2013-ENST-0076

EDITE - ED 130

Doctorat ParisTech

THÈSE

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « ELECTRONIQUE et COMMUNICATIONS »

présentée et soutenue publiquement par

Kaijie ZHOU

le 5 December 2013

Technique d'accès pour la communication machine- à-machine dans LTE/LTE-A

Directeur de thèse : **Navid NIKAEIN**
Co-encadrement de la thèse : **Christian BONNET**

THÈSE

Jury

M. Petar POPOVSKI, Titre, Unité de recherche, Ecole	Président
M. Xavier LAGRANGE, Titre, Unité de recherche, Ecole	Rapporteur
M. Mischa DOHLER, Titre, Unité de recherche, Ecole	Rapporteur
M. Raymond KNOPP, Titre, Unité de recherche, Ecole	Examinateur
M. Dusit NIYATO, Titre, Unité de recherche, Ecole	Examinateur

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

to my parents

Abstract

Machine type communications (MTC) is seen seen as a form of data communication that does not necessarily require human interaction. It provides ubiquitous connectivity between machines, enabling creation of the so-called Internet-of-Things (IoT). However, it is challenging to accommodate MTC in LTE as a result of the specific characteristics and requirements of MTC, massive number of connected devices, small and sporadic payload transmissions, power constraint devices, and quality-of-service (QoS) requirements.

The aim of this thesis is to propose mechanisms and optimize the access layer techniques for MTC in LTE. In particular, our research deal with the uplink channel access and downlink reception. For uplink access, we propose two methods to improve the performance of random access: a packet aggregation method and a Transmission Time Interval (TTI) bundling scheme. With packet aggregation, a UE triggers a random access when the number of packets in the buffer reaches a given threshold. This packet aggregation method can be used to lower the packet loss rate and energy consumption at the expense of latency increase. The TTI bundling method is used to reduce the latency caused by unsuccessful random access. Concretely, a UE sends multiple preambles in consecutive subframes to improve the successful rate for random access. With this method, the latency can be greatly reduced when load is not high.

In order to further reduce the uplink latency for MTC and enable massive number of connected devices, we propose a new method, denoted as contention based access (CBA). With CBA, a UE sends packets on the randomly selected common resources granted by the base station. Therefore, the uplink latency can be greatly reduced by bypassing the redundant signaling in random access and the latency induced by the regular scheduling. We employ MU-MIMO detection at the eNB side to identify the C-RNTIs of collided UEs such that dedicated resource can be allocated for those UEs in the next subframes. We also present a resource allocation method for CBA to guarantee the delay constraint under minimal resource allocation.

For downlink reception, we study the discontinuous reception (DRX) mechanism in LTE which is designed to save power for MTC device. With DRX, a MTC device saves power at the expense of latency increase, which may not be acceptable for real time MTC applications. We propose two methods to analyze the DRX performance for MTC applications: the first with the Poisson distribution and the second with the uniform and Pareto distribution for sporadic traffic, respectively. With our models, the power saving factor and wake up latency can be accurately estimated for a given choice of DRX parameters, thus allowing to select the ones presenting the optimal tradeoff.

Acknowledgements

First and most, I would like to thank my advisers, Prof. Navid Nikaein and Prof. Christian Bonnet. Prof. Navid Nikaein shows great patience and everlasting support throughout my study. He has taught me, both theoretically and practically, how to carry out research in wireless communications. He also spent plenty of time to help me improve my writting and presentation skills, which would be a great fortune for my future career. It is my great honor to have the chance to work with Prof. Navid Nikaein. I also great appreciate for the support and understanding from Prof. Christian Bonnet. I would also like to thank Prof. Raymond Knopp for his help in designing the contention based access method and Prof. Thrasyvoulos Spyropoulos for his guidance in Semi-Markov chain modeling.

I would also like to thank my jury members: Prof. Xavier Lagrange, Prof. Mischa Dohler, Prof. Dusit Niyato, Prof. Petar Popovski and Prof. Raymond Knopp, for their efforts and valuable comments, which improve the quality of this thesis.

Then, I want to thank all my friends at Eurecom. Without them, I cannot imagine how could it be possible for me to finish this long journey. I had lots of happy and memorable time with my friends during the past three years. Many thanks to Lusheng Wang, whom I shared the previous office with and helped a lot when I just arrived at Eurecom. Many thanks to Rui Min, Xueliang Liu, XuRan Zhao, Jinbang Chen, Xiaolan Sha, Heng Cui, Xiaohu Wu, Shengyu Liu, Jinyuan Chen, Xinping Yi, Haifan Yin, Qianrui Li, JinJing Zhang, Imran Latif, Tania Villa, Ayse Unsal, Ankit Bhamri, Bilel Ben Romdhanne, Tien-Thinh Nguyen, and Ngoc Duy Nguyen, for their kind help and happiness brought to me. I also would like to thank Robin Rajan Thomas and Miltiadis Filippou for the wonderful soccer time we share in Antibes.

Last but not least, I would thank my family for their constant love and support during the past years. Special mentions goes to my wife, Jie Shen, who always understand and support all my important decisions.

Contents

List of Abbreviations

MTC	Machine type communications
IoT	Internet-of-Thing
QoS	Quality-of-Service
TTI	Transmission Time Interval
PRACH	Physical Random Access Channel
UE	User Equipment
CBA	Contention Based Access
MU-MIMO	Multi User- Multiple Input and Multiple Output
DRX	Discontinuous Reception
ICT	Information and Communications Technology
RFID	Radio-Frequency Identification
LTE	Long Term Evolution
H2H	Human to Human
ITS	Intelligent Transport System
UMTS	Universal Mobile Telecommunications System
GSM	Global System for Mobile
WiMAX	Worldwide Interoperability for Microwave Access
3GPP	3rd Generation Partnership Project
EU	European Union
OAI	OpenAirInterface
OFDMA	Orthogonal Frequency Division Multiple Access
SON	Self-Organizing Networks
VoIP	Voice over IP
e-MBMS	Enhanced-Multimedia Broadcast Multicast Services
CA	Carrier Aggregation
CoMP	Coordinated Multiple Point Transmission and Reception
EPC	Evolved Packet Core
PDN	Packet Data Network
EPS	Evolved Packet System
P-GW	PDN Gateway
S-GW	Serving Gateway
MME	Mobility Management Entity
UTRAN	UMTS Terrestrial Radio Access Network
eNB	evolved NodeB
DeNB	Donor eNB
RN	Relay Node
RRC	Radio Resource Control

PRB	Physical Resource Block
CQI	Channel Quality Indicator
SC-FDMA	Single Carrier Frequency Division Multiple Access
PAPR	Peak-to-Average Power Ratio
PDCP	Packet Data Convergence Protocol
RLC	Radio Link Control
MAC	Medium Access Control
ARQ	Automatic Repeat reQuest
PDU	Protocol Data Unit
SDU	Service Data Unit
PUSCH	Physical Uplink Shared Channel
PUCCH	physical Uplink Control Channel
PDSCH	physical Downlink Shared Channel
PDCCH	Physical Downlink Control Channel
PHICH	Physical HARQ Indicator Channel
PCFICH	Physical Control Format Indicator Channel
PAN	Personal Area Network
ETSI	European Telecommunications Standards Institute
GERAN	GSM EDGE Radio Access Network
ERP	Enterprise Resource Planning
CRM	Customer Relationship Management
ARPD	Low Average Return Per Device
M2M	Machine to Machine
SMS	Short Message Service
SR	Scheduling Request
RACH	Random Access Channel
RAR	Random Access Response
C-RNTI	Cell- Radio Network Temporary Identifier
RA-RNTI	Random Access- Radio Network Temporary Identifier
DCF	Distributed Coordinated Functions
SG	Schedule Grant
MCS	Modulation and Coding Scheme
DRB	Data Radio Bearer
NDI	New Data Indicator
GBR	Guaranteed Bit Rate
D2D	Device-to-Device
CDF	Cumulative Distribution Function

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Motivation

Machine type communications (or machine to machine communications), for example: remote monitoring, smart city management, and e-health, plays an important role in the information and communications technology (ICT). Fig.1.1 shows the evolution of machine type communications (MTC); we can see that machine type communications evolves from Radio-frequency identification (RFID) to Internet of Things (IoT), where anything that can benefit from network connection is connected. Moreover, MTC is now evolving towards digital society, where life efficiency, in terms of economy, mobility, environment, living and governance, is obtained based on the intelligent management, integrated ICTs, and active citizen participation. Of course, Operators are pushing for digital society as the mobile market is saturating and just seeking higher data rates will not create new revenue; they are seeking a paradigm shift to create new revenue. As shown in Fig.1.2 [1], the number of MTC devices is 8-9 times larger than human population, and among which only 50 million machines are connected, thus there is great potential for MTC. There are two types of techniques to accommodate MTC: wired communication and wireless communications. Compared to wired communication technology, wireless communication has some advantages to enable MTC: mobility, ease of deployment, and robustness [2]. Specifically, Long Term Evolution (LTE) is seen as an promising technique to enable MTC due to its large coverage, low latency, and high spectral and energy-efficiency.

However, different from human to human (H2H) communications which the current cellular networks are primarily designed for, MTC has some specific characteristics and requirements:

- Some MTC applications, such as pressure control in oil-pipeline, traffic safety in intelligent transport system (ITS), and virtual and augmented-reality, requires very short latency, which might be much lower than those for conventional voice and Internet traffic.

- Many MTC devices are powered by battery. For this type of MTC device, low energy consumption is extremely important.

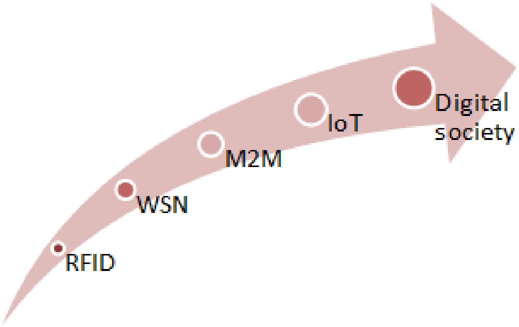
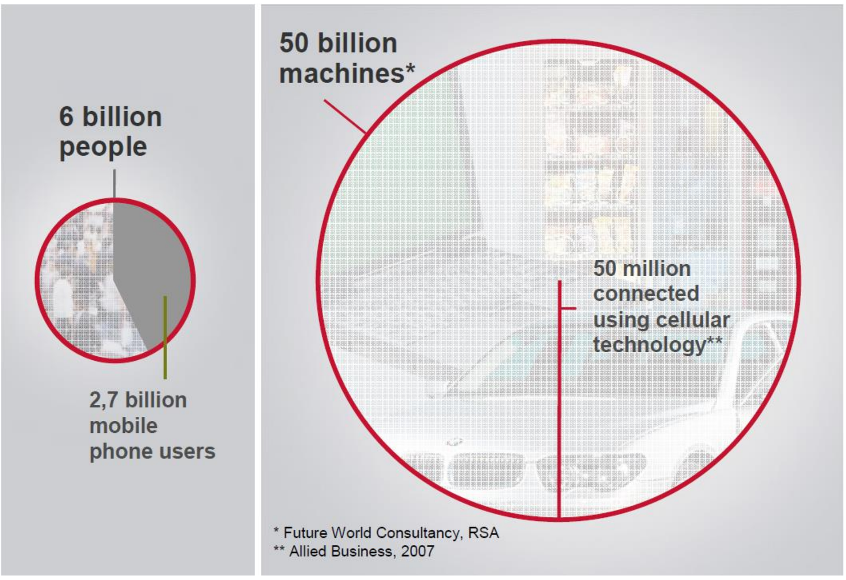


Figure 1.1: Evolution of machine type communications


6 billion people
2,7 billion mobile phone users
50 billion machines*
50 million connected using cellular technology**

* Future World Consultancy, RSA
** Allied Business, 2007

Figure 1.2: Potential of MTC

- In some cases, there are massive simultaneous transmissions for certain MTC application (e.g. public safety and remote monitoring), which may incur huge signaling and/or traffic overhead to the cellular networks. For example, future networks shall support up to 30000 MTC devices in one cell, which is orders of magnitude more than today's requirements [3].
- The data payload is usually small for some MTC applications (e.g. smart metering and cargo tracking). Efficient delivery of these small size packets is crucial to save precious spectrum resource.

