3 HARQ transmissions, and it can takes integer values from 1 to 8.
8. $mac-ContentionResolutionTimer$: It is the timer for contention resolution, and it can take the following values: 8, 16, 24, 32, 40, 48, 56, 64 subframes.

After retrieving the parameters from the network, especially those broadcasted by the cell in System Information Block Type 2 (SIB2), the terminal starts RACH procedure whose steps are as follows:

1. *Random Access Preamble Transmission (Msg1)*: The first step consists in transmitting a randomly chosen preamble. This step allows the eNB to estimate the transmission timing, i.e. Timing Alignment (TA), of the terminal that will

be used to adjust the uplink synchronization. Note that the terminal transmits the preamble by assuming that TA is zero.

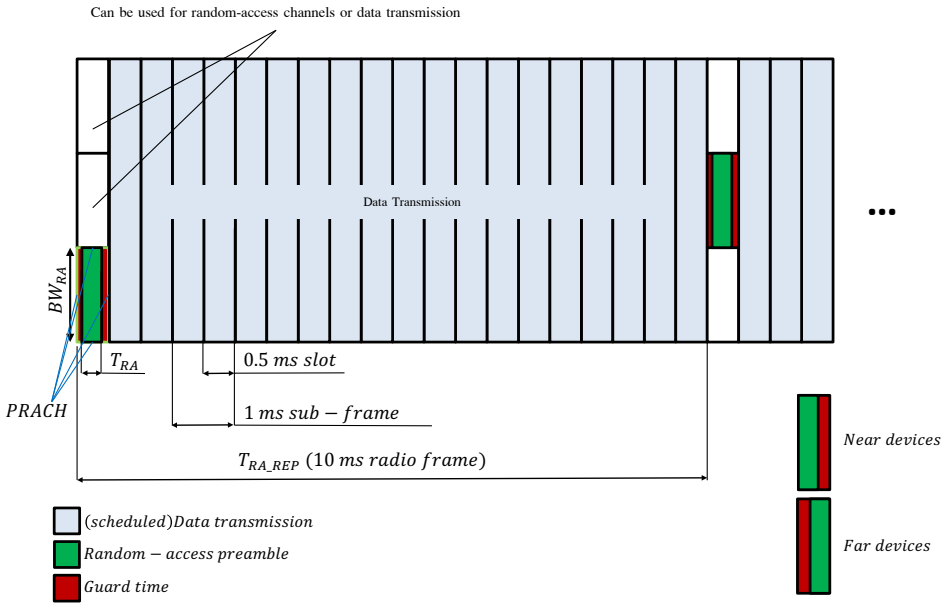

Figure 3.7 PRACH and RAOs illustration

The time-frequency resources in which the preamble is transmitted is known as the Physical RACH (PRACH), see Fig. 3.7. It should be noted that the random access transmission takes place in a specific sub-frame [55], named as a random access slot. The random access resources in LTE and LTE-A are determined in terms of Random Access Opportunities (RAOs), which are equal to the number of frequency bands in each random access slot multiplied by the number of reserved random access preambles [54]. As the preamble is randomly chosen, we may encounter the case that more than one terminal choose the same preamble, and therefore a collision will take place. However, collisions detection by eNBs is not always possible and depends on the cell size [56]. When the cell size is large enough, more than twice the distance corresponding to the maximum delay spread, the collision detection may be possible. If so, the eNB does not respond to the corresponding terminals, meaning that a collision has taken place. In the case where the cell is small, the collision detection is not possible, and therefore the terminals corresponding to the concerned preamble will wait until the reception of the last message of RACH procedure, i.e. *Msg4*, to know that a collision happened. Another important objective of this step is to adjust the power transmission of the terminal, which is achieved by the power ramping factor that is Power Ramping Step (PRS) (defined in equation 3.3). For the first

time of preamble transmission, all the terminals in the cell will transmit with the same power. The received power level of the signals transmitted by terminals close to the base station, i.e. eNB, would be enough to be detected, while this level for those far from the eNB may not be sufficient to be detected. In the latter situation, these terminals will retransmit the preamble with a power level $PRS\,dB$ higher than the one used in the precedent attempt. The advantage of this technique is that each terminal uses the power level that ensures that the signal is well detected by the eNB, without wasting any additional power.

2. *Random Access Response Reception (Msg2)*: Once the random access preamble is transmitted, the terminal, User Equipment (UE) or MTC, monitors the Physical Downlink Shared CHannel (PDCCH) to receive a Random Access Response (RAR) message during the RAR window. Thus, the RAR message is sent by the eNB on PDCCH, and identified by a Random Access_Radio Network Temporary Identifier (RA_RNTI) associated with the PRACH in which the random access preamble is transmitted. The RA_RNTI is obtained as follow:

$$RA\_RNTI = 1 + t_{id} + 10 \times f_{id} \tag{3.1}$$

where t_{id} is the index of the first subframe of the specified PRACH ($0 \leq t_{id} < 10$) and f_{id} is the index of the specified PRACH within that subframe ($0 \leq f_{id} < 6$) [49]. When using FDD, the f_{id} is equal to 0 and, therefore, the RA_RNTI is specified by the subframe number plus 1, i.e. $1 \leq RA\_RNTI \leq 10$. For Non-contention based RACH procedure, the terminal supposes that the RACH procedure successfully finished by the successful reception of RAR message, while the terminal with contention based continues to the third step.

3. *RRC Connection Request (Msg3)*: After the successful reception of Msg2 and adjusting the uplink synchronization, the UE sends the Msg3 containing its ID and the RRC connection request using the UpLink-Shared Channel (UL-SCH) assigned to the UE in the step 2.

4. *RRC Connection Setup (Msg4)*: This step helps in solving access problems when more than one terminal use the same resources (the same preamble and the same PRACH) while successfully receiving the second message (Msg2). Indeed, the terminals, in this case, share the same temporary identifier (TC-RNTI). Each terminal receiving the downlink message compares the identity in the message with the one transmitted in the third step. Only the terminal that observes a match between the two identities will declare that the random access

procedure has been successfully finished. However, the other terminals restart the RACH procedure. Note that the last two steps, i.e. $Msg3$ and $Msg4$, have a twofold objective: to request RRC connection and to solve the problem when more than one terminal choose the same preamble and the eNB successfully decode this preamble.

3.4.5 Power Consumption in the RACH Procedure

As stated earlier, power consumption is very critical for efficient deployment of MTC, especially in case of Massive MTC. The RACH procedure represents one of the most energy consuming procedures in the MTC device lifecycle. Formally speaking, the preamble transmission power can be expressed as follows [57]:

$$P_{PRACH} = min\{P_{CMAX}, PRTP + PL\} \tag{3.2}$$

where, P_{CMAX} is the maximum UE transmit power as specified in [58], PL is the Path Loss, explained below. It is worth noting that the maximum value of P_{CMAX} is $23\,dBm$, as specified by 3GPP. $PRTP$ is the Preamble Received Target Power, which is the perceived power level of the PRACH preamble when reaching the eNB. This power is given by the following equation [54]:

$$PRTP = PIRTP + \Delta_{prmbl} + (n_{tr} - 1) * PRS \tag{3.3}$$

where $PIRTP$ is the Power Initial Received Target Power, representing the initial values by which the PRACH preamble is transmitted for the first time, and it takes the values between $(-120\,dBm)$ and $(-90\,dBm)$ with a step (2), i.e. $PIRTP = \{-120, -118, ..., -90\}\ dBm$. Δ_{prmbl} is the preamble format based offset, and its value depends on the preamble format, where $\Delta_{prmbl} = 0\,dB$ for the preamble format 0. n_{tr} is the current number of preamble transmissions. PRS is the Power Ramping Step, which is the power ramping factor, and it can take the following values $\{0, 2, 4, 6\}\,dB$ [48]. PRS represents the open loop power control during the RACH procedure, wherein the UE increases its transmit power by $PRS\,dB$ in the next time when the preamble transmission fails. Regarding the pathloss PL, it is the downlink pathloss calculated in the UE in a dB unit. Pathloss can be defined as the signal attenuation between the transmitter and the receiver as a function of the propagation distance and other parameters, such as the environment and the frequency [59, 60]. As there is pathloss,

the UE should compensate this attenuation so that the signal would reach the receiver with the desired power level.

3.4.6 System Architecture

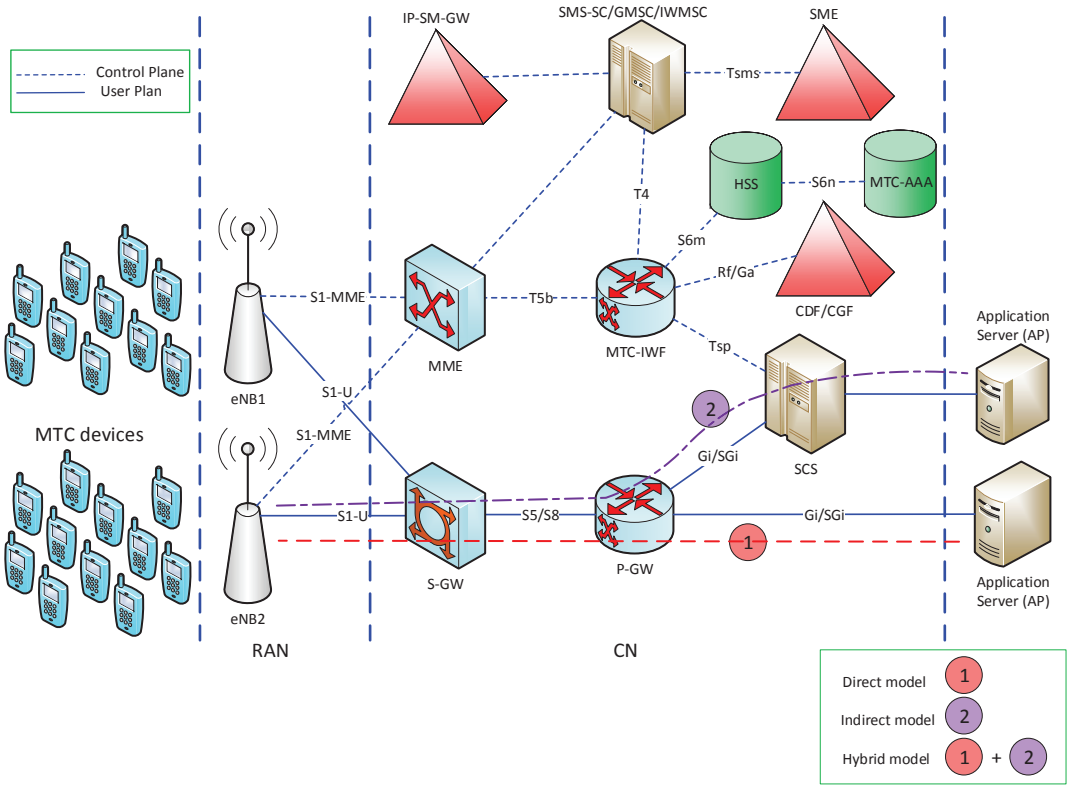

Figure 3.8 3GPP Architecture for Machine-Type Communication[4]

Fig. 3.8 illustrates 3GPP architecture for MTC [4]. It consists of three main domains: MTC domain, communication network domain, and MTC application domain. MTC application domain comprises MTC servers, which are under the control of the mobile network operator or a third party. Two new entities related to MTC communication have been recently added to the 3GPP architecture: MTC InterWorking Function (MTC-IWF) and Services Capability Server (SCS). As shown in the figure 3.8, there are three ways for establishing a communication between MTC servers and MTC devices: direct model, indirect model, and hybrid model [4]. In the direct model, a MTC server connects directly to the operator network in order to perform user plane communications directly with the UE without using any SCS. In the indirect model, the MTC server connects indirectly through the services of a SCS to the operator network. The hybrid model is when the direct and indirect models are used simultaneously.

Network Elements

The 3GPP network elements supporting the indirect and hyprid models of MTC are: MTC-IWF, Home Subscriber Server (HSS), PDN-GateWay (P-GW), MME, and others like MTC-Authentication, Authorization and Accounting (MTC-AAA) and Short Message Service-Service Centre (SMS-SC). In the following, a description of some of network elements is introduced, focusing on the ones supporting the indirect and hyprid models of MTC.

1. evolved Node B (eNB): It is LTE RAN node that interacts with the UEs via the Uu interface, connecting them to the network and the internet. However, the eNBs are interconnected with each other via the X2 interface, and they are connected to the Evolved Packet Core (EPC) through S1 interface. The eNB is responsible for many functions, such as [49]:
   (a) Radio Resource Management functions: Radio Bearer Control, Radio Admission Control, Connection Mobility Control, and resource scheduling (dynamic allocation of resources to the UEs in the uplink and the downlink).
   (b) Compression and encryption of IP header for user data stream.
   (c) Selection of MME when it is not determined from the information provided by a UE requesting network attachment.
   (d) Routing user plan data towards Serving-GateWay (S-GW).
   (e) Scheduling and transmission of paging messages (originated from MME) and broadcast information (originated from MME or Operations & Maintenance - O&M).
2. MTC-IWF: It hides the internal topology of the Public land mobile network (PLMN) and relays or translates signaling protocols used over Tsp (a reference point used by a SCS to communicate with the MTC-IWF related control plane signaling) in order to invoke specific functionality inside the PLMN. There are one or more instances of MTC-IWF in the Home-PLMN (HPLMN) and it can be a standalone entity or a functional entity of another network element, with the ability to connect to one or more SCSs [4]. This entity hosts the following functions [61]:
   (a) Termination of reference points: Tsp (used by the SCS to communicate with MTC-IWF related control plane signaling), S6m (used by the MTC-IWF to interrogate HSS).
   (b) Supporting the ability to authorize the SCS before the establishment of communication with 3GPP network.

(c) Supporting the ability to authorize control plane requests from a SCS.

(d) Supporting the reception of device trigger request from SCS, the reporting of the acceptance/non-acceptance of the device trigger request, and also reporting the status (i.e. success, failure, or unconfirmed outcome) to the SCS.

3. HSS: It is the main database, containing subscription-related information. It is the repository of all permanent user data, such as:
   (a) User identification, numbering and addressing information.
   (b) User security information, i.e. network access control information for authentication and authorization.
   (c) User profile information.
4. MME: It is a control plane entity for all mobility related operations, authentication, bearer management in Evolved Packet System (EPS). It is also the termination point of the Non Access Stratum (NAS) signaling. Some of the MME functions are *i*) NAS signalling and signalling security, *ii*) P-GW and S-GW selection, and *iii*) MME selection for handover with MME change [62].
5. P-GW: It is the gateway terminating the SGi interface towards the Packet Data Network (PDN). P-GW is the exit/entry point of UE traffic, providing the connectivity between the UE and the external PDN. It should be noted that the UE may be connected to many P-GWs simultaneously in order to provide the UE with the access to multiple PDNs at the same time. This entity hosts many functions, such as the allocation of IP address to the UEs [62].
6. S-GW: It is the gateway terminating the SGi interface towards Evolved Universal Terrestrial Radio Access Network (E-UTRAN). At a given time point, there is a single S-GW [62]. S-GW hosts many functions, such as: *i*) local mobility anchor point for inter-eNB handover, *ii*) mobility anchor for inter-3GPP mobility, *ii*) and packet routing and forwarding.
7. SCS: This entity connects MTC application servers to the 3GPP network so as to enable them to communicate through specific services, defined by 3GPP, with MTC and MTC-IWF. The SCS can be connected to one or more MTC-IWFs and it is controlled by the operator of the HPLMN or by a third party [61].

3.4.7 MTC Communication Scenarios

3GPP has defined two scenarios for MTC communication [41]. In the first scenario, the MTC devices communicate with one or more MTC servers, where the MTC server

can be controlled either by the network operator or by a third party, as depicted in Fig. 3.9. However, The MTC devices in the second scenario, illustrated in Fig. 3.10, communicate with each other without any intermediate MTC server. It should be noted that the second scenario is not considered in the release 12 of the specification.

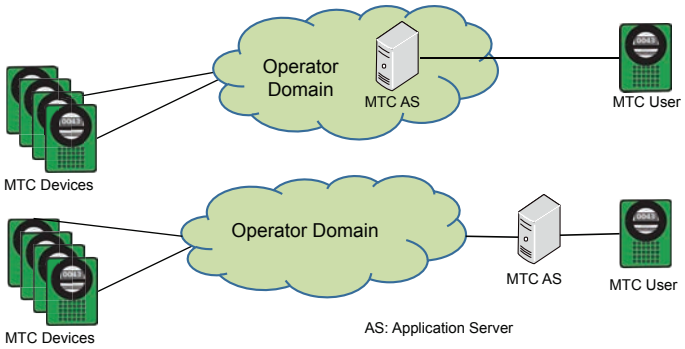

Figure 3.9 MTC devices communicating with MTC server, which is in the operator domain (top) and out of the operator domain (down)

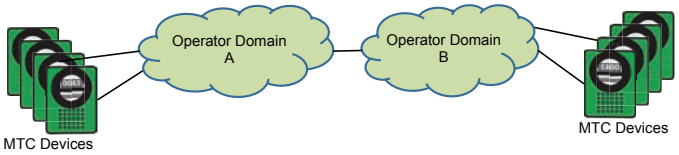

Figure 3.10 MTC devices communicating with each other directly without intermediate MTC server

3.5 Overload and Congestion Control Methods

Having described the global context of MTC in LTE, we will focus on methods and mechanisms proposed to control the overload and congestion in LTE. The existing methods can be classified into three broad classes: (i) RAN dedicated methods, where the decision and the execution of the control are done by the RAN part; (ii) RAN and CN methods, where the decision is done by one or more of CN nodes and the execution is done mainly by the RAN part; (iii) CN dedicated methods, where the decision and the execution of the control are done by the CN part.

3.5.1 RAN Congestion Control Methods

As illustrated in Fig. 3.11, the RAN congestion and overload control methods can be classified into two broad categories: Control without traffic discrimination and Control with traffic discrimination.

Figure 3.11 Classification of Congestion Control Methods

Control Without Traffic Discrimination

The methods under this category try to alleviate/eliminate the RAN congestion without splitting the traffic of H2H and M2M. Mechanisms falling in this category are:

1. Increasing the capacity of the system: It is one of most efficient mechanisms to cope with the congestion and system overload. In general, the capacity of a communications system can be increased either by increasing the available spectrum, by improving the spectrum efficiency, or by increasing the number of cells in the system. The first two ones have been used in the LTE/LTE-A systems by increasing the spectrum, comparing to 2th Generation (2G) and 3th Generation (3G), for the first possibility, and by using higher-order Quadrature Amplitude Modulation (QAM), e.g. 32-QAM, 64-QAM, or even 256-QAM (expected to be applied in the future for 5G), for the second possibility. The spectrum efficiency can be achieved by other ways, such as 3D Multiple-Input Multiple-Output (MIMO) [63, 64]. Another way to increase the system resources, more specifically the resources available for contention access, is to introduce codewords [65], where a codeword consisting of many consecutive RA slots is used instead of using just one RA slot. In order to further increase the capacity of the system, new cells can be introduced in the system, which is the notion of small cells [66, 67]. Beside the increasing of the capacity of the system, the station's cooperative can be effectively used in order to cope with the congestion. This can be done by shifting the traffic from one cell having higher load to another one having lower load, since there is overlap between the cells (i.e., between macro cells and small cells) as illustrated by Fig. 3.12. The authors

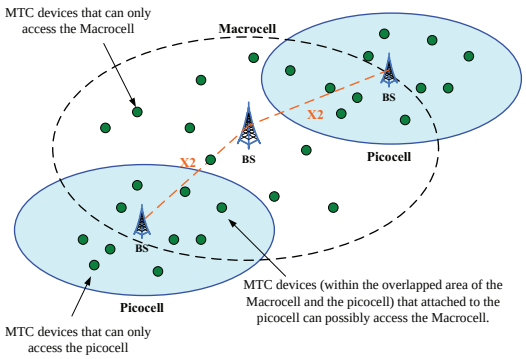

Figure 3.12 Cooperative between picocells and macrocells [5]

in [5] proposed a mechanism based on the cooperative between the picocells and macro cells to adjust the Access Class Barring (ACB) factor of the ACB method, explained later, for improving the performance. Although the proposed

mechanism significantly improved the access delay, it is infeasible for MTC devices with no mobility feature (as it is applied only to the overlapped region). It also did not consider the priorities among devices. However, increasing the system's capacity is not always possible. Therefore, other methods to remedy the concerned problem are needed.

2. System parameter tuning: In this mechanism, some of system's parameters are changed in order to cope with the congestion [68]. For example, parameters related to RACH procedure can be changed, such as the maximum number of preamble transmission, the size of RAR window, etc. The advantage of such solution is to keep the specification of 3GPP unchanged. However, it is not adaptive to the dynamic change of MTC traffic, because the change of the system's parameters takes time (the time to receive the parameters broadcasted by the eNB and apply them by the UE). Moreover, it is not always a good solution to change the parameters of the system. As illustrated in [6], increasing the number of preamble transmission increases the collision probability and decreases the success probability, i.e. the case becomes worse. This result is not trivial, as the success probability logically should increase as the terminals will have more chance to get access by increasing the number of preamble transmission. However, this behavior is not explained neither by 3GPP nor in the literature. This behavior explained in section 6.3.3. More specifically, Fig. 6.15 is a good explanation of the behavior in question, where the number of successful MTC devices degrades when increasing the number of MTC devices.

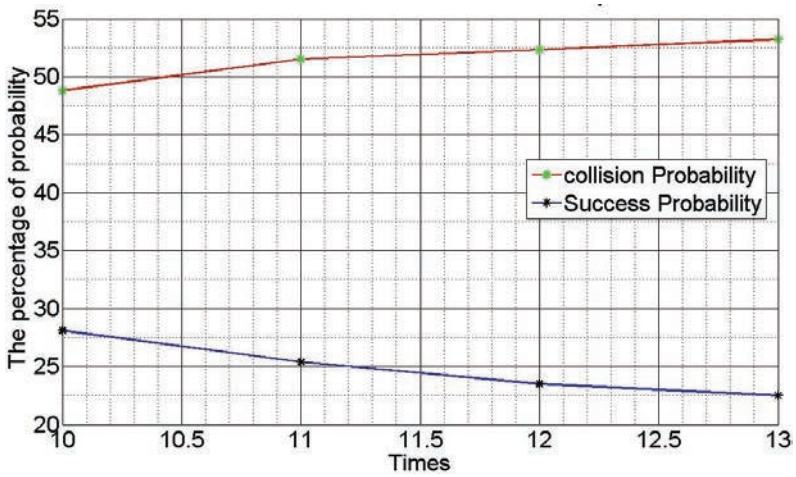

Figure 3.13 Success and collision probabilities as a function of the number of preamble transmissions [6]

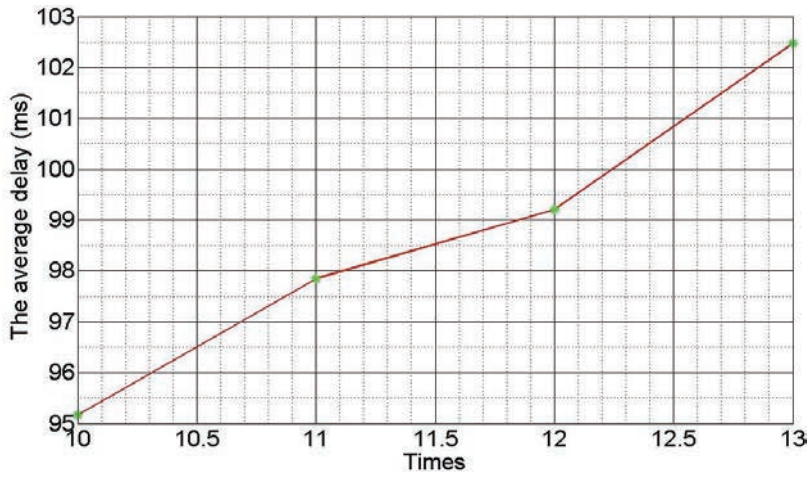

Figure 3.14 Average access delay as a function of the number of preamble transmission [6]

3. Station's cooperative: By exploiting the overlap between two adjacent cells, this mechanism tries to shift the traffic from one eNB having more traffic to another one having less traffic by the cooperation between the two cells. An example of such mechanism is proposed in [69]. This solution is a partial one as it is not applicable to the whole cell, only the overlapped regions among the eNBs can benefit from it. Moreover, it is inadequate for the MTC devices with no-mobility feature.

Note that the mechanisms in this category may not be used alone, but rather with other mechanisms discriminating between H2H and M2M devices, such as station's cooperative along with ACB [5, 69].

Control with traffic discrimination

Different from the first category, the methods under this category try to remedy the overload and congestion by discriminating between the M2M and H2H traffic. Under this category, many sub-categories can be found:

1. **Traffic Spreading**: This mechanism tries to spread the MTC devices over a long period of time, alleviating the contention on the RACH. Mechanisms that can be found in this sub-category are:
   (a) Backoff Indicator (BI) adjustment: This type of methods tries to apply longer BI, improving the access performance for delay tolerant devices [70, 68]. In the RACH procedure, for example, if there is collision in the transmission of the preamble or the RAR message been received, the MTC device draws out a randomly generated value between 0 and the value of BI and, then, it does

backoff during time equal to this value. Figs. 3.15 and 3.16 illustrate the success and collision probabilities and the average access delay for different values of BI (the results are provided by 3GPP [6]). The results clearly demonstrate that the performance is better by increasing the value of BI, but it comes at the price of delay. This method is effective under normal load, but its performance degrades when a large number of MTC devices tries to access/attach the network at the same time [71]. Note that BI adjustment has been well analyzed [72, 73, 70, 74].

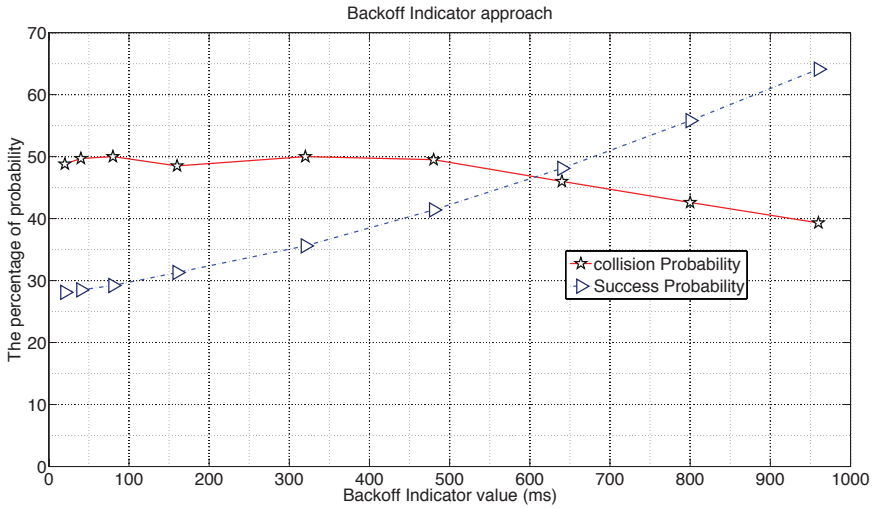

Figure 3.15 Success and collision probabilities of Backoff Indicator method for 30000 MTC devices distributed over 10 seconds [6]

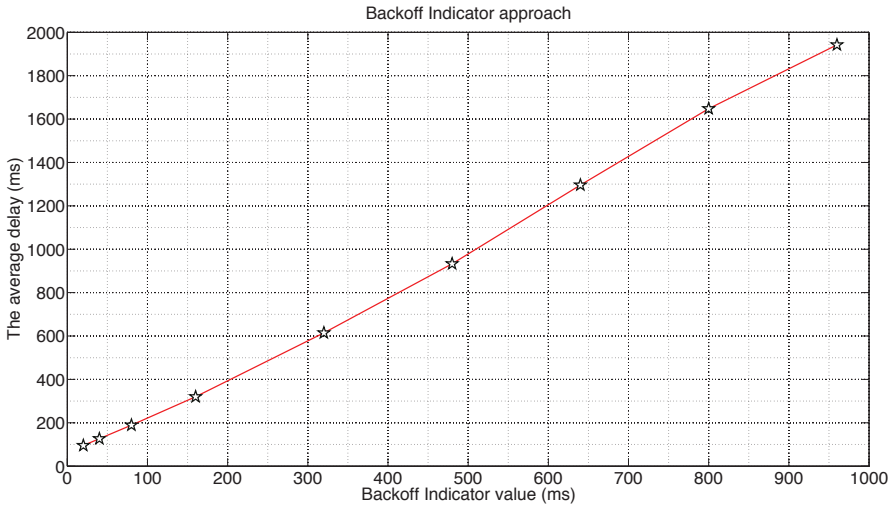

Figure 3.16 Average access delay of Backoff Indicator method for 30000 MTC devices distributed over 10 seconds [6]

(b) p-persistent: In this method, each MTC device is assigned a predefined value p. When a MTC device needs to attach/access the network, i.e. trying to do RACH procedure, it draws out a randomly generated value in the interval $[0, 1]$. If the generated value is smaller than the value p, the MTC device can start the RACH procedure by transmitting the preamble. Otherwise, the device does backoff [6]. Figs. 3.17 and 3.18 depict the results of simulation done in [6]. By decreasing the p-persistent value, there is a limited improvement on the successful probability and the collision probability (slightly decrease), while the delay quickly increases as more MTC devices do backoff when decreasing the value p.

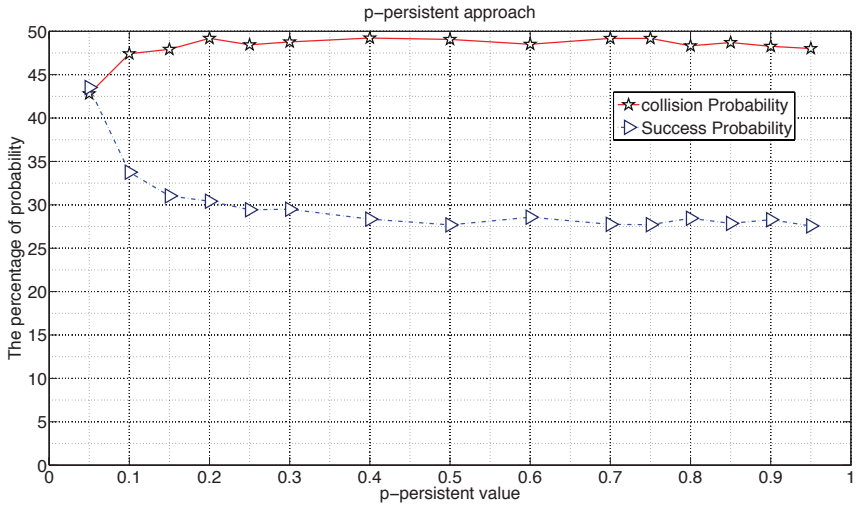

Figure 3.17 Success and collision probabilities of p-Persistent method for 30000 MTC devices distributed over 10 seconds [6]

(c) Wait timer adjustment: When failing to receive the RAR message, for example, the MTC device has to wait for a period defined by the parameter "wait timer" [68]. The simulation results depicted in Fig. 3.19 clearly demonstrate that the performance will be improved by increasing the wait timer value, but this improvement comes at the price of delay, as illustrated in Fig. 3.20. Nevertheless, this method can be applied for delay tolerant MTC devices.

From the above presented simulation results, we conclude that a tradeoff between the success probability and the delay is unavoidable.

2. **Exploiting the MTC's Features**: The methods under this sub-category try to exploit some features of MTC devices, such as no mobility feature, in order to enhance the performance. This feature can be exploited in many ways, such as disabling the signaling due to Tracking Area Update (TAU), grouping the MTC devices (as discussed later in the category of grouping), and exploiting

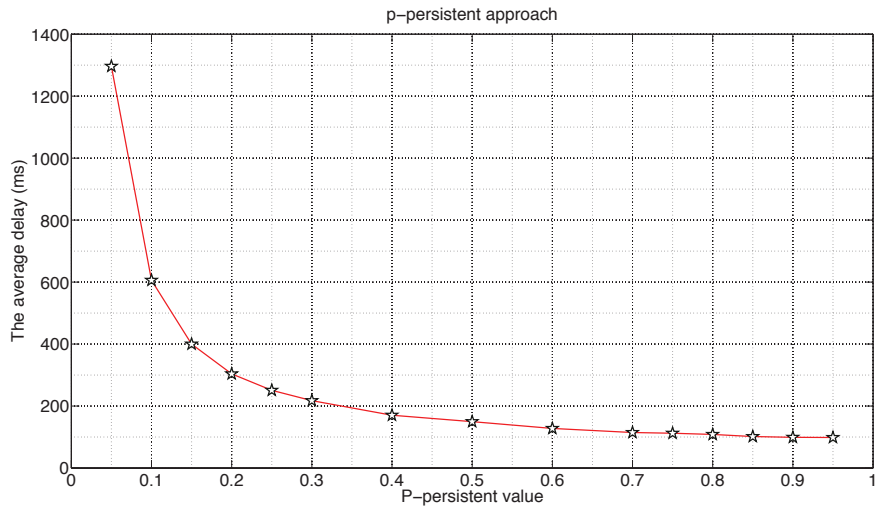

Figure 3.18 Average access delay of p-Persistent method for 30000 MTC devices distributed over 10 seconds [6]

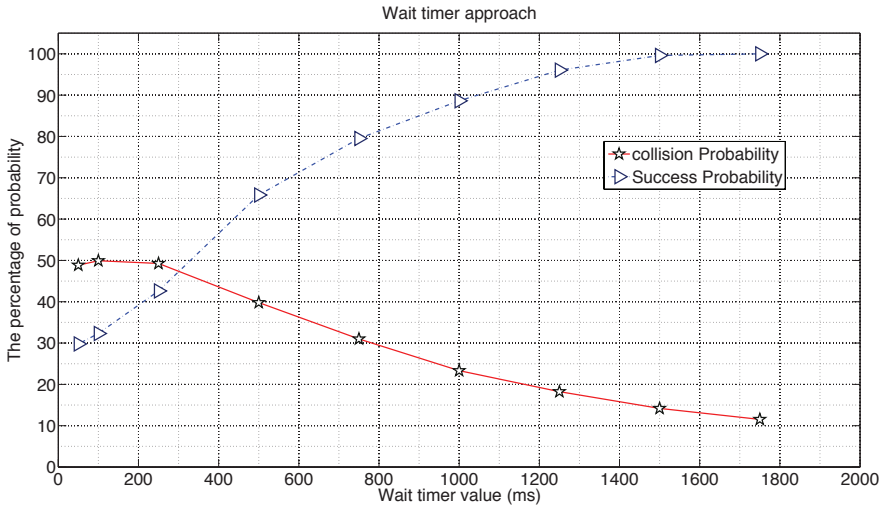

Figure 3.19 Success and collision probability of wait timer adjustment for 30000 MTC devices distributed over 10 seconds [6]

the fixed location of MTC devices to enhance the RACH procedure [75]. The authors in [75] used the fixed-location information to avoid the confusion when two or more MTC devices send the same preamble and the eNB successfully received it. In this method, the device compares the TA value in the response message with its own TA value. If there is a match, the device continues to the next step of RACH procedure, otherwise it does backoff and restarts the RACH procedure. In spite of the improvement, the method has a problem that is the MTC devices with the same distance from the eNB have the same TA. Moreover, the number of MTC devices having the same TA increases with the

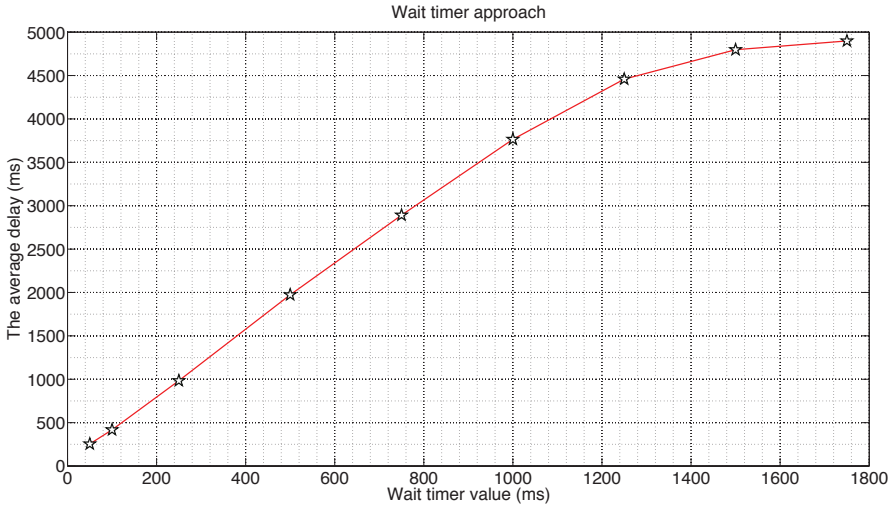

Figure 3.20 Average access delay of Backoff wait timer adjustment for 30000 MTC devices distributed over 10 seconds [6]

increasing distance from the eNB. However, the proposed method still has the superiority compared to the conventional RACH procedure.