These characteristics and requirements impose great challenges to the current cellular (UMTS, GSM, LTE) networks. Therefore, further improvements and optimizations are needed in order to enhance the performance for MTC applications. 3GPP, ETSI, and some other associations (e.g. WiMAX forum and WiFi Alliance) are working on standardizations for MTC. 3GPP released several specifications and reports (e.g. 3GPP TS 22.368, 3GPP TR 37.868 and 3GPP TR 36.888) to study and analyze the MTC applications in cellular networks (GSM, UMTS, LTE). The European Union (EU) commission has granted several projects to develop MTC, for example lola (http://www.ict-lola.eu/), smartsantander (http://www.smartsantander.eu/), sensei (http://www.sensei-project.eu/), and exalted (http://www.ict-exalted.eu/). Some vendors, for example Ericsson and Huawei, are creating their own visions, programs and initiatives, such as Ericsson's "50 billion connected devices", to drive development of MTC portfolio [5]- [6]. Moreover, how to accommodate MTC is also considered as an important topic in the future cellular networks (5G) as shown in Fig.1.3. Recently, EU commission granted Fresh 50 million EU research grants to develop 5G: metis (https://www.metis2020.com/) and 5GNow (http://www.5gnow.eu/).

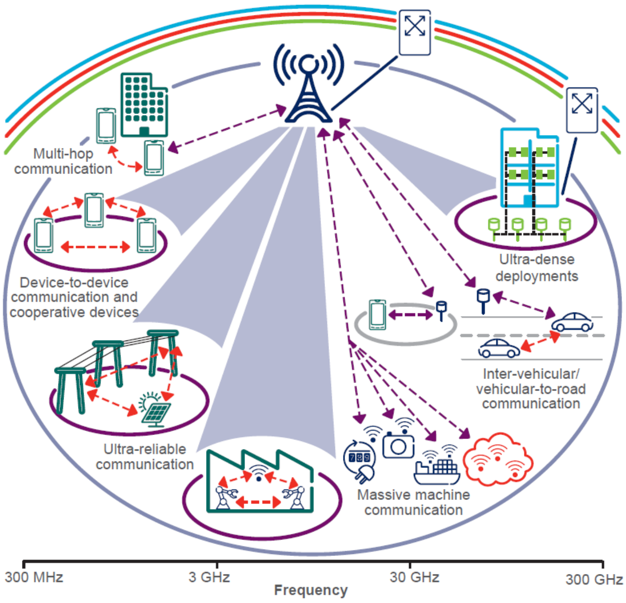

Multi-hop communication
Device-to-device communication and cooperative devices
Ultra-reliable communication
Massive machine communication
Inter-vehicular/ vehicular-to-road communication
Ultra-dense deployments
300 MHz
3 GHz
30 GHz
300 GHz
Frequency

Figure 1.3: 5G radio access solution

The aim of this thesis is to design and optimize the uplink channel access and downlink reception for MTC applications in LTE/LTE-A. Specifically, for MTC uplink transmission, we firstly propose a packet aggregation method to reduce the packet loss rate and power as well as a TTI bundling scheme to reduce uplink channel access latency. Moreover, for the MTC devices which are uplink synchronized, we propose a contention based access (CBA) method which further reduce the uplink access latency. For MTC downlink reception, we provide two methods to analyze the discontinuous reception (DRX) mechanism in LTE. With our model the optimal DRX parameter which saves the power at most while satisfying the delay constraint can be selected.

The following section provides a summary of the contributions of my thesis.

1.2 Contributions

1.2.1 Random Access Optimization for Machine Type Communication in LTE

Random access is used as the primary uplink channel access mechanism in LTE for MTC. However, it suffers from several problems: (1) the preamble collision rate is very high for heavy loaded network, which is typical for MTC with massive devices. (2) Due to backoff and the signaling overhead of random access, a UE has to wait for large amount of time to re-start a random access after an unsuccessful random access, which might not be acceptable for real-time MTC applications. To solve the first problem of random access, we propose a packet aggregation method. With the proposed method, a UE does not trigger a random access for every arrived packet. Instead, it triggers a random access until the number of packets in its buffer reaches a certain threshold. It is obvious that our method reduces preamble collision rate at the expense of extra latency which is used to accumulate some amount of packets. We design a method to set the optimal packet aggregation number in order to achieve the best tradeoff between latency increase and collision rate decrease. The another advantage of packet aggregation is energy saving: a UE reduces number of transmissions when packet aggregation is used. Therefore, we also propose an energy saving method: a UE aggregates packets as many as possible until its delay constraint reaches delay limit. To address the second problem of random access, we propose a transmission time interval (TTI) bundling method. With this method, a UE sends several randomly selected preambles in consecutive subframes (TTIs). If one of these preambles is correctly received by eNB and without collision, then the random access is successful, which reduces latency caused by unsuccessful random access. However, the preamble collision rate increases with the number of bundling TTIs since each UE has to trigger more transmissions, which may in turn increase latency. To address this problem, we provide a TTI bundling number selection mechanism to find the optimal TTI bundling number which minimizes the latency. The result was published in

- Zhou, Kaijie; Nikaein, Navid, " Packet aggregation for machine type communications in LTE with random access channel", in the proceedings of IEEE Wireless Communications and Networking Conference (WCNC), April, 2013, Shanghai, China.

and is currently under review

- Zhou, Kaijie; Nikaein, Navid, " Random access with TTI bundling for machine type communications in LTE", submitted to IEEE transactions on wireless communications.

and will be submitted as one part of

- Zhou, Kaijie; Nikaein, Navid, " Random access packet aggregation for machine type communications in LTE", under preparation.

1.2.2 Contention Based Access

To further remove the redundant signaling in random access, we propose a contention based access (CBA) method for uplink synchronized MTC device. The main feature of contention based access is that UE is not assigned with dedicated resource. Instead, the resource is allocated for all or a group of UEs. A UE randomly selects its resource and sends a data packet on it. As the CBA resources are allocated for all or a group of UEs, collision happens when multiple UEs within a group select the same resource. To address the problem of collision, in our method each UE sends its identifier, cell radio network temporary identifier (C-RNTI), along with the data on the randomly selected resource. MU-MIMO detection is used at the eNB side to decode packets: the highly protected C-RNTIs from different UEs might be decoded even if they are sent on the same resource. If the C-RNTI of a UE is successfully decoded while its data payload is lost, dedicated resource is allocated for that UE by eNB. We elaborate the implementation issues of CBA in the OpenAirInterface(OAI) software-defined radio platform [7] designed for LTE.

In addition to that, we also design a resource allocation scheme for CBA. With this method, eNB can assign minimum amount of resource to meet the delay requirement for real-time MTC applications.

The results were published in

- Zhou, Kaijie; Nikaein, Navid; Knopp, Raymond; Bonnet, Christian, " Contention based access for machine-type communications over LTE", IEEE 75th Vehicular Technology Conference, May, 2012, Yokohama, Japan.
- Zhou, Kaijie; Nikaein, Navid; Knopp, Raymond, "Dynamic resource allocation for machine-type communications in LTE/LTE-A with contention-based access", IEEE Wireless Communications and Networking Conference (WCNC), April, 2013, Shanghai, China.

1.2.3 Discontinuous Reception Analysis and Optimization

Lowering the power consumption is among the primary requirements for MTC applications because most of the MTC devices are powered by battery. To achieve this, discontinuous reception (DRX) is employed in LTE/LTE-A network. With DRX, a UE only turns on the receiver at some pre-defined time points while sleeps at others. However, DRX mechanism provides a power saving capability at the expense of an extra delay. Therefore it is preferred that the DRX parameters are selected such that the power saving is maximized while the application delay constraint is satisfied.

We provide two methods to analyze the detailed DRX mechanism in LTE/LTE-A. The first method deals with MTC applications with Poisson traffic. With this method, one can calculate the power saving factor and latency for a given DRX parameter set, which can be used to select the suitable DRX parameter. The second method is applicable to analyze the DRX performance with sporadic traffic. Based on this method, we also provide a simple method to find the optimal DRX parameter which maximizes the power saving factor while maintaining the latency requirement. The result was published in

- Zhou, Kaijie; Nikaein, Navid; Spyropoulos, Thrasyvoulos, " LTE/LTE-A discontinuous reception modeling for machine type communications", IEEE Wireless Communications Letters, vol.2, no.1, pp.102-105, 2013.

and will be submitted as one part of

- Zhou, Kaijie; Nikaein, Navid; Spyropoulos, Thrasyvoulos, "DRX modeling and optimization for sporadic machine type communications", under preparation.

In addition to the above contributions which comprise this thesis, some work on traffic modeling for MTC is carried out.

- Navid Nikaein, Markus Laner, Kaijie Zhou, Philipp Svoboda, Dejan Drajic, Milica Popovic, Srdjan Krco, "Simple traffic modeling framework for MTC", 10th International Symposium on Wireless Communication Systems (ISWCS), 27-30 August 2013, Ilmenau, Germany.

1.3 Organization of the Thesis

The rest of this thesis is organized as following as shown in Fig.1.4:

Chapter 2 provides the background on the network architecture and protocol stack for LTE highlighting the latency budget for random access and uplink scheduling as well as power consumption, and machine type communications.

Chapter 3 elaborates the two proposed methods to improve the performance of random access, in particular packet aggregation and TTI bundling.

Chapter 4 presents the contention based access (CBA) method for low-latency uplink channel access as well as an incremental resource allocation method to guarantee the delay constraint.

Chapter 5 proposes a DRX modeling method for MTC application with Poisson traffic and a DRX analysis and optimization method for MTC application with sporadic traffic.

Chapter 6 concludes the research work carried out during this thesis and presents the future work.

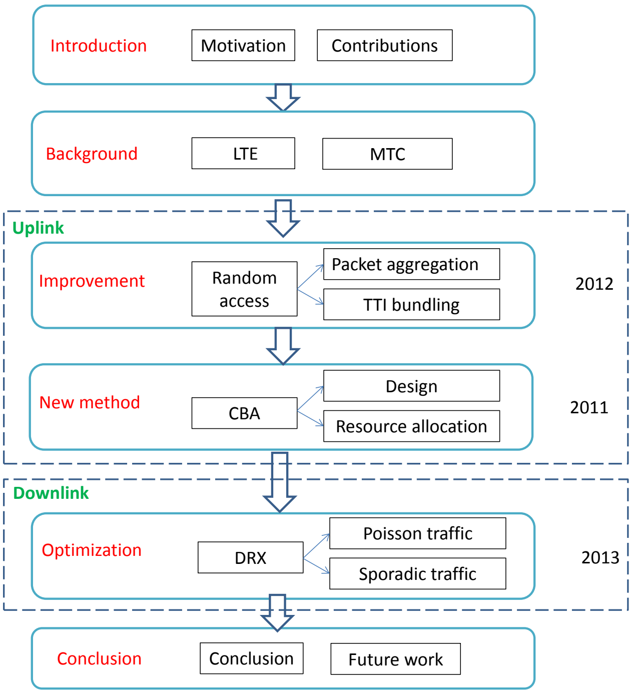


Figure 1.4: Organization of the thesis

CHAPTER 2
Background

The long-term evolution (LTE) is a highly flexible radio interface defined by 3rd Generation Partnership (3GPP) to provide high rate access. Among the various applications provided by LTE, machine type communications, is one of the most promising applications.

In this chapter, we give a brief introduction to the LTE/LTE-A system as well as machine type communication.

2.1 LTE

To meet the need of increasingly traffic, 3GPP standardization body developed the LTE technology. The LTE system is expected to greatly improve the throughputs, spectral efficiency and user experience. Moreover, LTE is also aimed to support IP-based traffic with end-to-end QoS guarantee. Voice traffic is supported mainly as VoLTE (Voice over LTE), which enables the integration with other multimedia services [8]. To satisfy performance requirements in Rel-8, LTE employs a set of techniques: orthogonal frequency division multiple access (OFDMA), multiple-input and multiple-output (MIMO), flatter-IP core network , self-organizing networks (SON), etc. Here are some features of LTE [9]

- Peak data rate with 20 MHz channel
- Downlink rate: 100 Mbps
- Uplink rate: 50 Mbps
- Up to 200 active users in a cell with 5MHz
- Less than 5 ms latency (user plane)
- Mobility support
- e-Multimedia Broadcast Multicast Services (e-MBMS)
- Spectrum flexibility

- End-to-end QoS
- All-IP, and IPv6 support

Furthermore, to meet the IMT-Advanced requirements coined by ITU, the 3GPP is developing the LTE-A technology. Some of the new techniques of LTE-Advanced include: carrier aggregation (CA), coordinated multiple point transmission and reception (CoMP), enhanced SON, and enhanced multiple antenna technologies.

2.1.1 Network Architecture of LTE/LTE-A

The network architecture of LTE is shown in Fig.2.1 [9]- [10], which is composed of the core network, known as Evolved Packet Core (EPC), and the access network known as evolved universal terrestrial radio access network (E-UTRAN). A UE connects the E-UTRAN through the LTE-Uu interface. The EPC provides the IP connectivity between the UE and the Packet Data Network (PDN) through the SGi interface.

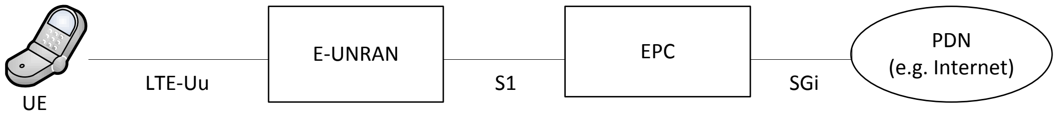


Figure 2.1: Network Architecture of LTE/LTE-A [9]- [10]

Fig.2.2 shows the architecture of E-UTRAN [11]. It can be seen that E-UTRAN is composed of only eNB making LTE flatter compared to previous cellular systems. The main functionalities of eNB include: radio access networking, radio resource management, and connectivity to the EPC [9]. The eNBs are inter-connected among each other through the X2 interface, and are connected to MME/S+P-GW through the S1 interface. The main functions of mobility management entity (MME) include: non-access stratum (NAS) signaling, roaming and handover management, and evolved packet system (EPS) bearer control. There are two gateways in EPC: PDN gateway (P-GW) and serving gateway (S-GW). The S-GW is the connection point between EPC and eNB; it serves UE by routing the IP packets. The P-GW is the connection point between EPC and PDN; it routes the packet from and to PDN [10].

As shown in Fig.2.3, LTE-A introduces the relay node (RN) to the access network architecture [11]. By the use of relay technique, the coverage can be extended and the throughput of hotspots can also be increased. A relay node is wirelessly connected to donor eNB (DeNB) through the Un interface. While a RN is acting as a UE from the perspective of DeNB, it has the role of regular eNB from the perspective of UEs. In this sense, a RN has a dual protocol stack with the data backhauling support.

2.1.2 Protocol Stack

Fig.2.4 shows the protocol stack of control plane and user plane [11]. The protocols running between the UE and the MME are known as the NAS protocols operating on the control-

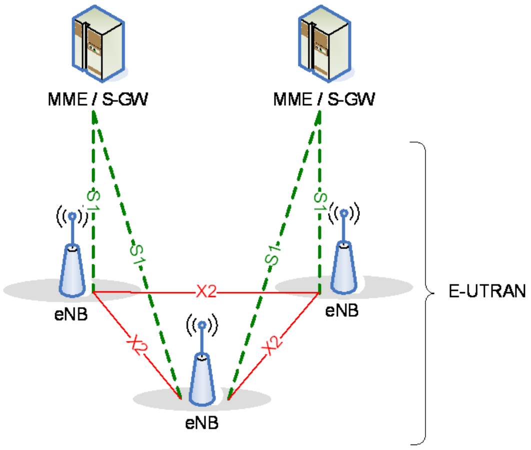


Figure 2.2: Architecture of E-UTRAN [11]

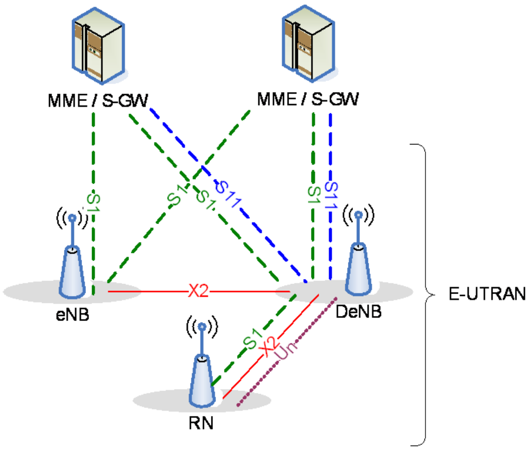


Figure 2.3: E-UTRAN Architecture supporting RNs [11]

plane. The main functions of the NAS protocols include: EPS bearer management, authentication, security control, paging and mobility management.

Layer 3

The RRC protocol is referred to as 'Layer 3' in the access stratum (AS) protocol stack. The main functions of RRC protocol include [12]:

- Broadcast of system information
- RRC connection control
- Inter-RAT mobility management

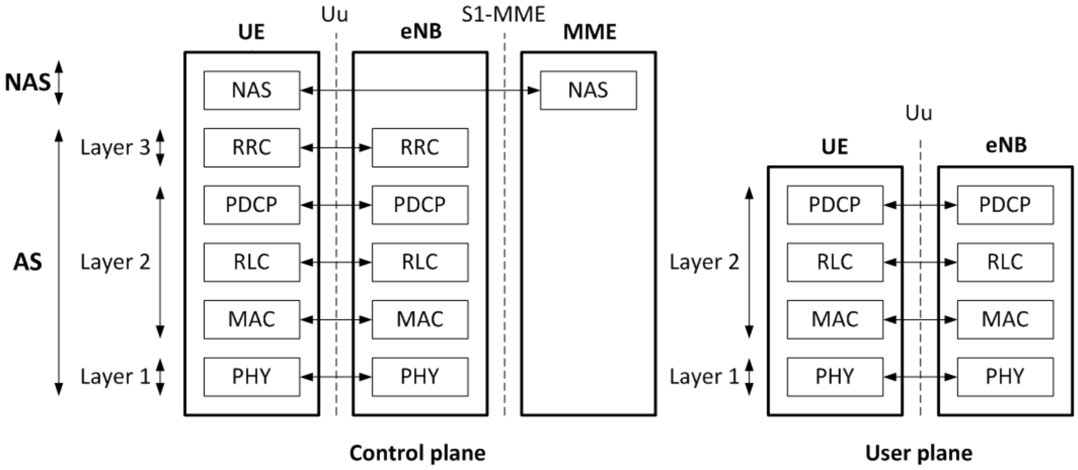


Figure 2.4: LTE Protocol stack [11]

- Measurement configuration and reporting
- Generic protocol error handling
- Support of self-configuration and self-optimization
- Other functions, for example: support for E-UTRAN sharing and transfer of dedicated NAS information.

A UE is in RRC_CONNECTED state if the RRC connection has been established. Otherwise, the UE is in the RRC_IDLE state. Usually, a UE uses random access to send the RRC connection request information to establish RRC connection between UE and eNB.

Layer 2

The layer 2 protocol stack has three sub-layers: packet data convergence protocol (PDCP), radio link control (RLC) and medium access control (MAC). Concretely, the functions of three sub-layers are [11]:

- The PDCP layer: For the RRC message of the control plane, the main functions of the PDCP are integrity protection and ciphering/deciphering. While for the IP packet of the user plane, the main functions of PDCP are: header compression and decompression, reordering and retransmission for handover, discarding uplink user plane data when timeout, and ciphering and deciphering.
- The RLC layer: In order to fit the size indicated by MAC layer, the RLC layer fragment or concatenate the PDCP protocol data units (PDU). The RLC layer also reorders the RLC PDUs which are received out of sequence due to hybrid ARQ (HARQ) operation in the MAC layer. Moreover, when acknowledged mode (AM) is enabled, the RLC

layer retransmits RLC PDUs if error happens. The other functions of RLC layer include: duplicate detection, protocol error detection, RLC re-establishment, etc.

- The MAC layer: One of the main function of MAC layer is multiplexing and demultiplexing. Applying for the scheduling policy, multiplexing is performed across different UEs and logic channels to meet the different QoS requirements among UEs and different priorities among logic channels. The functionality of demultiplexing entity is to extract the MAC service data units (SDU) for logic channels from MAC PDUs. The other important function of MAC layer is HARQ. A UE can use up to eight HARQ processes in parallel to reduce the waiting time for ACK/NACK. For uplink, the HARQ is synchronous, i.e., the retransmission happens 8 ms after the initial transmission. While for downlink, the HARQ is asynchronous, i.e., the time instant for retransmission is flexible.