3. **Prediction**: As expected from its name, prediction-based methods try to estimate the incoming RA attempts in order to cope with the congestion and system overload. It should be noted that this sub-category does not solve the congestion's problem itself, but it is used as a basic step to take an appropriate action, i.e. to remedy the congestion. So, it can be used with other methods, such as p-persistent or ACB scheme. As an example, the authors in [76] introduced a mechanism, named Fast Adaptive Slotted-ALOHA (FASA), for this purpose. In FASA, the statistics of consecutive idle and collision RA slots are used to estimate the number of active devices for bursty traffic. The simulation results show that the proposed method well estimates the number of active devices when the number of devices is not big, e.g. 500 devices. However, it fails in the estimation when there is a higher number of devices, e.g. 1000 devices [76]. Another example on the prediction of the arrival rate is the work in [77]. The proposed method is a combination of two propositions: a mechanism applied before accessing the RACH procedure and traffic prediction. In this mechanism, a MTC device, which is about to send an access request, randomly generates a value δ from 0 to 1. If δ is inferior or equal to α, a value broadcasted by the network, the device will proceed the RACH procedure, otherwise, it will check whether the network is stable. If the MTC passes it, it will retry the RACH procedure in the next slot, or it will abandon the procedure if it fails to pass it.

Regarding the second proposition, the traffic prediction is used to improve the performance. Simulation results show the superiority of the method compared to Backoff scheme.

4. **Grouping of MTC devices**: In the following, the mechanisms try to solve the problem on a basis of groups of MTC devices. Grouping of MTC devices can be achieved based on many metrics:
   (a) Radio Resources: When sharing the same resources, H2H and M2M devices will experience the same collision probability in case of collision. So, it is a good idea to separate the resources into two groups (one for H2H and another for M2M devices) in order to avoid the effect of M2M on H2H traffic. There is also a possibility for semi-separation, wherein one group is dedicated to H2H devices and another shared between H2H and M2M devices [78]. In general, there are three ways to divide the radio resources:
      i. Divide the preambles: the available preambles are divided into two groups in a static way, such as separate RACH resources for MTC [45]. The drawback of the static allocation of RACH resources is that when there is congestion in one type of traffic, e.g. non-MTC traffic, and there is low load of MTC traffic. In this case, the non-MTC devices can not benefit from the resources of MTC, even if the resources are non-utilized. In [73], three cases were analyzed, where the number of preambles for MTC/UE devices are 53/1, 50/4, and 46/8 for the parameters specified in [79]. Results show that, compared to the case when there is no preambles' separation, this method can reduce the impact on H2H devices, and slightly reduce the H2H performance. However, it can not improve the performance of H2H devices to a satisfactory level in the case of high level of RACH congestion [73]. In the case when the network has the capability to foresee the period when the access load will surge with MTC' traffic, additional preambles can be allocated to cope with this load, i.e. dynamic separation. Dynamic allocation of RACH resources is a good example on this type of separation [45].
      ii. Divide the Physical RACH (PRACH) opportunities: The same idea of the first mechanism can be applied on the PRACH resources, where the separation can be also static or dynamic. In [73], it was shown that the dynamic allocation of RACH resources can be beneficial to cope with the RACH congestion. However, its disadvantage is that the improvement is reduced in the case of high traffic load. Generally, RACH separation can

be enforced in the time domain, frequency domain, or both. Some ways to separate the physical RACH are:

A. **Assigning PRACH occasions in the frequency domain all the time**: In this case, one frequency band, for example, is assigned to H2H devices and another frequency band is assigned to M2M devices. Some methods fall in this type are separate RACH resources and dynamic allocation of RACH resources, which are supported by the 3GPP [45].

B. **Slotted Access**: Access slots, i.e. Access Grant Time (AGT), are assigned to the MTC devices (MTC groups), and each MTC device (MTC group) can access the network only at its dedicated access slot [74]. It can be found in many proposed methods in the literature, e.g. [80, 81].

C. **Assign PRACH resources to a group of MTC devices during certain interval**: Normally, this mechanism is used when the MTC server needs information from the MTC devices, or the MTC server is aware when the MTC devices have data to send. So, when paging the concerned MTC devices, the network can inform them the resources that can be used during certain interval, i.e. group paging [82–84]. Group paging is agreed by 3GPP as one of the candidate solutions to solve the problem of RAN overload [82].

iii. Divide the preambles and the PRACH occasions: This mechanism is also possible. For example, a group of MTC devices can be assigned one frequency band (out of many available ones) and a group of preambles (out of many available ones) during certain interval. Other divisions may be also possible.

(b) Geographical Position: As expected from the name, MTC devices are grouped into multiple groups based on the geographical location. To connect the members of the group to the network, there are many possibilities:

i. Assign dedicated resources every certain time, for example, to each group, i.e. slotted access, if the MTC devices have time controlled feature.

ii. Select a MTC device, in each group, to act as a cluster head.

iii. Apply small cells in the macro cell, e.g. each group can connect to a small cell [80].

The cluster head, when used, acts as an aggregation point, where it aggregates the traffic from the members of the group and relays it to the macro base station. The authors in [80] used the first and third possibilities, i.e. small cells slotted access mechanism are applied. The work in [85] used the second concept, i.g. each group has a cluster head. Furthermore, they shown that the reuse of the random access resources among the clusters is feasible. The reuse of the resources can be seen as a very good advantage of grouping the MTC devices based on the geographical location. It is worth noting that many other works on the clustering concepts were proposed, such as [86, 87].

(c) QoS characteristics and requirements: In this case, grouping of MTC devices is done based on many metrics/features, such as low priority access, low mobility, small data transmission, etc. The advantage of this type of grouping is that the MTC devices in each group have the same (or similar) characteristics. Therefore, in the case of congestion, a group with low priority access, for example, can be denied from access the network without having any penalty for other MTC devices having different priority (i.e., belonging to another group). The work in [88] is an example of such grouping. In this work, the authors proposed a method to group the MTC devices based on two metrics: packet arrival rate γ_i, and maximum tolerable jitter δ_i. The larger the value of γ_i is, the higher the priority of the cluster (group) is. The disadvantage of this method is that it can not handle a mass of MTC devices, as it did not take into account the number of MTC devices in each group [81]. Furthermore, it does not solve the problem of RACH congestion as the parameters are sent to the network after the RACH preamble transmission stage. Another mechanism belonging to this category of grouping is Access Class Barring (ACB), which is agreed by the 3GPP. The legacy ACB mechanism, originally designed for H2H devices, has 16 classes, i.e. classes $0-15$. The first ten classes, i.e. Classes $0-9$, are for normal UE, and the normal UE (H2H device) is assigned to one of theses classes randomly [89]. In the recent releases of 3GPP, ACB mechanism is extended by adding one or more new classes for MTC devices, and thus an individual ACB factor can be assigned for each of theses classes [69]. The drawback of legacy ACB is that there is no mechanism to differentiate the service quality between the classes as the ACB factors are common for the classes $0-9$ [1]. As mentioned in [85], the ACB can solve the congestion's problem in case of low duty cycle traffic if the MTC devices can tolerate long access delay, while it can not resolve the

problem in case of high duty cycle traffic. In LTE, when ACB takes effect, the ACB information is broadcasted in the System Information Block Type 2 (SIB2). This SIB2 contains information about the allowed ACB classes as well as two parameters (for the barred classes): $ac\_BarringFactor$ and $ac\_BarringTime$ [48]. The possible values for $ac\_BarringFactor$ rang from 0.0 tp 0.95, while the values for $ac\_BarringTime$ rang from $4s$ to $512s$. The UE is not barred if a random number uniformly drawn from the interval $[0,1)$ is less than $ac\_BarringFactor$. Otherwise, it is barred. It is mentioned in [74] that the ACB can effectively alleviate/avoid the RACH congestion, compared to other push based RAN overload control approaches, in the case of network failure event where all roaming UEs will start access attempts to attach to another PLMN simultaneously. ACB has been well studied by 3GPP [74, 72, 73]. Moreover, there are many works in the literature about the improvement of ACB, such as [90, 5, 1, 81], etc. In [1], an analysis of ACB is done, and it was shown that a tradeoff between the success probability and access delay is necessary. However, the legacy ACB does not introduce a mechanism to calculate the ACB barring factor. Here, we can mention two methods. The first one is proposed in [90], based on Proportional Integrative Derivative (PID) controller, which is RAN and CN solution. In this method, the proposed PID controller is used to update the barring probability of ACB method according to the current load in the network. The second one is presented in [5]. In this method, small cells have been introduced with macro cells, and the ACB parameters have been cooperatively decided by the base stations in the overlapped regions. Simulation results show that there is about 30% improvement in the average delay compared to the ordinary ACB. However, there is a degradation in the throughput.

3.5.2 RAN and CN Congestion Control Methods

As mentioned before, in this kind of solutions, one of the CN nodes decides or sends information about the congestion to the RAN part, which in turn takes the action, i.e. admission control. The methods under this category can be classified into two broad categories:

Individual Based Control

In this category, the admission control is applied on a MTC device basis. Extended Wait Timer (EWT) is a good example of such mechanisms [68]. For example, if the message Msg3 of the RACH procedure indicates that it comes from a delay tolerant MTC devices, an Information Element (IE) could be included in the response in order to bar these MTC devices.

Group Based Control

Solutions under this category are applied to a group of MTC devices, where the grouping is done based on some metric. In the Extended Access Barring (EAB), MTC devices are categorized into classes. When there is congestion, a CN node, e.g. MME, sends to the RAN node, i.e. eNB, information about the classes to be barred. The difference from the ACB is that all the MTC devices in the barred class are barred from the access to the network in the case of EAB. However, in the ACB the MTC devices are barred based on the probability, i.e. some MTC devices can access while the others can not even thought they belong to the same class. The work in [90] is another example, where a PID controller is used to calculate the barring probability for ACB, which is updated according to the load in the network. Simulation results show the superiority of the proposed method compared to the legacy ACB. Congestion control based on common information, e.g. MTC devices send information to the same destination, is a good mechanism to tackle with the congestion problem. The authors in [91] introduce the concept of bulk handling in order to group (squeeze) many messages having common information. Tracking Area Update (TAU) message has been introduced as an example of common information sent by MTC devices. The results of the proposed method show that there is about 73% of reduction in the size of exchanged messages compared to the case where there is no grouping. The drawback of bulk is that it is infeasible for intolerable applications as the method proposes to hold back the messages for a pre-defined timeout or until a number of messages arrives. Instead of holding back the messages during certain time, a connection can be initiated between the sender and the receiver, e.g. eNB and MME, with the creation of profile between the two nodes [92]. In this case, the ID of the profile represents the common information. Upon receiving a signaling message having the common information represented by the profile ID, the node, e.g. eNB, sends a message containing the profile ID and other information of the original message that is not common to the destination, e.g. MME. Simulation results show the superiority of the profile-ID compared to simple

grouping. Although the bulk signaling shows the superiority compared to Profile-ID in the CN overload, it fails in the RAN overload as the MTC messages are grouped only when they are successfully received by the eNB [92]. Moreover, bulk signaling experiences more transmission delay as it holds back the arrival messages for a certain time. The concept of virtual Bearer is introduced in [93] to alleviate the RAN and CN overload and congestion. The main idea is to let a group of MTC devices located in a certain location of the cell to utilize the same bearer instead of creating an individual bearer for each MTC device. Although this mechanism reduces the control plane traffic, compared to the case without using virtual bearer, the RACH congestion problem remains, as the authors did not mention how the MTC devices acting as an aggregator access the network.

3.5.3 CN Congestion Control Methods

This type of solutions is totally applied by one or more of CN nodes. The work introduced in [92] is an example of such methods, where the proposed mechanism tries to alleviate the CN overload and congestion by reducing the amount of singaling messages exchanged when triggering low mobility MTC devices. This solution proposes to exclude the MME node when triggering this kind of MTC devices. Another example is the control done by the MTC-IWF node. As mentioned in [61], in case of overload, e.g. because of trigger requests sent by SCS, the MTC-IWF may inform the SCS about this overload. The SCS in turn may implement backoff timer, when the SCS does not initiate Tsp requests to the MTC-IWF, until the expiration of the timer. The authors in [94] propose that MTC-IWF controls the device triggering rate coming from MTC servers by the aid of a CN network entity, such as MME, that computes the allowed device triggering rate.

3.6 Conclusion

M2M, or MTC, is considered as the corner stone of the Internet-of-Things (IoT) vision. However, deploying this kind of devices in the current cellular mobile networks will not be achieved without causing congestion and system overload. After presenting a general introduction on MTC in the wireless networks, with a focus on 4G networks, a comprehensive survey of the overload and congestion control methods in the literature along with a classification have been presented. These mechanisms are classified into three broad classes: RAN, RAN and CN, and CN overload and congestion control

methods. The first class is further divided into two broad classes: without and with discrimination between M2M and H2H traffic, which are also sub-divided into many sub-classes. The second one, i.e. RAN and CN methods, is classified into two classes: individual-based and group-based solutions, where under each class there are many solutions. Finally, the CN methods are introduced.

Although lots of work have been proposed to remedy the problem in question, a great effort is still required in order to accommodate M2M applications in the current and future cellular mobile networks, more specifically LTE and beyond networks. With the focus on network optimization and power efficiency, the next chapters will introduce many congestion control mechanisms and algorithms based on two main axis: traffic prediction and Group Paging (GP) optimization. Besides, a powerful tool helping in validate the congestion control methods will be also introduced.

Chapter 4

Traffic Prediction and Network optimization

For many applications, M2M traffic is highly synchronized, which may cause congestion in the network's access. To model this kind of M2M traffic, 3GPP proposed the use of Beta distribution. In this model, a large number of MTC devices (e.g., 30000) will be activated and enter the network during a short period, e.g. 10 s. In such a case, static assignment of resources may not represent a good option. The reason is that the resources will be underutilized when there is a little traffic, while an overload will occur when existing a large number of arrivals. Therefore, dynamic allocation of resources will be needed. However, the dynamicity can be done only when the traffic is known, or at least estimated. In this chapter, a traffic prediction model is proposed [17], focusing on highly synchronized MTC traffic (i.e., Beta model). This chapter is organized as follows. Firstly, a state of the art of traffic prediction in the context of M2M will be introduced, and then the proposed method will be detailed.

4.1 State of the Art

Traffic prediction can be considered as one of the most efficient ways to remedy the problem of congestion and system overload. Indeed, once the traffic is known, or at least estimated (or predicted), the resources can be efficiently and dynamically allocated, thus optimizing the network resources.

In the literature, there are many works attacking the aforementioned problem. In [95], the authors proposed stabilization algorithm, which is based on the pseudo Bayesian algorithm [96], trying to control the access of users to the network by changing the transmission probability as a function of the number of estimated users at each

time. Despite its performances, this method is proposed for Poisson traffic, modeling the behavior of UEs, and thus it is traffic-dependent. However, Poisson traffic does not represent the traffic of MTC applications. As mentioned before, M2M traffic is modeled as either Uniform distribution (for a realistic traffic) or Beta distribution (for synchronized traffic). Other works on estimating the network status (i.e., the number of active users at each time) are Q+-Algorithm [97], stable control procedure [98], and Fast Adaptive Slotted-ALOHA (FASA) [76, 99]. FASA method exploits the history of many consecutive RA slots so as to estimate the number of active users at each time. By simulations, the authors proved the superiority of FASA compared to Q+-Algorithm, whose throughput suffers due to fluctuations in the estimation. Although FASA method is an efficient one, it is proposed for just one channel, where the authors mentioned that the method can be extended to multichannel case by borrowing the idea of [95]. However, the extension of FASA to multichannel case can not be done directly, as proved in section 4.2. Regarding the stable control procedure [98], it is proposed to stabilize multichannel slotted ALOHA system with MTC traffic load. Despite its performances, demonstrated by computer simulations for Poisson traffic, this proposed method does not well work under a heavy traffic, such as Beta distribution-based traffic load. This issue is proved in section 4.3.

In order to overcome the aforementioned problems a novel method, namely Multi-Channel Slotted ALOHA-Optimal Estimation (MCSA-OE), is proposed. The advantage of MCSA-OE is that it is traffic independent and well works under heavy traffic, such as Beta traffic. Before detailing the proposed method, the drawback of FASA and stable control procedure will be introduced firstly.

4.2 Why can not FASA be directly generalized to multi - channel?

FASA method uses the statistics of consecutive idle and collision slots in order to accelerate the tracking process of network status. The authors use the simplified Interrupted Poisson Process (IPP) [100] as an arrival model. Actually, IPP was firstly suggested to simulate overflow traffic. After estimating the network status, FASA method adjusts its parameters based on drift analysis in order to ensure the stability of the method.

In spite of its good performances, FASA method can not be extended to multichannel case, where it is initially designed for one channel case. Generally speaking, to prove

the invalidity of an algorithm, it is sufficient to prove the invalidity of one of its cases. This is what would be done in the following.

Firstly, let's copy the equations used by FASA to estimate the number of arrivals (the equations 17-19 of [99]). Let K_t denote the access outcomes in the past consecutive slots, where $K_0 = 0$ for $t = 0$. For $t > 0$, K_t is given by the following equation [99]:

$$K_t = \begin{cases} -\min\{K_{0,t-1}, k_m\} & ; \text{if } Z_{t-1} = 0 \\ 0 & ; \text{if } Z_{t-1} = 1 \\ -\min\{K_{c,t-1}, k_m\} & ; \text{if } Z_{t-1} = c \end{cases}$$

where, $K_{0,t-1}$ and $K_{c,t-1}$ are the numbers of consecutive idle and collision slots up to the slot $(t-1)$. Z_t is the access outcome at the time t and takes the values 0, 1, or c whether the slot is idle, successful, or collided, respectively, while k_m is the maximum number of consecutive slots. However, the estimated value K_{t+1} is given by [99]:

$$K_{t+1} = \begin{cases} -\min\{\mid K_t \mid +1, k_m\} & ; \text{if } K_t < 0, Z_t = 0 \\ -1 & ; \text{if } K_t \geq 0, Z_t = 0 \\ 0 & ; \text{if } Z_t = 1 \\ 1 & ; \text{if } K_t \leq 0, Z_t = c \\ \min\{\mid K_t \mid +1, k_m\} & ; \text{if } K_t > 0, Z_t = c \end{cases}$$

Regarding the estimation of the number of arrivals, it is given by [99]:

$$\hat{N}_{t+1} = \begin{cases} \max\{1, \hat{N}_t - 1 - h_0(v) \mid K_{t+1} \mid^v\} & ; \text{if } Z_t = 0 \\ \hat{N}_t & ; \text{if } Z_t = 1 \\ \hat{N}_t + \frac{1}{e-2} + h_0(v)(K_{t+1})^v & ; \text{if } Z_t = c \end{cases}$$

where $h_0(v)$ is a function of v that guarantees the right direction of the estimation.

In the following, the case where there is no traffic is taken, i.e. there would be many consecutive idle slots. From the precedent equations, it can be clearly remarked that the absolute value of K_t becomes large, and thus the absolute value of K_{t+1} becomes large also, as K_t is negative. Therefore, the estimated number of arrivals $\hat{N}_{t+1}$ will be equal to one, as the term $h_0(v) \mid K_{t+1} \mid^v$ will be large. As a result, the estimated number of arrivals is equal to one, while the channel is idle.

Now, the idea of [95] will be borrowed in order to extend FASA to multichannel case, as suggested by the authors. Firstly, the concerned equation of [95] is copied here

for the sake of clarification.

$$\hat{U}_{t+1} = \sum_{r=1}^{R} \hat{U}_{r,t+1}$$

where R is the number of channels, $\hat{U}_{r,t+1}$ is the estimated number of devices for the channel r ($\hat{U}_{r,t+1} = \hat{N}_{t+1}$), and $\hat{U}_{t+1}$ is the estimated number of devices for R channels. As it can be seen from the above discussion, $\hat{U}_{r,t+1}$ is equal to 1 when the channel is idle. If $R = 54$, for example, the estimated value $\hat{U}_{t+1}$ is equal to 54, though all the channels are idle. Fig. 4.1 illustrates the performance of FASA for multi-channel system under heavy traffic load, where there are 30000 MTC devices and 1000 UEs with Beta and Poisson distribution based traffic loads, respectively. The simulation parameters are summarized in Table 4.1, while the parameters of FASA are $v = 1$ and $k_m = 16$. We observe from this figure that the estimated number of arrivals $\hat{U}(t)$ remains equal to R at the start of the time. By increasing the number of arrivals, the estimated value $\hat{U}(t)$ increases with large steps, moving away from the true value. Therefore, the extension of FASA from one channel to multi-channel can not be done directly. The same case can be found for the equation of Q+-algorithm used in [76].

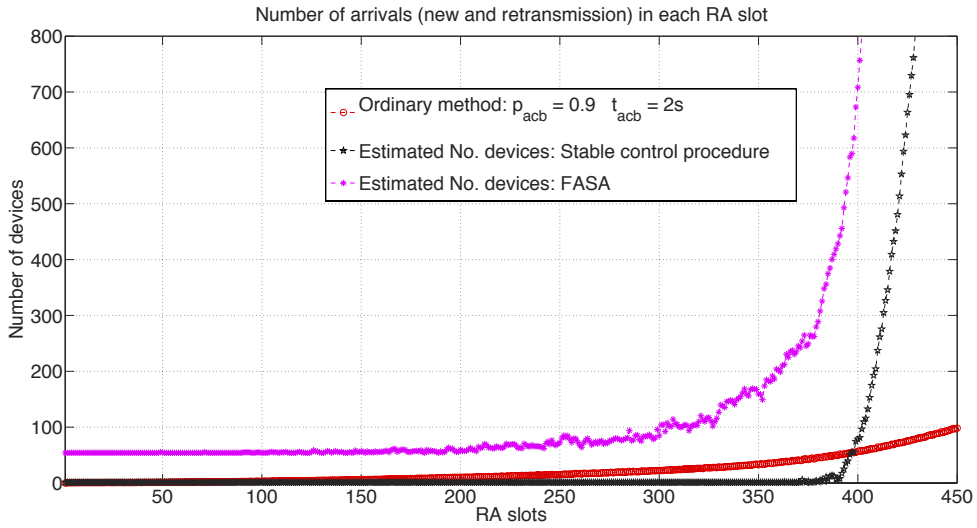

Figure 4.1 The behavior of stable control procedure and FASA for 30000 MTC devices following Beta distribution, $\alpha = 3$, $\beta = 4$, and 1000 UEs following Poisson distribution

4.3 Why is not the stable control procedure adequate for high traffic load?

In the following, the focus will be on the work presented in [98], named stable control procedure, to prove its inefficiency for high traffic load, e.g. Beta traffic. Indeed, although stable control procedure has been proposed for MTC with Poisson traffic load, it is found that this method can not be adequate for high traffic load, e.g. Beta distribution based traffic load. This method uses the following model to estimate the number of arrivals at each RA slot [98]:

$$Z(t+1) = \max\{1, Z(t) + \Delta Z(t)\}$$

$$\Delta Z(t) = \sum_{i=1}^{R} (aI(v_i(t) = 0) + bI(v_i(t) = 1) + cI(v_i(t) \geq 2))$$

where $Z(1) = 1$, $v_i(t)$ is the number of devices attempting to access the channel i at the time t. $I(v_i(t) = 0)$, $I(v_i(t) = 1)$, and $I(v_i(t) = c)$ are th indicator functions of the channel events when it is idle, successful and collided, respectively. Note that the indicator function is defined as follows:

$$I(x = \alpha) = \begin{cases} 1 & ; \text{if} x = \alpha \\ 0 & ; \text{if} x \neq \alpha \end{cases}$$

The parameters of the method are set as follows; $a = -1$, $b = -1$, and $c = 2/(e-2)$. Let $R_I(t)$, $R_S(t)$, and $R_C(t)$ denote the number of channels, where $(v_i(t) = 0)$, $(v_i(t) = 1)$, and $(v_i(t) \geq 2)$, respectively, at the time t. Thus, the equation of $\Delta Z(t)$ can be written as:

$$\Delta Z(t) = aR_I(t) + bR_S(t) + cR_C(t) = -R_I(t) - R_S(t) + \left(\frac{2}{e-2}\right) R_C(t) \tag{4.1}$$

where $R_I(t) + R_S(t) + R_C(t) = R$, which is the total number of available preambles. At the start of the time, there is a little traffic (when Beta distribution based traffic load is applied), and thus $R_I(t) + R_S(t) > \left(\frac{2}{e-2}\right) R_C(t)$. This means that $\Delta Z(t)$ will be negative, and therefore $Z(t+1) = 1$. When $\Delta Z(t)$ is equal to zero, we have,

$$-R_I(t) - R_S(t) + \left(\frac{2}{e-2}\right) R_C(t) = 0 \tag{4.2}$$

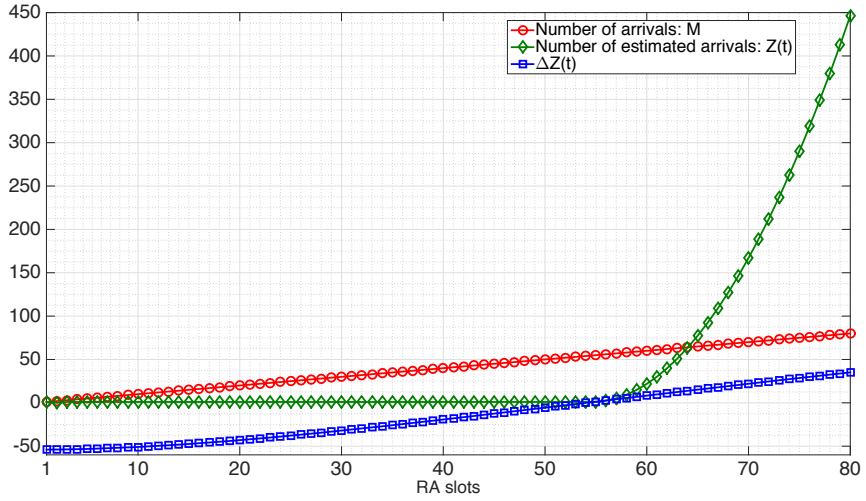

Figure 4.2 The behavior of stable control procedure in the case where the number of arrivals increases by one at each time

but $R_I(t)+R_S(t)=R-R_C(t)$. By substituting in the equation 4.2, we find

$$R_C(t)=\left(\frac{e-2}{e}\right)R \tag{4.3}$$

It is worth noting that the collision probability can be approximated by $\hat{P}_C(t)=\frac{R_C(t)}{R}$, and thus $\hat{P}_C(t)=\frac{e-2}{e}=1-2e^{-1}$. We know that the collision probability can be written as $P_C=1-P_I-P_S=1-e^{-\frac{M}{R}}-\frac{M}{R}e^{-\frac{M}{R}}$. Therefore, we find that this collision probability corresponds to the case when the number of arrivals $M(t)$ is equal to the number of channels R. Therefore, $\Delta Z(t)$ is equal to zero once $M(t)=R$, and at this point $Z(t+1)$ will start to become more than 1, as illustrated in Fig. 4.2. This can be true when $R=1$, i.e. for one channel, while it is not true for multi-channel, especially when $R=54$ for LTE and LTE-A networks. Fig. 4.1 is an additional proof of this result, where the estimated number of devices $\hat{M}(t)$ remains 1 until $M(t)$ reaches the number of channels R that is equal to 54 for the considered parameters. By increasing the number of devices, the estimated value $\hat{M}(t)$ augments with large steps, moving away from the true value.

4.4 Multi-Channel Slotted ALOHA - Optimal Estimation (MCSA - OE):

In this section, the proposed method, namely Multi-Channel Slotted ALOHA-Optimal Estimation (MCSA-OE), is introduced. The main objective of the this method

is to estimate the number of arrivals in each RA slot, and thus tries to optimize the network, i.e. maximizing the capacity of Multi-Channel Slotted ALOHA . The proposed mechanism is composed of two parts: estimation and fitting. Firstly, the estimation of the idle probability and the incoming number of devices are done. Then, fitting these values is done in order to avoid the large fluctuations in the estimation. Note that the two steps are done at the same time, i.e. at each RA slot. In the following, these two stages would be elaborated.

4.4.1 Estimation and Fitting of Idle Probability

Let R denote the number of available preambles in each RA slot. After having transmitted by the terminals, we assume that the evolved Node B (eNB) computes the number of successful, collided, and idle preambles. It should be noted that the assumption "no capture effect" is made when detecting the preambles by eNB. The capture effect means that it is possible, sometimes, for the eNB to distinguish the preambles when they are chosen by more than one terminal [68]. Let t be the time of the RA slot, where $t = 0, 1, 2, ...$ Generally, the number of preambles at any time can be written as:

$$N_S(t) + N_I(t) + N_C(t) = R \tag{4.4}$$

where $N_S(t)$ is the number of preambles that are successfully received by the eNB, $N_I(t)$ is the number of preambles that are idle, and $N_C(t)$ is the number of preambles that are used by more than one device, MTC or UE, at the time t. Dividing the equation (4.4) by the number of preambles R, we obtain:

$$\frac{N_S(t)}{R} + \frac{N_I(t)}{R} + \frac{N_C(t)}{R} = \hat{P}_S(t) + \hat{P}_I(t) + \hat{P}_C(t) = 1 \tag{4.5}$$

where $\hat{P}_S(t)$, $\hat{P}_I(t)$, and $\hat{P}_C(t)$ are the estimated success, idle, and collision probabilities, respectively, at the time t. For hereunder, we will be interested only by the estimated idle probability. The reason to choose the idle probability for the estimation is the feasibility to estimate the number of arrivals, whereas it is highly difficult to expect the number of arrivals from the other probabilities. It is worth noting that the idle probability $P_I(t)$ can be written as [101]:

$$P_I(t) = \left(1 - \frac{1}{R}\right)^{M(t)} \simeq e^{-\frac{M(t)}{R}} \tag{4.6}$$

where $M(t)$ is the number of arrivals at the time t. Therefore, $M(t)$ can be estimated from the equation (4.6) as follows:

$$\hat{M}(t) = R\ln\left(\frac{1}{\hat{P}_I(t)}\right) \tag{4.7}$$

It is clear from the equation (4.7) that small changes in $\hat{P}_I(t)$ will be accompanied with large changes in the estimated number of devices $\hat{M}(t)$, hence fitting procedure is needed. Algorithm 1 is used to fit the idle probability, where ΔP_I and ΔM are used to avoid the large fluctuations in the estimation. Regarding the fitting, it is inspired

Algorithm 1 Estimation and fitting of the idle probability $\hat{P}_I(t)$ with the estimation of the number of devices $\hat{M}(t)$

$\Delta P_I \leftarrow 1/R$
if $t \leq 2$ **then**
  $\hat{P}_{I_n}(t) \leftarrow \hat{P}_I(t)$
else

$$\hat{P}_{I_n}(t) \leftarrow \left(2 * \hat{P}_{I_n}(t-1) + \hat{P}_{I_n}(t-2) - \hat{P}_{I_n}(t-3)\right)/2 \tag{4.8}$$

end if
if $|\hat{P}_{I_n}(t) - \hat{P}_I(t)| \leq \Delta P_I$ **then**
  $\hat{P}_{I_n}(t) \leftarrow \hat{P}_I(t)$
  $\hat{M}(t) \leftarrow R * \ln(1/\hat{P}_{I_n}(t))$
else
  if $\hat{P}_{I_n}(t) - \hat{P}_I(t) > 0$ **then**
    $\alpha \leftarrow \left(\hat{P}_{I_n}(t) - \hat{P}_I(t)\right) - \Delta P_I$
    $\hat{P}_{I_n}(t) \leftarrow \hat{P}_I(t) - \alpha$
  else
    $\alpha \leftarrow -\left(\hat{P}_{I_n}(t) - \hat{P}_I(t)\right) - \Delta P_I$
    $\hat{P}_{I_n}(t) \leftarrow \hat{P}_I(t) + \alpha$
  end if
  $\hat{M}(t) \leftarrow R * \ln\left(1/\hat{P}_{I_n}(t)\right)$
end if

from the Infinite Impulse Response (IIR) filter [102], a well known digital filter. The IIR filter can be expressed as:

$$y[t] = \sum_{k_1=0}^{P} b_{k_1} x[t-k_1] - \sum_{k_2=1}^{Q} b_{k_2} y[t-k_2] \tag{4.9}$$

where, b_{k_1} and b_{k_2} are the feedforward and feedbackward filter coefficients, and P and Q are the feedforward and feedbackward filter orders, respectively. As seen from Algorithm 1, the following equations have been used to fit the idle probability

$$\hat{P}_{I_n}(t) = \hat{P}_I(t) - \left(\hat{P}_{I_n}(t-1) + 0.5\hat{P}_{I_n}(t-2) - 0.5\hat{P}_{I_n}(t-3) - \hat{P}_I(t)\right) \mp \Delta P_I$$

$$= 2\hat{P}_I(t) - \hat{P}_{I_n}(t-1) - 0.5\hat{P}_{I_n}(t-2) + 0.5\hat{P}_{I_n}(t-3) \mp \Delta P_I \quad (4.10)$$

or

$$\hat{P}_{I_n}(t) = b_0\hat{P}_I(t) - \sum_{k_2=1}^{3} b_{k_1}\hat{P}_{I_n}(t-k_1) \mp \Delta P_I \quad (4.11)$$

From the equations 4.9 and 4.11, we find that the model used for fitting is the IIR filter with $P = 0$ and $Q = 3$, where the coefficients obtained empirically. In fact, as the estimation of the number of arrivals is instantaneous and depends on the statistics of the number of preambles that are idle, the estimated value should be limited to avoid the large fluctuations. Therefore, the changes of the estimated idle probability is limited by the smallest change, which is $\Delta P_I = 1/R$.