Layer 1

The LTE physical layer is designed to provide connectivity between the base station and UE. By the use of OFDM and MIMO techniques, LTE physical layer can support peak rate 100 Mbps for downlink and 50 Mbps for uplink. The transmission bandwidth for LTE ranges from 1.4 MHz to 20 MHz, which improves the spectrum efficiency and supports different UE's capabilities [9]. The minimum resource used for transmission is referred to as a physical resource block (PRB). A PRB has both a time and frequency dimension. Concretely, a PRB is composed of 12 subcarriers (180 KHz) for the duration of one subframe (1 ms). Depending on the cyclic prefix length, there are 14 (normal) or 12 (extended) OFDM symbols in one subframe.

Fig.2.5 [13] shows the signal processing procedure for uplink physical channel: (1) after channel coding, the codewords[1] are firstly scrambled; (2) the scrambled bits are modulated according to modulation scheme indicated by eNB (the modulation scheme for PUSCH can be QPSK, 16 QAM or 64 QAM); (3) the modulated bits are then mapped to different layers (the modulated bits of a codeword can be mapped to one or two layers); (4) at each layer, discrete Fourier transform (DFT) based precoding is applied to turn the time domain signal into the frequency domain signal; (5) the complex valued symbols are precoded to support spatial multiplexing; (6) the precoded symbols are mapped to time-frequency resource; (7) finally, time-domain single carrier frequency division multiple access (SC-FDMA) signal is generated for each antenna port. It has to be noted that SC-FDMA is used for uplink transmission. The main benefit of SC-FDMA is the lower peak-to-average power ratio (PAPR) of its transmitted signal, which reduces the cost of UE and improves transmit power efficiency [13]. For uplink, there are three types of channels: physical uplink shared channel (PUSCH), physical uplink control channel (PUCCH), and physical random access channel (PRACH). PUSCH is used to carry data as well as control information (channel quality indicator (CQI), scheduling request (SR), etc.). While PUCCH can also be used to send control information when the PUSCH is not available, and PRACH is used to send preamble for random access [13].

[1] A codeword is defined as the output of each channel coding stage for a single transport block from the MAC layer [14].

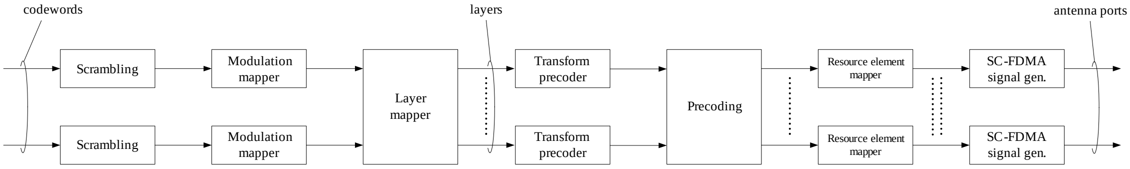


Figure 2.5: Overview of uplink physical channel processing [13]

Fig.2.6 shows the signal processing procedure for downlink physical channel, which is similar to uplink (transform precoding is not needed since it is specific to SC-FDMA) [13]. It has to be noted that for downlink transmission, OFDM is employed to improve capacity. Moreover, for downlink precoding, spatial multiplexing, beamforming and transmission diversity techniques can be applied. Regardless of multicast and broadcast channel, there are four types of channels in downlink: physical downlink shared channel (PDSCH), physical downlink control channel (PDCCH), physical hybrid-ARQ indicator channel (PHICH), and physical control format indicator channel (PCFICH). (1) PDSCH is used for data transmission. (2) PDCCH is used to carry downlink control information (DCI). The DCI carries uplink and downlink resource allocation information, power control commands, index of modulation and channel coding scheme (MCS), etc. (3) PHICH is used to send ACK/NACK information in response to the uplink transmission. (4) PCFICH is used to send control format indicator (CFI). The CFI is used to indicate the amount of OFDM symbols used to send DCI in one subframe.

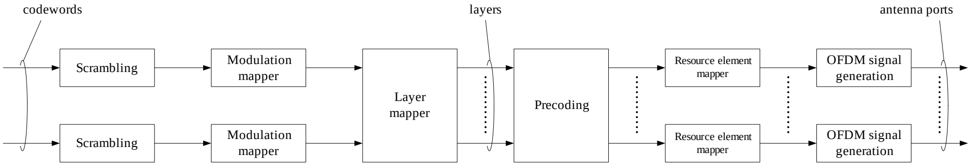


Figure 2.6: Overview of downlink physical channel processing [13]

2.1.3 Latency and Power Consumption in LTE

In this thesis, we mainly deal with latency and power consumption. For example, we optimize the DRX mechanism to reduce power consumption while maintaining delay constraint. We use packet aggregation method to reduce packet loss rate and power consumption at the expense of latency increase. Hence, here we introduce the latency and power consumption in LTE.

Latency

For real time MTC application, such as collision avoidance in intelligent transport system, low latency is very an important issue. The latency in LTE includes: control plane latency and user plane latency. The *control plane latency* is defined as the latency for a UE to enter to a state where it can send or receive data [15]. In LTE, this control plane latency is related to transition time from RRC_IDLE state to RRC_CONNECTED state and from discontinuous reception state (sleep state) to the active state [16].

When a UE is at the RRC_IDLE state and need to send a packet, it has to transfer to the RRC_CONNECTED state. Usually, a UE uses random access to set up RRC connection. Fig. 2.7 shows the contention-based random access procedure in LTE: (1) a UE sends a randomly selected preamble; (2) if this preamble is correctly received by eNB, eNB sends the random access response (RAR) to allocate resource for the next round transmission; (3) a UE send the RRC connection request message (L3 message) on the allocated resource; (4) the eNB sends the contention resolution message to acknowledge the correctly received message. It has to be noted that if multiple UEs send the same preamble in the first step, they will be allocated with same resource in the second step. As a result, the RRC connection request messages are sent on the same resource, which might not be decoded by eNB. To conclude, when random access is used, the transition time from RRC_IDLE state to RRC_CONNECTED state (control plane latency) depends on the amount of preamble resource, channel condition, traffic load, etc.

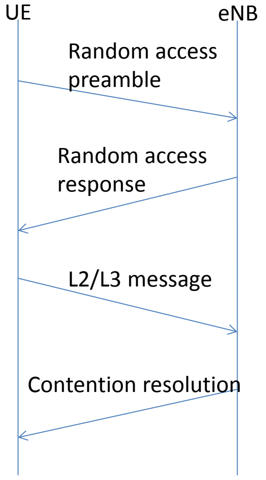


Figure 2.7: Contention based random access in LTE

With discontinuous reception (DRX) mechanism, as shown in Fig.2.8, a UE transfers to the active state to monitor the physical downlink control channel (PDCCH) at some pre-defined time instant [2]. Therefore, the transition time from sleep state to active state (control plane latency) is determined by the DRX parameters, especially the DRX cycle. The value of long DRX cycle in LTE ranges from 10 ms to 2560 ms. Therefore, the maximum latency caused by DRX is up to 2559 ms, which might be unacceptable for real time MTC application.

The *user plane latency* is defined as the duration for a user plane data packet being available at the IP layer of the sender and the availability of this packet at IP layer of the receiver [15]. For uplink transmission, if the scheduling request (SR) resource is available, a UE uses regular scheduling to send a data packet. Fig. 2.9 shows the uplink scheduling mechanism in LTE: (1) a UE sends the SR information on the physical uplink control channel (PUCCH); (2) a

[2] We will explain the DRX mechanism with more details in Chapter 5

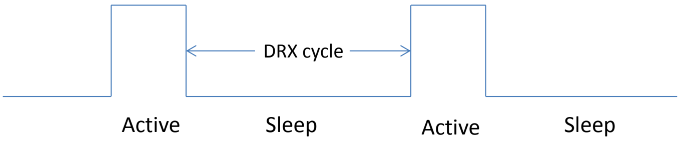


Figure 2.8: DRX mechanism in LTE

eNB decodes the SR packet and allocates resource for that UE; (3) a UE sends its buffer state report (BSR) on the allocated resource; (4) a eNB allocates suitable amount of resource for the UE according to its BSR information; (5) UE sends the data packet. Assuming that the SR period is 5ms and the eNB processing time for SR, BSR and data packets is 3ms, the latency for this uplink scheduling is 22.5ms. Therefore, by the use of regular scheduling, the user plane latency is mainly determined by the period of resource, HARQ retransmission delay, UE and eNB processing time, channel condition, etc. However, if a UE is not allocated with scheduling request resource, it has to use random access to apply for resource: it sends the SR information through random access. In this case, the user plane latency is mainly determined by amount of preamble resource, traffic load, channel condition, etc.

Therefore, the latency we treated in Chapter 3 and Chapter 4 are user plane latency (a UE is always RRC connected), while in Chapter 5 we mainly consider the control plane latency.

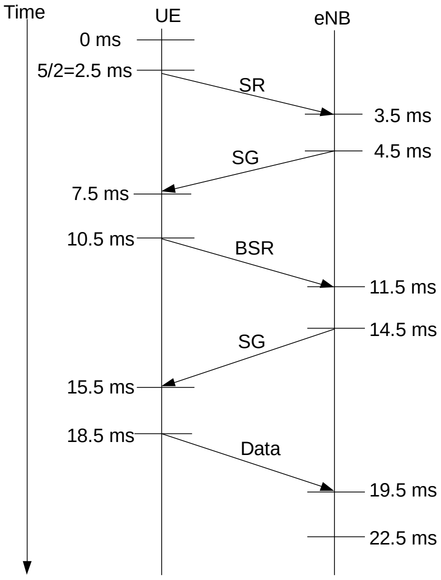


Figure 2.9: Uplink packet scheduling in LTE

Power consumption

Since most MTC devices are powered by battery, therefore reducing power consumption is also a very important point for MTC. To save power, if there is no transmission or reception for a period of RRC inactivity timer, a UE releases its RRC connection to enter the RRC_IDLE state as shown in Fig.2.10 [17]- [18]. In the RRC_IDLE state, a UE consumes the least power. It sleeps for most of the time while just periodically wakes up to receive the system information. The wake up period is determined by the paging cycle. The larger paging cycle is, the larger latency is introduced and the more power is saved.

When a UE is at the RRC_CONNECTED state, if DRX is enabled, a UE sleeps most of the time. It enters the active state when DRX cycle timer expires as shown in Fig.2.8 and Fig.2.10 [17]- [18]. Therefore, with DRX mechanism, a UE can also save power. However, when DRX is enabled, a UE still can send the uplink data. Therefore, it consumes more power than the idle since it has to maintain the uplink synchronization with eNB. If DRX is not enabled, a UE is always in the active state: it monitors PDCCH for every subframe (1ms), which hence consumes more power than the idle state and the sleep state. While in this active state, most power is consumed by transmission (TX) and reception (RX) [19]. For reception, the RX radio frequency (RF) power consumption is independent on the downlink rate, but depends on the receiving signal strength. As for the receiver baseband, the channel estimation, equalization, and turbo decoding consume most of the power. At the transmitter side, the power used for TX RF is also independent on TX rate, but depends on the energy efficiency of power amplifier (PA). While for the baseband of TX, turbo encoding consumes most of the power. Moreover, the power consumed by turbo decoding depends on the packet size [19].

In our packet aggregation method proposed in Chapter 3, a UE does not send a preamble for every single packet. Instead, it triggers a random access for multiple packets, which hence saves power.

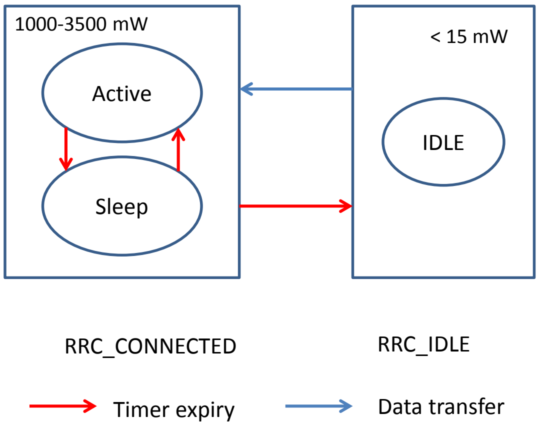


Figure 2.10: UE state transition [17]- [18]

2.2 Machine Type Communications

Machine type communication MTC (or Machine to Machine (M2M) communications) is seen as a form of data communication that does not necessarily require human interaction [20]. In parallel to the evolving of mobile communication systems, the application of machine type communications (MTC) has been growing very fast as well, for example: remote monitoring/control, intelligent transport system (ITS), e-health, fleet-tracing, smart grid, etc. One of the real deployment is *ekobus* [21], where the public vehicles are used to monitor environmental parameters and to provide traffic information. It is predicted the MTC promises huge market growth with expected 50 billion connected devices by 2020 [22]. However, different from the conventional human-to-human (H2H) communications, such as voice or web surfing, MTC has some specific characteristics and requirements, which requires significant improvements in the wireless communication system.

MTC is a very active area under discussion in 3GPP for integration within the LTE/LTE-A framework and more generally within European Telecommunications Standardization Institute (ETSI) M2M adhoc group. It would not be surprising to find more and more work related to the constraints imposed by M2M in 3GPP RAN groups in the coming years. MTC poses many interesting problems with respect to traffic modeling and the related PHY/MAC procedures (HARQ, adaptive coding and modulation, MAC layer scheduling, etc.). Another key aspect is the design of low -layer signaling which allows for extremely short acquisition times for event -driven traffic and switching between idle, sleep, and active states.

In this section, we provide a brief introduction to MTC. We introduce the network architecture, applications, and benefits of MTC and then discuss the challenges of MTC and the efforts to address these challenges.

2.2.1 Network Architecture for Machine Type Communications

Fig. 2.11 [23] demonstrates the M2M network architecture proposed by ETSI, which includes the device and gateway domain as well as network domain.

The device and gateway domain includes the following components [23]:

- M2M device: a device which runs M2M applications by the use of M2M service.
- M2M area network: connects M2M device and M2M Gateway.

  Usually personal area network (PAN) is used as M2M Area network such as: Zigbee, Bluetooth, IEEE 802.15, etc.
- M2M Gateway: a gateway acts as a proxy between M2M devices and Network domain; it collects and transmits information from M2M device.

The network domain consist of the following elements [23]:

- Access network: provides communications between core networks and m2M device and gateway domain.

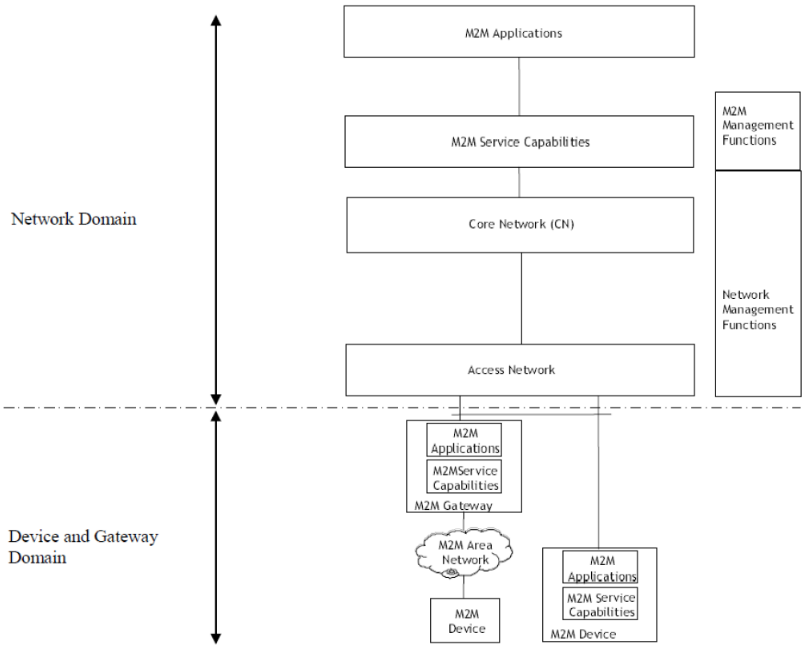


Figure 2.11: High level architecture for M2M [23]

Examples of access network include: wireless local area network (WLAN), WiMax, E-UTRAN, UTRAN, GSM edge radio access network (GERAN), and etc.

- Core networks:
  - IP connectivity and interconnection (with other networks)
  - Service and network control functions
  - Core networks can be 3GPP core network, or ETSI core network, or others.
- M2M service capabilities:
  - M2M functions to M2M applications
  - Use core network functionalities
  - interfaces to simplify and optimize application development
- M2M applications: use M2M service capabilities
- M2M management functions: includes functions to manage M2M service capabilities in the network domain
- Network management functions: includes functions to manage the Access and core networks.

2.2.2 Machine Type Communication Applications

There are various MTC applications with distinct characteristics. [20] presents a non exhaustive list of MTC applications. For each MTC application, there are unique characteristics. However, there are four basic components which are common to each MTC application [24]: (1) collection of data; (2) Transmission of selected data through a communication network; (3) assessment of the data; (4) response to the available information.

Table 2.1: Examples of MTC applications [20]

Service Area	MTC applications
Security	Surveillance systems Backup for landline Control of physical access (e.g. to buildings) Car/driver security
Tracking & Tracing	Fleet Management Order Management Pay as you drive Asset Tracking Navigation Traffic information Road tolling Road traffic optimization/steering
Payment	Point of sales Vending machines Gaming machines
Health	Monitoring vital signs Supporting the aged or handicapped Web Access Telemedicine points Remote diagnostics
Remote Maintenance/Control	Sensors Lighting Pumps Valves Elevator control Vending machine control Vehicle diagnostics
Metering	Power Gas Water Heating Grid control Industrial metering
Consumer Devices	Digital photo frame Digital camera eBook

2.2.3 Benefits of Machine Type Communications

There are lots of advantages to use machine type communications. For example, in case of an accident, the MTC module in a car can send this emergency information to backend system. The backend system distributes this information to other cars which are close to the location of an accident. With this information, other cars can take proper actions (brake, change direction, etc). To sum up, the benefits provided by MTC include [25]- [26]:

- **Increase efficiency for business and government**. By the use of MTC, business process can be optimized to improve efficiency and save cost. For example, with the help of MTC application, automated business process management, enterprise resource planning (ERP) and streamlining of customer relationship management (CRM) can greatly increase the IT efficiency for companies and public organizations. The related MTC applications include: remote monitoring, information collection and distribution, and asset tracking, etc.

- **Help products gain or maintain a competitive edge**. By the use of MTC, a manufacture provides connectivity, which facilitate the addition of new features to products and services. For example, a consumer electronics manufacturer can equip a digital camera with wireless communication module to enable the cloud storage solution; a car manufacturer provides the intelligent transportation system (ITS) functionality by adding wireless communication modules.

- **Enabling companies to comply with regulation**. MTC can also help to address the requirements of regulations in some countries. For example: in EU, auto manufacturers are required to have cellular connectivity to enable emergency call system.

- **Saving the planet**. MTC can be used to save the energy consumption, thus creating a more sustainable society. For example, real time data can be used to reduce energy for heating, lightning, and air conditioning.

2.2.4 Challenges for Machine Type Communications

The current cellular networks (2G/3G/LTE) are mainly designed for human-to-human communications (H2H) with a continuous flow of information, at least in terms of the timescales needed to send several IP packets (often large for user plane data). Therefore, the signaling overhead is manageable. However, MTC application has quite different characteristics and requirements. Table 2.2 lists some differences between H2H and M2M [27].

Table 2.2: Difference between H2H and M2M [27]

Topics	H2H	M2M
Density	Maybe not that much compared to M2M potential	M2M outnumbers human end users by order of magnitude.
Data Volume	Most traffic is downlink and requires significant amounts of bandwidth.	Uplink dominant traffic of small size, except for case of video surveillance.
Battery	Rechargeable	Be capable of auto-generating power or be self-sustaining for long periods
Delay	Tolerable to some extent	Some real time applications (urgent/emergency) have little tolerance.
Revenue	Good	Low Average Return Per Device (ARPD).
Value chain	Well defined	To be created.
Reachability	Satisfying	Might require much longer dormant period to minimize signaling overhead and to save power.

With these significant differences from H2H communication, MTC imposes great challenges to the current cellular networks. Therefore, some optimizations are needed in order to meet the specific features of MTC [28]:

1. Massive transmission

   For some MTC application (e.g. public safety and surveillance), usually there are numerous MTC devices transmitting simultaneously. This feature may impose huge signaling and/or traffic overhead to the cellular networks, which requires enhancements to the channel access mechanism (random access), bandwidth request method, and higher layer protocol.

2. Small burst transmission

   The MTC traffic is mainly of small size (e.g. 10 bytes). For this type of small data transmission, more efficient scheduling mechanism and/or frame structure might be needed in order to avoid signaling overhead.

3. Sporadic traffic

   Most MTC traffic is uplink dominant. Therefore, the MTC downlink traffic is more likely to be sporadic. This feature enables sleep/idle mode improvement which saves signaling overhead and power consumption.

4. Extremely low latency

   Some MTC application requires very low latency (e.g. alarms, and pipeline pressure sensor). In order to meet the requirement of low latency, changes to signaling procedure, frame structure, and scheduling mechanism might be needed.

5. Low/No mobility

   Many MTC applications involve low mobility devices (e.g. water/electric metering, video surveillance, and remote payment). For these MTC applications, some simplifications to the protocol may be needed in order to save signaling overhead. For example, the procedure related to handover is not needed for these applications. Some other procedure (e.g. channel estimation, and timing alignment) can also be simplified.