4.4.2 Fitting the Estimated Number of Devices

The second step of the mechanism is to fit the estimated number of devices $\hat{M}(t)$. This step is necessary when the number of devices is large. As mentioned before, the idle probability can be expressed as $P_I(t) = e^{-M(t)/R}$. By taking the variation of $P_I(t)$, we obtain:

$$\frac{\Delta P_I(t)}{\Delta M(t)} = -\frac{1}{R}e^{-\frac{M(t)}{R}} \quad (4.12)$$

$$\mid \Delta M(t) \mid = R\Delta P_I(t)e^{\frac{M(t)}{R}} \quad (4.13)$$

where $\Delta P_I(t)$ is set to a fixed value, i.e. $\Delta P_I(t) = \Delta P_I$. The number of arrivals in the case when $P_I(t) = \Delta P_I$ is equal to:

$$\hat{M}(t) = R\ln\left(\frac{1}{\Delta P_I}\right) = R\ln(R) = \ln\left(R^R\right) \quad (4.14)$$

By substituting in the equation 4.13, we find:

$$\mid \Delta M(t) \mid = e^{\frac{1}{R}\ln\left(R^R\right)} = R \quad (4.15)$$

Therefore, the changes in the estimated number of arrivals will be limited to the value R. Algorithm 2 is proposed for fitting $\hat{M}(t)$, and is similar to that of $\hat{P}_I(t)$. Indeed, it follows the same IIR filter principle. The output of the algorithm are $\hat{P}_{I_F}(t)$ and $\hat{M}_F(t)$, which are the estimated values of $P_I(t)$ and $M(t)$, respectively.

Algorithm 2 Fitting of the estimated number of devices $\hat{M}(t)$ in each RA slot

$\Delta M(t) \leftarrow e^{\hat{M}(t)/R}$
if $\Delta M > R$ **then**
  $\Delta M \leftarrow R$
end if
if $t \leq 2$ **then**
  $\hat{M}_F(t) \leftarrow \hat{M}(t)$
else

$$\hat{M}_F(t) \leftarrow \left(2 * \hat{M}_F(t-1) + \hat{M}_F(t-2) - \hat{M}_F(t-3)\right)/2 \tag{4.16}$$

end if
if $|\hat{M}_F(t) - \hat{M}(t)| \leq \Delta M$ **then**
  $\hat{M}_F(t) \leftarrow \hat{M}(t)$
  $\hat{P}_{I_F}(t) \leftarrow e^{-\hat{M}_F(t)/R}$
else
  if $\hat{M}_F(t) - \hat{M}(t) > 0$ **then**
    $\alpha \leftarrow \left(\hat{M}_F(t) - \hat{M}(t)\right) - \Delta M$
    $\hat{M}_F(t) \leftarrow \hat{M}(t) - \alpha$
  else
    $\alpha \leftarrow -\left(\hat{M}_F(t) - \hat{M}(t)\right) - \Delta M$
    $\hat{M}_F(t) \leftarrow \hat{M}(t) + \alpha$
  end if
  $\hat{P}_{I_F}(t) \leftarrow e^{-\hat{M}_F(t)/R}$
end if

4.5 Performance Evaluation

4.5.1 System Model

In order to evaluate the proposed mechanism, the traffic model 2 (as proposed by the 3GPP group [45]) is considered, i.e. the arrival of devices follows Beta distribution with $\alpha = 3$ and $\beta = 4$. It is assumed that there is just one eNB, and there are 30000 MTC devices as well as 1000 UEs activated according to Beta and Poisson distributions, respectively, during $10s$ for the both. The ACB mechanism will be applied only on the

MTC devices, while the UEs directly pass to the RACH procedure. Regarding ACB parameters, $acb\_barringTime$ will be fixed to the value $2s$ for the proposed method and for the methods with which the comparison will be done, while $acb\_barringFactor$ is set to 0.9 only for the ordinary method, i.e. when applying the ACB mechanism with fixed parameters. Additionally, $acb\_barringFactor$ will be dynamically adjusted according to the number of arrivals for the best acknowledgment case (i.e. the number of arrivals in each RA slot is well known) and for the proposed algorithm MCSA-OE. The best acknowledgment case is used as a benchmark in the comparison. The reason to choose the aforementioned parameters is to show how MCSA-OE behaves under such a worse case, wherein the number of cumulative arrivals (new and retransmission) for the ACB mechanism, with the considered values, reaches up to 300 arrivals.

4.5.2 Network simulation tool

The simulation in this thesis has been conducted by C++-based discrete-event network simulator. More specifically, this tool has been implemented to simulate the RACH procedure. Regarding the control-plane latency analysis of the RACH procedure, it is determined as in Table B.1.1.1-1 in [3] and summarized in the appendix B.1. This network simulator is a proprietary one, and it was tested on the group paging case, by comparing the results obtained by our tool with the ones presented by 3GPP[82].

4.5.3 Simulation Results

Regarding the parameters used in the simulation, they are summarized in Table 4.1. However, the evaluation of the algorithm will be done on two parts. The first part is for showing the performance of MCSA-OE when the $acb\_BarringFactor$ is fixed, i.e. MCSA-OE is applied only to estimate the number of arrivals. In the second part, MCSA-OE is used to dynamically adjust the $acb\_BarringFactor$ based on the estimated number of arrivals.

Performance with fixed ACB parameters:

In this case, a fixed value for $acb\_BarringFactor$ is used to see how the algorithm behaves. Fig. 4.3 illustrates the number of devices in each RA slot. It is clear that the proposed algorithm MCSA-OE well works until certain value and then the estimated number of arrivals will fluctuate around this value, which represents the limitation of the algorithm. To explain this, we recall that $\Delta P_I = 1/R$. Thus, when the estimated

Table 4.1 simulation parameters

Parameter	setting
Traffic model for MTC devices	Beta distribution (α=3, β=4)
Traffic model for UEs	Poisson distribution
Average number of MTC devices	30000
Average number of UEs	1000
Distribution period	10 s
Cell Bandwidth	5MHz
PRACH config index	6
The total number of preambles in a random access slot	54
Maximum number of RARs that can be carried in one response message	3
Size of random access response window in sub-frame unit	5
mac-ContentionResolutionTimer	48
Backoff Inicator	20 ms
HARQ retransmission probability for Msg3 and Msg4 (non-Adaptive HARQ)	10%
Maximum number of HARQ transmission for Msg3 and Msg4 (non-Adaptive HARQ)	5
acb_BarringFactor	0.9
Acb_BarringTime	2 s

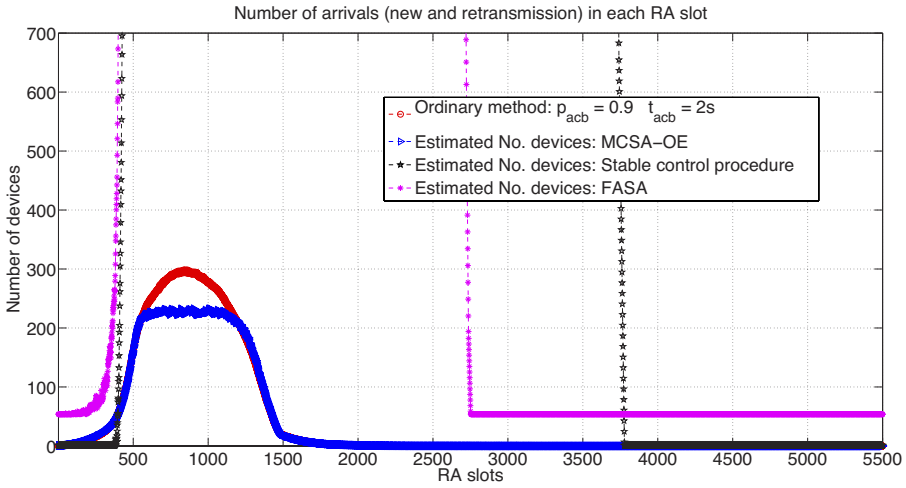

Figure 4.3 Number of devices in each RA slot with static ACB

idle probability $\hat{P}_I(t)$ reaches ΔP_I, the estimated number of devices reaches the limited value, which is $\hat{M}(t) = R\ln(1/\Delta P_I) = R\ln(R)$ or $\hat{M}(t) = \ln(R^R)$. For the considered parameters, we have $\hat{M}(t) = \ln(54^{54}) \simeq 215$, and hence $\hat{M}(t)$ fluctuates around this value. Regarding the stable control procedure, it is demonstrated, again, that this

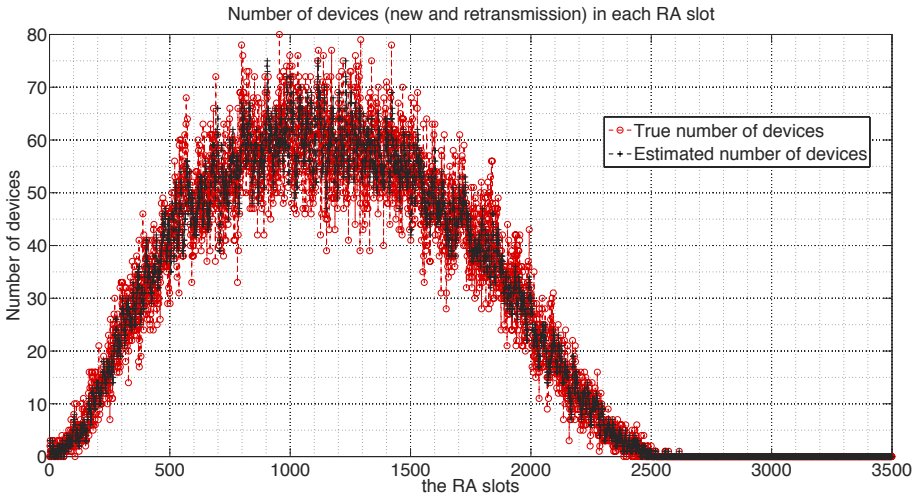

Figure 4.4 Performance of MCSA-OE for one experiment

method can not track the number of arrivals under heavy traffic load. Fig. 4.3 also proves that FASA can not be directly generalized to multi-channel slotted ALOHA case.

Performance with dynamic ACB:

In this case, the algorithm will be evaluated with dynamic ACB, i.e. *acb_Barring Factor* is adjusted according to the estimated number of arrivals and sent to the devices at the end of each RA slot. Fig. 4.4 shows the true and the estimated number of devices in each RA slot for just one experiment. We clearly see that MCSA-OE well tracks the number of devices. The performance of MCSA-OE compared to the ordinary method (ACB mechanism with fixed parameters) and the best acknowledgment case is depicted in Fig. 4.5. We remark from this figure that the behavior of MCSA-OE merely tends to that of the best acknowledgment. This is clearly observed in Fig. 4.6, where MCSA-OE mostly reaches the maximum capacity. It is worth noting that the maximum capacity can be reached when the number of arrivals $M(t)$ is equal to the number of channels R, and thus the success probability is equal to $P_S(t) = \left(\frac{M(t)}{R} \times e^{-\frac{M(t)}{R}}\right) = e^{-1} \simeq 0.37$. The success probability and the average access delay for the considered methods are presented in Table 4.2. We remark that the performance of MCSA-OE is similar to that of the best acknowledgment case with a slightly augmentation in the access delay for MTC devices, but with a little difference regarding the UEs. These results validate the ones shown in Fig. 4.6, wherein the behavior of the proposed algorithm MCSA-OE is similar to the best acknowledgment case.

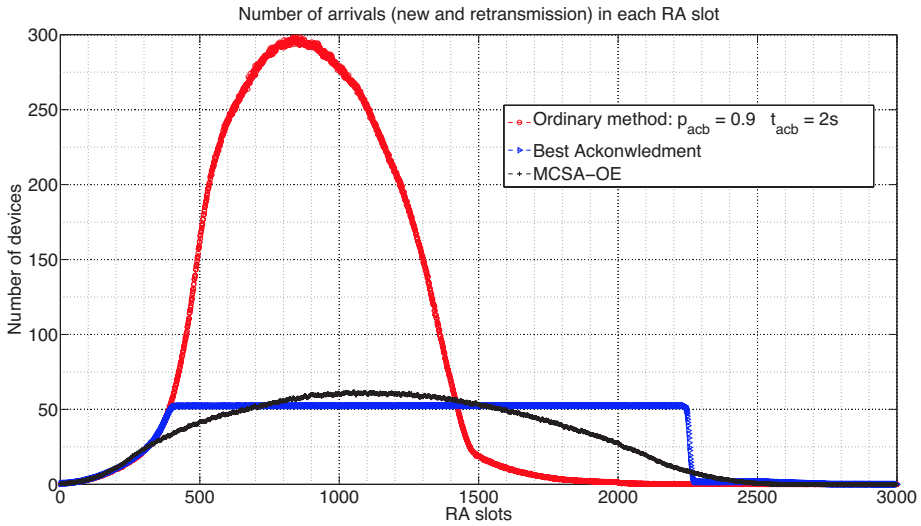

Figure 4.5 Number of devices in each RA slot with dynamic ACB

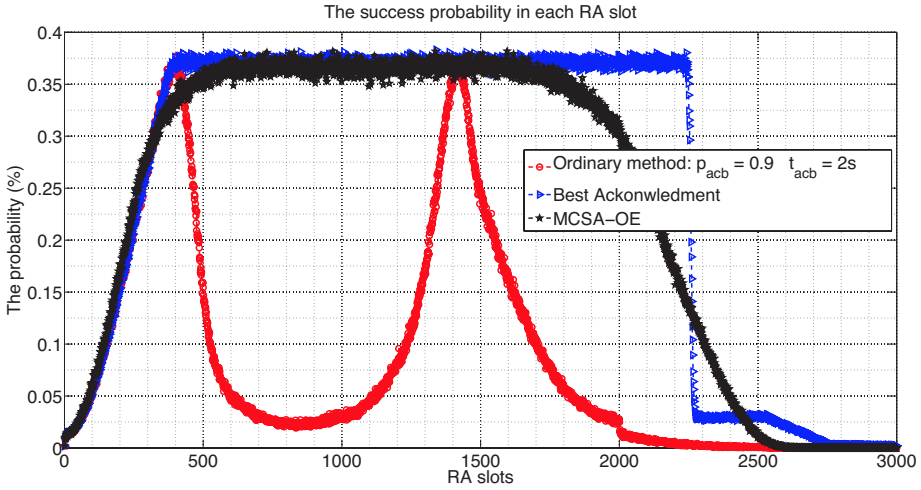

Figure 4.6 Success probability in each RA slot with dynamic ACB

Table 4.2 Success probability and average access delay for the considered methods

	Success Probability(%)		Average Delay(ms)	
	MTC	UE	MTC	UE
Ordinary ACB	33.88	61.74	426.7	43.9
Best Acknowledgment	95.17	99.05	1844.91	52.66
MCSA-OE	94.72	98.43	1878.24	53.46

4.6 Conclusions

One of the most efficient ways to control the congestion and system overload is knowing, or at least estimating, the number of arrivals at each time. By knowing the number of arrivals, the network's parameters can be easily adjusted so that the network capacity will be fully exploited. In this chapter, a state of the art on the concerned issue is introduced, and then followed by the proposed method, namely Multi-Channel Slotted ALOHA-Optimal Estimation (MCSA-OE). This method is presented to estimate the number of arrivals in each RA slot, and thus to better control the congestion in the network. Our proposition has been tested on two parts ; estimating the number of arrivals and adjusting the $acb\_BarringFactor$ of the ACB mechanism according to the estimated number of arrivals. Simulation results of the first part have proved that MCSA-OE well tracks the number of arrivals as long as they are smaller than or equal to $\ln(R^R)$. By dynamically adjusting the $acb\_BarringFactor$ of the ACB mechanism based on the estimated number of arrivals, MCSA-OE behavior tends to that of the best acknowledgment case, as proved by simulation. Therefore, MCSA-OE achieves of best case regarding the success probability and the average access delay with high traffic load. Furthermore, MCSA-OE does not depend on the traffic model, and thus it could well work with any traffic model.

Chapter 5

RACH Procedure: a General Model

Congestion and system overload is an inherent problem of a random access system, such as slotted ALOHA. Slotted ALOHA system started in the early 1970s, where it was initially used for resolving the contention of one common channel. After that, it was extended to multichannel slotted ALOHA system, where a terminal can randomly choose a channel among the available ones. In order to evaluate the system, Poisson traffic has been widely used. Although this traffic is a valid one for simulating the behavior of human being, it is not valid for M2M applications, where the behavior is different. Aiming at evaluating the performance of the network under such type of devices, 3GPP proposed two types of traffic; Uniform distribution as a realistic scenario and Beta distribution as synchronized one, i.e. the case when massive MTC devices try to attach the network.

However, any solution addressing the congestion problem raised by MTC should be tested with Beta-based traffic model. The reason to choose Beta traffic is that it is considered as a highly synchronized traffic, and thus it can be considered as the worst case. The aim of this chapter is to introduce a general scheme that models the RACH procedure. This model, namely General Recursive Estimation (GRE) [19, 18], is traffic-independent and general one. The proposed model helps in evaluating the congestion control methods targeting the RAN part.

5.1 State of the Art

Driven by the success of M2M-based applications (such as intelligent transport services), one study foresees that there would be roughly 50 billion of MTC devices by

2020 [11], while others expect to be 1000 Internet-connected devices for each person by 2040 [12]. However, the number of MTC devices might exceed the expectations, especially after the introduction of the new type of wireless communication using the visible light, namely Visible Light Communication (VLC) [103–106], which may open new horizons for new services. Although the traffic generated per MTC device is somewhat little, the aggregated traffic of tens of thousands of MTC devices (in each cell) will put very high pressure not only on the RAN part, but also on the CN part. The main problem in the RAN part comes from the fact that the current cellular mobile networks are designed and optimized for H2H communication, featured by their relative low numbers. Therefore, when a large number of MTC devices tries to get access to the network, via the Random Access Channel (RACH) procedure, the congestion at the RACH stage would surely occur. Therefore, the RACH performance may become a bottleneck of the wireless access network. The higher the congestion of the network is, the lower the number of UEs to get access is, and the lower the resource utilization is.

In order to evaluate the network performance under different access intensities and to show the effectiveness of the congestion control methods for MTC applications, one should first define a good traffic model that characterizes the behavior of this type of devices. 3GPP has identified two traffic models: Uniform Distribution (over 60 s) as a realistic scenario (non-synchronized traffic), and Beta Distribution (over 10 s) as an extreme scenario (synchronized traffic) [45]. In the literature, there are many works trying to model the traffic of MTC applications. The authors in [107] provide a real traffic of MTC applications, where the traffic tends to be periodic during the days of the week. In [108], the authors proposed Coupled Markov Modulated Poisson Processes (CMMPP) framework to model the traffic of MTC, which reflects in a more accurate way the behavior of devices. However, this model is more complex and traffic-dependent, as the parameters of the proposed scheme have to be established for each real traffic MTC application [109].

Regarding the modeling of RACH procedure, there are many works in the literature, such as [110–115, 84, 116, 76]. Some of them tried to estimate the number of arrivals at each time, and then to adjust the network's parameters, without introducing a model for the RACH procedure [76, 99]. Others proposed an analytical model for RACH procedure, but either they did not take into considerations the network's constraints, i.e. the number of responses in the RACH procedure is limited, [116], or they proposed a proprietary solution for certain traffic model [84, 115, 112]. However, it is stated in [115] that the errors could reach up to 200%. The authors in [110] proposed an

analytical model for the separation of preambles between M2M and H2H, while in [111] the authors introduced an analysis of slotted access scheme. In [112], The authors proposed to dynamically allocate resources to accommodate the number of arrivals in the case of Group Paging. The focus of the authors in [113] is on the design of new RACH procedure for MTC communication. However, by looking at the steps of the proposed scheme, it can be found that it is a combination of RACH with ACB. The authors in [114] introduced an analytical model of access reservation for LTE. However, it will be shown in the next section that this model does not accurately capture the behavior of RACH procedure. The authors in [117] propose to model the network with MTC application by Beta/M/1 queue model, where the focus is on the overall system performances. Moreover, the authors assume that the service time (the RACH procedure) is following an exponential distribution, which is not realistic and limits the accuracy of the model. Besides, the authors in [117] give just a high level value of the delay and drop, while our model allows to estimate MTC-related metrics, such as the success probability and access delay. Another limitation of this work is the fact that the number of transmission attempts has no limit (no queue limit), which is not realistic as each number of transmission attempt for a packet is limited and after reaching this limit the packet is dropped. In our model, we take into consideration these limitations. However, none of the above mentioned works have introduced a clear analytical model for RACH procedure. Therefore, a general analytical model for modeling the performance of RACH procedure will be proposed, and it will be applied to MTC traffic with Beta Distribution, as an example. The choice of MTC with Beta traffic is motivated by the high load resulting in the network, causing high congestion and system overload. Thus, activating MTC devices based on Beta distribution represents the worst case for the RACH procedure. The aim of the proposed scheme is to help validating new solutions that address the problem of congestion when activating the MTC devices according to Beta distribution, or other traffic models. The advantage of the proposed analytical model is simple, accurate (by report to the above-cited models), and general one, i.e. it can be applied with any other traffic model. Regarding the consistency with the congestion control methods, our model can be directly applied with all the methods that control the number of new arrivals, such as Access Class Barring (ACB), slotted access, and Extended Access Barring (EAB). Therefore, our model is aligned with the LTE/MTC standard, and it does not require any changes to the assumption already taken by the 3GPP on MTC traffic.

5.2 The Proposed Model: General Recursive Estimation (GRE)

5.2.1 System Model

In the present study, it is assumed that the MTC devices are activated according to Beta distribution [45], i.e. there will be just one type of traffic. All the MTC devices will be activated during certain interval I_β. It is also assumed that all the MTC devices fall within the coverage of just one eNB. Regarding the channel resources, the eNB reserves R random access preambles. Generally, for each preamble transmission the MTC device could take up to $\lceil (T_{RAR}+W_{RAR}+W_{BO})/T_{RA\_REP} \rceil T_{RA\_REP}$ sub-frames before retrying the transmission of preamble, where T_{RAR}, W_{RAR}, W_{BO}, and T_{RA_REP} are waiting time before the start of Random Access Response (RAR) window, size of RAR window, size of backoff window, and interval between two consecutive RA slots, respectively, as illustrated in Fig. 6.6. Therefore, the number of RA slots required in our study, i.e. for Beta Distribution, will be equal to:

$$I_{ra} = \left\lceil \frac{I_\beta}{T_{RA\_REP}} \right\rceil + (N_{PT_{max}} - 1) \left\lceil \frac{\Psi}{T_{RA\_REP}} \right\rceil \tag{5.1}$$

where I_β is the interval of Beta Distribution, in a sub-frame unit, $N_{PT_{max}}$ is the maximum number of preamble transmission. Note that this interval can be adapted to any other traffic by changing the value I_β. In the current study, the MTC devices will be considered as successful ones if they receive the message Msg2 of the RACH procedure, while ignoring the effect of messages Msg3 and Msg4 [84]. The reason of this assumption is the following. The probability of preamble retransmission due to the messages Msg3 and Msg4 is equal to [84]:

$$P_{ret} = p_f^{N_{HARQ}} + \sum_{j=0}^{N_{HARQ}-1} p_f^j (1-p_f) p_f^{N_{HARQ}} \tag{5.2}$$

where p_f is the retransmission probability of Msg3 and Msg4, and N_{HARQ} is the maximum number of Hybrid automatic repeat request (HARQ) transmissions of Msg3 and Msg4. Assuming that $p_f = 0.1$ and $N_{HARQ} = 5$ [45], this probability is equal to $P_{ret} = 2 \times 10^{-5}$, which is a negligible one. If there are 30000 MTC devices, for example, less than one MTC will retransmit the preamble.

5.2.2 Analytical Model

In order to calculate the success, idle, and collision probabilities of the preambles transmission, balls and bins problem is used [118]. Let R and M be the number of bins (or preambles) and balls (or MTC devices), respectively. Let p be the probability that a ball falls in a bin, which is equal to $(1/R)$ as all the bins are considered to have the same probability. Generally, the probability that k balls fall in a bin, noted w, is equal to:

$$Pr(N_w = k) = \binom{M}{k} (p)^k (1-p)^{M-k} \tag{5.3}$$

where $\binom{m}{k}$ is k-combinations and it is equal to:

$$\binom{m}{k} = \frac{m!}{k!\,(m-k)!}$$

The probabilities that (none of the balls fall/one ball falls) in the bin w represent the idle and success probabilities, respectively, and they are equal to;

$$P_S(i) = \binom{M_i}{1} \left(\frac{1}{R}\right)^1 \left(1-\frac{1}{R}\right)^{M_i-1} \approx \frac{M_i}{R} e^{-\frac{M_i}{R}}$$

$$P_I(i) = \binom{M_i}{0} \left(\frac{1}{R}\right)^0 \left(1-\frac{1}{R}\right)^{M_i} = \left(1-\frac{1}{R}\right)^{M_i} \approx e^{-\frac{M_i}{R}}$$

M_i is the number of MTC devices at the time (i), or the RA slot (i) for the considered problem. Moreover, the collision probability is the complement of the success and idle probabilities;

$$P_C(i) = 1 - P_I(i) - P_S(i)$$

where this equation is not well positioned in [114]. The number of successful MTC devices, which is equal to the number of preambles/bins chosen by only one MTC/ball, is equal to;

$$M_S(i) = R \times P_S(i) = M_i e^{-\frac{M_i}{R}} \tag{5.4}$$

However, the total number of MTC devices M_i includes the ones whose preambles are transmitted for the first, second, ..., and the $N_{PT_{max}}$-th time, and thus the precedent equation becomes;

$$M_S(i) = \sum_{n=1}^{N_{PT_{max}}} M_i[n] e^{-\frac{M_i}{R}} \tag{5.5}$$

As the probability to detect the n^{th} preamble transmission by the eNB is equal to $(p_n = 1 - e^{-n})$ rather than (1) [45], the precedent equation will be written by:

$$M_S(i) = \sum_{n=1}^{N_{PT_{max}}} M_{S,n}(i) = \sum_{n=1}^{N_{PT_{max}}} M_i[n] p_n e^{-\frac{M_i}{R}} \tag{5.6}$$

One of the main constraints imposed by the network consists in the fact that no more than N_{ACK} responses can be sent, after the preamble transmission, even if the number of successful preambles is more than N_{ACK}. Note that $N_{ACK} = N_{RAR} \times W_{RAR}$, where N_{RAR} is the number of responses per a RAR message. Therefore, the number of successful MTC devices for the nth preamble transmission is:

$$M_{S,n}(i) = \begin{cases} M_i[n] p_n e^{-\frac{M_i}{R}} & ;\text{if } \sum_{n=1}^{N_{PT_{max}}} M_i[n] p_n e^{-\frac{M_i}{R}} \leq N_{ACK} \\ \dfrac{M_i[n] p_n e^{-\frac{M_i}{R}}}{\sum_{n=1}^{N_{PT_{max}}} M_i[n] p_n e^{-\frac{M_i}{R}}} N_{ACK} & ;\text{otherwise} \end{cases} \tag{5.7}$$

Based on the analysis in section 6.3.2, we have:

$$\begin{cases} x_a(i) = i + \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rceil = i + \left\lceil \frac{\Gamma}{T_{RA\_REP}} \right\rceil \\ x_{bc}(i) = i + \left\lceil \frac{\Gamma}{T_{RA\_REP}} \right\rceil + k \\ x_d(i) = i + \left\lfloor \frac{T_{RAR} + W_{RAR} + W_{BO}}{T_{RA\_REP}} \right\rfloor + 1 = i + \left\lfloor \frac{\Psi}{T_{RA\_REP}} \right\rfloor + 1 \end{cases} \tag{5.8}$$

where $x_a(i)$, $x_{bc}(i)$, and $x_d(i)$ are the order of the RA slots (a), (bc), and (d), respectively, within the backoff interval W_{BO} relative to the preamble transmission at the RA slot (i), as illustrated in Fig. 6.6, $k = 1, 2, ..., K_{max}$, whereas K_{max} is equal to;

$$K_{max} = \left\lfloor \frac{W_{BO} - \alpha_a W_{BO}}{T_{RA\_REP}} \right\rfloor$$

Alternatively, the proportions of the collided MTC devices whose backoff timers expire and retransmit their preambles at the RA slots (a), (bc), and (d) are equal to:

$$\begin{cases} \alpha_a = \dfrac{\lceil \Gamma / T_{RA\_REP} \rceil T_{RA\_REP} - \Gamma}{W_{BO}} \\ \alpha_{bc} = \dfrac{T_{RA\_REP}}{W_{BO}} \\ \alpha_d = \dfrac{\Psi - T_{RA\_REP} \lfloor \Psi / T_{RA\_REP} \rfloor}{W_{BO}} \end{cases} \tag{5.9}$$

From the equations (5.8) and (5.9) and Fig. 6.6, we can conclude that the number of MTC devices retransmitting their preambles for the n^{th} time is equal to:

$$M_i[n] = \sum_{j=i-k_2}^{i-k_1} \alpha_j M_{C,n-1}(j) \qquad ;\text{for } n = 2 : N_{PT_{max}} \tag{5.10}$$

where $M_{C,k}(j)$ is the number of collided MTC devices corresponding to the preamble transmission at the RA slot (j) for the k^{th} time, α_j can be α_a, α_{bc}, or α_d, while k_1 and k_2 are equal to;

$$k_1 = \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rceil \tag{5.11}$$

$$k_2 = \left\lceil \frac{T_{RAR} + W_{RAR} + W_{BO}}{T_{RA\_REP}} \right\rceil + 1 \tag{5.12}$$

Note that k_1 and k_2 are directly obtained from $x_a(i)$ and $x_d(i)$, respectively. Regarding the equation (5.10), it can be written as follows:

$$M_i[n] = \alpha_a M_{C,n-1}(i-k_1) + \alpha_d M_{C,n-1}(i-k_2) +$$

$$\sum_{k=i-k_2+1}^{i-k_1-1} \alpha_{bc} M_{C,n-1}(k) \qquad ;\text{for } n = 2 : N_{PT_{max}} \tag{5.13}$$

For $(n = 1)$, the number of MTC devices, i.e. $M_i[1]$, will be the value determined by Beta distribution (for our study) or by any traffic model.

5.2.3 Beta Distribution

Let M be the total number of MTC devices in the cell. By assuming that all the MTC devices will be activated, according to Beta distribution, between $(t = 0)$ and $(t = T)$, the expected number of arrivals in the random access opportunity (i) is given by the following equation:

$$M_i[1] = M \int_{t_i-1}^{t_i} p(t)dt \tag{5.14}$$

where t_i is the time of the RA opportunity (i), and the distribution $p(t)$ follows Beta distribution:

$$p(t) = \frac{t^{\alpha-1}(T-t)^{\beta-1}}{T^{\alpha+\beta-1} Beta(\alpha,\beta)} \; ; \; \alpha > 0, \beta > 0 \tag{5.15}$$

where $Beta(\alpha,\beta)$ is Beta function, and it is given by:

$$Beta(\alpha,\beta) = \frac{(\alpha-1)!(\beta-1)!}{(\beta+\alpha-1)!} \tag{5.16}$$

It should be noted that $\int_0^T p(t)dt = 1$, and the values α and β are set to be 3 and 4, respectively, for MTC Beta traffic [45]. In order to find the expected number of arrivals at each RA slot, the authors in [108] propose to use modulated Poisson process to find the number of new arrivals at each time. However, a simpler way, which is based on approximating the integration in the equation (5.14) through the trapezoidal rule [119], will be used in the proposed method. The trapezoidal rule is given by the following equation;

$$\int_a^b f(x)dx = (b-a)\left[\frac{f(a)+f(b)}{2}\right]$$

Therefore, the equation (5.14) can be written by:

$$M_i[1] = M(t_{i+1}-t_i)\frac{p(t_i)+p(t_{i+1})}{2} \tag{5.17}$$

As the interval between two consecutive RA slots is equal to T_{RA_REP}, we set $t_{i+1}-t_i = T_{RA\_REP}$, and therefore:

$$M_i[1] = \frac{M \times T_{RA\_REP}}{2T^{\alpha+\beta-1}Beta(\alpha,\beta)}\left[t_i^{\alpha-1}(T-t_i)^{(\beta-1)} + t_{i+1}^{\alpha-1}(T-t_{i+1})^{(\beta-1)}\right] \tag{5.18}$$

Equation (5.18) represents the number of new arrivals according to Beta distribution.

After determining the number of MTC devices for each preamble transmission at the RA slot (i) (equations 5.13 and 5.18), the number of successful MTC devices can be calculated by the equation (5.7), where the number of collided MTC devices is:

$$M_{C,n}(i) = M_i[n] - M_{S,n}(i)$$

It should be noted that the proposed analytical model GRE can be applied for another traffic models, where the only change that should be made in the model is the number of new arrivals, i.e. $M_i[1]$, while the rest of the model remains unchanged.