6. Low power consumption

   Most MTC device are powered by battery or self-generating power. Therefore, in order to prolong the lifetime for the MTC device, lower power consumption is crucial. In UMTS and LTE, discontinuous reception is designed for power saving.

7. Low cost

   As ARPD is low for MTC, lowering the cost of MTC is extremely important to enable the deployment of MTC applications. Lowering cost may be attained by reducing the operational power consumption, simplifying the transceiver, etc.

2.2.5 Work towards Machine Type Communications

In order to address the challenges imposed by various kinds of MTC applications, lots of efforts have been made by academia and industry. 3GPP and ETSI are the two most active associations investigating the requirements, application scenario, and network architecture. Table 2.3 lists the standards proposed by ETSI and 3GPP for MTC [27].

Table 2.3: MTC Standards of 3GPP and ETSI

	Specification description	Specification reference
3GPP	Study on facilitating machine to machine communication in 3GPP systems	3GPP TR 22.868
	Service requirements for Machine-Type Communications	3GPP TS 22.368
	System improvements for Machine-Type Communications	3GPP TR 23.888
	Feasibility study on the security aspects of remote provisioning and change of subscription for Machine to Machine (M2M) equipment	3GPP TR 33.812
	Security aspects of Machine-Type Communications	3GPP TR 33.868
	RAN Improvements for Machine-type Communications	3GPP TR 37.868
	GERAN improvements for Machine-Type Communications	3GPP TR 43.868
	Study on provision of low-cost Machine-Type Communications User Equipments (UEs) based on LTE	3GPP TR 36.888
ETSI	Machine-to-Machine communications; Functional architecture	ETSI TS 102 690
	Machine-to-Machine communications; service requirements	ETSI TS 102 689
	Smart metering	ETSI TR 102 692
	eHealth	ETSI TR 102 732
	Connected consumer	ETSI TR 102 857
	Automotive	ETSI TR 102 898
	City automation	ETSI TR 102 897
	M2M Definitions	ETSI TR 102 725
	Threat analysis and counter measures to M2M service layer	ETSI TR 102 167
	Re-use of 3GPP nodes by M2MSC layer	ETSI TR 101 531

In Release 10, 3GPP specifies some general requirements for machine type communications, such as overload control, and addressing, and identifies system improvement for MTC. In Release 11 and beyond, 3GPP mainly considers the continuous enhancement for MTC, for example: low cost MTC, network selection and steering. ETSI defines Functional architecture and service requirements for MTC and specifies some usage scenarios.

Besides 3GPP and ETSI, there are some other associations working on the standardization for MTC. For example [28], IEEE 802.16p (WiMax) investigates the network architecture and requirements for MTC as well as some optimizations to the air interface for MTC. The WiFi Alliance is working on promoting Wi-FI as the preferred wireless technology for smart grid and health care. The GSM Association defines a set of GSM based embeded modules to solve the operational issues for MTC as well as some usage scenarios in vertical market.

In addition to the standardization for MTC, there are also literatures covering various aspects for MTC. There are mainly three types of work:

1. System design for MTC

   As we introduced, MTC has some specific characteristics which is different from H2H communications. Therefore, some improvements on the air-interface are needed in order to accommodate MTC. Typically, there are massive MTC devices in a cell with various traffic characteristics and QoS requirements. Hence, how to avoid the huge signaling overload so as to save resource and meet the QoS requirements for MTC applications becomes crucial. There are lots of literature working on this topic. For example, reference [29]- [32] proposed several methods to improve the performance of random access (random access is the main uplink channel access method for MTC). References [33]- [34] presented group based cooperative access methods for MTC. References [35]- [37] introduced several resource allocation schemes for MTC. Moreover, therefore some other literatures considering other optimizations for MTC. Reference [38] analyzed the power consumption for distributed M2M video sensors. Reference [39] investigated symbol error rate for mobile-relay-based M2M system with double Nakagami-m fading channel. Reference [40] studied the feasibility of cognitive M2M communication on cellular bands.

2. Usage scenario for MTC

   There are numerous applications for MTC. However, each application has some specific requirements. Therefore, some optimizations are needed in order to adapt to the specific requirements for certain M2M applications. References [42]- [45] propose the application scenarios of M2M in the field of smart grid. References [46]- [47] investigate issues of Home M2M: architectures, requirements, design and implementation.

3. Security issues for MTC

   Security is also very important issue for MTC. [48] presents the implementation of a two post-quantum public-key for power limited M2M system. [49] investigates the security risks of M2M by analyzing two cases: a Zoomback and a vehicle security module. [50] proposes a application-layer security compression mechanism for M2M communications by the use of SMS.

2.3 Summary

In this chapter, firstly we introduce the LTE/LTE-A system from the perspective of network architecture and protocol stack. Then, we present the control- and data-plane latency as defined by 3GPP and power characteristics in LTE.

In the second part, we present the network architecture and application scenarios for MTC. With various MTC applications, great benefit can be attained. However, due to the significant differences between H2H and M2M communication, there are also lots of challenges imposed by MTC. To address these challenges, several associations are working the standardization for MTC. Moreover, there are also numerous of literatures proposing methods in order to optimize the performance for MTC.

CHAPTER 3

Random Access Optimization

3.1 Introduction

In LTE, random access procedure is mainly used during RRC connection set up, uplink synchronization, handover between cells, and scheduling request. There are two types of random access: contention based access initiated by UEs and contention free random access coordinated by eNBs. Random access is crucial to support efficiently machine type communication in LTE. The main reason are explained below.

- Random access is mainly used to send scheduling request (SR) to eNB as the regular uplink channel scheduling method is not efficient for MTC application.

  Several problems arise when using regular uplink scheduling mechanism for machine type communications. First, to provide collision free transmission for SR, eNB reserves resources for each UE at certain subframes. For sporadic MTC traffic, such resource reservation mechanism becomes inefficient due to the sporadic nature of MTC traffic pattern. For example: assuming the expected period for an uniformly distributed MTC traffic is 500ms, if the eNB reserves SR resource for that traffic in every subframe, only 1 of the 500 resource is used, which causes a significant waste of resources. On the other hand, if we want to save resource, the resource allocation periodicity could be set as 500 ms, which indicates that a UE has to wait 250 ms on average to access the resource. Second, SR period increases with the number of UE (RRC connected) in a cell. In LTE, the maximum amount of resource for SR transmission in one subframe is 36. Supposing there are 1000 MTC devices in a cell, the SR period increases to $1000/36 =$ 28ms, which also increases the latency. By the use of random access, a UE sends a preamble on the *common* random access resource to request resource from eNB, which eliminates the problems of the regular scheduling.

- Some MTC devices may reside in RRC_IDLE state most of the time to save the signaling overhead and power consumption. For these types of MTC devices, they use random access to send RRC connection request to eNB such that RRC connection can be established between UEs and eNB.

- There are some other cases for the usage of random access: a MTC device uses random access to get uplink synchronized with eNB even if it is in the RRC connected state; a MTC device uses random access to recover from radio link failure, etc.

Therefore, it can be seen that random access is essential for machine type communications in LTE, and play a key role in achievable performance of MTC applications.

However, there are several challenges to use random access for machine type communications in LTE. It is known that the preamble used for random access is not UE specific, therefore collision happens when multiple UEs using the same preamble. Thus, the collision rate becomes very high when there are massive number of UEs in a cell accessing the channel simultaneously. For example, suppose that the total number of UE in a cell is 1000, the packet transmission probability for a UE in one subframe is 0.03 and the available number of preamble is 64, then collision probability is 0.9997, which indicates that (1) most preambles cannot be received by eNB and (2) huge latency is introduced. The second problem is that a UE usually has to wait certain amount of time before starting another random access if the initial random access fails, which greatly increases the latency. For example, assuming the backoff window size is 50, the average backoff time is 25ms before starting the next random access, which in turn introduces an extra latency.

There have been numerous works in the literature addressing the performance of random access. Reference [52] presents a resource allocation scheme for spatial multi-group random access in LTE. Authors in [53] investigate the collision probability of random access method used for MTC application and provide a model to derive the collision probability, the success probability, and the idle probabilities of UE. Reference [54] proposes a joint massive access control and resource allocation schemes to minimize total energy consumption of the M2M system in both flat-fading and frequency selective fading channel. Authors in [55] propose a collision resolution method for random access based on the fixed timing advance information for massive fixed-location MTC in LTE. Reference [56] introduces a new code-expanded method for random access in LTE. With the proposed method, the amount of available contention resource is expanded, which therefore can reduce the collision rate in random access. Reference [57] introduces a cooperative access class barring scheme for global stabilization and access load sharing. In their method, each MTC groups are assigned with specific access class barring to differentiate access priorities. Authors in [58] propose a massive access management method to provide QoS guarantee for MTC devices. In this method, UEs are grouped due to their quality-of-service (QoS) requirements: UEs with deterministic QoS requirement are reserved with resource while UEs with soft QoS guarantees are opportunistically scheduled to improve resource utilization efficiency. Reference [59] analyzes the throughput and access latency for random access in an OFMDA system. However, retransmission users are not considered in their model, which does not complies with the random access mechanism in LTE [60]. Authors in [61] present a prioritized random access scheme to provide QoS for different classes of MTC devices in LTE-A networks, where the different access priorities is achieved by using different backoff procedures. Reference [3] discusses the possible solutions for random access overload control, which includes: access class barring schemes, separate RACH resource for MTC, dynamic allocation of RACH resource, MTC specific backoff scheme, slotted access, and pull based scheme. However, no detailed method is provided there.

To improve the reliability of random access, we propose a packet aggregation method for random access. With our method, a UE does not start a random access for every arrived packet. Instead, it triggers a random access when the number of packets in the buffer reaches a certain threshold. In the example described above if we set the packet aggregation threshold to 5, then the collision probability is reduced to 0.21, which is much lower than the original value 0.9997. Note that the higher the threshold for the packet aggregation, the smaller the collision rate becomes. However, the preamble collision rate is reduced at the expense of extra waiting time used to buffer packets, which may not be desirable for some real time MTC applications. In order to avoid large latency for real-time MTC applications, we firstly derive the packet loss rate and channel access latency as functions of amount of aggregated packet using a Semi-Markov process model [62]. With the derived results, the optimal amount of aggregated packets which maintains the packet loss rate requirement while minimizing the latency can be selected. It has to be mentioned that the another benefit to use the packet aggregation method is to *save the power* by reducing the number of random access transmissions. We also propose a power saving method, where a UE aggregates packets as long as the delay constraint is satisfied.

To reduce the latency caused by unsuccessful random access, we propose a Transmission Time Interval (TTI) bundling scheme. The idea of TTI bundling is introduced in LTE Rel. 8 to improve the uplink coverage for VoIP application [63]. In that method [63], a UE sends a VoIP packet through a bundle of several subsequent TTIs before receiving the HARQ from eNB, which eliminates the latency caused by retransmissions and improves the QoS for VoIP application [64]. Inspired by this, we apply the idea of TTI bundling to random access. With the proposed TTI bundling scheme, a UE sends multiple preambles in several consequent TTIs/subframes. Hence, a random access is successful if one of the preambles is correctly received by eNB and without collision, which eliminates the latency caused by unsuccessful random access. It is obvious that achieving low uplink channel access latency by the use of the TTI bundling method is beneficial to real time MTC applications, for example oil/gas pipeline monitoring, sensor-based fire alarm and automotive collision detection.