5.3 Performance Evaluation

The proposed model has been implemented using C++ network simulator (see section 4.5.2). The parameters of RACH procedure are taken as specified by Table 6.2.2.1.1 in [45], and they are detailed in Table 5.1. The simulations were developed based on Monte-Carlo approach, where 350 experiments have been used to average the results.

Notations	Definition	Values
α, β	The parameters of Beta Distribution	3, 4
I_β	The interval of Beta Distribution	$10 * 1000$
M	Average number of MTC devices in the cell	30000
R	Total number of preambles in a random access slot	54
BI	Backoff indicator in a sub-frame unit	20
$N_{PT_{max}}$	Maximum number of preamble transmission	10
N_{RAR}	Maximum number of RARs that can be carried in one response message	3
T_{RAR}	Processing delay required by the eNB in order to detect the transmitted preamble in a sub-frame unit	2
W_{RAR}	The size of the random access response window in a sub-frame unit	5
N_{ACK}	Maximum number of MTC devices that can be acknowledged within the RAR window	$N_{ACK} = N_{RAR} \times W_{RAR}$
$PRACH_{config\_indx}$	PRACH configuration index	$PRACH_{confic\_indx} = 6$
T_{RA_REP}	The interval between two consecutive Random Access (RA) slots	5
W_{BO}	Backoff window size	$BI + 1$
p_n	Preamble detection probability for the n-th preamble transmission	$p_n = 1 - e^{-n}$
T_{CRT}	Contention Resolution timer	48
p_{HARQ_RET}	HARQ retransmission probability for Msg3 and Msg4 (non-adaptive HARQ)	10%
N_{HARQ}	Maximum number of HARQ TX for Msg3 and Msg4 (non-adaptive HARQ)	5

Table 5.1 Basic simulation parameters

5.3.1 Performance Metrics

In order to show the performance of our proposed model GRE, the following metrics will be considered: i) the total number of MTC devices at each RA slot, ii) the number of successful MTC devices at each RA slot, iii) the collision probability, iv) the success probability, v) the average number of preamble transmission, vi) the average access delay, v) and the Cumulative Distribution Function (CDF) of preamble transmission. The total number of MTC devices is given by $M_i = \sum_{n=1}^{N_{PT\_max}} M_i[n]$, while the number of successful MTC devices is given by $M_S(i) = \sum_{n=1}^{N_{PT\_max}} M_{S,n}(i)$. Regarding the collision probability, it can be defined as the number of collided RAOs, during the

interval I_{ar}, to the total number of reserved RAOs $I_{ar} \times R$, and it is given by the following equation:

$$P_C = \frac{\sum_{i=1}^{I_{ra}} (R - M_S(i) - Re^{-\frac{M_i}{R}})}{R \times I_{ra}} \tag{5.19}$$

The success probability is equal to the number of successful MTC devices within the maximum number of preamble transmissions to the total number of MTC devices, and it is given by the following equation:

$$P_S = \frac{\sum_{i=1}^{I_{ra}} M_S(i)}{M} \tag{5.20}$$

Regarding the average number of preamble transmissions, it is equal to the total number of preamble transmissions for all the MTC devices successfully accessed the network divided by the total number of successful MTC devices, and it is given by:

$$PRM_{avg} = \frac{\sum_{i=1}^{I_{ra}} \sum_{n=1}^{N_{PT_{max}}} n M_{S,n}(i)}{\sum_{i=1}^{I_{ra}} \sum_{n=1}^{N_{PT_{max}}} M_{S,n}(i)} \tag{5.21}$$

Concerning the average access delay T, it is the total access delay for all the MTC devices successfully finished the RACH procedure divided by the number of success MTC devices, and it is given by the following equation;

$$T = \frac{\sum_{i=1}^{I_{ra}} \sum_{n=1}^{N_{PT_{max}}} D_n M_{S,n}(i)}{\sum_{i=1}^{I_{ra}} \sum_{n=1}^{N_{PT_{max}}} M_{S,n}(i)} \tag{5.22}$$

where D_n is given by;

$$D_n = \frac{T_{RA\_REP}}{2} + (n-1)\left\lceil \frac{T_{RAR} + W_{RAR} + W_{BO}/2}{T_{RA\_REP}} \right\rceil T_{RA\_REP} + T_{RAR} + \frac{W_{RAR}}{2}$$

Note that the first term $T_{RA\_REP}/2$ is the average waiting time for the next RA slot, while $W_{RAR}/2$ is the average waiting time for the RAR message. Let ω be the number of preamble transmissions to access the network for the MTC devices successfully finished the RACH procedure. The CDF of preamble transmission, noted by $CPT(\omega)$, is the ratio between the number of MTC devices whose number of preamble transmission is

less than or equal to (ω) and the total number of preamble transmission for all the MTC devices successfully accessed the network. *CPT(ω)* is given by the following equation:

$$CPT(\omega) = \frac{\sum_{i=1}^{I_{ra}} \sum_{n=1}^{\omega} M_{S,n}(i)}{\sum_{i=1}^{I_{ra}} \sum_{n=1}^{N_{PTmax}} M_{S,n}(i)} \tag{5.23}$$

5.3.2 Results

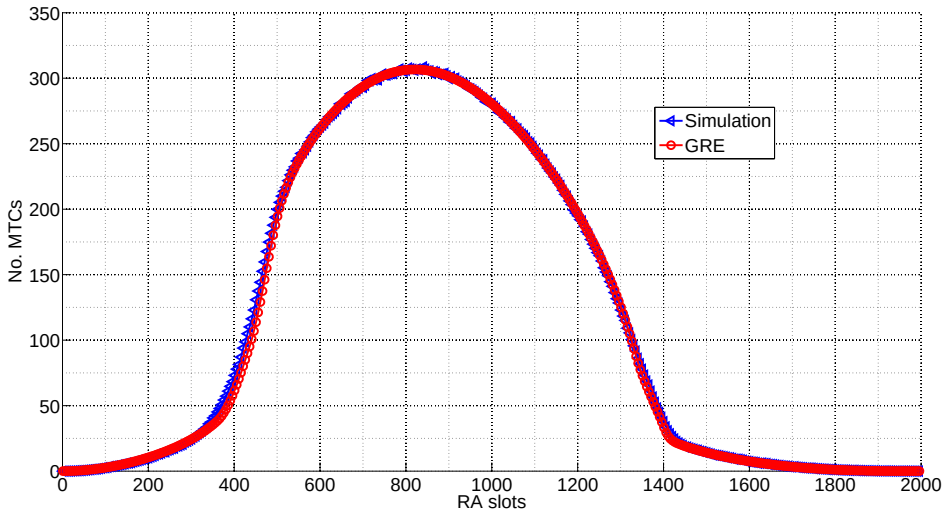

Figure 5.1 Total number of MTC devices in each RA slot

Fig. 5.1 shows the total number of MTC devices at each RA slot. From this figure, it is clear that the proposed analytical model GRE gives an accurate approximation of Beta distribution. However, there is a small difference between the simulation and the analytical model when the total number of MTC devices is of the order of the number of preambles. This difference is clearer in Figs. 5.2 and 5.3, which represent the number of successful MTC devices at each RA slot for 15000 and 30000 MTC devices, respectively. Generally, the accuracy of GRE is achieved except the regions where the total number of MTC devices is of the order of the number of preambles. This difference comes from the fact that the network cannot send back responses to more than N_{ACK} MTC devices even if the number of successful preambles is more than N_{ACK}. Therefore, when a lot of trails is taken, the mean value will never reach the value N_{ACK} as it is the maximum allowed one, while in the analytical model the value N_{ACK} is used directly when the number of success preambles is more than N_{ACK}. This case is clear in Figs. 5.2 and 5.3, where the upper part of the analytical

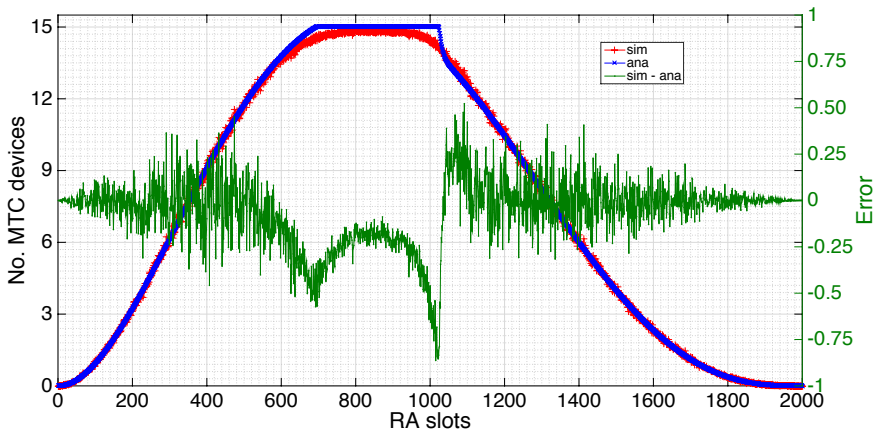

Figure 5.2 Number of successful MTC devices in each RA slot: M = 15000

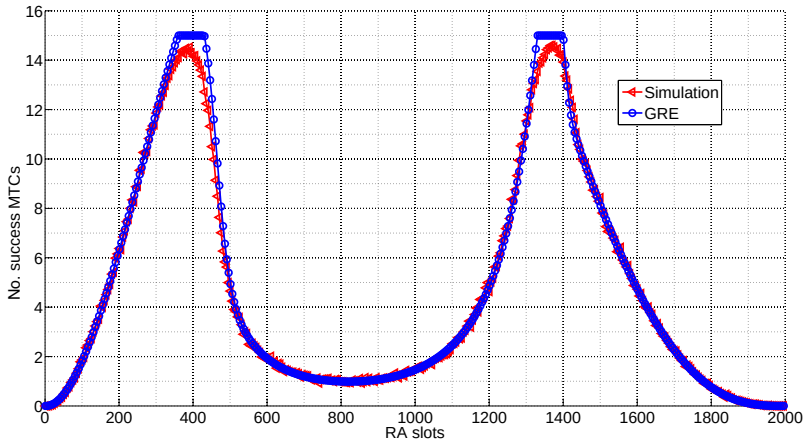

Figure 5.3 Number of successful MTC devices in each RA slot: M = 30000

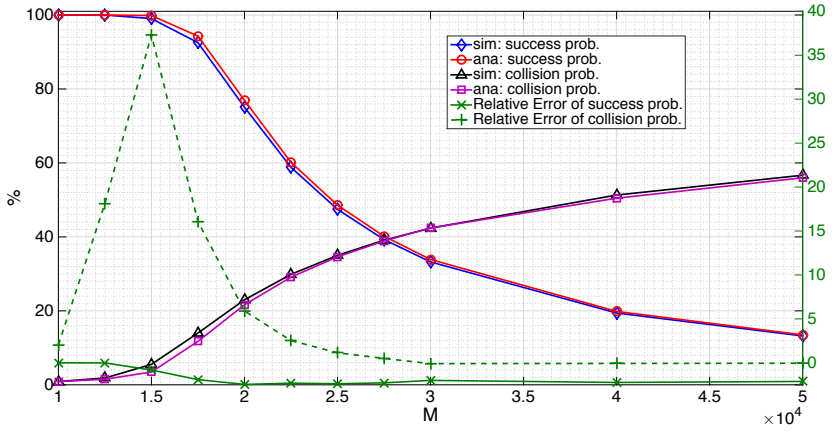

Figure 5.4 Success and collision probabilities

model is shown as if it is cut. In spite of this difference between the simulation and the analytical model, the results in Figs. 5.4, 5.5, and 5.6 indicate that the proposed

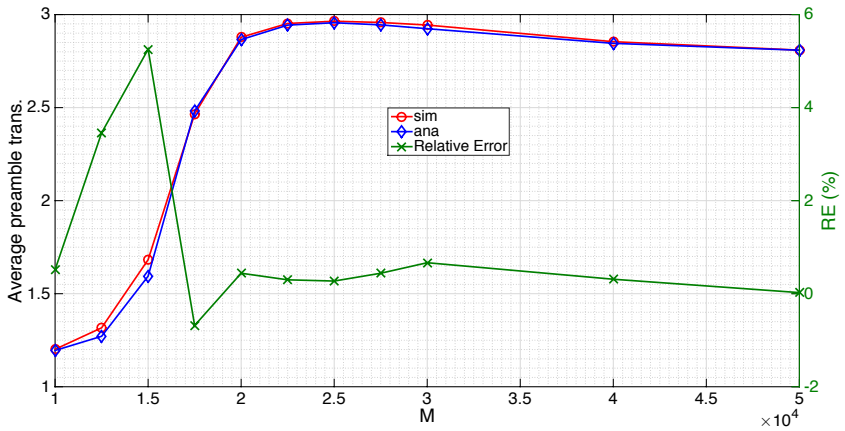

Figure 5.5 Average number of preamble transmission

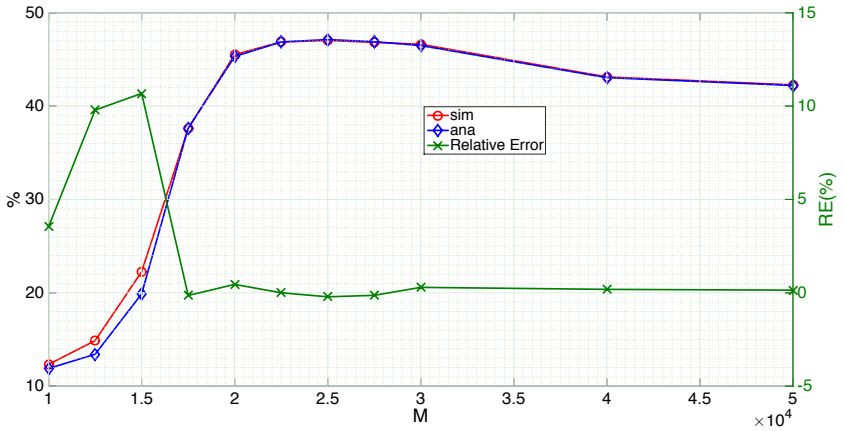

Figure 5.6 Average access delay

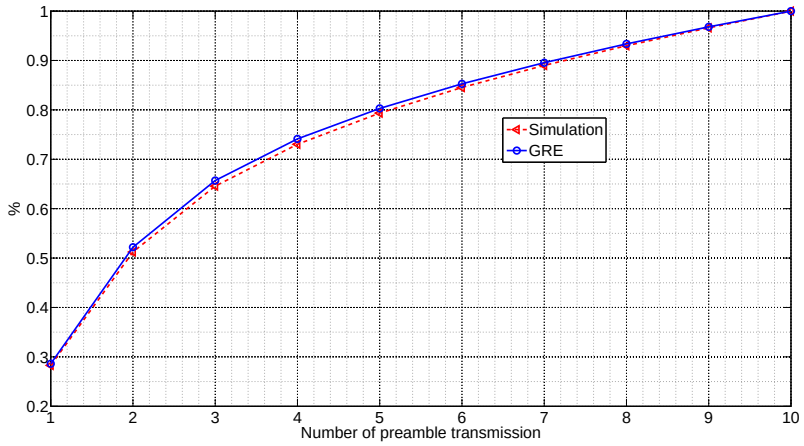

Figure 5.7 CDF of preamble transmission: M = 30000

analytical model has an accurate approximation regarding the success and collision probabilities. Fig. 5.4 illustrates the success and collision probabilities, and also the

Relative Errors (RE) which is $RE = 100 \times (sim - ana)/sim$, where sim and ana are simulations and analytical results, respectively. From this figure, it can be clearly seen that the analytical model, generally, gives a good approximation with low RE. However, this RE becomes pretty large for certain values of the total number of MTC devices, where it reaches its maximum for $M = 15000$. Figs. 5.5 and 5.6 show the average number of preamble transmissions and the average access delay, respectively. From these figures, we observe that our analytical model is very accurate when the number of MTC devices is relatively large (more than 15000 MTC devices), where the RE is less than one percent. The worst RE value in our model reaches 40% , which is far from the RE obtained in [115] (i.e., 200%). The effectiveness of our analytical model is further proved by Fig. 5.7 that illustrates the CDF of preamble transmission. It should be noted that the analytical model GRE is valid for any traffic model, where only the number of new arrivals at each RA slot, i.e. $M_i[1]$, will be adapted to the traffic model while the rest of the model will be unchanged.

5.4 Conclusions

Beta distribution is one of the traffic models proposed by 3GPP to represent the behavior of synchronized MTC applications. However, any solution addressing the congestion problem raised by MTC devices should be first tested with Beta-based traffic model. The reason is that Beta traffic can be considered as the worst case for the network. After introducing a state of the art about the work already exist in the literature on the modeling the RACH procedure, a general analytical model, namely General Recursive Estimation (GRE), is proposed and tested on Beta-activated traffic. The effectiveness of the proposed model (i.e., GRE) is proven through computer simulations, where many metrics have been considered, such as success and collision probabilities, the average number of preamble transmission, and the CDF of preamble transmission. Simulation results show the accuracy of the proposed model GRE when the number of devices is large (more then 15000 MTC devices), achieving a relative error less than one percent on the average number of preamble transmission and the average access delay and less than two percent on the success probability. Moreover, the proposed analytical model can be used to model not only Beta-based traffic model, but also any other traffic models, since the only required change concerns the number of new arrivals whereas the rest remains unchanged.

Chapter 6

Resources Management and Power Optimization

The current cellular mobiles networks are designed and optimized to deal with a relatively small number of terminals in each cell. At most, the expected number of UEs in each cell will be equal to the number of population in the concerned cell. However, the case is totally different for MTC applications, where it is expected that the number of M2M devices would hit 50 billion, i.e. nearly sevenfold. Adding this additional huge number of devices in the current cellular mobile networks would put a very high pressure not only on the RAN part, but also on the CN part, causing congestion and system overload. Paging method is one of the proposed methods to remedy the congestion problem. However, this method arises new problems, such as the long period to page a large number of MTC devices in just one cell. For example, paging 30000 MTC devices may take e.g. 11 seconds, which represents a relatively long period. This issue is overcome by introducing the Group Paging (GP) method [45]. In this method, a large number of MTC devices can be paged by only one paging message, where all the concerned MTC devices will be addressed by a unique ID, namely Group ID (GID).

In spite of the good performance of GP method, its performances dramatically decrease when increasing the number of MTC devices being paged. Therefore, GP suffers from many problems, such as low access success probability, high latency associated with high power consumption. To overcome its disadvantages, we devised two new methods. The first one, namely Controlled Distribution of Resource (CDR) [20], is proposed to improve the performance of GP method in case where the MTC devices are in the $RRC\_CONNECTED$ state but there is no uplink synchronization. The main idea of this method is to distribute the resources to the MTC devices based on

their IDs, i.e. Cell-Radio Network Temporary Identifier (C-RNTI). This method highly improves the performance of GP methods, as shown later in this chapter, where it is congestion-free, i.e. the success probability is 100%, in addition to other advantages like the reduction in the access delay.

The second method, named as Traffic Scattering For Group Paging (TSFGP) [21], tackle with the congestion problem regardless the state of the terminal. The main idea of TSFGP is to scatter the MTC devices being paged during the available interval, instead of leaving them to start the RACH procedure just after receiving the paging message. Compared to GP method, TSFGP highly improves the performances, where the success probability is highly increased and the average access delay and the average number of preambles are highly reduced. Moreover, an improvement of this method, dubbed as Further Improvement-TSFGP (FI-TSFGP) [22], is introduced. This method FI-TSFGP gives the flexibility to change the network's parameters without affecting the quality of the method, which is not the case for TSFGP. Compared to GP and other methods, FI-TSFGP highly outperforms these methods in terms of success and collision probabilities, average access delay, average number of preamble transmission, and ultimately energy efficiency.

6.1 State of the Art

As mentioned before, Paging and Group Paging (GP) are methods proposed by 3GPP in order to tackle with the congestion problem. Basically, GP method well behaves when the number of MTC devices being paged are relatively small. On the contrary, its behavior degrades when increasing the number of MTC devices being paged.

Many methods have been proposed in the literature in order to address the problem of congestion and system overload. Firstly, in the GP method the MTC devices will start the RACH procedure just after receiving a paging message. However, this method is not practical in the presence of a high number of MTC devices as it will take a long period to page all the devices, e.g. paging 36000 MTC devices will take 11.25s. Existing solution of this problem in the literature consists in organizing the MTC devices in groups, where each one is represented by a Group ID (GID). Therefore, the MTC devices in one group can be paged by only one paging message. Upon receiving a paging message that includes their GID, the MTC devices start the RACH procedure. Again, this group paging method still has many disadvantages, such as the low value of resources' utilization as the optimal value is about 20 % [84]. Another disadvantage is

the access success probability that rapidly decreases when the size of the group exceeds a certain threshold. For example, the success probability degrades to about 20 % and the collision probability to more than 70 % for certain network's configuration [84].

In the literature, many methods have been proposed to improve the performance of GP method. The authors in [115] and [84] proposed an analytical model for GP method, wherein they tried to capture the behavior of GP method. The authors in [120] propose to repeat the group paging interval, i.e. Consecutive Group Paging (CGP), so that the MTC devices not succeeded in the first GP interval will try to access the network in the subsequent GP interval(s). However, CGP performances are worse than the classical Group Paging [82] for certain configurations. In [121], and similar in spirit to the idea of [70], the authors proposed enforcing some backoff time on new transmission attempts before the first preamble transmission, i.e. Pre-BackOff (PBO). Results presented in [121] show the superiority of PBO method by report to the classical GP method.

6.2 Controlled Distribution of Resources (CDR) for MTC devices

6.2.1 System Model

In this study, it is assumed that each group contains M MTC devices in a paging area comprising N cells [82]. These devices are uniformly distributed over the cells, and thus each cell contains, on average, M/N MTC devices. It is assumed, in the proposed solution, that all the MTC devices are in the RRC_CONNECTED_OUT_OF_SYNC state (see section 3.4.3) and, therefore, all the MTC devices have C-RNTI. This means that the MTC devices have RRC context, but there is no synchronization between the devices and the network (i.e., there is no uplink transmission). When there is data to be sent or a request from the network to get information, the MTC device must first perform the random access procedure. Moreover, it is assumed that the M MTC devices are assigned by a Group ID (GID). The network reserves R dedicated random access resources and utilizes the GID in order to page the concerned devices to access the network simultaneously. Upon receiving a paging message containing the GID, the MTC device in the proposed solution shall start the random access procedure by sending certain preamble in a certain random access slot determined by the C-RNTI of the MTC device itself.

6.2.2 The Proposed Mechanism: CDR

Instead of leaving the MTC devices to randomly choose the resources, the available ones will be distributed as follows (for the simplicity, we assume that there is only one frequency band in each random access slot):

— Assign dedicated random access resources to the group during a certain interval, namely group access interval. In the CDR, this interval depends on the number of MTC devices and the number of reserved preambles. However, as mentioned in [84], this interval in the ordinary group paging depends on the maximum number of preamble transmission.

— Assign a pair of values (σ,ω) to each MTC device, where "σ" is the order of the random access slot during the group access interval and "ω" is the index of the preamble in this random access slot.

Thereby, the number of required random access slots in CDR method is obtained as follows:

$$N_{slots} = \left\lceil \frac{M/N}{R} \right\rceil \tag{6.1}$$

where, R is the number of reserved preambles and $\lceil x \rceil$ denotes the upper integer value of x.

Because of the limitation of the number of Random Access Responses (RARs) in each RAR message, the number of preambles reserved for the group in the CDR should be inferior or equal to the maximal number of MTC devices N_{ACK} that can be acknowledged within the Random Access Response window (W_{RAR}), where N_{ACK} is given by:

$$N_{ACK} = N_{RAR} \times W_{RAR} \tag{6.2}$$

and N_{RAR} is the maximum number of RARs that can be carried in a response message.

For instance, if we consider the following parameters: $W_{RAR} = 4$, $N_{RAR} = 1$, $R = 4$, and $M/N = 12$; we obtain $N_{ACK} = 4$. In this case, the number of required slots is $N_{slots} = 3$. Table 6.1 depicts the attribution of MTC devices in the reserved slots. From this table, we clearly remark that there is no any waste in the resources, which is an optimal situation. However, the worst case takes place when there is only one MTC device in the last slot (see Table 6.2). Generally, the relation between the number of MTC devices and the number of preambles can be expressed as follows;

$$M/N = h \times R + k \tag{6.3}$$

Note that h and k are integer values, where k can take the values 0 (for the optimal case), 1 (for the worst case), ..., and $(R-1)$.

RA Slot 1		RA Slot 2		RA Slot 3	
MTC devices	The pair of values (a, b)	MTC devices	The pair of values (a, b)	MTC devices	The pair of values (a, b)
MTC 1	(1, 1)	MTC 5	(2, 1)	MTC 9	(3, 1)
MTC 2	(1, 2)	MTC 6	(2, 2)	MTC 10	(3, 2)
MTC 3	(1, 3)	MTC 7	(2, 3)	MTC 11	(3, 3)
MTC 4	(1, 4)	MTC 8	(2, 4)	MTC 12	(3, 4)

Table 6.1 The attribution of MTC devices in the reserved slots in the optimal case

RA Slot 1		RA Slot 2		RA Slot 3	
MTC devices	The pair of values (a, b)	MTC devices	The pair of values (a, b)	MTC devices	The pair of values (a, b)
MTC 1	(1, 1)	MTC 5	(2, 1)	MTC 9	(3, 1)
MTC 2	(1, 2)	MTC 6	(2, 2)		
MTC 3	(1, 3)	MTC 7	(2, 3)		
MTC 4	(1, 4)	MTC 8	(2, 4)		

Table 6.2 The attribution of MTC devices in the reserved slots in the worst case

Before going in the details, it is worth to recall that the proposed solution CDR is dedicated to the case where the MTC devices are in the RRC_CONNECTED_OUT_OF_SYNC state and, thus, the MTC devices have, among other things, the Group ID (GID) and C-RNTI. Regarding the attribution of the values "σ" and "ω", one and trivial solution consists in directly assigning the pair of values once the MTC device joins the group. However, this requires additional signaling overhead to be exchanged with each MTC device, and it is also not dynamic in cases like the change of the number of preambles reserved for the group. Instead of the direct assignment, the CDR mechanism depends on the C-RNTI that can take values in the interval $(003D-FFF3)_{hex} = (0000\ 0000\ 0011\ 1101 - 1111\ 1111\ 1111\ 0011)_{bin}$ [54]. The method CDR proposes to allocate a subinterval of values to the group. As an example, we can reserve the values $V_I = (0000\ 1111\ \text{xxxx}\ \text{xxxx})$, which corresponds to the interval $(0000\ 1111\ 0000\ 0000 - 0000\ 1111\ 1111\ 1111)_{bin}$, where x is either "0" or "1". This interval can contain up to $2^8 = 256$ MTC devices, i.e. a particular group can comprise at most 256 MTC devices. If we take the parameters in the Table 6.3, the number of MTC devices ($M/N = 256$) and the number of reserved preambles ($R = N_{ACK} = 15$), then the number of required slots is:

$$N_{slots} = \left\lceil \frac{M/N}{R} \right\rceil = \left\lceil \frac{256}{15} \right\rceil = \left\lceil 17.0667 \right\rceil = 18$$

or

$$M/N = 17 \times 15 + 1$$

which represents the worst case. In the following, it will be assumed that there is a group of 256 MTC devices, and the devices having the $V_{129} = (0000\,1111\,1000\,0001)_{bin} = 129_{dec}$ will be used so as to illustrate the the proposed method CDR.

The pair of values (σ, ω) can be obtained directly from the C-RNTI of each MTC device by the following steps:

— Sending the mask MSK to all the MTC devices. This mask can be sent through the paging message used to inform the MTC devices to start the RACH procedure. In our example, the mask MSK is equal to $(0000\,1111\,0000\,0000)_{\text{bin}}$.

— Each MTC device performs the logical operation XOR between the mask and its C-RNTI as follows:

$$PID = MSK \oplus V_I \tag{6.4}$$

For the concerned device whose ID is equal to 129_{dec}, the PID will be;

$$PID = 0000\,1111\,0000\,0000 \oplus 0000\,1111\,1000\,0001 = 1000\,0001 \tag{6.5}$$

— The pair of values (σ, ω) can be obtained using the following equations:

- $\sigma = 1 + \lfloor PID/R \rfloor$, which represents the number of the random access slot and takes values in $[1, N_{slots}]$. Here, $\lfloor x \rfloor$ represents the lower integer value of x. For the device V_{129}, this value is $\sigma = 1 + \lfloor 129/15 \rfloor = 1 + \lfloor 8.6 \rfloor = 9$
- $\omega = 1 + (PID \; mod \; R)$, which denotes the identity of the preamble in the random access slot specified by "σ", and it takes values in $[1, R]$. The value ω for the device V_{129} is $\omega = 1 + (129 \; mod \; 15) = 10$.

For our example V_{129}, the pair of values (σ, ω) is equal to (9, 10). This means that the MTC device whose PID is equal to 129 will transmit the 10^{th} preamble in the 9^{th} slot.

At the network side, the eNB can determine the C-RNTI for each MTC device as follows:

— The eNB can know the pair of values(σ, ω) based on which preamble is used and in which random access slot the preamble was transmitted. Following a general form of notation, it is known that $PID/R = y + z/R$, where $y = \lfloor PID/R \rfloor$, which is an integer value, and $z = (PID \; mod \; R)$. By using the equations of "σ" and "ω", we can obtain that $y = (\sigma - 1)$ and $z = (\omega - 1)$. Hence, $PID = (\sigma - 1) \times R + (\omega - 1)$, which can be used at the network side to obtain the PID of the corresponding MTC device.

For the considered device V_{129}, we have;

$$PID = (\sigma - 1) \times R + (\omega - 1) = (9 - 1) \times 15 + (10 - 1) = 129$$

— In order to obtain the C-RNTI of the MTC devices, an XOR is performed between the mask and the PID, i.e. C_RNTI = PID $\oplus$ MSK. Therefore, the $C - RNTI$ for the considered device is equal to:

$$\begin{aligned} C\_RNTI = PID \oplus MSK &= (1000\ 0001)_{bin} \oplus (0000\ 1111\ 0000\ 0000)_{bin} \\ &= (0000\ 1111\ 1000\ 0001)_{bin} \end{aligned}$$

6.2.3 Performance Evaluation

The proposed solution CDR is implemented by using MATLAB. For the sake of comparison, the simulation and numerical results in [82] and [84] are used.

Notations	Definitions	Values
M/N	Average number of MTC devices in each cell	10 ~ 1000
R	The total number of preambles in a random access slot	15
N_{RAR}	Maximum number of RARs that can be carried in one response message	3
W_{RAR}	Size of random access response window in sub-frame unit	5
N_{ACK}	Maximum number of MTC devices that can be acknowledged within the RAR window	$N_{ACK} = N_{RAR} \times W_{RAR}$
N_{RA_SLOT}	Interval between two successive random access slots in sub-frame unit	$N_{RA\_SLOT} = 5$ if PRACH configuration Index = 6
N_{PR_D}	Processing time required by an eNB in order to detect the transmitted preamble in sub-frame unit	2

Table 6.3 Basic simulation parameters

Performance Metrics

The collision probability, the access success probability, the average access delay, the statistics of access delay, and the resource utilization are used as performance metrics in order to evaluate the CDR mechanism. The collision probability is defined as the ratio between the number of collided RAOs (two or more MTC devices use the same preamble and the random access slot) and the total number of reserved RAOs (with or without random access attempts). Note that, in CDR, there is no collision as each MTC device chooses a different preamble and random access slot than all other MTC devices do. Therefore, the collision probability is always equal to zero. The access

success probability is the probability to successfully complete the RACH procedure within the maximal number of preamble transmission. In the proposed solution, the access success probability is 100% as there is no collisions during the transmission of the preamble. Furthermore, there is no need of the messages Msg3 and Msg4 of the RACH procedure as the eNB can know exactly which MTC device transmits which preamble and in which random access slot. Therefore, the random access procedure in the proposed solution becomes like contention-free random access procedure. With these two advantages, i.e. the needless for Msg3 and Msg4 and also contention-free advantage, power consumption will be highly reduced. We argue this by the close relation between the number of preamble transmissions and the power consumption. Note that reducing the power consumption of power-limited devices (i.e., MTC devices) is one of the main issues to be overcome for the efficient deployment of this type of devices, and it is also one of the main targets of 5G system. Let D_k be the access delay for the k^{th} MTC device, where $k \in [1, M/N]$. This delay consists of the time required to transmit the preamble $(1+(\sigma_k-1)\times T_{RA\_SLOT})$ and the time required to receive the RAR message $(T_{PR\_D}+\lceil\frac{\beta_k}{N_{RAR}}\rceil)$ that consists of the processing delay in the eNB and the waiting time for the RAR message. The access delay D_k can be expressed, then, as follows:

$$D_k = 1+(\sigma_k-1)\times T_{RA\_SLOT}+T_{PR\_D}+\lceil\frac{\beta_k}{N_{RAR}}\rceil \tag{6.6}$$

where σ_k and β_k are the pair of values (σ,β) of the k^{th} MTC device. Let d be the delay for the RACH procedure between the first random access slot in the paging access interval and the completion of the RACH procedure. The statistics of access delay is defined as the Cumulative Distribution Function (CDF) of the delay for the RACH procedure between the first random access slot and the completion of the RACH procedure for the successfully accessed MTC devices. The CDF of access delay, denoted by $C(d)$, can be expressed as:

$$C(d) = \frac{\sum_{n=1}^{d} MTC_n}{M/N} \tag{6.7}$$

where MTC_n is the MTC device whose access delay is no more than n. The average access delay $\overline{D}$ is the total access delay for all successfully accessed MTC devices divided by the total number of successfully accessed MTC devices, and it can be

expressed as:

$$\overline{D} = \frac{\sum_{i=1}^{M/N} T_i}{M/N} \tag{6.8}$$

where T_i is the access delay for the i^{th} MTC device. The utilization of RAOs, denoted by U, is the ratio between the number of successfully accessed MTC devices and the number of reserved RAOs, and it can be expressed as:

$$U = \frac{M/N}{N_{slots} \times R} \tag{6.9}$$

Recall that the access success probability of the CDR is always 100%.

Results

Figs. 6.1 and 6.2 illustrate the access success probability and the collision probability, respectively, as a function of the number of MTC devices in each cell. Regarding the collision probability (Fig. 6.2), it is always 0% for CDR method, while it increases from 0.18% (for $M/N = 10$) to 72.8% (for $M/N = 1000$) for the ordinary GP method. The access success probability of the CDR (Fig. 6.1) is always 100%, while it decreases from 100% (for $M/N = 10$) to 21.3% (for $M/N = 1000$) for the ordinary GP method.

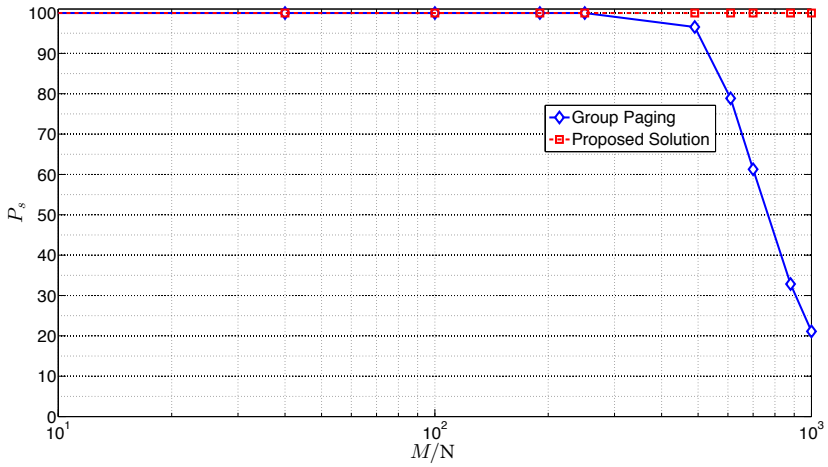

Figure 6.1 Access Success Probability

This can be explained by the fact that each MTC device in the CDR chooses a unique preamble and a random access slot than the other devices, while the MTC device in the ordinary group paging randomly chooses a preamble that can engender collision, leading to the degradation of access success probability when there is a large number of MTC devices, i.e. high contention on the RACH resources.

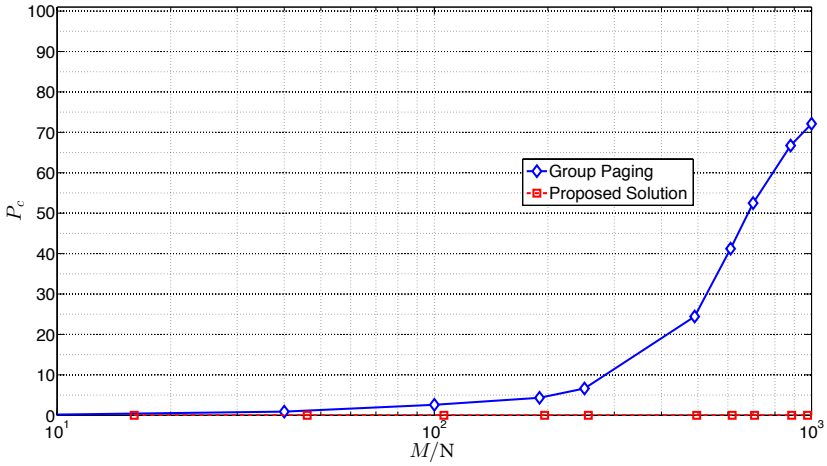

Figure 6.2 Collision Probability

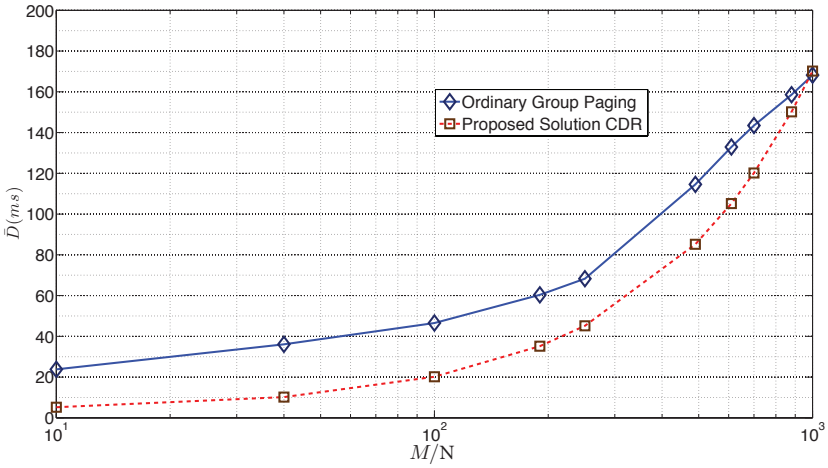

Figure 6.3 Average Value of Access Delay

Fig. 6.3 shows the average value of access delay for both the ordinary group paging and the CDR. The figure clearly demonstrates that the delay in the CDR is less than in the ordinary group paging. However, this delay increases as the number of MTC devices increases and becomes greater than the delay for the ordinary group paging when $M/N = 1000$. To better interpret the results of Fig. 6.3, this figure is divided into two regions: the first region where the delay of the CDR is less than that of the ordinary group paging and the second one is the contrary. The results of the first region of the figure can be explained by two factors: (i) no need for preamble retransmission in the CDR and ; (ii) the absence of the messages Msg3 and Msg4 in the RACH procedure (i.e., the RACH procedure in the CDR became like contention-free RACH procedure). Therefore, the delay in the proposed solution depends only on the time required to transmit the preamble and to receive the message Msg2. In addition, as there is no

need for preamble retransmission, the power consumption is reduced, which is a very important issue for the MTC devices with limited power resources. In addition to what is mentioned, there is also a gain in the resources of the network which comes from the absence of the messages Msg3 and Msg4. As mentioned in [68], the message Msg4 (in the downlink) consumes 4 Control Channel Elements (CCE) for one UE. The CCE consists of 36 Resource Element (RE), where the RE is one $15\,kHz$ subcarrier in the frequency domain and one symbol in the time domain. For example, if we have 1000 MTC devices, the gain in the resources only in the downlink is $1000 \times 4 = 4000$ CCEs which represents about $126\,Kbytes$ if each symbol represents 7 bits. The results of the second region is explained by the fact that the access success probability for the CDR is always equal to 100%. Therefore, the larger the number of MTC devices is, the larger the average access delay is. However, the access success probability decreases as the number of MTC devices increases in the ordinary group paging. Regarding the

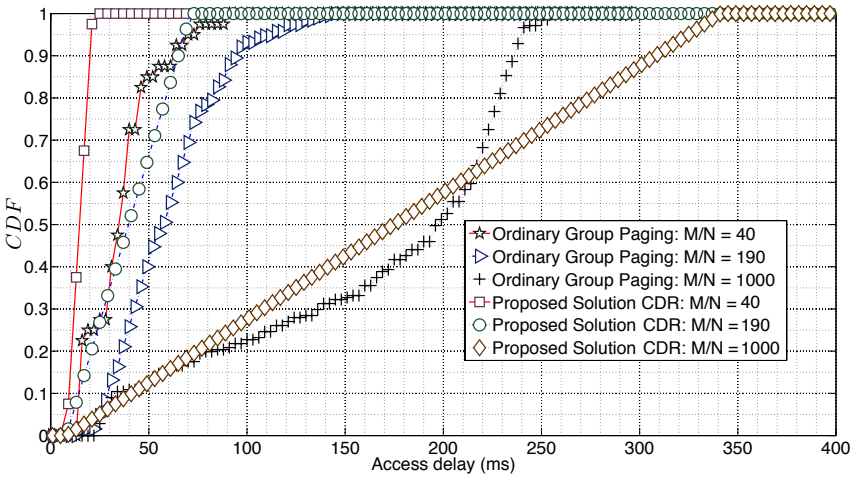

Figure 6.4 CDF of Access Delay

CDF of access delay, Fig. 6.4 clearly shows that the CDF of the CDR is always linear. The figure also illustrates that the maximum delay of the CDR when $M/N = 1000$ is larger than that of the ordinary group paging. This can be argued by the fact that the paging access interval, in the ordinary group paging, depends on the maximal number of preamble transmissions and non on the number of MTC devices. In contrast, the paging access interval in the CDR depends on the number of MTC devices and, thus, the maximum delay increases by increasing the number of devices. Fig. 6.5 illustrates the resource utilization for the CDR. It is clear that the resource utilization increases as the number of MTC devices increases and it is always larger than 50%, whereas it is mentioned in [84] that the optimal value of resource utilization for the ordinary

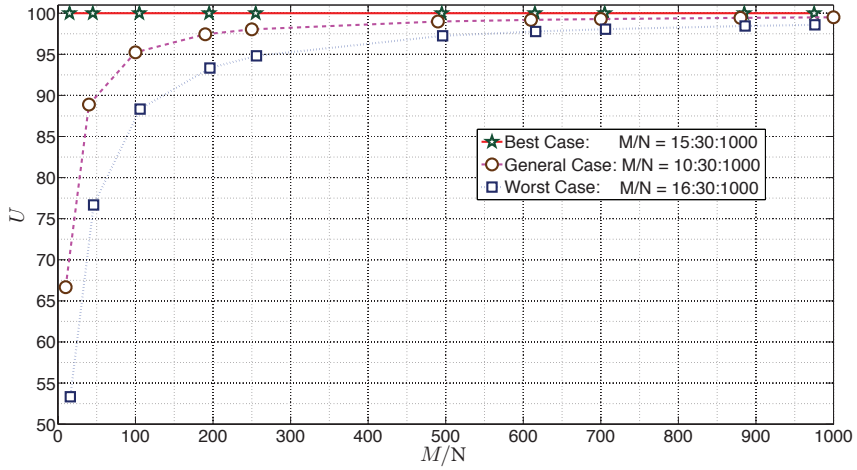

Figure 6.5 Resource Utilization of the CDR

group paging is approximately 20%. A summary of comparison between the proposed solution CDR and the ordinary group paging is given in Table 6.4.

	The ordinary group paging	The proposed solution	M/N
Collision Probability (%)	0.18	0	10
	5.32	0	100
	72.8	0	1000
Access Success Probability (%)	100	100	10
	100	100	100
	21.3	100	1000
Average Access Delay (ms)	23.99	5.2	10
	39.4	20.2	100
	168.16	170.2	1000

Table 6.4 Comparison between CDR and ordinary GP method

6.3 Traffic Spreading For Group Paging (TSFGP)

6.3.1 System Model:

Like in the precedent study, we consider a group of M MTC devices uniformly distributed over N cells is considered, i.e. M/N MTC devices in each cell. Again, it is assumed that each eNB reserves R RA resources for the contention access. After transmitting the paging message, addressed by the Group ID (GID), the MTC devices will start the RACH procedure with a probability p_{act}, instead of leaving them to start all at the same time. The objective behind using the probability p_{act} is to ensure that

the MTC devices starting the RACH procedure, in each RA slot, have access to the channel with a success probability that matches the network capacity. In other words, the objective is that the number of successful MTC devices in each RA slot is equal, at most, to the number of MTC devices that can be acknowledged during the RAR window W_{RAR}. Recall that the number of RAR responses during RAR window is equal to:

$$N_{ACK} = N_{RAR} W_{RAR}$$

6.3.2 Another Vision of Group Paging

The authors in [84] introduced a good analysis of the Group Paging (GP) method. However, in this section an alternative analysis of the GP will be introduced in order to well understand the proposed method TSFGP. After receiving the paging message, all the group members, i.e. M MTC devices (by assuming that there is only one cell), will start the contention-based RACH procedure in the first available RA slot. After transmitting the preambles, there is a part of MTC devices that successfully transmitted the preambles, while the preambles of the others would be collided, not collided but not detected by eNB, or not collided, detected by the eNB, but not indicated by the RAR message. The numbers of successful and collided MTC devices after the first preamble transmission are equal to [84]:

$$M_{i,s} = M_{1,s} = \begin{cases} M e^{-\frac{M}{R}} p_1 & \text{if } M e^{-\frac{M}{R}} p_1 \leq N_{ACK} \\ N_{ACK} & \text{otherwise} \end{cases} \tag{6.10}$$

$$M_{i,c} = M_{1,c} = M - M_{1,s} \tag{6.11}$$

where (i) is the order of the RA slot within the GP interval and p_1 is the detection probability for the first preamble transmission. Generally, for the nth preamble transmission, the probability p_n is equal to $p_n = 1 - e^{-n}$. After finishing the RAR window, all the MTC devices that did not receive a response, i.e. the $M_{1,c}$ MTC devices, suppose that a collision has occurred. Therefore, they will do backoff and then restart the RACH procedure by transmitting the preamble once the backoff timer expires. As the backoff time follows a uniform distribution, the collided MTC devices will be uniformly distributed over the next slots during the backoff interval W_{BO}. The number of MTC devices retransmitting their preambles for the next time, in a certain RA slot, is equal to the part of slots (named as α_a, α_{bc} and α_d), from the backoff interval, falling before this RA slot multiplied by the number of collided MTC devices. In the following, we will calculate the position of the RA slots falling within the backoff interval relative to

the preamble transmission at the RA slot (i), and the corresponding proportions, i.e. α_a, α_{bc} and α_d, of the MTC devices whose backoff timers expire and retransmit their preamble at these RA slots. The first RA slot that falls within the backoff window (as illustrated in Fig. 6.6) will be the one that comes just after the finish of RAR window. It will be at the position:

$$x_a(i) = i + \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rceil \tag{6.12}$$

where $x_a(i)$ is the order of the first RA slot within backoff window, relative to the

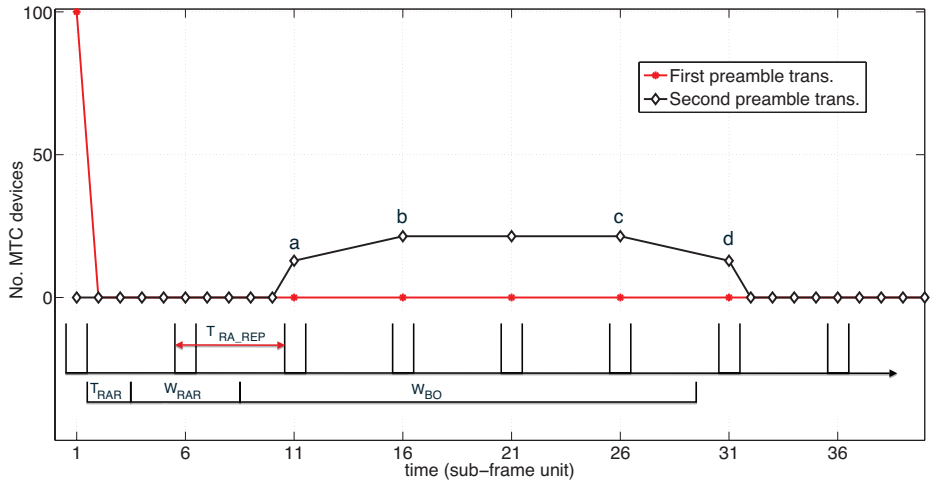

Figure 6.6 Number of MTC devices at each RA slot for the first and second preamble transmission for $R = 54$, and $M/N = 100$

preamble transmission at the RA slot (i), T_{RAR} is the processing delay at the eNB, and T_{RA_REP} is the interval between two consecutive RA slots. The proportion of the MTC devices whose backoff timers reach zero and hence retransmit their preambles at the RA slot (a) is equal to the time of the slot (a), in a sub-frame unit, minus the duration before the start of the backoff window (normalized by W_{BO}):

$$\alpha_a = \frac{[1 + (x_a(i) - 1)\, T_{RA\_REP}] - [1 + (x_a(i) - 1)\, T_{RA\_REP}]}{W_{BO}}$$

or

$$\alpha_a = \frac{\left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rceil T_{RA\_REP} - (T_{RAR} + W_{RAR})}{W_{BO}} \tag{6.13}$$

Regarding the RA slots from (b) to (c), they will be just after the RA slot (a), i.e.:

$$x_{bc}(i) = x_a(i) + k = i + \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rceil + k \tag{6.14}$$

where, $k = 1, 2, ..., K_{max}$. K_{max} represents the number of RA slots from the backoff window that fall between the slots (b) and (c). It is equal to (see the appendix A.2 for the proof):

$$K_{max} = \left\lfloor \frac{W_{BO} - \alpha_a W_{BO}}{T_{RA\_REP}} \right\rfloor \tag{6.15}$$

However, the proportion of MTC devices that retransmit their preambles at these RA slots is equal to:

$$\alpha_{bc} = \frac{T_{RA\_REP}}{W_{BO}} \tag{6.16}$$

The rest of collided MTC devices will transmit their preambles at the last RA slot within the backoff window, i.e. the RA slot (d). This slot will be just after the last one of the RA slots (bc), i.e.:

$$\begin{aligned} x_d(i) &= i + \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rceil + K_{max} + 1 \\ &= i + \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rceil + \left\lfloor \frac{W_{BO} - \alpha_a W_{BO}}{T_{RA\_REP}} \right\rfloor + 1 \\ &= i + \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rceil + \left\lfloor \frac{W_{BO}}{T_{RA\_REP}} - \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rceil + \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rfloor + 1 \end{aligned}$$

or

$$x_d(i) = i + \left\lfloor \frac{T_{RAR} + W_{RAR} + W_{BO}}{T_{RA\_REP}} \right\rfloor + 1 \tag{6.17}$$

and the proportion of MTC devices in this case is equal to:

$$\begin{aligned} \alpha_d &= 1 - \alpha_a - \alpha_{bc} K_{max} \\ &= 1 - \frac{Q T_{RA\_REP} - (T_{RAR} + W_{RAR})}{W_{BO}} - \\ &\qquad \frac{T_{RA\_REP}}{W_{BO}} \left\lfloor \frac{W_{BO}}{T_{RA\_REP}} - Q + \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rfloor \end{aligned}$$

where $Q = \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rceil$. Therefore, α_d is equal to:

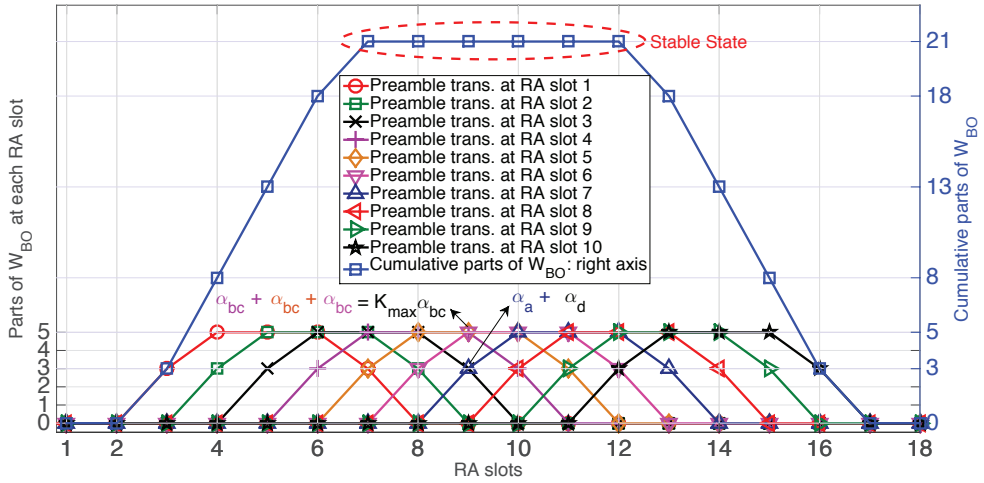

Figure 6.7 Cumulative parts of W_{BO} for each RA slot for MTC devices transmitting their preambles for the second time, where $W_{BO} = 21$ and $T_{RA\_REP} = 5$

$$\alpha_d = \frac{T_{RAR} + W_{RAR} + W_{BO}}{W_{BO}} - \frac{T_{RA\_REP}}{W_{BO}} \left\lfloor \frac{T_{RAR} + W_{RAR} + W_{BO}}{T_{RA\_REP}} \right\rfloor \tag{6.18}$$

It is worth noting that $\alpha_a + K_{max}\alpha_{bc} + \alpha_d = 1$. Accordingly, the numbers of MTC devices retransmitting their preambles for the second time are equal to:

$$M_{retrans} = \begin{cases} M_{1,c} \times \alpha_a & ;\text{for RA slot } a \\ M_{1,c} \times \alpha_{bc} & ;\text{for RA slots } bc \\ M_{1,c} \times \alpha_d & ;\text{for RA slot } d \end{cases} \tag{6.19}$$

By assuming that each RA slot experiences the same number of new arrivals, the number of successful and collided MTC devices will be the same as given by the equations (6.10), (6.11), and (6.19), and they will generate the same graphic as illustrated in Fig. 6.6. Therefore, the number of collided MTC devices at each RA slot will be the sum of the contribution of each RA slot, as illustrated in Fig. 6.7. From this figure, we clearly see that when the number of new arrivals at each RA slot is the same, we come up to a situation where the number of MTC devices retransmitting their preambles is constant. This implies that the number of successful and collided MTC devices at each RA slot will be constant too.

6.3.3 Analytical Model

The key idea behind the proposed scheme FI-TSFGP is to scatter the MTC devices of a group being paged over the available interval rather than leaving them to start

the contention-based RACH procedure all at once. Generally speaking, the number of MTC devices at the RA slot (i) can be written by the following equation:

$$M_i = \sum_{n=1}^{N_{PT_{max}}} M_i[n] \tag{6.20}$$

where $N_{PT_{max}}$ is the maximum number of preamble transmissions, and $M_i[n]$ is the number of MTC devices transmitting their preamble for the n^{th} time in the RA slot (i). The number of successful MTC devices at the RA slot (i) is equal to [84, 122]:

$$M_{i,s}[n] = \begin{cases} M_i[n]e^{-\frac{M_i}{R}}p_n & ;\text{if } \sum_{n=1}^{N_{PT_{max}}} M_i[n]e^{-\frac{M_i}{R}}p_n \leq N_{ACK} \\ \frac{M_i[n]e^{-\frac{M_i}{R}}p_n}{\sum_{n=1}^{N_{PT_{max}}} M_i[n]e^{-\frac{M_i}{R}}p_n} N_{ACK} & ;\text{otherwise} \end{cases}$$

However, the network can not send back responses to more than N_{ACK} MTC devices even if the number of successful MTC devices is more than N_{ACK}. Hereafter, the focus will be on the case where the number of successful MTC devices is less than or equal to N_{ACK}, i.e. $\sum_{n=1}^{N_{PT_{max}}} M_i[n]e^{-\frac{M_i}{R}}p_n \leq N_{ACK}$. Accordingly, the number of successful MTC devices at the RA slot (i) could be written as:

$$M_{i,s}[n] = M_i[n]e^{-\frac{M_i}{R}}p_n \tag{6.21}$$

Let M_{arv} denote the number of new arrivals at each RA slot, which represents the value $M_i[1]$, and therefore the number of successful and collided MTC devices will be:

$$M_{i,s}[1] = M_i[1]e^{-\frac{M_i}{R}}p_1 = M_{arv}e^{-\frac{M_i}{R}}p_1 \tag{6.22}$$

$$M_{i,c}[1] = M_{arv} - M_{i,s}[1] = M_{arv}(1 - e^{-\frac{M_i}{R}}p_1) \tag{6.23}$$

From Fig. 6.7, it is clear that when the total number of MTC devices, and consequently the number of successful MTC devices, is stable (i.e., merely constant), the cumulative parts of W_{BO} is equal to W_{BO}. Therefore, the collided MTC devices, engendered from the precedent RA slots, whose backoff timers expire and retransmit the preamble for the $(n+1)^{th}$ time at the current RA slot, i.e. $M_i[n+1]$, will be equal to the number of collided MTC devices at the current RA slot transmitting their preambles for the n^{th} time, i.e. $M_{i,c}[n]$. This means that $M_{i,c}[n] = M_i[n+1]$. For example, the number of

MTC devices transmitting their preamble for the second time is equal to:

$$M_i[2] = \sum_{h=i-H_1}^{i-H_2} \alpha_h M_{h,c}[1] \tag{6.24}$$

where

$$H_1 = \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rceil$$

$$H_2 = \left\lfloor \frac{T_{RAR} + W_{RAR} + W_{BO}}{T_{RA\_REP}} \right\rfloor + 1$$

Note that H_1 and H_2 are deduced directly from $x_d(i)$ and $x_a(i)$, respectively. α_h can be one of the following values: α_a, α_{bc}, and α_d. As the system is in the stable state, both the number of collided MTC devices transmitting their preamble for the first time ($M_{h,c}[1]$) and the total number of MTC devices (M_h) are constant. Note that $M_{h,c}[1]$, which is equal to $M_{i,c}[1]$, is given by the equation (6.23). Then, the equation (6.24) becomes:

$$M_i[2] = M_{i,c}[1] \times \sum_{h=i-H_1}^{i-H_2} \alpha_h \tag{6.25}$$

As the cumulative parts of W_{BO} is equal to W_{BO}, we deduce from Figs. 6.6 and 6.7 that:

$$\sum_{h=i-H_1}^{i-H_2} \alpha_h = \alpha_a + K_{max}\alpha_{bc} + \alpha_d = 1 \tag{6.26}$$

and thus $M_i[2] = M_{i,c}[1]$. The numbers of collided and successful MTC devices transmitting their preamble for the second time are equal to:

$$M_i[2] = M_{i,c}[1] = M_{arv}(1 - e^{-\frac{M_i}{R}} p_1)$$

$$M_{i,s}[2] = M_i[2] e^{-\frac{M_i}{R}} p_2 = M_{arv}(1 - e^{-\frac{M_i}{R}} p_1) e^{-\frac{M_i}{R}} p_2$$

$$\begin{aligned} M_{i,c}[2] &= M_i[2] - M_{i,s}[2] \\ &= M_{arv}(1 - e^{-\frac{M_i}{R}} p_1)(1 - e^{-\frac{M_i}{R}} p_2) \\ &= M_{arv} \prod_{k=1}^{2} (1 - e^{-\frac{M_i}{R}} p_k) \end{aligned}$$

By induction, we find that:

$$\begin{aligned} M_i[n] &= M_{i,c}[n-1] \\ M_{i,s}[n] &= M_{arv} \prod_{k=1}^{n-1} (1 - e^{-\frac{M_i}{R}} p_k) e^{-\frac{M_i}{R}} p_n \\ M_i[n+1] &= M_{i,c}[n] = M_{arv} \prod_{k=1}^{n} (1 - e^{-\frac{M_i}{R}} p_k) \end{aligned} \tag{6.27}$$

or

$$M_i[n] = M_{i,c}[n-1] = M_{arv} \prod_{k=1}^{n-1} (1 - e^{-\frac{M_i}{R}} p_k) \tag{6.28}$$

Therefore, the total number of MTC devices at each RA slot, in the stable state, is equal to:

$$M_i = \sum_{n=1}^{N_{PT_{max}}} M_i[n] = M_{arv} \sum_{n=1}^{N_{PT_{max}}} \prod_{k=1}^{n-1} (1 - e^{-\frac{M_i}{R}} p_k) \tag{6.29}$$

The equation (6.29) can be written by the following form (see the appendix A.1 for the demonstration):

$$M_i = M_{arv} \sum_{m=0}^{N_{PT_{max}}-1} \alpha_m e^{-\frac{m M_i}{R}} \tag{6.30}$$

where α_m is:

$$\alpha_m = \sum_{t=1}^{N_{PT_{max}}-m} (-1)^m \underbrace{\sum_{k_1=1}^{t} \sum_{k_2=k_1+1}^{t+1} \dots \sum_{k_m=k_{m-1}+1}^{t+m-1}}_{\text{m times}} p_{k_1} p_{k_2} \dots p_{k_m} \tag{6.31}$$

However, the exponential function can be approximated by the following equation [123]:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \dots \tag{6.32}$$

Applying this approximation to the equation (6.30), we find that:

$$\frac{M_i}{M_{arv}} = \sum_{m=0}^{N_{PT_{max}}-1} \alpha_m - \sum_{m=0}^{N_{PT_{max}}-1} m \alpha_m \frac{M_i}{R} + \sum_{m=0}^{N_{PT_{max}}-1} m^2 \alpha_m \frac{M_i^2}{2R^2} \tag{6.33}$$

or,

$$\left(\sum_{m=0}^{N_{PT_{max}}-1} m^2\alpha_m\right) M_i^2 - 2\left(\frac{R^2}{M_{arv}} + R\sum_{m=0}^{N_{PT_{max}}-1} m\alpha_m\right) M_i + 2R^2 \sum_{m=0}^{N_{PT_{max}}-1} \alpha_m = 0 \tag{6.34}$$

This equation is a second order one for M_i, which can be solved easily. After obtaining the total number of MTC devices in the stable state, M_i, the number of successful MTC devices can be calculated by the following equation:

$$M_{i,s} = \sum_{n=1}^{N_{PT_{max}}} M_{i,s}[n] \tag{6.35}$$

where $M_{i,s}[n]$ is given by the equation (6.27).

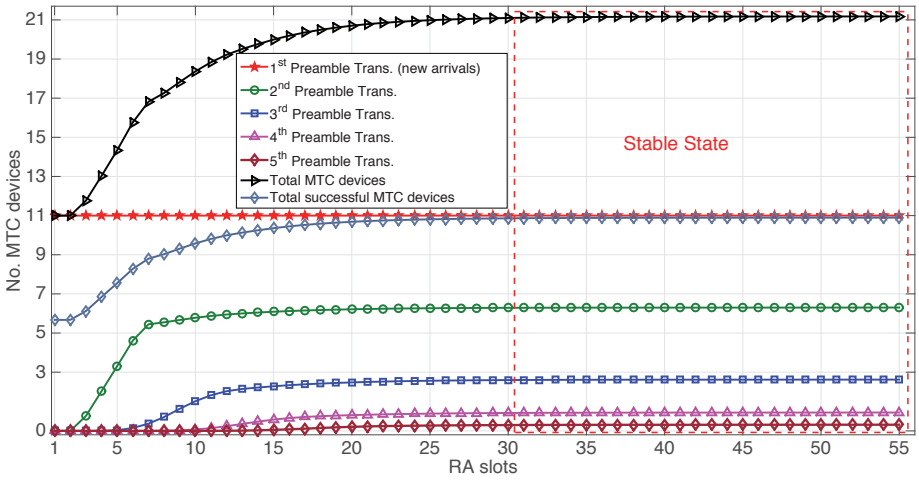

Figure 6.8 Number of MTC devices for each preamble transmission as well as the number of total and successful MTC devices in each RA slot; $R = 54$, $N_{PT_{max}} = 5$

Fig. 6.8 shows the number of MTC devices transmitting their preambles for the i^{th} time, and also the total number of arrivals and the number of successful MTC devices ($N_{PT_{max}} = 5$). It is worth noting that the calculated value by the equation (6.34) is for the case when the number of arrivals is stable.

Figs. 6.9 - 6.11 and 6.10 - 6.12 illustrate the true and the approximated values of the total number of MTC devices (equation 6.34) and the number of successful MTC devices (equation 6.35), respectively. These figures include the results for TSFGP as well FI-TSFGP for the sake of comparison. Moreover, different values of R, M_{arv},

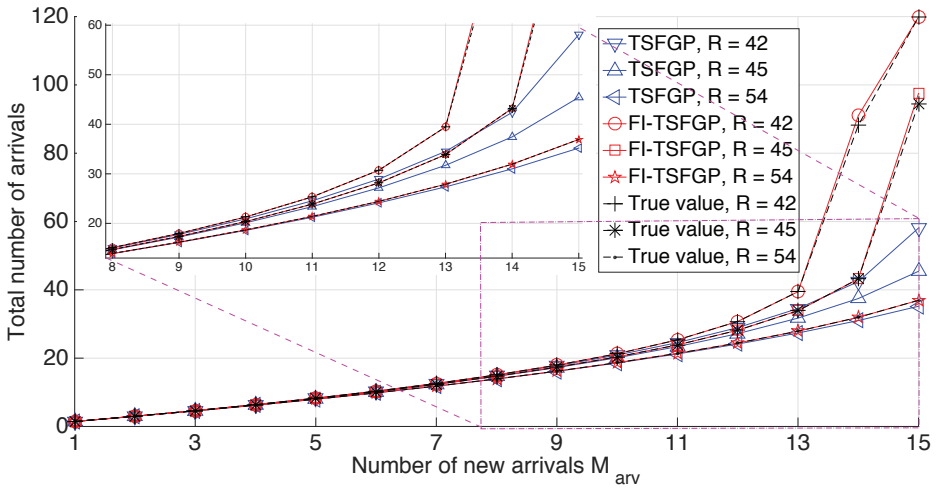

Figure 6.9 The total number of arrivals in the stable state as function of the number of new arrivals M_{arv} for different number of preambles; $N_{ACK} = 15$ and $N_{PT_{max}} = 10$

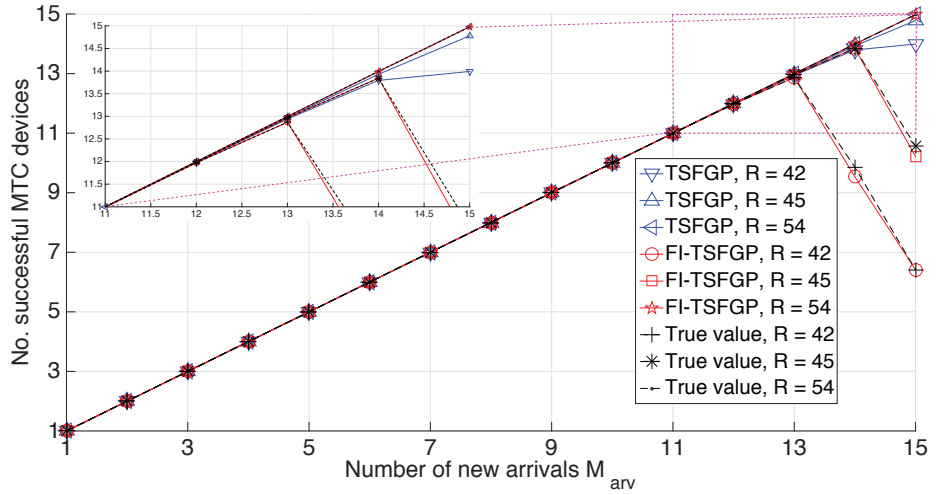

Figure 6.10 Number of successful MTC devices in the stable state as a function of the number of new arrivals M_{arv} for different numbers of preambles; $N_{ACK} = 15$ and $N_{PT_{max}} = 10$

and $N_{PT_{max}}$ were considered. From these figures, it is clear that TSFGP method generally gives a good estimation of the total number and also the number of successful MTC devices. However, TSFGP fails to estimate the intended values for certain configurations, e.g. $R = 42$, $M_{arv} = 15$ and $N_{PT_{max}} = 10$ (Figs. 6.9 and 6.10). To cope with this shortcoming, FI-TSFGP uses an iterative operation as illustrated in **Algorithm 3**, where δ is the tolerated error. Note it is assumed that the value calculated by the equation (6.34) is the initial guess of the total number of MTC devices in the stable state.