It has to be noted the methods proposed in this chapter are mainly designed for the scenario where they are massive number of MTC devices and the traffic arrival rats is not very small (we assume the packet interval is in magnitude of the 100 ms on average). Here we provide two examples:

- Auto pilot [66]- [67]

  As shown in Fig. 3.1, cars in auto pilot system are equipped with M2M devices. These M2M devices send information to the backend system, which is used to detect collision and to avoid collisions. The backend system sends notifications to cars which are close to the location of a collision. These information can be used by automatic system to take proper actions (brake, change direction, etc).

  The cars in an auto pilot system is required to send information about time, position and velocity. The packet interval 100 ms on average. Moreover, for high speeds cars, they are required to send packets every meter, which is the resolution of GPS. In this sense, assuming the speed is 160 km/h (44.5 m/s), the number of packets that a car send in one second is about 45.

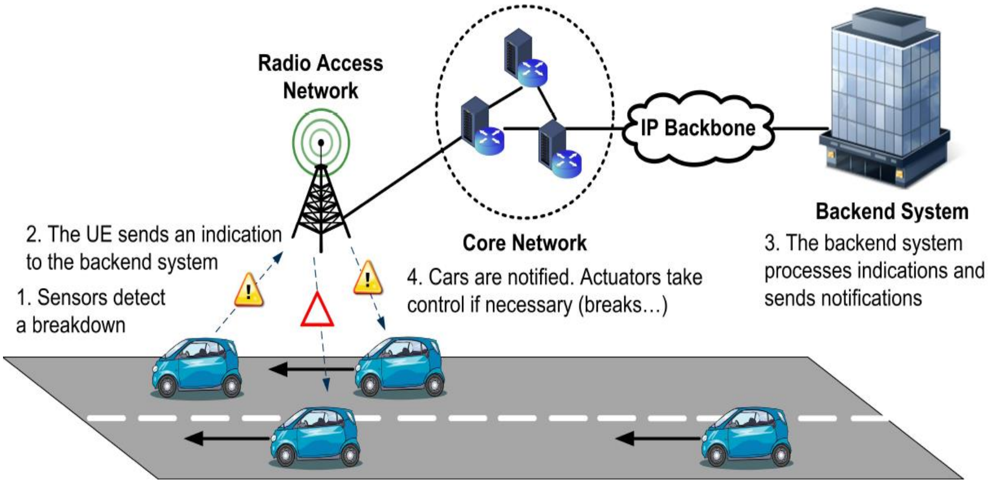

Radio Access Network
IP Backbone
Backend System
Core Network
2. The UE sends an indication to the backend system
1. Sensors detect a breakdown
4. Cars are notified. Actuators take control if necessary (breaks…)
3. The backend system processes indications and sends notifications

Figure 3.1: Auto pilot

- Virtual game [66]- [67]

  Another example for MTC application is the virtual race. For example, in a virtual bicycle race game using real bicycles, opponents are not required to be on the same locations. At the beginning of the game, the length of this game is set, for example 10 km or 20 km. The bicycles are equipped with sensors. These sensors are used to send information about location, temperature, speed, etc. These information is exchanged between opponents to show them their state in the current game.

  Assuming the typical speed of bicycle is around 50km/h, and the packet interval 100 ms, then the resolution of a bicycle's location is around 1.4m.

The reminder of this chapter is organized as following. In Section 3.2, we briefly introduce the random access mechanism in LTE. In Section 3.3, we present the packet aggregation method for random access. Section 3.4 provides the TTI bundling method for random access. Section 3.5 discusses the usage of the proposed method and section 3.6 concludes this chapter.

3.2 Random Access in LTE

The random access mechanism is specified in [60], which includes two types of random access in LTE: contention free random access and contention based random access. Firstly, we introduce the contention free random access.

3.2.1 Contention Free Random Access

The contention free random access is mainly used for handover between cells in order to avoid call drop. The procedure for contention free random access is shown in Fig.3.2. Firstly, the eNB allocates dedicated preambles for UEs. Secondly, a UE sends the allocated preamble on the random access resource. Finally, the eNB sends back the random access response message which includes time advance information, temporary C-RNTI and uplink grant. It can be seen that with contention free random access, a UE can get uplink synchronized with eNB in a short duration, which enables smooth handover between cells. However, contention free random access is not applicable to machine type communications. The reason is very similar to that for the regular scheduling: firstly, if a dedicated preamble is allocated for each UE, due to the limitation of preamble resource in each subframe, the period for a UE to get the dedicated preamble increases with the number of UE. Secondly, for the MTC with sporadic traffic, allocating a dedicated preamble for each UE causes huge waste as a UE may not have traffic to send when the resource is allocated. Therefore, only the contention based random access is suitable for machine type communications in LTE.

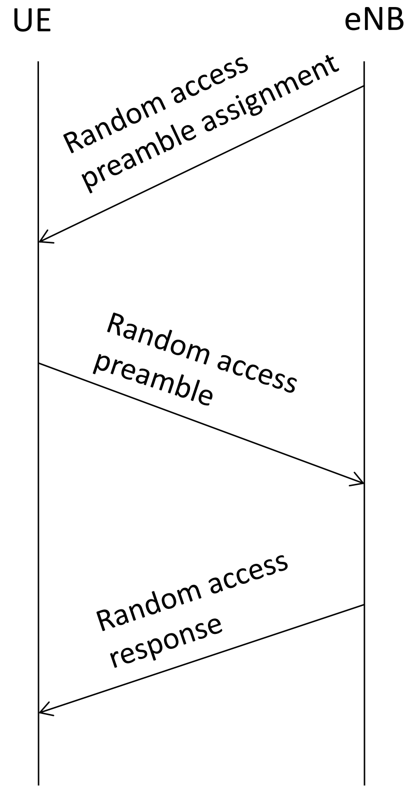


Figure 3.2: Contention free random access in LTE

3.2.2 Contention Based Random Access

The contention based random access is used for UE's state transition from RRC_IDLE to RRC_CONNECTED, recovering from radio link failure, uplink synchronization, sending scheduling request (SR), etc. Recall the contention based random access described in the last chapter (shown in Fig. 3.3), it includes four steps: (a) transmission of a random access preamble; (b) reception of random access response (RAR); (3) transmission of L2/L3 message (SR, RRC connection request); (4) reception of contention resolution.

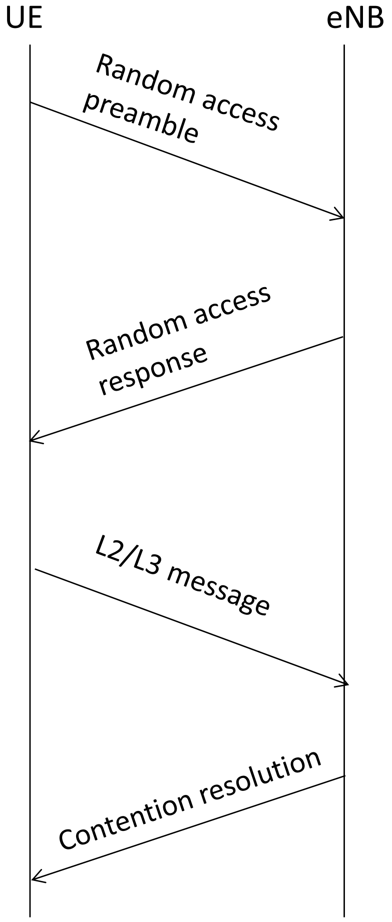


Figure 3.3: Contention based random access in LTE

Random Access Preamble Transmission

A UE select one of the $64 - N_c$ random access preambles randomly, where N_c is the number of preambles reserved for contention free random access. In LTE, Zadoff-Chu sequences [68] is employed for uplink random access preamble transmission due to its low peak-to-average power ration (PAPR) which is important for power limited uplink transmission of UE. The ZC sequence of odd-length N_{ZC} is [9]

$$x_u(n) = \exp[-j\frac{\pi un(n+1)}{N_{ZC}}], n \in [0, N_{ZC} - 1] \tag{3.1}$$

where u is the ZC sequence root index and it should be prime with N_{ZC}.

The ZC sequences have ideal cyclic autocorrelation property which can be given by

$$r_{uu}(\sigma) = \sum_{n=0}^{N_{ZC-1}} x_u(n)x_u^*(n+\sigma) = \delta(\sigma) \tag{3.2}$$

where N_{ZC} is the length of ZC sequence, $x_u(n+\sigma)$ is the cyclic shift version of $x_u(n)$ with shift σ. The random access preambles are obtained from a ZC sequence with different cyclic shifts. Specifically the number of preambles per ZC sequences is

$$N_p = \lfloor \frac{N_{ZC}}{N_{CS}} \rfloor \tag{3.3}$$

where N_{CS} is the cyclic shift size. In FDD-LTE, N_{ZC} and N_{CS} is 839 and 13 respectively, and therefore the number of available preambles per ZC sequence is 64 (including preambles for contention based and contention free random access).

To inform eNB about the packet size of L2/L3 message, the preambles used for contention based access are divided into two subgroups: Random Access Preambles group A and Random Access Preambles B. A UE whose L2/L3 message size is larger than the *messageSizeGroupA* which is configured by eNB selects a preamble from Random Access Preambles B; otherwise it uses preambles in Random Access Preambles group A [60].

The transmission power for random access preamble is

$$P_{rc} = P_t + L + \delta_p + (C_p - 1)R_s \tag{3.4}$$

where P_t is the target received power of the preamble, L is the path loss, δ_p is configured by eNB, C_p is a counter for preamble transmission, R_s is the power ramping step [60].

Random Access Response Reception

After sending the random access preamble, a UE decodes the physical dedicated control channel (PDCCH) with random access-radio network temporary identifier (RA-RNTI) in the random access response window to receive the random access response (RAR) message.

The start point of random access response window and its length *ra-ResponseWindowSize* is configure by eNB through RRC message. The earliest random access response window starts at the subframe that is at the end of the preamble transmission plus two subframes. However, it is also common that there is a slight delay for the random access window (4 or 5 ms). Fig.3.4 shows an example for a random access response window, where its size is 3ms with a delay of 2 ms [9].

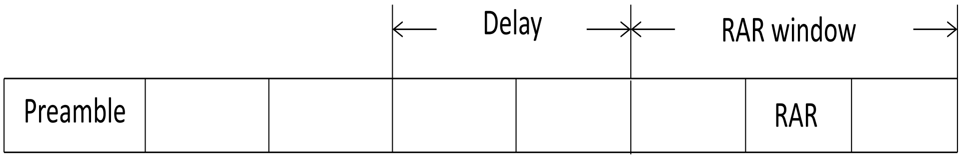


Figure 3.4: Random access response window

The RA-RNTI is computed as:

$$RA - RNTI = 1 + t_{id} + 10f_{id} \tag{3.5}$$

where t_{id} is the index of the first subframe of the specified physical random access channel (PRACH) ($0 \leq t_{id} < 10$), and f_{id} is the index of the specified PRACH within that subframe, in ascending order of frequency domain ($0 \leq f_{id} < 6$).