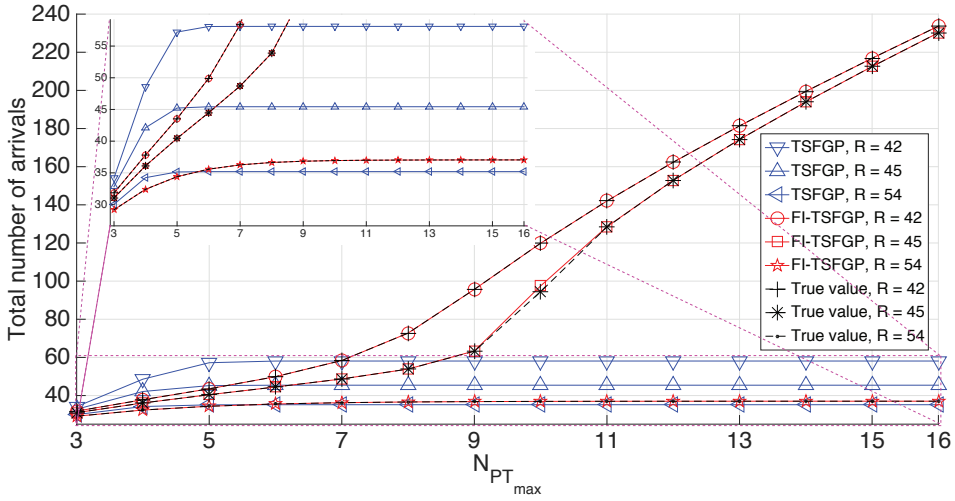

Figure 6.11 Total number of arrivals in the stable state as a function of the number of preamble transmissions $N_{PT_{max}}$; $M_{arv} = N_{ACK} = 15$

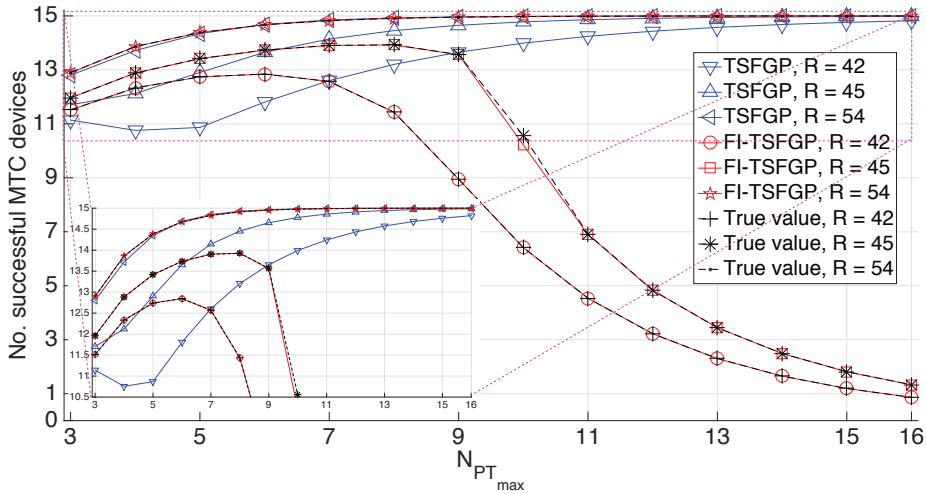

Figure 6.12 Number of successful MTC devices in the stable state as a function of the number of preamble transmissions $N_{PT_{max}}$; $M_{arv} = N_{ACK} = 15$

Returning to Figs. 6.9, 6.10, 6.11, and 6.12, we observe that FI-TSFGP has a great impact when both the number of new arrivals M_{arv} and the number of preamble transmissions $N_{PT_{max}}$ are large. This is attributable to the improvement obtained via the iterative operation. Further, these four figures reveal that FI-TSFGP can be applied for any configuration, while TSFGP is valid for certain configurations. Therefore, FI-TSFGP gives the flexibility to change the network's parameters, e.g. increasing $N_{PT_{max}}$ for increasing the available interval. Regarding the number of successful MTC devices, we remark that, for a fixed value of $N_{PT_{max}}$, the relationship between the number of new arrivals (M_{arv}) and the number of successful MTC devices is roughly linear

Algorithm 3 Iteration operation for further improvement of the approximated value of M_i

1: $M_{i_{guess}} \leftarrow$ (the solution of equation (6.34))
2: $M_{i_{current}} \leftarrow M_{i_{guess}}$
3:
$$M_{i_{new}} \leftarrow M_{arv} \sum_{m=0}^{N_{PT_{max}}-1} \alpha_m e^{-\frac{mM_{i_{current}}}{R}}$$
4: **while** $|M_{i_{new}} - M_{i_{current}}| > \delta$ **do**
5:
$$M_{i_{current}} \leftarrow M_{i_{new}}$$
$$M_{i_{new}} \leftarrow M_{arv} \sum_{m=0}^{N_{PT_{max}}-1} \alpha_m e^{-\frac{mM_{i_{current}}}{R}}$$
6: **end while**
7: $M_{i_{current}} \leftarrow M_{i_{new}}$

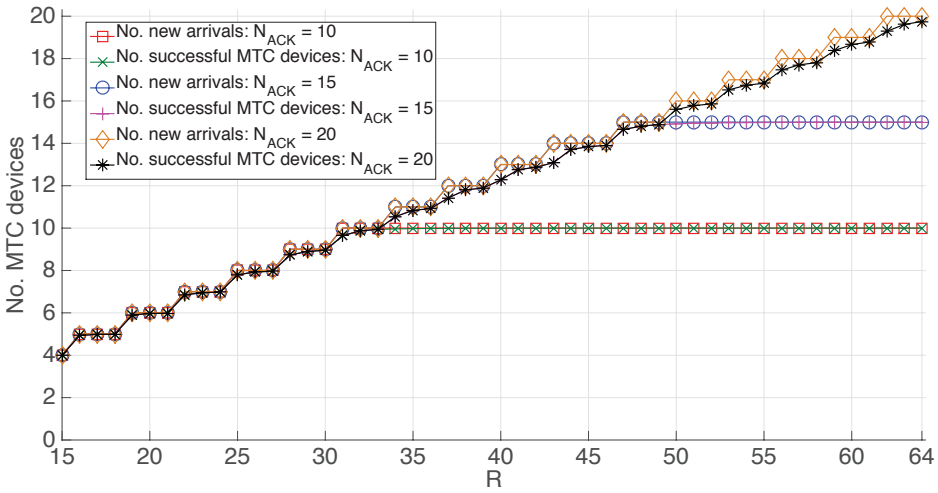

Figure 6.13 Number of new arrivals that maximizes the number of successful MTC devices and the corresponding number of successful MTC devices as a function of the number of preambles for different values of N_{ACK}; $N_{PT_{max}} = 10$

as long as M_{arv} is smaller than a certain value, which is equal to ($M_{arv} = 13$) when $R = 42$. Thus, the best number of new arrivals for a certain configuration will be the value that maximizes the number of successful MTC devices as illustrated in Fig. 6.13. This figure is highly important since it illustrates the optimal number of new arrivals M_{arv} for a given number of preambles and certain values of N_{ACK}. From this figure, we see that the number of new arrivals (and consequently the number of successful MTC devices) grows as the number of available preambles increases. Moreover, this

relationship could be approximated to a linear one. However, when the number of available preambles exceeds certain value ($R = 50$ when $N_{ACK} = 15$), the improvement becomes minimal. In this case, it is more appropriate to choose ($R = 50$) for a better utilization of resources.

In order to go further inside, the total number of MTC devices and the number of successful MTC devices at the stable state are plotted for different values of the number of new arrivals M_{arv}, the number of available preambles R, and the maximum number of preamble transmissions $N_{PT_{max}}$. The results are shown in Figs. 6.14, 6.15, 6.16, and 6.17. For a fixed value of M_{arv} (which is equal to 15), Figs. 6.14 and 6.15 show how the total number of MTC devices and also the number of successful MTC device vary for different values of R and $N_{PT_{max}}$. From these two figures, it is clear that the number of successful MTC devices is generally near to the capacity of the value that the network can support, i.e. the number of responses during the RAR window, where the total number of arrivals is somewhat reasonable compared to the number of available preambles. However, there is a certain region where the number of successful MTC devices dramatically drops, accompanying by a dramatic increase of the total number of MTC devices. Therefore, this region (that can be viewed as a collapse region) of configuration must be avoided in order to maintain high throughput of the network. This behavior explain the simulation results approved by 3GPP on increasing the maximum number of preamble transmissions $N_{PT_{max}}$ [6], where it is mentioned that increasing $N_{PT_{max}}$ may make the network worse under overload network conditions. However, it is not mentioned neither in that study nor in the literature why the behavior becomes worse when increasing $N_{PT_{max}}$, while Figs. 6.14 and 6.15 are well interpretation of this behavior. Figs. 6.16, and 6.17 show the total number of MTC devices and the number of successful MTC devices for different values of M_{arv} and R, where $N_{PT_{max}}$ is fixed to be 10. These figures also show a similar behavior, where the number of successful MTC devices linearly increases as the number of new arrivals does, except for certain region. The behavior in the concerned region is different, where a collapse takes place when increasing one (or the two) of the concerned parameters. Therefore, from the last four figures, it can be concluded that triple values should be jointly and carefully determined in order to maintain a good throughput. This demonstrates that not only increasing the maximum number of preamble transmission may make the network worse, but also the number of preambles and the number of new arrivals too.

Taking into account these results, it is better to activate, at each RA slot, a number of MTC devices less than or equal to N_{ACK}, instead of leaving all the members

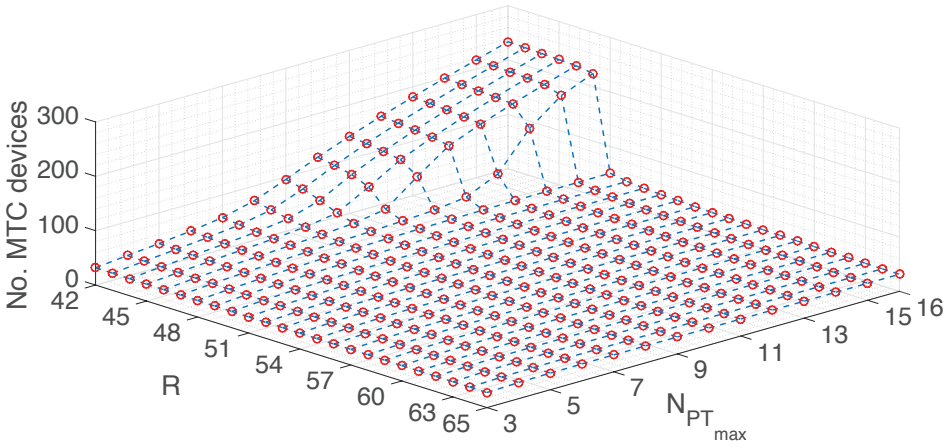

Figure 6.14 Total number of MTC devices as a function of R and $N_{PT_{max}}$, where $M_{arv} = 15$.

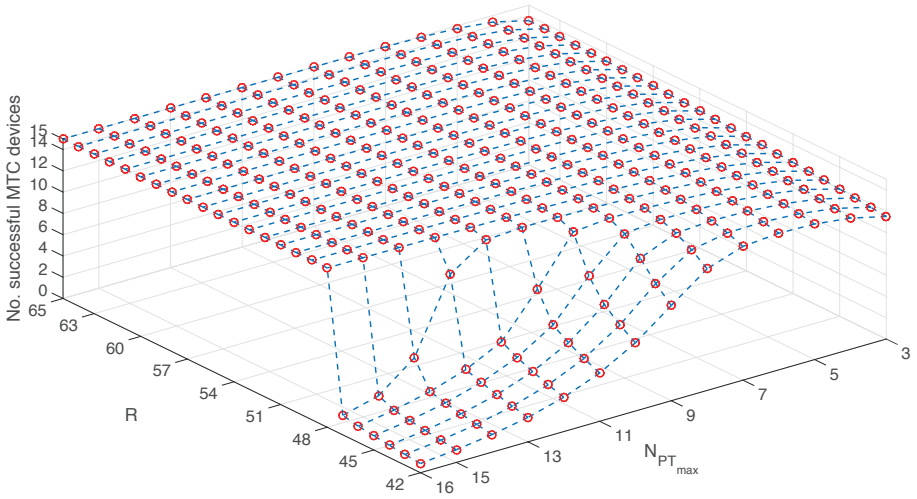

Figure 6.15 Number of successful MTC devices as a function of R and $N_{PT_{max}}$, where $M_{arv} = 15$.

of the group to start the RACH procedure all at the same time. If we need to uniformly distribute (M/N) MTC devices over I_{max} RA slots, there will be, on average, $(M/N)/I_{max}$ MTC devices at each RA slot, where I_{max} is given by the following equation [84]:

$$I_{max} = 1 + (N_{PT_{max}} - 1) \left\lceil \frac{T_{RAR} + W_{RAR} + W_{BO}}{T_{RA\_REP}} \right\rceil$$

In order to make sure that there will be, on average, M_{arv} MTC devices at each RA slot, we then distribute the devices over a virtual interval containing $I_{V_{max}}$ RA slots,

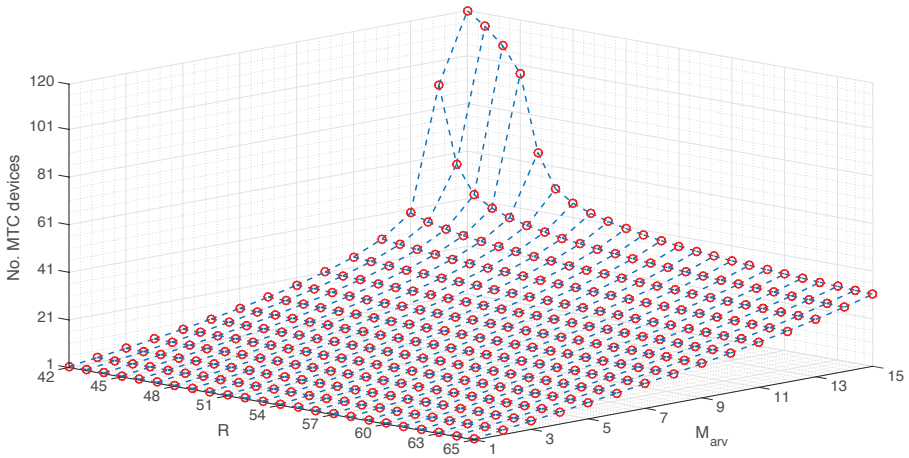

Figure 6.16 Total number of MTC devices as a function of R and $Marv$, where $N_{PT_{max}} = 10$.

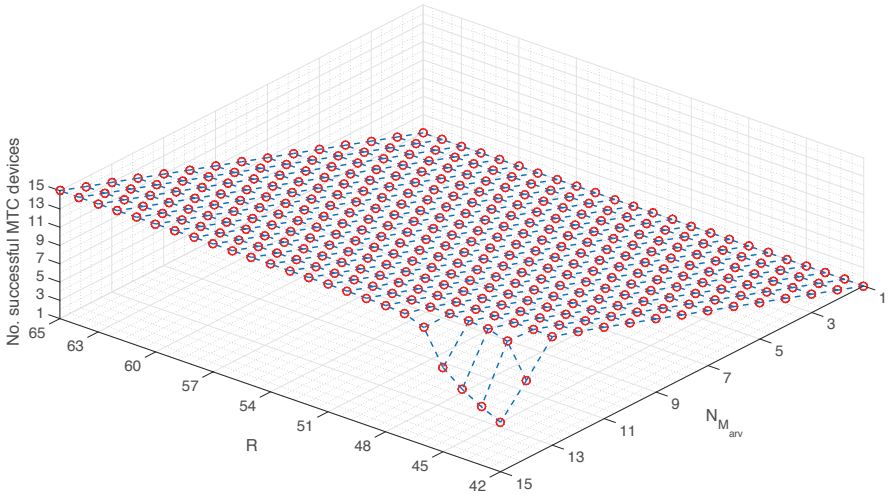

Figure 6.17 Number of successful MTC devices as a function of R and $Marv$, where $N_{PT_{max}} = 10$.

where;

$$I_{V_{max}} = \left\lceil \frac{(M/N)}{M_{arv}} \right\rceil \tag{6.36}$$

Now, each MTC device randomly generates an integer value in the interval $[1, I_{V_{max}}]$. This value represents the RA slot in which the MTC would start the contention-based RACH procedure. Note that the generated values follow the uniform distribution. If the generated value falls within the interval $[1, I_{max}]$, then this device will start the RACH procedure in this RA slot, otherwise, it goes out and returns to the inactive state. The objective behind this technique is to directly determine whether a MTC will proceed the RACH procedure or not. Thus, this technique will avoid leaving the

MTC devices to attempt the transmission at each RA slot, e.g. like the $p-$persistent mechanism [68]. As the MTC devices are uniformly distributed over the available RA slots, increasing the number of RA slots, i.e. increasing I_{max}, will further improve the performance of FI-TSFGP, where the optimal performance would be achieved when $I_{max} = I_{V_{max}}$. It is worth noting that GP performance can not be improved by increasing the number of RA slots (I_{max}), as all the MTC devices start the RACH procedure at the first available RA slot, by supposing that the number of preamble transmissions $N_{PT_{max}}$ is fix.

Assuming that the M MTC devices are uniformly distributed over N cells, the FI-TSFGP mechanism could be deployed in a real environment as follows:

1. The network (i.e., eNB) sends the paging message to the intended MTC devices, containing the number of MTC devices to be paged (M/N) and indicating the maximum number of new arrivals, M_{arv}, that the network can support at each RA slot.
2. When receiving the paging message, the MTC device can calculate the virtual interval $I_{V_{max}}$, via the equation (6.36), using the values in the received message.
3. Regarding the interval I_{max}, it can be either calculated by the MTC device using the parameters broadcasted by the network or explicitly sent in the paging message.
4. After obtaining the values I_{max} and $I_{V_{max}}$, the MTC device generates an integer value $k \in [1, I_{V_{max}}]$.
5. If $k <= I_{max}$, the MTC device starts the contention-based RACH procedure at the k^{th} RA slot. Otherwise, it disconnects and goes back to the inactive state.

Looking at the precedent steps of FI-TSFGP, it becomes apparent that FI-TSFGP is simple to be applied in practice. Compared to the original Group Paging (GP) method, FI-TSFCP incurs at MTC devices only minimal additional calculations as per steps 2), 4), and 5).

6.4 Performance Evaluation

In order to evaluate the performance of FI-TSFGP, C++-based discrete events simulator is built. In the simulation, a group of MTC devices ranging from (10) to (5000) has been considered. Regarding the parameters of RACH procedure, they are summarized in Table B.1. For the sake of simplicity, the pathloss remains constant and is the same for all the MTC devices. Regarding the parameters of power consumption

Table 6.5 Basic simulation parameters

Notations	Definition	Values
M/N	Average number of devices in each cell	10~5000
R	Total number of preambles in a random access slot	54
BI	Backoff indicator in a sub-frame unit	20
$N_{PT_{max}}$	Maximum number of preamble transmissions	16
N_{RAR}	Maximum number of RARs that can be carried in one response message	3
T_{RAR}	Processing delay required by the eNB in order to detect the transmitted preamble in a sub-frame unit	2
W_{RAR}	Size of random access response window in a sub-frame unit	5
N_{ACK}	Maximum number of MTC devices that can be acknowledged within the RAR window	$N_{ACK} = N_{RAR} \times W_{RAR}$
$PRACH_{config\_indx}$	PRACH configuration index	$PRACH_{config\_indx} = 6$
$T_{RA\ REP}$	The interval between two consecutive RA slots	5
W_{BO}	Backoff window size	$BI + 1$
p_n	Preamble detection probability for the n-th preamble transmission	$p_n = 1 - e^{-n}$
T_{CRT}	Contention Resolution timer	48
p_{HARQ_RET}	HARQ retransmission probability for Msg3 and Msg4 (non-adaptive HARQ)	10%
N_{HARQ}	Maximum number of HARQ TX for Msg3 and Msg4 (non-adaptive HARQ)	5
Parameters relative to CGP		
C_{max}	Number of paging cycles	7
$N_{PT_{max}}$	Maximum number of preamble transmissions	3
Parameters relative to PBO		
W_{PBO}	Pre-backoff window in a sub-frame unit	240
$N_{PT_{max}}$	Maximum number of preamble transmissions	8
Parameters relative to Power consumption		
P_{CMAX}	Maximum transmit power	$0\ dBm$
$PIRTP$	Preamble initial received target power	$-120\ dBm$
Δ_{prmbl}	Preamble format based offset	$0\ dB$
PRS	Power Ramping Step	$0\ dB$
PL	Pathloss	$98\ dBm$
P_0	Power consumption in the inactive state	$0\ mW$
P_1	Power consumption when the UE is waiting for RA slot, and also when it is in backoff	$-37\ dBm$
P_2	Power consumption when the UE receives (or expected to receive) the messages Msg2 and Msg4	$-22\ dBm$
P_3	Power consumption when the UE transmits a signal, such as preamble transmission	$P_3 = P_{PRACH}$

(Table B.1), P_2 is taken to be P_3 for the first time of preamble transmission and P_1 is about 30 times less than P_2. The power ramping factor is set to be zero (i.e., it will be nullified). As the average waiting time for the first available RA slot is equal to $T_{RA\_REP}/2 = 2.5ms$, the average power consumption during this period is equal to $2.5 \times 10^{0.1 \times * P_1}\ mW$, which is added for all the MTC devices at the start of the simulation. In order to show how the proposed method behaves, FI-STFGP will be compared with the Group Paging (GP), Consecutive Group Paging (CGP) [120], and Pre-BackOff (PBO) methods [121]. The main idea of CGP is to repeat the paging interval many times so that the MTC devices not succeeded in the first paging interval will try to access in the next paging interval and so on. Note that the number of paging cycles, i.e. C_{max}, and $N_{PT_{max}}$ are chosen to be equal to (7) and (3), respectively, as these values maximize the performance of CGP [120]. Regarding the PBO method [121], all the

members of the intended group will do backoff before the first preamble transmission. The value of pre-backoff for PBO method is chosen to be equal to $W_{PBO} = 240$ and the corresponding maximum number of preamble transmissions is $N_{PT_{max}} = 8$. It is worth to recall that FI-TSFGP tries to activate, at each RA slot during the available interval, the number of MTC devices that maximizes the performance. However, PBO tries to spread the MTC devices during a certain interval regardless the size of the intended group.

6.4.1 Performance Metrics

The metrics considered to evaluate the performance of the four above-mentioned schemes are: success, collision, and drop (only for FI-TSFGP) probabilities, average access delay, average number of preamble transmissions, resource utilization, CDF of both preamble transmission and access delay, and power consumption. The success probability is defined as the number of MTC devices successfully finished their RACH procedure within the maximum number of preamble transmissions, normalized by the total number of MTC devices (activated and non-activated, for FI-TSFGP, ones). The collision probability is the ratio between the number of collided RAOs and the total number of available RAOs. Since M_{arv} MTC devices will be activated at each RA slot for FI-TSFGP, there will be a part of MTC devices that will not be activated when $(M/N) > I_{max}M_{arv}$. Thus, the drop probability is equal to:

$$P_d = \begin{cases} \dfrac{(M/N) - I_{max}M_{arv}}{(M/N)} & ; \text{if}\,(M/N) > I_{max}M_{arv} \\ 0 & ; \text{otherwise} \end{cases} \tag{6.37}$$

Regarding the average access delay, it represents the total access delay for all the MTC devices, which successfully finished the RACH procedure, between the first preamble transmission and the completion of the random access procedure (within the maximum number of preamble transmissions) divided by the total number of successful MTC devices [82, 45]. The average number of preamble transmissions is the total number of preamble transmissions of all the MTC devices successfully finished the RACH procedure, divided by the number of successful MTC devices. For CGP scheme, this time will be the sum of the access delay in the current paging interval plus the time of the precedent paging intervals, and the same thing for the average number of preamble transmissions. Let (r) be the number of preamble transmissions, then the CDF of preamble transmission can be defined as the number of MTC devices successfully finished their RACH procedure by (r) times or less of preamble transmissions divided

by the total number of successful MTC devices, and it is given by the following equation:

$$CDF_R(r) = \frac{\sum_{i=1}^{I_{max}} \sum_{n=1}^{r} M_{i,s}[n]}{\sum_{i=1}^{I_{max}} \sum_{n=1}^{N_{PT_{max}}} M_{i,s}[n]} \tag{6.38}$$

Let (d) be the access delay for the RACH procedure between the first attempt and the completion of the RACH procedure. The CDF of access delay can be defined as the number of MTC devices successfully finished the RACH procedure before the time (d) and the total number of successful MTC devices. It is given by the following equation:

$$CDF_D(d) = \frac{\sum_{t=1}^{d} M_{s,t}}{\sum_{t=1}^{T_{max}} M_{s,t}} \tag{6.39}$$

where $M_{s,t}$ is the number of successful MTC devices whose access delay is equal to (t), and T_{max} is the maximum access delay that is equal to the time of the paging interval in a sub-frame unit, i.e. $T_{max} = 1 + (I_{max} - 1) * T_{RA\_REP} + T_{RAR} + W_{RAR}$. The Resource Utilization (RU) can be defined as the ratio of the total number of successful MTC devices to the total available RAOs, and it can be given by the following equation:

$$RU = \frac{\sum_{i=1}^{I_{max}} \sum_{n=1}^{N_{PT_{max}}} M_{i,s}[n]}{I_{max} R} \tag{6.40}$$

Regarding the power consumption, four values are considered: the power consumption of successful, collided, dropped (just for FI-TSFGP), and the total number of MTC devices. The Power consumption of successful/collided/dropped MTC devices is the mean power consumption of the MTC devices successfully accessed the network/collided/dropped, respectively. These parameters will be calculated for GP method, and then generalized for FI-TSFGP. Usually, the power consumption of the successful MTC devices consists of the following parts (it is assumed that the device needs n preamble transmissions before a successful attempt):

1. The power consumption when the device is waiting for the first RA slot, which is equal to $(T_{RA\_REP}/2)P_1$.
2. The power consumption when the device is transmitting the preamble for the first $(n-1)$ times, and collision occurs. This power is equal to the one consumed

in the following steps; transmitting the preamble (P_3), waiting for the RAR window ($T_{RAR}P_1$), during the RAR window ($W_{RAR}P_2$), and during the backoff and waiting for the next RA slot that is equal to;

$$\left(\left\lceil \frac{1+T_{RAR}+W_{RAR}+\frac{W_{BO}}{2}}{T_{RA\_REP}} \right\rceil T_{RA\_REP}-1-T_{RAR}-W_{RAR}\right)P_1$$

Note that $W_{BO}/2$ is the average time of the backoff as the backoff timer can expire at any time during the backoff window.

3. The power consumed during the n^{th} preamble transmission (the successful transmission); which is equal to $P_3+T_{RAR}P_1+(W_{RAR}/2)P_2$. It is worth noting that $(W_{RAR}/2)P_2$ is the average power consumption during the RAR window as the MTC device can receive the RAR message at any sub-frame during the RAR window.
4. The power consumed for the messages Msg3 and Msg4 of the RACH procedure (by ignoring the effect of Msg3 and Msg4 retransmission [84]), which is the power consumed during the processing of the message Msg2 ($T_{p_{Msg2}}P_2$), the power consumed during the transmission of Msg3 (P_3), the power consumed when the MTC is waiting for the acknowledgment (ACK) of Msg3 ($T_{HARQ}P_2$), the power consumed when receiving the ACK of Msg3 (P_2), the power consumed when receiving the Msg4 (P_2), the power consumed after receiving Msg4 and before transmitting the ACK of Msg4 ($T_{HARQ}P_2$), and finally the power consumed for transmitting the ACK of Msg4 (P_3).

Accordingly, the power consumption for the MTC devices successfully accessed the network is:

$$W_S=\frac{T_{RA\_REP}}{2}P_1+(n-1)\Bigg(P_3+T_{RAR}P_1+W_{RAR}P_2+$$

$$\left(\left\lceil \frac{1+T_{RAR}+W_{RAR}+W_{BO}/2}{T_{RA\_REP}} \right\rceil T_{RA\_REP}-1-T_{RAR}-W_{RAR}\right)P_1\Bigg)+P_3+$$

$$T_{RAR}P_1+(W_{RAR}/2)P_2+T_{p_{Msg2}}P_2+P_3+T_{HARQ}P_2+P_2+T_{HARQ}P_2+P_3 \quad (6.41)$$

or

$$W_S=\left(T_{RA\_REP}/2+(n-1)\left\lceil \frac{1+T_{RAR}+W_{RAR}+W_{BO}/2}{T_{RA\_REP}} \right\rceil T_{RA\_REP}\right.$$

$$+T_{RAR}-(n-1)W_{RAR}-(n-1)\Big)P_1+\Big(1+(n-1/2)W_{RAR}$$
$$+2T_{HARQ}+T_{p_{Msg2}}\Big)P_2+(n+2)P_3 \tag{6.42}$$

Regarding the power consumption of the failed MTC devices, it can be deduced directly from the average power consumption of the successful MTC devices, wherein the number of preamble transmissions is the maximum allowed one and there is no transmission of Msg3 and Msg4. Therefore, it is equal to:

$$W_F=\frac{T_{RA\_REP}}{2}P_1+(N_{PT_{max}}-1)\Bigg(P_3+T_{RAR}P_1+W_{RAR}P_2$$
$$+\Bigg(\left\lceil\frac{1+T_{RAR}+W_{RAR}+W_{BO}/2}{T_{RA\_REP}}\right\rceil T_{RA\_REP}-1$$
$$-T_{RAR}-W_{RAR}\Bigg)P_1\Bigg)+P_3+T_{RAR}P_1+W_{RAR}P_2 \tag{6.43}$$

The average power consumption of the dropped MTC devices (only for FI-TSFGP) is equal to;

$$W_D=\frac{T_{RA\_REP}}{2}P_1+(I_{max}-1)T_{RA\_REP}P_0$$

The Power consumption for the total number of MTC devices is the mean power consumption of all the MTC devices, i.e. successful, failed, and dropped, and it is given by the following equation:

$$W=\frac{M_SW_S+M_FW_F+M_DW_D}{M_S+M_F+M_D} \tag{6.44}$$

where M_S, M_F, and M_D are the number of successful, failed, and dropped MTC devices. For the power consumption of FI-TSFGP, it is sufficient to add the value $(k-1)T_{RA\_REP}P_1$, where $k\in[1,I_{max}]$, as the MTC device is waiting for its RA slot identified by the value k.

6.4.2 Results

Fig. 6.18 illustrates the success probability of the four considered methods, i.e. GP, CGP, PBO, and FI-TSFGP. It is clear, from the figure, that CGP introduces an important improvement, compared to GP, when the number of MTC devices in the group is moderate (nearly until 2500). However, the behavior of CGP becomes

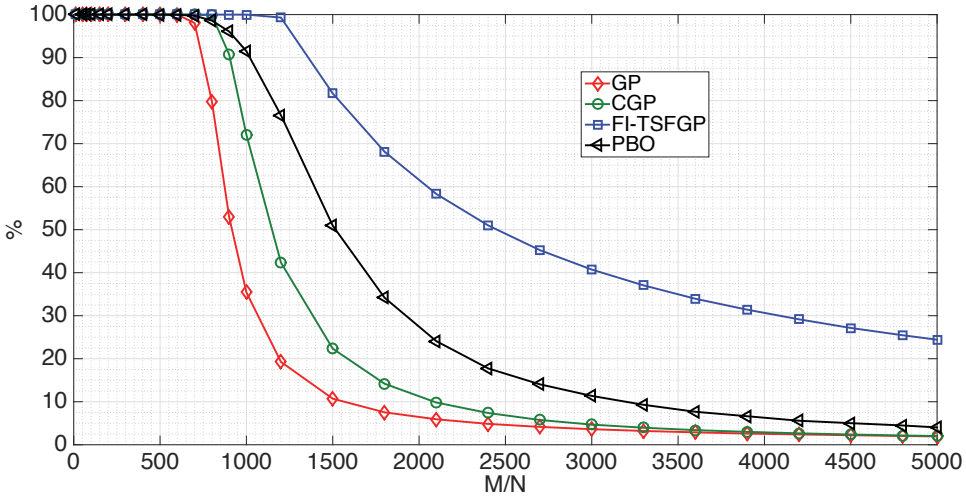

Figure 6.18 Success probability for the considered methods

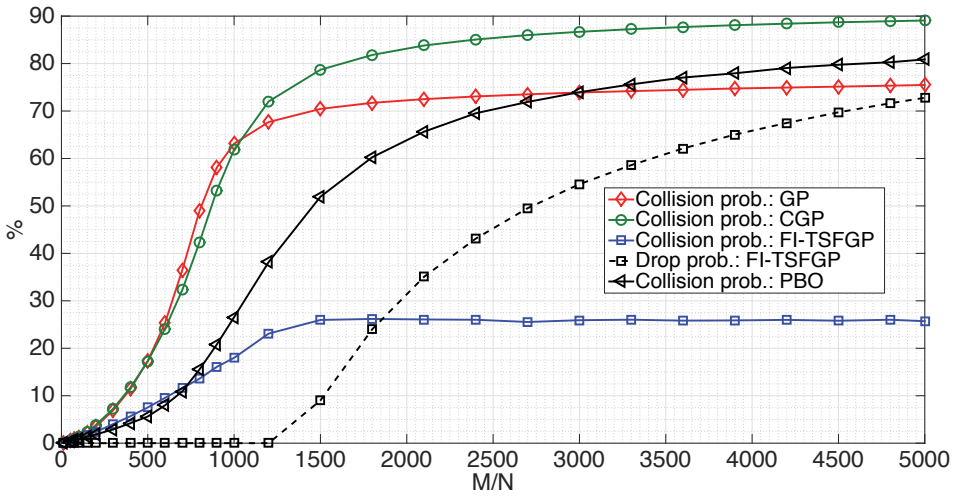

Figure 6.19 Collision and drop probabilities for the considered methods

similar to that of GP when the number of MTC devices becomes larger than 2500, for the considered parameters. For PBO, the figure shows that it outperforms both GP and CGP, regardless the size of the group. However, the success probability becomes small when there is a large number of MTC devices. Concerning FI-TSFGP, it is clear that there is a large improvement, even when the number of MTC devices in each group is large. Note that the success probability for FI-TSFGP is more than 20% when the number of MTC devices is large (e.g., 5000), while it is less than 5% for PBO and the other methods. Furthermore, the collision probability of FI-TSFGP, as illustrated in Fig. 6.19, slightly increases as the number of MTC devices increases, and then remains roughly stable below 30%, while this probability is more than 70% for

GP and more than 85% for CGP. This means that FI-TSFGP achieves a degradation to about the third. Comparing with FI-TSFGP, it can be observed that the collision probability of PBO is similar for a small size of group, with a little improvement brought by PBO. By increasing the number of MTC devices, the collision probability of PBO keeps increasing, and it becomes even worse than GP for a large number of MTC devices (more than 3000 MTC devices for the considered parameters). It

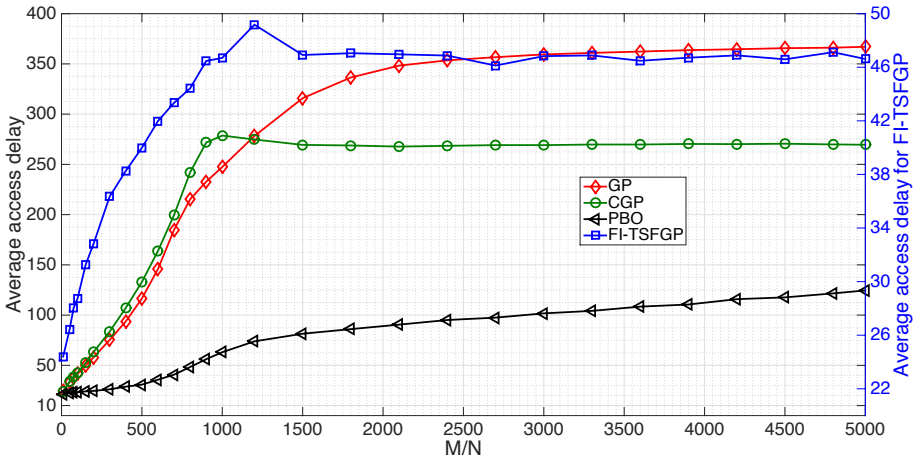

Figure 6.20 Average access delay for the considered methods

is illustrated that CGP achieves an important improvement regarding the average access delay (Fig. 6.20) and the average number of preamble transmissions (Fig. 6.21) by report to GP. However, there is some price on this improvement, given that a part of MTC devices has access delay and number of preamble transmissions higher than that in GP method, (Figs. 6.22 and 6.23). Regarding PBO, it improves these performances by considerably reducing the average access delay and the average number of preamble transmissions. Again, it can be seen that the average access delay and the average number of preamble transmissions of FI-TSFGP are similar to that of PBO for a small group size, with a small improvement brought by PBO. However, FI-TSFGP's performances become better when increasing the number of MTC devices. An important observation can be also seen from Figs. 6.20 and 6.21, wherein the average (access delay/preamble transmission) of FI-TSFGP becomes constant after certain size of the group (more than 1500 for the considered parameters), while these values are increasing with the number of MTC devices for PBO. Regarding the CDF of (access delay/preamble transmission), it can be seen that PBO outperforms GP and CGP regardless the number of MTC devices, while it slightly outperforms FI-TSFGP only for the case of small group sizes. However, FI-TSFGP outperforms all the considered methods, including PBO, for larger group sizes, where the achieved gain can reach

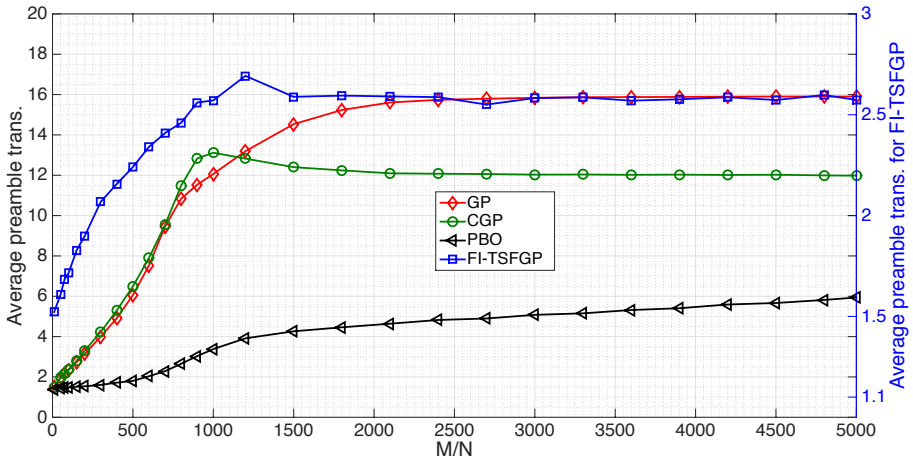

Figure 6.21 Average preamble transmission for the considered methods

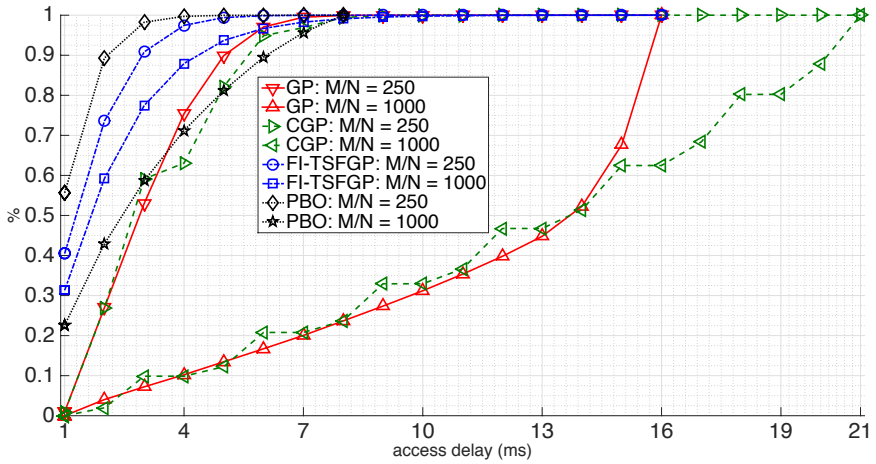

Figure 6.22 CDF of Preamble transmissions

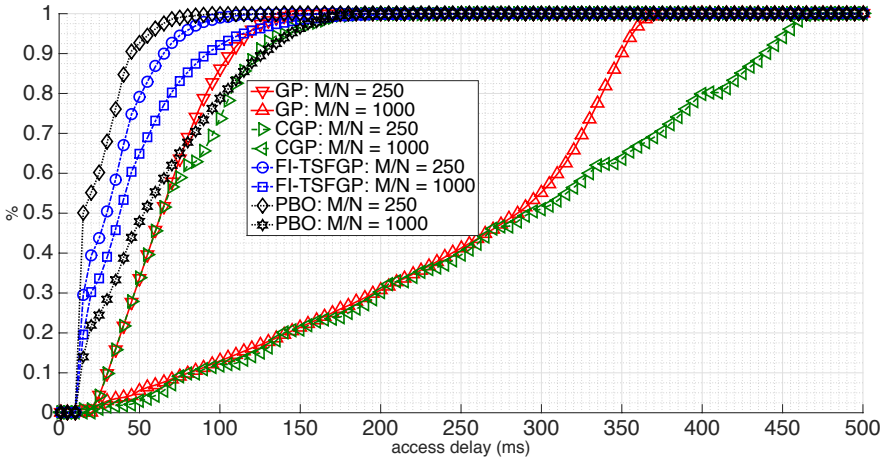

Figure 6.23 CDF of access delay

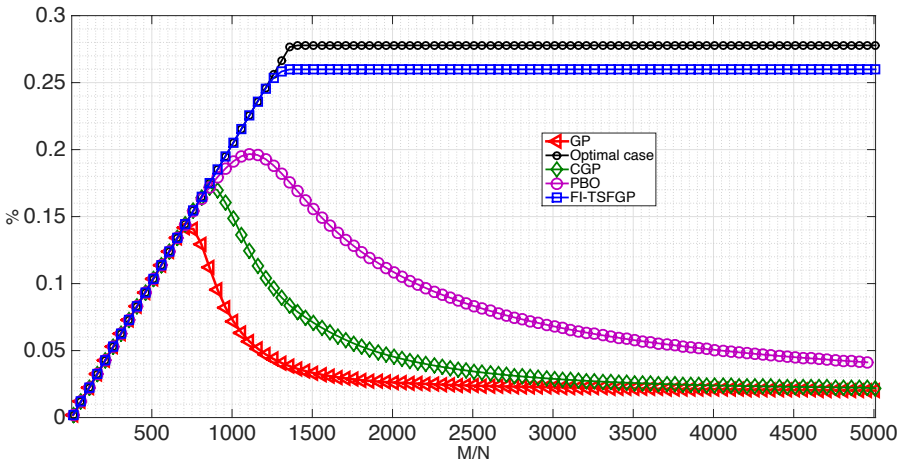

Figure 6.24 Resource utilization for the considered methods

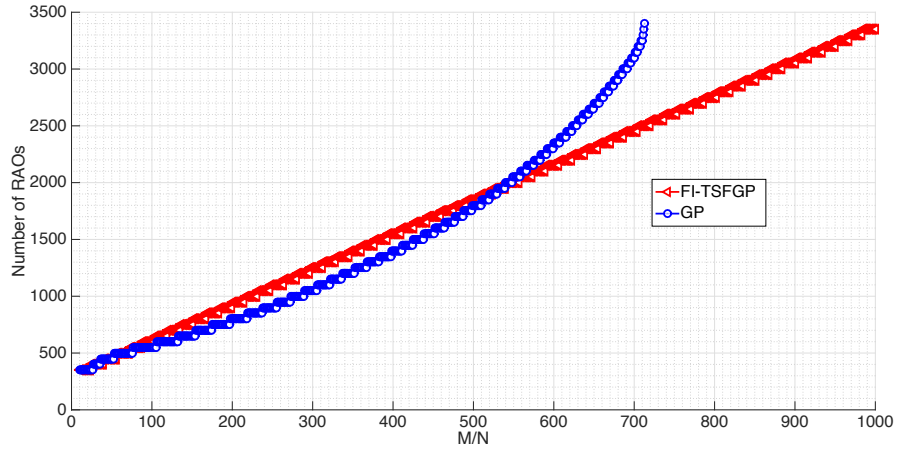

Figure 6.25 Minimum resources required in order to achieve 90% of success probability

more than 15% for CDF of preamble transmission and more than %40 for CDF of access delay. It should be noted that the number of preamble transmissions needed to access the network and thus the time required to get access have a close relation with the power consumption. Therefore, FI-TSFGP introduces a large reduction of the power consumption (as shown later), which is a very important achievement, especially for those with a limited power resources. Furthermore, FI-TSFGP largely reduces the access delay, which is an important issue for the time-critical MTC applications, for example.

Looking at the resource utilization, Fig. 6.24 shows again that the CGP achieves some improvement when the number of MTC devices is somewhat moderate, while the behavior becomes nearly the same as of GP when the number of MTC devices is large. As for the precedent performance metrics, PBO method has a better resource

utilization, compared to GP and CGP. But, this utilization decreases when exceeding the number of MTC devices after 1000 M/N. FI-TSFGP achieves a high percentage of resource utilization, similar to that of the ideal case. Note that the latter represents the situation where the total number of arrivals engenders a number of successful MTC devices that is equal to N_{ACK}, i.e. the number of MTC devices that the network can acknowledge within the RAR window. Furthermore, it can be remarked that there is a small difference between FI-TSFGP and the ideal case, which is about 2.5% when the number of MTC devices is large, while it is more than 20% for GP, CGP, and PBO. From Fig. 6.24, it can be observed that FI-TSFGP maintains a stable resource utilization regardless the size of the group. This means that FI-TSFGP achieves a constant number of successful MTC devices whatever the group size, while the other methods fail to do that. Another improvement gained by FI-TSFGP is the minimum resources to achieve 90% of success probability. Fig. 6.25 shows the relationship between the required resources to achieve 90% of success probability and the number of MTC devices. From this figure, it can be seen that the required resources for FI-TSFGP is more than that for GP, when the number of MTC devices is small. To better explain this behavior, we return to Fig. 6.8, where we clearly see that the number of successful MTC devices, at the start of the group paging interval, is not equal to that value in the stable state. Therefore, when there is a small number of MTC devices, the average number of successful MTC devices at each RA slot will be relatively low (compared to the reserved resources). Generally, the higher is the number of MTC devices, the higher is the average number of successful MTC devices. However, the relation shown in Fig. 6.25 can be approximated to an exponential one for GP, and a linear one for FI-TSFGP. This advantage is very important, as we can achieve the same percentage of success with much more less of resources, especially with the existence of a very large number of MTC devices.

Figs. 6.26 and 6.27 illustrate the power consumption of the successful MTC devices and that of the total number of MTC devices, and the power consumption of the failed MTC devices and the dropped (only for FI-TSFGP) ones, respectively. From Fig. 6.26, it can be observed that the power consumption of successful MTC devices for CGP is smaller than that of GP. This is expected as the average number of preamble transmissions of CGP is smaller than that of GP in the presence of a large number of MTC devices. However, GP outperforms CGP when considering the power consumption of failed and total number of MTC devices (Figs. 6.26 and 6.27). This can be argued by the fact that the collision probability and the total number of preamble transmissions are larger for CGP, where the total number of preamble transmissions is 21 for CGP

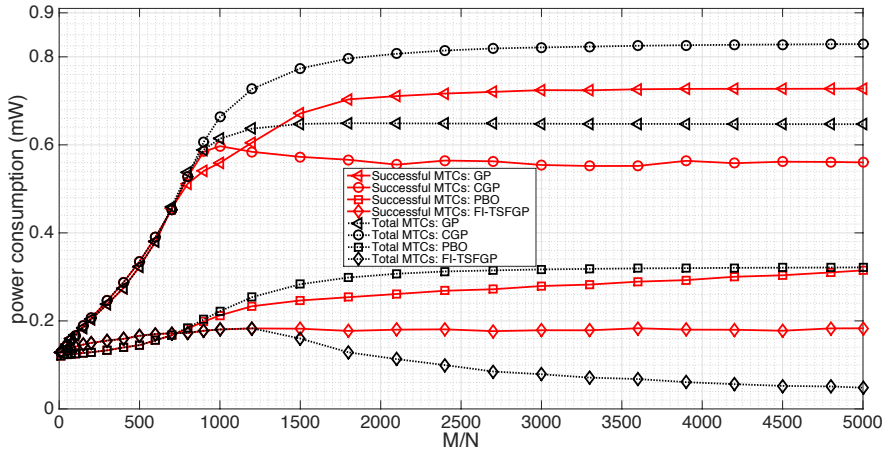

Figure 6.26 Power consumption of the successful MTC devices and that of the total MTC devices

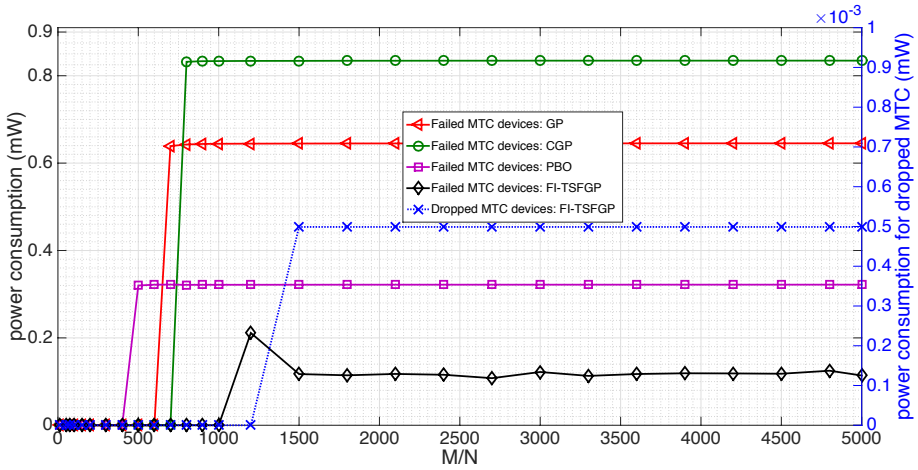

Figure 6.27 Power consumption of the failed and dropped MTC devices

and 16 for GP, as shown in Fig. 6.22. For the same reasons, PBO highly outperforms GP and CGP regarding the three considered values of power consumption, and thus highly conserves the energy. Regarding FI-TSFGP, it highly outperforms both GP and CGP for all the considered values. Compared with PBO, the power consumption of FI-TSFGP is similar for small group sizes (with a small difference), while FI-TSFGP outperforms PBO for all the considered values when increasing the number of MTC devices. From Fig. 6.26, it is clear that the power consumption for successful MTC devices is less than $0.2\,mW$ for FI-TSFGP, while it is about $0.55\,mW$ for CGP, more than $0.70\,mW$ for GP, and about $0.30\,mW$ for PBO for a large number of MTC devices. Another important improvement is the average power consumption of the total MTC devices. From Fig. 6.26, it is observed that the average power consumption for GP,

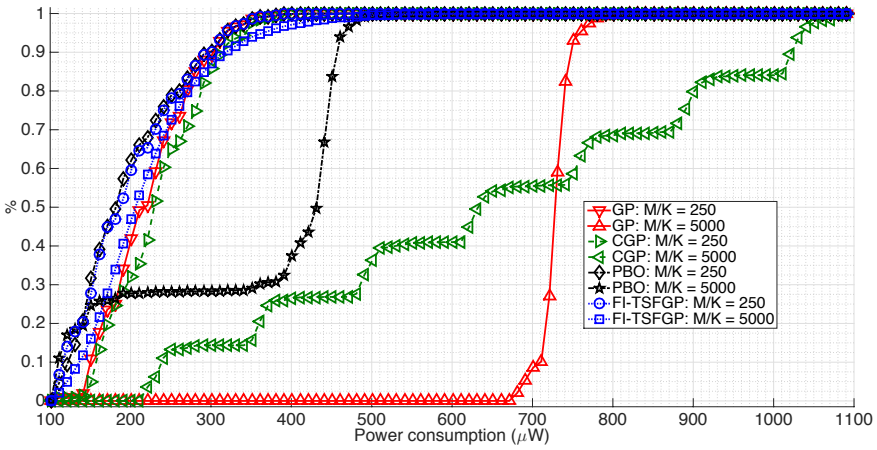

Figure 6.28 CDF of power consumption for the successful MTC devices

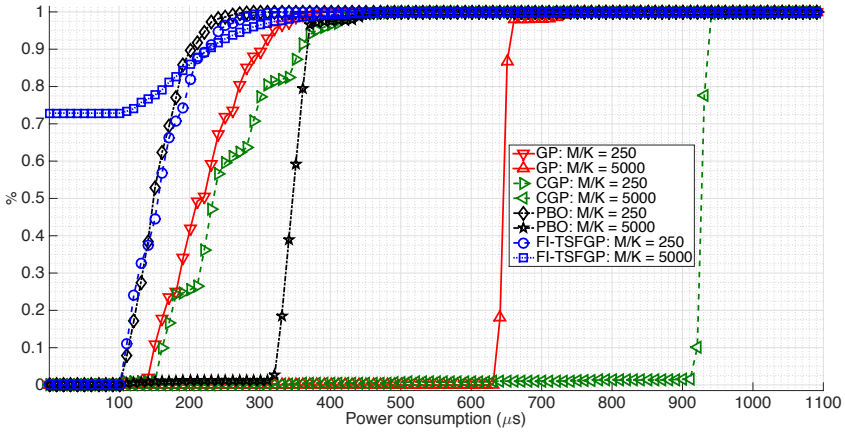

Figure 6.29 CDF of power consumption for the total number of MTC devices

CGP, and PBO increases as the number of MTC devices increases, and then it becomes merely stable (or so slowly increases) when the number of MTC devices becomes large. These values is about $0.65\,mW$ for GP, more than $0.80\,mW$ for CGP, and about $0.3\,mW$ for PBO. However, the average power consumption for FI-TSFGP firstly increases as the number of MTC devices increases, and then it decreases. The decreasing behavior of the average power consumption for FI-TSFGP can be justified by the fact that there is a part of MTC devices that are dropped, i.e. they come back to inactive state. Indeed, when go idle, the dropped devices consume a very small amount of power, compared to the activated ones, and their numbers would be increased when increasing the number of MTC devices. Therefore, the average power consumption of the total number of MTC devices logically decreases as the number of MTC devices increases.

To further show the effectiveness of FI-TSFGP, Figs. 6.28 and 6.29 illustrate the CDF of the power consumption for the successful MTC devices and the total number

of MTC devices, respectively. From Fig. 6.28, the superiority of PBO compared to GP and CGP is clear, while it introduces some improvement by report to FI-TSFGP only for small group sizes. For a large number of MTC devices (e.g., 5000), it can be clearly seen that more than 90 % of the MTC devices consume only 0.3 $mWatt$ for FI-TSFGP method, while they consume more than 0.7 $mWatt$ for the GP method, more than 1 $mWatt$ for the CGP method, and more than 0.4 $mWatt$ for PBO method.

Besides the superiority of FI-TSFGP, shown by Fig. 6.29 for the CDF of power consumption of the total number of MTC devices, a specific behavior to the FI-TSFGP method can be observed. This behavior is that the percentage of the total number of MTC devices consuming certain power level augments as the number of MTC devices increases. This benefit of FI-TSFGP can be justified by the fact that the network activates certain number of MTC devices, while the others go back to the inactive state in which the MTC devices consume the minimum power level. Therefore, the higher number of MTC devices is, the higher percentage of MTC devices consuming a certain power level is. Taking into account the fact that the number of MTC devices is naturally large, it can be concluded that the proposed method FI-TSFGP outperforms the other methods for all the considered parameters. Based on the aforementioned results, especially the ones concerned the power consumption, it can be said that the proposed method FI-TSFGP is very attractive for battery-limited MTC devices deployment.

6.5 Conclusions

Although there are significant benefits in M2M services, their deployment in the current cellular mobile networks is faced by many challenges. Some of these challenges are the scarce resources and the power consumption of a myriad of devices, accompanying by congestion and system overload. Group Paging (GP) method was proposed to tackle with the problem of congestion in LTE and LTE-A networks. In spite of its performances, GP is characterized by high power consumption and low resources utilization in the presence of a large number of MTC devices. In order to overcome the aforementioned problem, we proposed two new methods; Controlled Distribution of Resource (CDR), and Further Improvement-TSFGP (FI-TSFGP).

The CDR mechanism, which is an improvement of GP method, is intended for the case where the MTC devices are in the RRC_CONNECTED OUT_OF_SYNC state. its main advantage is that the RACH procedure becomes a contention-free one. Besides reducing the collisions and increasing channel access probabilities, CDR allows

to reduce the power consumption, which is a very important issue for MTC devices with limited power resources. Moreover, the resource utilization of CDR varies between 50% (for the worst case) and 100% (for the best case), while it is 20%, at most, for the ordinary GP.

However, the disadvantage of CDR is that it is dedicated for the MTC devices in the RRC_CONNECTED_OUT_OF_SYNC state, and ignores the MTC devices that are in the IDLE state. Therefore, FI-TSFGP was proposed to overcome the aforementioned problem and improve the performance of GP too. FI-TSFGP has been evaluated for a relatively large number of MTC devices (5000 MTC devices). It has outperformed both the Group Paging (GP) and the Consecutive Group Paging (CGP) methods, for all the considered metrics. Compared with PBO, FI-TSFGP method has a similar performance for a low number of MTC devices, while it outperforms PBO for a large number of MTC devices. Besides the access delay and the average number of preamble transmissions improvements, FI-TSFGP highly reduces the power consumption for both the successful MTC devices and also for the total number of MTC devices, which is one of the key objectives of $5G$ systems. Moreover, FI-TSFGP maintains a stable resource utilization when existing a large number of MTC devices, meaning that the number of successful MTC devices is maintained regardless the group size. Finally, FI-TSFGP achieves the same percentage of success probability for MTC with a much more less of resources, preserving thus the network resources, which can be used by Non-MTC devices, for example.

Chapter 7

Conclusions and Perspectives

M2M applications will enter nearly every sector, imaginable and unimaginable (by the time being), in our everyday life. Few examples of M2M are Smart grid, industrial and environment monitoring, eHealth, and Intelligent Transport System (ITS). This diversity of M2M applications will surely lead to real deployment of IoT vision and also bring the smart city vision to the light. As a result, M2M has been taking a lot of attention of many operators and vendors, since it represents a new and very attractive income. Unfortunately, M2M deployment can not be done before overcoming challenging issues facing the introduction of M2M in the current cellular mobile networks, which represent the most relevant network candidate to host M2M-based services. This lack of possibility comes from the fact that the current networks are designed and optimized for H2H deployment, which is highly different from M2M deployment. The first difference is the presence of a very large number of MTC devices in each cell ($100K$ devices per m^2 in 4G networks and $1M$ devices per m^2 in 5G networks), while the number of H2H devices is, at most, equal to the population's number on the earth (7.4 billion). Besides, traffic behavior of MTC is different from that of H2H, such as the ratio (uplink/downlink) which is higher for M2M whereas it is the contrary for H2H. The power consumption is another concern, since M2M devices are battery-equipped.

In this thesis, many algorithms and methods were proposed and explored in order, on one hand, to tackle with the congestion problem and, on the another hand, to optimize the resource utilization and power consumption.

7.1 Results Obtained During the Thesis

The contributions of the thesis were divided into three parts:

- **An algorithm for traffic prediction:** This algorithm, namely Multi-Channel Slotted ALOHA-Optimal Estimation (MCSA-OE), allows for a better prediction, compared to other methods in the literature, of the current number of arrivals. This prediction, in turn, is used to throttle (or control) the number of MTC that may arrive in the next period, which further permits to better utilization of resources. This algorithm is composed of two phases: *i*) estimation and fitting the idle probability, *ii*) fitting the estimated number of arrivals. The proposed algorithms were tested on two phases: *i*) the case when the network parameters (more specifically the parameters of ACB method) are fix, *ii*) the case when the parameters of ACB method are updated according to the estimated number of arrivals. The obtained results were encouraging. The proposed method MCSA-OE well tracks the number of arrivals as long as they are smaller than or equal to $\ln\left(R^R\right)$, where R is the number of available channels. Note that this limitation would not be reached when applying MCSA-OE with some congestion control method like Access Class Barring (ACB). Moreover, this method has better performances regarding many parameters like the access success probability and resource utilization. They demonstrate the ability of MCSA - OE to well estimate the number of arrivals even in heavy traffic cases like Beta traffic.
- **A general model for RACH Procedure:** The objective of this contribution was to model the performance of the RACH procedure. The proposed model, namely General Recursive Estimation (GRE), relies on the recursive estimation of the number of arrivals at each time, and thus determining all the relevant parameters, such as the success and collision probabilities, the access delay, and the number of preamble transmission. The advantage of GRE is that it is traffic-independent. Therefore, it can be applied with any traffic model, such as Poisson and Uniform distributions. However, the model was tested on Beta traffic model, since it is considered as the worst traffic model for M2M. Simulation results showed that the proposed analytical model is accurate, where the relative error is less than one percent on the average access delay and average number of preamble transmissions, and less than two percent on the success probability. Another advantage of GRE is that it is complied with the LTE/MTC standard. Thus, there is no need for any change to the assumption already taken by the 3GPP on MTC traffic.
- **Group paging optimization and power efficiency:** In this context, two methods were proposed: *i*) Controlled Distribution of Resource (CDR), *ii*) Further Improvement-TSFGP (FI-TSFGP). The CDR method permits to uniquely

distribute the resources to the MTC devices such that the collision probability is zero, i.e. the RACH procedure becomes like contention-free, and consequently the success probability is always 100%. The reduction of power consumption in CDR method is achieved by reducing the number of attempts to get access the network to the minimum, where each device will send an attach request only once. Regarding FI - TSFGP, it scatters the members of the intended group during the available GP interval, leading to reduction in the collision probability and augmentation in the access success probability. Compared to GP and two other methods, FI - TSFGP introduces high improvement regarding all the considered parameters, such as the collision and success probabilities, the average access delay and the average number of preamble transmission, and the power consumption. For example, the collision probability is reduced from more than 60%, for GP method, to less than 30% for FI-TSFGP, while there is a gain to about 20% on the success probability and that for lager group sizes (e.g., 5000 MTC devices). Another example, there is a reduction in the power consumption to more than twice for the successful MTC devices, and more than five times regarding the total number of MTC devices. Not only FI-TSFGP method introduces the aforementioned improvements, but also it gives an efficient way tp determine the network's parameters that maximize the performance of the network.

7.2 Perspectives

To extend our works, the future research directions will cover the following axis:

- **Propose a comprehensive model of congestion control:** The proposed methods in the literature so far tackle the congestion issue in the RAN, CN, or partially in the two parts. However, we would like to build a general model, for the problem in question, that includes not only RAN and CN parts, but also the traffic coming from MTC servers which is in the form of triggering. Taking into account these three parts, the envisioned model will consider the combination of many traffic models, such as Beta and uniform distributions, so that the congestion and resource management problems would be investigated. The final goal of the envisioned model would be network's optimization as a whole, not just a partial solution.
- **Propose a 5G - complied MAC protocol:** In this direction, MAC protocol that obeys 5G specifications and IoT-friendly at the same time will be investigated.

The random access in 5G may be somewhat different, since a new technology, mmWave beamforming, would be used. Although there are some works on these issues, such as [124, 125], there is still a lot of work to do on the random access in the context of 5G. Moreover, M2M devices should take a special attention as they may not use all the supported 5G technologies because of the constraints on the power and the cost of devices.

- **Propose a method of grouping based on dynamic clustering:** So far, the proposed methods in the literature, in the context of M2M, are done by the assumption that the devices are fix. However, an advanced model should be introduced in order to group not only the MTC devices with a fixed location, but also the moving ones. The envisioned model would rely on both Device - to - Device (D2D) communication and the communication to cellular networks.

Publications from the thesis

- International Journals:
  1. *O. Arouk*, A. Ksentini, and T. Taleb, **"Group paging-based energy saving for massive MTC accesses in LTE and beyond networks,"**, in Selected Areas in Communications, IEEE Journal on , vol.PP, no.99, pp.1-1
  2. *O. Arouk* and A. Ksentini, **"General Model for RACH Procedure Performance Analysis,"** IEEE Commun. Letter, vol.PP, no.99, pp.1-1.
- International Conferences:
  3. *O. Arouk*, A. Ksentini, and T. Taleb, **"Performance analysis of RACH procedure with Beta Traffic-Activated Machine-Type- Communication,"** in IEEE Global Communications Conference (GC 2015): Adhoc and Sensor Networks Symposium (GC' 15 - Adhoc and Sensor Networks), San Diego, USA, Dec. 2015.
  4. *O. Arouk*, A. Ksentini, and T. Taleb, **"Group paging optimization for Machine-Type-Communications,"** in IEEE ICC 2015 - Ad-hoc and Sensor Networking Symposium (ICC'15 (09) AHSN), London, United Kingdom, Jun. 2015, pp. 8128-8133.
  5. *O. Arouk*, and A. Ksentini, **"Multi-Channel slotted aloha optimization for Machine-Type-Communication,"** in 17th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM'14), Montreal, Canada, Sep. 2014, pp. 119-125.
  6. *O. Arouk*, A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, **"On improving the group paging method for Machine-type-Communications,"** in IEEE ICC 2014 - Ad-hoc and Sensor Networking Symposium (ICC'14 AHSN), Sydney, Australia, Jun. 2014, pp. 484-489.

Bibliography

[1] U. Phuyal, A. T. Koc, M.-H. Fong, and R. Vannithamby, "Controlling access overload and signaling congestion in M2M networks," in *Signals, Systems and Computers (ASILOMAR), 2012 Conference Record of the Forty Sixth Asilomar Conference on*, 2012, pp. 591–595.

[2] 3GPP TS 36.211 V11.1.0, "Physical Channels and Modulation," December 2012.

[3] 3GPP TR 36.912, "Feasibility study for Further Advancements for E-UTRA (LTE-Advanced)," September 2012.

[4] 3GPP TS 23.682 V11.4.0, "Architecture enhancements to facilitate communications with packet data networks and applications," June 2013.

[5] S.-Y. Lien, T.-H. Liau, C.-Y. Kao, and K.-C. Chen, "Cooperative Access Class Barring for Machine-to-Machine Communications," *Wireless Communications, IEEE Transactions on*, vol. 11, no. 1, pp. 27–32, 2012.

[6] 3GPP R2-113083, "RAN overload handling," MediaTek Inc., RAN2#74, 2011.

[7] "World's First Connected Toothbrush Will Keep Cavities Away," http://mashable.com/2014/01/05/kolibree-connected-toothbrush/.

[8] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599–616.

[9] IEEE 802.16ppc-10/0002r7, "Machine to Machine (M2M) Communication Study Report," March 2010.

[10] A. Maeder, P. Rost, and D. Staehle, "The challenge of m2m communications for the cellular radio access network," in *11th WÃŒrzburg Workshop on IP: Joint ITG and Euro-NF Workshop" Visions of Future Generation Networks"(EuroView2011)*, 2011.

[11] Changwei, Li, "Telco development trends and operator strategies," *WinWin Magazine*, no. 13, pp. 19–22, July 2012.

[12] SoftBank CEO: the average person will have 1,000 internet-connected devices by 2040. [Online]. Available: https://www.techinasia.com/softbank-son-iot-1000-devices-2040/

[13] "The Internet of Everything Global Private Sector Economic Analysis," http://www.cisco.com/web/about/ac79/docs/innov/IoE-Economy-FAQ.pdf, 2013.

[14] C. Kim, A. Soong, M. Tseng, and X. Zhixian, "Global Wireless Machine-to-Machine Standardization," *Internet Computing, IEEE*, vol. 15, no. 2, pp. 64–69, 2011.

[15] 3GPP TR 36.888 V12.0.0, "Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE," June 2013.

[16] M. Beale, "Future challenges in efficiently supporting M2M in the LTE standards," in *Wireless Communications and Networking Conference Workshops (WCNCW), 2012 IEEE*, April 2012, pp. 186–190.

[17] O. Arouk and A. Ksentini, "Multi-channel Slotted Aloha Optimization for Machine-type-communication," in *Proceedings of the 17th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, ser. MSWiM '14. New York, NY, USA: ACM, 2014, pp. 119–125. [Online]. Available: http://doi.acm.org/10.1145/2641798.2641802

[18] O. Arouk, A. Ksentini, and T. Taleb, "Performance Analysis of RACH procedure with Beta Traffic-Activated Machine-Type-Communication," in *IEEE Global Communications Conference (GC 2015): Adhoc and Sensor Networks Symposium (GC' 15 - Adhoc and Sensor Networks)*, San Diego, USA, dec 2015.

[19] O. Arouk and K. Adlen, "General model for rach procedure performance analysis," *Communications Letters, IEEE*, vol. PP, no. 99, pp. 1–1, 2015.

[20] O. Arouk, A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb, "On improving the group paging method for machine-type-communications," in *Communications (ICC), 2014 IEEE International Conference on*, June 2014, pp. 484–489.

[21] 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.

[22] O. Arrouk, A. Ksentini, and T. Taleb, "Group Paging-based Energy Saving for Massive MTC Accesses in LTE and Beyond Networks," *Selected Areas in Communications, IEEE Journal on*, vol. PP, no. 99, pp. 1–1, 2016.

[23] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "A first look at cellular Machine-to-Machine traffic: large scale measurement and characterization," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 1, pp. 65–76, 2012.

[24] "oneM2M Use cases collection," September 2013.

[25] ETSI TR 102 898 V1.1.1, "Machine to Machine communications (M2M); Use cases of Automotive Applications in M2M capable networks," April 2013.

[26] ETSI TR 102 857 V1.1.1, "Machine-to-Machine communications (M2M); Use Cases of M2M applications for Connected Consumer," August 2013.

[27] ETSI TR 102 691 V1.1.1, "Smart Metering Use Cases," May 2010.

[28] ETSI TR 102 935 V1.1.1, "Impact of Smart Grids on M2M platform," September 2012.

[29] gemalto. http://www.gemalto.com/.

[30] ETSI TR 102 732 V2.1.1, "Use Cases of M2M applications for eHealth," September 2013.

[31] ETSI TR 102 898 V1.1.1, "Use Cases of Automotive Applications in M2M capable networks," April 2013.

[32] H. Schulzrinne, H. Tschofenig, A. Newton, and T. Hardie, "LoST: A Protocol for Mapping Geographic Locations to Public Safety Answering Points," in *Performance, Computing, and Communications Conference, 2007. IPCCC 2007. IEEE Internationa*, April 2007, pp. 606–611.

[33] J. N. Al-Karaki and A. E. Kamal, "Routing techniques in wireless sensor networks: a survey," *Wireless Communications, IEEE*, vol. 11, no. 6, pp. 6–28, 2004.

[34] "Focus group on m2m service layer," http://www.itu.int/en/ITU-T/focusgroups/m2m/Pages/default.aspx.

[35] Machine to Machine Communications. http://www.etsi.org/technologies-clusters/technologies/m2m.

[36] IEEE 802.16's Machine-to-Machine (M2M) Task Group. http://www.ieee802.org/16/m2m/.

[37] "oneM2M," http://www.onem2m.org.

[38] oneM2M, Why Join? http://www.onem2m.org/whyjoin.cfm.

[39] Leading ICT Standards Development Organizations Launch oneM2M. http://www.etsi.org/news-events/news/401-news-release-24-july-2012.

[40] A. Kunz, K. LaeYoung, K. Hyunsook, and S. S. Husain, "Machine type communications in 3GPP: From release 10 to release 12," in *Globecom Workshops (GC Wkshps), 2012 IEEE*, 2012, pp. 1747–1752.

[41] 3GPP TR 22.368 V12.2.0, "Service requirements for Machine-type Communications (MTC)," March 2013.

[42] 3GPP TR 21.905 V12.0.0, "Vocabulary for 3GPP Specifications," June 2012.

[43] 3GPP TR 23.888 V11.0.0, "System improvements for Machine-Type-Communications (MTC)," September 2012.

[44] 3GPP R2-102781, "Paging and downlink transmission for MTC," CATT, RAN2#70, May 2010.

[45] 3GPP TR 37.868 V11.0.0, "Study on RAN improvement for Machine-type-Communications," September 2012.

[46] 3GPP R2-104873, "Comparing Push and Pull based Approaches for MTC," Institute for Information Industry (III), Coiler Corporation, RAN2#71, August 2010.

[47] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*.

[48] 3GPP TS 36.331 V12.7.0, "Radio Resource Control (RRC); Protocol specification," September 2015.

[49] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description," 2013.

[50] L. Sanguinetti and M. Morelli, "An Initial Ranging Scheme for the IEEE 802.16 OFDMA Uplink," *Wireless Communications, IEEE Transactions on*, vol. 11, no. 9, pp. 3204–3215, 2012.

[51] C.-L. Lin and S.-L. Su, "A robust ranging detection with MAI cancellation for OFDMA systems," in *Advanced Communication Technology (ICACT), 2011 13th International Conference on*, Feb 2011, pp. 937–941.

[52] H. Minn and X. Fu, "A new ranging method for OFDMA systems," in *Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE*, vol. 3, Nov 2005, pp. 6 pp.–.

[53] L. Sanguinetti, M. Morelli, and L. Marchetti, "A random access algorithm for lte systems," *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 1, pp. 49–58, 2013. [Online]. Available: http://dx.doi.org/10.1002/ett.2575

[54] 3GPP TS 36.321 V12.7.0, "Medium Access Control (MAC) protocol specification," September 2015.

[55] R.-G. Cheng, C.-H. Wei, S.-L. Tsao, and F.-C. Ren, "Rach collision probability for machine-type communications," in *Vehicular Technology Conference (VTC Spring), 2012 IEEE 75th*, May 2012, pp. 1–5.

[56] S. Sesia, I. Toufik, and M. Baker, *LTE: the UMTS long term evolution*. Wiley Online Library, 2009.

[57] 3GPP TS 36.213 V12.0.0, "Physical layer procedures," December 2013.

[58] 3GPP TS 36.101 V12.6.0, "User Equipment (UE) radio transmission and reception," December 2014.

[59] H. L. Bertoni, *Radio Propagation for Modern Wireless Systems*. Prentice Hall Professional Technical Reference, 1999.

[60] J. G. Andrews, A. Ghosh, and R. Muhamed, *Fundamentals of WiMAX: Understanding Broadband Wireless Networking*. Pearson Education, 2007.

[61] 3GPP TS 29.368 V11.3.0, "Tsp interface protocol between the MTC InterWorking Function (MTC-IWF) and Service Capability Server (SCS)," June 2013.

[62] 3GPP TS 23.401, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," December 2013.

[63] Y. Pan, Q. Luo, G. Li, Y. Zhao, and Z. Xiong, "Performance Evaluation of 3D MIMO LTE-Advanced System," in *Vehicular Technology Conference (VTC Fall), 2013 IEEE 78th*, 2013, pp. 1–5.

[64] E. Seidel, "3GPP LTE-A Standardisation in Release 12 and Beyond," Nomor Research GmbH, Munich, Germany, Tech. Rep., January 2013.

[65] H. Thomsen, N. K. Pratas, C. Stefanovic, and P. Popovski, "Code-expanded radio access protocol for machine-to-machine communications," *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 4, pp. 355–365, 2013. [Online]. Available: http://dx.doi.org/10.1002/ett.2656

[66] W. Geng, S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson, "M2M: From mobile to embedded internet," *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 36–43, 2011.

[67] Y. W. Blankenship, "Achieving high capacity with small cells in LTE-A," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, 2012, pp. 1680–1687.

[68] M.-Y. Cheng, G.-Y. Lin, H.-Y. Wei, and A. C.-C. Hsu, "Overload control for Machine-Type-Communications in LTE-Advanced system," *Communications Magazine, IEEE*, vol. 50, no. 6, pp. 38–45, 2012.

[69] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches," *Communications Magazine, IEEE*, vol. 51, no. 6, pp. 86–93, 2013.

[70] 3GPP R2-112863, "Backoff enhancements for RAN overload control," ZTE, RAN2#73, 2011.

[71] A. Lo, Y. W. Law, M. Jacobsson, and M. Kucharzak, "Enhanced LTE-advanced random-access mechanism for massive machine-to-machine (M2M) communications," in *27th World Wireless Research Forum (WWRF) Meeting*, 2011, pp. 1–5.

[72] 3GPP R2-113197, "Performance comparison of access class barring and MTC specific backoff schemes for MTC," May 2011.

[73] 3GPP R2-104662, "MTC simulation results with specific solutions," ZTE, RAN2#71, August 2010.

[74] 3GPP R2-112071, "Evaluation on push based RAN overload control schemes," 2011.

[75] K. Kab Seok, K. Min Jeong, B. Kuk Yeol, S. Dan Keun, K. Jae Heung, and A. Jae Young, "A Novel Random Access for Fixed-Location Machine-to-Machine Communications in OFDMA Based Systems," *Communications Letters, IEEE*, vol. 16, no. 9, pp. 1428–1431, 2012.

[76] H. Wu, C. Zhu, R. La, X. Liu, and Y. Zhang, "FASA: Accelerated S-ALOHA Using Access History for Event-Driven M2M Communications," *Networking, IEEE/ACM Transactions on*, vol. 21, no. 6, pp. 1904–1917, 2013.

[77] Z. Jiang and X. Zhong, "Fast Retrial and Dynamic Access Control Algorithm for LTE-Advanced Based M2M Network," in *AICT 2012, The Eighth Advanced International Conference on Telecommunications*, 2012, pp. 24–28.

[78] L. Ki-Dong, K. Sang, and Y. Byung, "Throughput comparison of random access methods for M2M service over LTE networks," in *GLOBECOM Workshops (GC Wkshps), 2011 IEEE*, 2012, pp. 373–377.

[79] 3GPP R2-104663, "[70bis#11] LTE: MTC LTE simulations," ZTE, RAN2#71, August 2010.

[80] T. Ang-Hsun, W. Li-Chun, H. Jane-Hwa, and L. Tzu-Ming, "Overload Control for Machine Type Communications with Femtocells," in *Vehicular Technology Conference (VTC Fall), 2012 IEEE*, 2012, pp. 1–5.

[81] J.-P. Cheng, C.-h. Lee, and T.-M. Lin, "Prioritized Random Access with dynamic access barring for RAN overload in 3GPP LTE-A networks," in *GLOBECOM Workshops (GC Wkshps), 2011 IEEE*, pp. 368–372.

[82] 3GPP R2-113198, "Further analysis of group paging for MTC," May 2011.

[83] 3GPP R2-104870, "Pull based RAN overload control," August 2010.

[84] C.-H. Wei, R.-G. Cheng, and S.-L. Tsao, "Performance Analysis of Group Paging for Machine-Type Communications in LTE Networks," *Vehicular Technology, IEEE Transactions on*, vol. 62, no. 7, pp. 3371–3382, 2013.

[85] S.-H. Wang, H.-J. Su, H.-Y. Hsieh, S.-p. Yeh, and M. Ho, "Random access design for clustered wireless machine to machine networks," in *Communications and Networking (BlackSeaCom), 2013 First International Black Sea Conference on*, 2013, pp. 107–111.

[86] L. Hyun-kwan, K. Dong Min, H. Youngju, Y. Seung Min, and K. Seong-Lyun, "Feasibility of cognitive machine-to-machine communication using cellular bands," *Wireless Communications, IEEE*, vol. 20, no. 2, pp. 97–103, 2013.

[87] J. Tzu-Chuan, W. Shih-En, and H. Hung-Yun, "Data-centric clustering for data gathering in machine-to-machine wireless networks," in *Communications Workshops (ICC), 2013 IEEE International Conference on*, 2013, pp. 89–94.

[88] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 66–74, 2011.

[89] G. V12.0.0, "Technical Specification Group Services and System Aspects; Service accessibility," 2013.

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

[91] T. Taleb and A. Kunz, "Machine type communications in 3GPP networks: potential, challenges, and solutions," *Communications Magazine, IEEE*, vol. 50, no. 3, pp. 178–184, 2012.

[92] T. Taleb and A. Ksentini, "On alleviating MTC overload in EPS," *Ad Hoc Networks*, no. 0, 2013.

[93] K. Samdanis, A. Kunz, M. I. Hossain, and T. Taleb, "Virtual bearer management for efficient MTC radio and backhaul sharing in LTE networks," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on*, Sept 2013, pp. 2780–2785.

[94] A. Ksentini, T. Taleb, X. Ge, and H. Honglin, "Congestion-aware MTC device triggering," in *Communications (ICC), 2014 IEEE International Conference on*, June 2014, pp. 294–298.

[95] S. Dongxu and V. O. K. Li, "Stabilized multi-channel aloha for wireless ofdm networks," in *Global Telecommunications Conference, 2002. GLOBECOM '02. IEEE*, Month Published.

[96] J. Tsitsiklis, "Analysis of a multiaccess control scheme," *Automatic Control, IEEE Transactions on*, vol. 32, no. 11, pp. 1017–1020, Nov 1987.

[97] D. Lee, K. Kim, and W. Lee, *Q+ -Algorithm: An Enhanced RFID Tag Collision Arbitration Algorithm*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, vol. 4611, ch. 3, pp. 23–32.

[98] O. Galinina, A. Turlikov, S. Andreev, and Y. Koucheryavy, "Stabilizing multichannel slotted aloha for machine-type communications," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, Month Published.

[99] H. Wu, C. Zhu, R. La, X. Liu, and Y. Zhang, "Fast adaptive s-aloha scheme for event-driven machine-to-machine communications," in *Vehicular Technology Conference (VTC Fall), 2012 IEEE*, Sept 2012, pp. 1–5.

[100] A. Kuczura, "The interrupted poisson process as an overflow process," *Bell System Technical Journal*, vol. 52, no. 3, pp. 437–448, 1973. [Online]. Available: http://dx.doi.org/10.1002/j.1538-7305.1973.tb01971.x

[101] C.-H. Wei, C. Ray-Guang, and S.-L. Tsao, "Modeling and estimation of one-shot random access for finite-user multichannel slotted aloha systems," *Communications Letters, IEEE*, vol. 16, no. 8, pp. 1196–1199, 2012.

[102] wikipedia. Infinite Impulse Response (IIR) Filter. [Online]. Available: http://en.wikipedia.org/wiki/Infinite_impulse_response

[103] I. Stefan and H. Haas, "Hybrid Visible Light and Radio Frequency Communication Systems," in *Vehicular Technology Conference (VTC Fall), 2014 IEEE 80th*, Sept 2014, pp. 1–5.

[104] S. Haruyama, "Advances in visible light communication technologies," in *Optical Communications (ECOC), 2012 38th European Conference and Exhibition on*, Sept 2012, pp. 1–3.

[105] "IEEE Standard for Local and Metropolitan Area Networks–Part 15.7: Short-Range Wireless Optical Communication Using Visible Light," *IEEE Std 802.15.7-2011*, pp. 1–309, Sept 2011.

[106] M. Bhalerao and S. Sonavane, "Visible light communication: A smart way towards wireless communication," in *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*, Sept 2014, pp. 1370–1375.

[107] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "A First Look at Cellular Machine-to-machine Traffic: Large Scale Measurement and Characterization," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 1, pp. 65–76, Jun. 2012. [Online]. Available: http://doi.acm.org/10.1145/2318857.2254767

[108] M. Laner, P. Svoboda, N. Nikaein, and M. Rupp, "Traffic models for machine type communications," in *Wireless Communication Systems (ISWCS 2013), Proceedings of the Tenth International Symposium on*, Aug 2013, pp. 1–5.

[109] M. Laner, N. Nikaein, P. Svoboda, M. Popovic, D. Drajic, and S. Krco, "8 - Traffic models for machine-to-machine (M2M) communications: types and applications ," in *Machine-to-machine (M2M) Communications*, C. A.-H. Dohler, Ed. Oxford: Woodhead Publishing, 2015, pp. 133 – 154. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9781782421023000083

[110] A. Pourmoghadas and P. Poonacha, "Performance analysis of a machine-to-machine friendly MAC algorithm in LTE-advanced," in *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*, Sept 2014, pp. 99–105.

[111] C.-H. Wei, R.-G. Cheng, and Y.-S. Lin, "Analysis of slotted-access-based channel access control protocol for lte-advanced networks," *Wireless Personal Communications*, pp. 1–19, 2015.

[112] R. Cheng, F. Al-Taee, J. Chen, and C. Wei, "A dynamic resource allocation scheme for group paging in lte-advanced networks," *Internet of Things Journal, IEEE*, vol. PP, no. 99, pp. 1–1, 2015.

[113] D. Wiriaatmadja and K. W. Choi, "Hybrid random access and data transmission protocol for machine-to-machine communications in cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 14, no. 1, pp. 33–46, Jan 2015.

[114] J. J. Nielsen, D. Kim, G. C. Madueño, N. K. Pratas, and P. Popovski, "A Tractable Model of the LTE Access Reservation Procedure for Machine-Type Communications," *CoRR*, vol. abs/1505.01713, 2015.

[115] C.-H. Wei, R.-G. Cheng, and S.-L. Tsao, "Modeling and Estimation of One-Shot Random Access for Finite-User Multichannel Slotted ALOHA Systems," *Communications Letters, IEEE*, vol. 16, no. 8, pp. 1196–1199, August 2012.

[116] A. Pourmoghadas and P. Poonacha, "Performance analysis of a machine-to-machine friendly MAC algorithm in LTE-Advanced," in *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*, Sept 2014, pp. 99–105.

[117] X. Jian, X. Zeng, Y. Jia, L. Zhang, and Y. He, "Beta/M/1 Model for Machine Type Communication," *Communications Letters, IEEE*, vol. 17, no. 3, pp. 584–587, March 2013.

[118] R. Motwani and P. Raghavan, "Algorithms and theory of computation handbook," M. J. Atallah and M. Blanton, Eds. Chapman & Hall/CRC, 2010, ch. Randomized Algorithms, pp. 12–12. [Online]. Available: http://dl.acm.org/citation.cfm?id=1882757.1882769

[119] wikipedia. Trapezoidal Rule. [Online]. Available: http://en.wikipedia.org/wiki/Trapezoidal_rule

[120] R. Harwahyu, R.-G. Cheng, and R. F. Sari, "Consecutive group paging for LTE networks supporting machine-type communications services," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on*, Sept 2013, pp. 1619–1623.

[121] R. Harwahyu, X. Wang, R. Sari, and R.-G. Cheng, "Analysis of group paging with pre-backoff," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, 2015.

[122] C.-H. Wei, G. Bianchi, and R.-G. Cheng, "Modeling and analysis of random access channels with bursty arrivals in ofdma wireless networks," *Wireless Communications, IEEE Transactions on*, vol. 14, no. 4, pp. 1940–1953, April 2015.

[123] wikipedia. Exponential function. [Online]. Available: http://en.wikipedia.org/wiki/Exponential_function

[124] C. Jeong, J. Park, and H. Yu, "Random access in millimeter-wave beamforming cellular networks: issues and approaches," *Communications Magazine, IEEE*, vol. 53, no. 1, pp. 180–185, January 2015.

[125] V. Desai, L. Krzymien, P. Sartori, W. Xiao, A. Soong, and A. Alkhateeb, "Initial beamforming for mmwave communications," in *Signals, Systems and Computers, 2014 48th Asilomar Conference on*, Nov 2014, pp. 1926–1930.

Acronyms

2G 2th Generation. 34

3G 3th Generation. 34

3GPP 3^{rd} Generation Partnership Project. 11

5G 5th Generation. 11

ACB Access Class Barring. 8, 34, 43, 67, 122

AGT Access Grant Time. 42

ARIB Association of Radio Industries and Businesses. 14

AS Application Server. 18

ATIS Alliance for Telecommunications Industry Solutions. 14

BI Backoff Indicator. 36

C-RNTI Cell-Radio Network Temporary Identifier. 22, 80

CCE Control Channel Elements. 89

CCSA China Communications Standards Association. 14

CDF Cumulative Distribution Function. 73, 86

CDR Controlled Distribution of Resource. 10, 79, 118, 122

CGP Consecutive Group Paging. 81, 119

CMMPP Coupled Markov Modulated Poisson Processes. 66

CN Core Network. i, 8, 12, 17

CP Cyclic Prefix. 21

E-UTRAN Evolved Universal Terrestrial Radio Access Network. 31

EAB Extended Access Barring. 45, 67

eCall emergency Call. 13

eNB evolved Node B. 30, 55

EPC Evolved Packet Core. 30

EPS Evolved Packet System. 31

ETSI European Telecommunications Standards Institute. 14

EWT Extended Wait Timer. 45

FASA Fast Adaptive Slotted-ALOHA. 40, 50

FDD Frequency Division Duplex. 19

FI-TSFGP Further Improvement-TSFGP. 10, 80, 118, 122

GID Group ID. 79–81, 83, 90

GP Group Paging. i, 9, 10, 47, 79, 80, 91, 105, 118, 119

GRE General Recursive Estimation. 10, 65, 78, 122

H2H Human-to-Human. i

H2M Human-to-Machine. 7

HARQ Hybrid automatic repeat request. 68

HPLMN Home-PLMN. 30

HSS Home Subscriber Server. 30

ICT Information and Communications Technology. 14

IE Information Element. 45

IIR Infinite Impulse Response. 56

IoT Internet-of-Things. i, 1, 7, 11, 46

IR Initial Ranging. 23

ITS Intelligent Transport System. i, 7, 13, 121

ITU-T International Telecommunication Union - Telecommunication sector. 14

LTE-A LTE-Advanced. i

LTE Long Term Evolution. i, 9, 12

M2H Machine-to-Human. 7

M2M Machine-to-Machine. i, 7, 11

MAC Media access control. 12, 19

MCSA-OE Multi-Channel Slotted ALOHA-Optimal Estimation. 10, 50, 54, 63, 122

ME Mobile Equipment. 15

MIB Master Information Block. 23

MIMO Multiple-Input Multiple-Output. 34

MME Mobility Management Entity. 17

MTC Machine-Type-Communication. i, 1, 7, 11

MTC-AAA MTC-Authentication, Authorization and Accounting. 30

MTC-IWF MTC InterWorking Function. 29

NAS Non Access Stratum. 31

OFDM Orthogonal Frequency-Division Multiplexing. 21

P-GW PDN-GateWay. 17, 30

PBCH Physical Broadcast Channel. 23

PBO Pre-BackOff. 81

PDCCH Physical Downlink Shared CHannel. 27

PDN Packet Data Network. 31

PID Proportional Integrative Derivative. 44

PLMN Public land mobile network. 30

PRACH Physical RACH. 26, 41

PS Public Safety. i

PSAP Public-Safety Answering Point. 13

QAM Quadrature Amplitude Modulation. 34

QoS Quality of Service. 16

RA Random Access. 24

RA_RNTI Random Access_Radio Network Temporary Identifier. 27

RACH Random Access Channel. i, 8, 11, 18, 66, 141

RAN Radio Access Network. i, 8, 12, 17

RAOs Random Access Opportunities. 26

RAR Random Access Response. 25, 27, 68

RARs Random Access Responses. 82

RB Resource Block. 19, 21

RE Resource Element. 21, 89

RRC Radio Resource Control. 9, 21

S-GW Serving-GateWay. 17, 30

SC-FDMA Single Carrier-Frequency Division Multiple Access. 21

SCADA Supervisory Control And Data Acquisition. 7

SCS Services Capability Server. 29

SDOs Standards Development Organizations. 14

SIB2 System Information Block Type 2. 25, 44

SMS-SC Short Message Service-Service Centre. 30

TA Timing Alignment. 25

TAU Tracking Area Update. 38, 45

TDD Time Division Duplex. 19

TIA Telecommunications Industry Association. 14

TSFGP Traffic Scattering For Group Paging. 80

TTA Telecommunications Technology Association. 14

TTC Telecommunication Technology Committee. 14

UE User Equipment. 11, 27

UL-SCH UpLink-Shared Channel. 27

VLC Visible Light Communication. 66

WSN Wireless Sensor Networks. 13

Appendix A

A.1 The proof of equation 6.30

In this section, we try to rewrite the equation (6.29). First of all, we have

$$W_i = \frac{M_i}{M_1} = \sum_{n=1}^{N_{PTmax}} W_i[n] = \sum_{n=1}^{N_{PTmax}} \left(\prod_{k=1}^{n-1} (1 - e^{-\frac{M_i}{R}} p_k) \right)$$

When varying n frome 1 to N_{PTmax}, we have

$$W_i[1] = 1$$

$$W_i[2] = 1 - e^{-\frac{M_i}{R}} p_1$$

$$W_i[3] = (1 - e^{-\frac{M_i}{R}} p_1)(1 - e^{-\frac{M_i}{R}} p_2) = 1 - (p_1 + p_2) e^{-\frac{M_i}{R}} + p_1 p_2 e^{-\frac{2M_i}{R}}$$

$$\begin{aligned} W_i[4] &= (1 - e^{-\frac{M_i}{R}} p_1)(1 - e^{-\frac{M_i}{R}} p_2)(1 - e^{-\frac{M_i}{R}} p_3) \\ &= 1 - (p_1 + p_2 + p_3) e^{-\frac{M_i}{R}} + (p_1 p_2 + p_1 p_3 \\ &\qquad + p_2 p_3) e^{-\frac{2M_i}{R}} - p_1 p_2 p_3 e^{-\frac{3M_i}{R}} \end{aligned}$$

$$\begin{aligned} W_i[5] &= (1 - e^{-\frac{M_i}{R}} p_1)(1 - e^{-\frac{M_i}{R}} p_2)(1 - e^{-\frac{M_i}{R}} p_3)(1 - e^{-\frac{M_i}{R}} p_4) \\ &= 1 - (p_1 + p_2 + p_3 + p_4) e^{-\frac{M_i}{R}} + (p_1 p_2 + p_1 p_3 + \\ &\quad p_1 p_4 + p_2 p_3 + p_2 p_4 + p_3 p_4) e^{-\frac{2M_i}{R}} - (p_1 p_2 p_3 + \\ &\quad p_1 p_2 p_4 + p_1 p_3 p_4 + p_2 p_3 p_4) e^{-\frac{3M_i}{R}} + \\ &\quad p_1 p_2 p_3 p_4 e^{-\frac{4M_i}{R}} \end{aligned}$$

$$\vdots$$

Now, if we try to make the sum for the similar terms, we can find that:

$$\begin{aligned} W_i &= W_i[1] + W_i[2] + W_i[3] + W_i[4] + W_i[5] + ... \\ &= \sum_{t=1}^{N_{PT_{max}}-0} (-1)^0 1 + \sum_{t=1}^{N_{PT_{max}}-1} (-1)^1 \left(\sum_{k_1=1}^{t} p_{k_1} \right) \\ &\times e^{\frac{-1\times M}{R}} + \sum_{t=1}^{N_{PT_{max}}-2} (-1)^2 \left(\sum_{k_1=1}^{t} \sum_{k_2=k_1+1}^{t+1} p_{k_1} p_{k_2} \right) e^{\frac{-2\times M}{R}} \\ &+ \sum_{t=1}^{N_{PT_{max}}-3} (-1)^3 \left(\sum_{k_1=1}^{t} \sum_{k_2=k_1+1}^{t+1} \sum_{k_3=k_2+1}^{t+2} p_{k_1} p_{k_2} p_{k_3} \right) \\ &e^{\frac{-3\times M}{R}} + ... \end{aligned} \tag{A.1}$$

From the equation (A.1), we can conclude that:

$$W_i = \sum_{m=0}^{N_{PT_{max}}-1} \sum_{t=1}^{N_{PT_{max}}-m} (-1)^m \times \underbrace{\sum_{k_1=1}^{t} \sum_{k_2=k_1+1}^{t+1} ... \sum_{k_m=k_{m-1}+1}^{t+m-1}}_{\text{m times}} p_{k_1} ... p_{k_m} e^{-\frac{mM_i}{R}} \tag{A.2}$$

Let α_m be equal to:

$$\alpha_m = \sum_{t=1}^{N_{PT_{max}}-m} (-1)^m \underbrace{\sum_{k_1=1}^{t} ... \sum_{k_m=k_{m-1}+1}^{t+m-1}}_{\text{m times}} p_{k_1} ... p_{k_m} \tag{A.3}$$

Therefore, we have:

$$W_i = \frac{M_i}{M_1} = \sum_{m=0}^{N_{PT_{max}}-1} \alpha_m e^{-\frac{mM_i}{R}} \tag{A.4}$$

which is equal to the equation (6.30).

A.2 The proof of K_{max} (equation 6.15)

In this appendix, we will try to prove whether the value K_{max} is true. It is given in the text by the following equation

$$K_{max} = \left\lfloor \frac{W_{BO} - \alpha_a W_{BO}}{T_{RA\_REP}} \right\rfloor$$

$$= \left\lfloor \frac{T_{RAR} + W_{RAR} + W_{BO}}{T_{RA\_REP}} \right\rfloor - \left\lceil \frac{T_{RAR} + W_{RAR}}{T_{RA\_REP}} \right\rceil$$

Let $T_{RAR}+W_{RAR}+W_{BO}=\Psi$ and $T_{RAR}+W_{RAR}=\Gamma$, and thus;

$$K_{max}=\left\lfloor\frac{\Psi}{T_{RA\_REP}}\right\rfloor-\left\lceil\frac{\Gamma}{T_{RA\_REP}}\right\rceil \tag{A.5}$$

Generally, the number of RA slots falling within the backoff (BO) window is equal to the time of last sub-frame in the BO window minus the time before starting the BO window (divided by the interval between two consecutive RA slots), and it is given by:

$$\begin{aligned}N_{RA}&=i+\left\lceil\frac{T_{RAR}+W_{RAR}+W_{BO}}{T_{RA\_REP}}\right\rceil-i-\left\lfloor\frac{T_{RAR}+W_{RAR}}{T_{RA\_REP}}\right\rfloor\\&=\left\lceil\frac{\Psi}{T_{RA\_REP}}\right\rceil-\left\lfloor\frac{\Gamma}{T_{RA\_REP}}\right\rfloor\end{aligned} \tag{A.6}$$

Depending on the values of Ψ and Γ, we have four cases:

1. Both Ψ/T_{RA_REP} and Γ/T_{RA_REP} are not integer values: in this case, we can write $\lceil\Psi/T_{RA\_REP}\rceil$ as $(\lfloor\Psi/T_{RA\_REP}\rfloor+1)$ and $\lfloor\Gamma/T_{RA\_REP}\rfloor$ as $(\lceil\Gamma/T_{RA\_REP}\rceil-1)$. Thus N_{RA} becomes:

$$N_{RA}=\left\lfloor\frac{\Psi}{T_{RA\_REP}}\right\rfloor+1-\left\lceil\frac{\Gamma}{T_{RA\_REP}}\right\rceil+1=K_{max}+2$$

   where the value 2 represents the RA slots $x_a(i)$ and $x_d(i)$.

2. Ψ/T_{RA_REP} is integer whereas Γ/T_{RA_REP} is not: in this case, the value α_d (equation **??**) is equal to zero and thus:

$$N_{RA}=\left\lfloor\frac{\Psi}{T_{RA\_REP}}\right\rfloor-\left\lceil\frac{\Gamma}{T_{RA\_REP}}\right\rceil+1=K_{max}+1$$

   where the value 1 represents the RA slot $x_a(i)$. Note that when x/y is an integer value, we have $x/y=\lceil x/y\rceil=\lfloor x/y\rfloor$.

3. Ψ/T_{RA_REP} is not integer whereas Γ/T_{RA_REP} is: in this case, the value α_a (equation 6.13) is equal to zero and thus:

$$N_{RA}=\left\lfloor\frac{\Psi}{T_{RA\_REP}}\right\rfloor+1-\left\lceil\frac{\Gamma}{T_{RA\_REP}}\right\rceil=K_{max}+1$$

   where the value 1 represents the RA slot $x_d(i)$.

4. Both Ψ/T_{RA_REP} and Γ/T_{RA_REP} are integer values: in this case, the values α_a and α_d are equal to zero and thus:

$$N_{RA} = \left\lfloor \frac{\Psi}{T_{RA\_REP}} \right\rfloor - \left\lceil \frac{\Gamma}{T_{RA\_REP}} \right\rceil = K_{max}$$

Therefore, the equation giving K_{max} is true.

Appendix B

B.1 Control-Plane Latency Analysis of RACH Procedure

In this section, the control plane latency related to the Random Access Channel (RACH) procedure would be introduced. Fig. B.1 provides the control plane flow when a terminal moves from IDLE to CONNECTED [3].

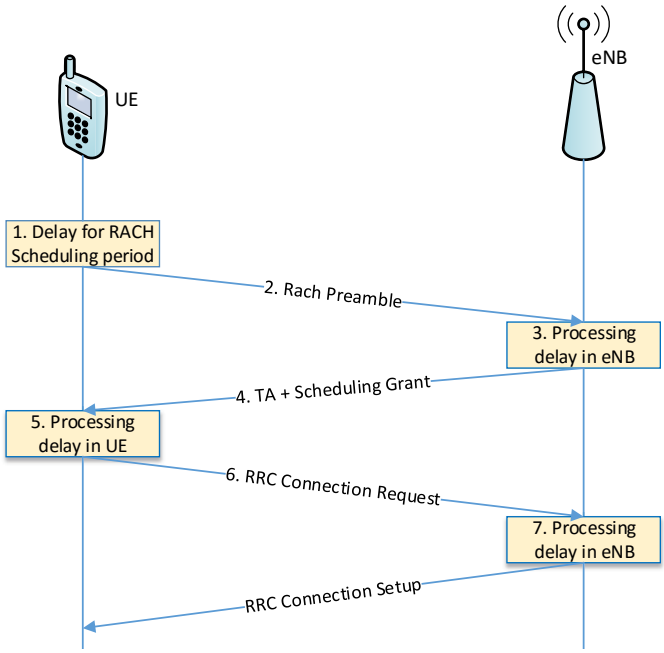

Figure B.1 C-plane activation procedure: RACH procedure [3]

Table B.1 Control plane latency analysis based on the procedure depicted in figure B.1

Component	Description	Minimum [ms]	Average [ms]
1	Average delay due to RACH scheduling period	0.5	2.5
2	RACH Preamble	1	1
3-4	the end RACH transmission and UE's reception of scheduling grant and timing adjustment)	3	5
5	UE Processing Delay (decoding of scheduling grant, timing alignment and C-RNTI assignment + L1 encoding of RRC Connection Request)	5	5
6	Transmission of RRC Connection Request	1	1
7	Processing delay in eNB (L2 and RRC)	4	4
8	Transmission of RRC Connection Set-up (and UL grant)	1	1