A UE identifies its RAR through the random access preamble identifier. The RAR message includes the identity of the detected preamble (random access preamble identifier), uplink channel synchronization information, resource allocation information for the subsequent L2/L3 message transmission, backoff indicator which instructs UEs to backoff for certain time before starting the next random access (the backoff time is uniformly selected over a period configured by eNB), temporary C-RNTI, etc [9]. If the UE does not receive a RAR after the random access response window, it starts a new preamble transmission.

It has to be noted that UEs which select the same preamble in step 1 find the same RAR message in this stage and hence they are allocated with the same resource. As a result, they will send the L2/L3 message on the same resource, which may not be received by eNB.

L2/L3 Message Transmission

In this step, UEs send the actual message for this random access procedure, which includes: RRC connection request, handover request, and scheduling request (SR). After sending the L2/L3 message, a UE will wait for the contention resolution message until the contention resolution timer expires.

As mentioned above, collision happens in this stage if UEs select same preamble in the first stage. To help eNB identify collision, the temporary C-RNTI which is allocated in stage 2 and either C-RNTI (for RRC_CONNECTED UE) or the 48-bit UE identity should be transmitted along with the L2/L3 message. It has to be noted that the C-RNTI is unique in the E-UTRAN domain and UE identity is globally unique. If error happens with the L2/L3 message, a UE uses HARQ to retransmit the L2/L3 message to combat the wireless channel error.

Contention Resolution Reception

eNB acknowledges the successfully decoded L2/L3 message through contention resolution message. The contention resolution message is addressed to either the C-RNTI or the temporary C-RNTI of the decoded L2/L3 message (the UE identity should be included in L2/L3 message in the latter case). Therefore, if a UE receives a contention resolution message addressed to itself but cannot find its identity, the UE can infer that collision happens for the previous L2/L3 message transmission and the packet sent by itself is not correctly received. Moreover, a random access is also considered as unsuccessful when the contention resolution timer expires. For an unsuccessful random access, a UE can trigger another random access when the backoff counter expires.

3.3 Packet Aggregation for Machine Type Communications with Random Access

In this section, we consider that UEs use random access to send SR to apply for resource from eNB. Therefore, the latency treated in this section is user plane latency. There are two types of scenarios analyzed in this section: single type of traffic and multiple types of traffic. The first scenario is related to the case where dedicated preamble resource is allocated for each type of MTC traffic. The second scenario corresponds to the case where UEs with different types of traffic share the common preamble resource.

3.3.1 Packet Aggregation for Random Access with Single Type of Traffic

As discussed above, the main problem of random access is that the collision becomes very high if there are massive UEs in a cell. To address this problem, we propose a packet aggregation method. With our method, a UE does not trigger a random access until the buffered packet reaches the given threshold [1]. As the number of transmission is reduced through packet aggregation, the preamble collision rate is also decreased. However, this method introduces an extra latency in order to accumulate multiple packets. Moreover, the packet loss rate might be increased due to the increased packet size, which also increases the latency. However, for the sake of simplicity, we ignore this effect in our method.

The standard slotted Aloha system also uses the kind of random access as it channel access method. However, there are some differences between the slotted Aloha and the random access in LTE:

1. The retransmission number for random access in LTE is finite. While for the standard Aloha system, the retransmission number is infinite, i.e., a UE keeps sending a packet until it is correctly received.

2. For the random access mechanism in LTE, a packet can be generated during the random access. This new generated packet can be delivered along with the existing packet by one transmission. Therefore, the number of random access attempts are reduced (**In section 3.3.3, we provide an example to demonstrate this effect**). However, for the slotted Aloha system, the new arrived packet are backlogged until the precedent packets are successfully transmitted.

3. For the random access in LTE, after sending a preamble, a UE has to wait for certain time to receive the random access response, to send the L2/L3 message, and to receive the contention resolution message. In contrast, the slotted Aloha do not have such mechanism. The effect of such mechanism is that packets can be generated during a random access. As we explained, these new generated packets are delivered along with existing aggregated packets through one transmission, which reduces the random access attempts.

[1] This threshold is set by eNB and sent to each UE.

The above differences make it less obvious to analyze the random access with the methods used for Aloha performance analysis, which motivates us to propose a method to address this problem.

Bianchi [69] proposed a discrete-time Markov chain model to analyze the distributed coordinated functions (DCF) of IEEE 802.11. In that model, it is assumed that each station is saturated with traffic, i.e., the transmission queue of each station is assumed to be always full. However, this assumption does not apply for our case where the traffic might be sporadic. To address this problem, here we use a Semi-Markov process model. In a Semi-Markov process model, each state has its sojourn time (or holding time), i.e., a state is kept for a certain amount of time. Therefore, it is intuitive to use this characteristic to model the non-saturated traffic: a station is kept at the idle state until a packet arrives. Here we introduce one type of state to represent the packet generation procedure. For a state which is used to represent packet generation procedure, its sojourn time can be calculated as $\frac{1}{\lambda}$, where λ is the packet arrival rate. Hence, by the use of Semi-Markov process model, we can extend Bianchi's model to analyze the scenario where the traffic is not saturated. Our method to handle the idle state is very similar to the methods used in [70]- [75], where they use idle state to the represent the case where there is no packet. If a packet arrives, it starts to backoff. Otherwise, it stays at the idle state.

To use the Semi-Markov process method, the assumptions are listed as following:

1. Similar to the assumption used in Bianchi's paper [69], here we also assume that each packet collides with constant and independent probability. The assumption is feasible when the backoff window and number of UE are large.

2. Regardless of the packet size, all the packets in a UE's buffer can be sent by one uplink transmission. This assumption handles the following problem: during the random access, it is possible that new packets are generated. With this assumption, these new packets are delivered with existing packets by one transmission. Therefore, when a UE re-starts at the initial state, there is no packet in its buffer. Due to the memoryless characteristic, the waiting time for the first packet is still $\frac{1}{\lambda}$, where λ is the packet arrival rate.

3. The packet arrival is Poisson distributed.

4. The random access channel is available in every subframe, which is related to random access resource configuration index 14 as specified 3GPP TS. 211 [13]. This assumption is applicable to the scenario where there are massive MTC devices generating high access rate, for example: thousands of sensors for MTC applications in smart city [65].

5. The probability τ that a station will attempt transmission in one subframe is constant across all subframes [69].

The Semi-Markov process model is shown in Fig.3.5, where

- aggregation (idle) state $S_{0,n}$, $n \in [0, N]$, means the random access is not started and there are n packets in the UE's buffer where N is the packet aggregation threshold;

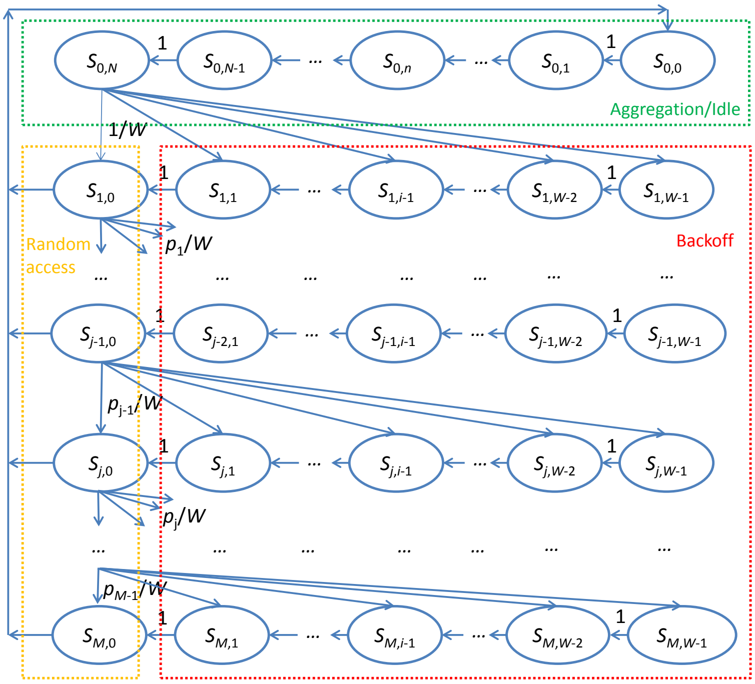


Figure 3.5: Semi Markov process model for random access with packet aggregation

- backoff state $S_{j,i}$, $j \in [1, M], i \in [1, W-1]$, means the the UE is performing the backoff with a counter size of i for the jth random access, where M is the random access limit and W is the maximum backoff counter size.
- random access state $S_{j,0}$, $j \in [1, M]$, means that the UE is performing the jth random access.

A UE transfers between states as follows:

1. When the UE is at state $S_{0,n}$, $n \in [0, N-1]$, for each arrived packet it transfers to state $S_{0,n+1}$.
2. When the UE is at state $S_{0,N}$, it selects a random number i which is uniformly distributed over $[0, W-1]$ and then transfers to state $S_{1,i}$.
3. When the UE is at state $S_{j,i}$, $j \in [1, M], i \in [1, W-1]$, it decrease its backoff counter by 1 after one subframe and transfers to state $S_{j,i-1}$.
4. When the UE is at state $S_{j,0}$, $j \in [1, M-1]$, it starts a random access. If the UE is allocated with some resource after the random access (the random access is successful), it sends the aggregated packet and transfers to state $S_{0,0}$. Otherwise, it increases the transmission counter by 1 and transfers to state $S_{j+1,i}$, where i is uniformly distributed over $[0, M-1]$.
5. When the UE is at state $S_{M,0}$, it performs the random access. If the random access is successful, it sends the aggregated packet on the allocated resource and transfers to state $S_{0,0}$. Otherwise, it drops the aggregated packet and transfers to state $S_{0,0}$.

Table 3.1: Symbols used in Section 3.3.1

Notation	Definition
d	overall latency for the first aggregated packet
d'	expected time used to deliver an aggregated packet
d_0	time used for one random access with packet aggregation
$E_{i,j}$	probability that a preamble cannot be decoded by eNB when it is sent by $i+1$ UEs for the jth random access
h_j	average state holding time for state $S_{j,0}$
M	transmission limit for random access
N	packet aggregation threshold
N_C	amount of preamble allocated for contention based random access
N_M	amount of MTC device in the cell
N_{max}	maximum allowed amount of aggregated packets
p_c	preamble collision probability
$p_{E,j}$	probability that a preamble cannot be detected by eNB when collision happens for the jth random access
p_j	unsuccessful probability for the jth preamble transmission
p'_j	error probability caused by wireless channel for the jth preamble transmission
Q_j	proportion of time that a UE is at state $S_{j,0}$
$S_{j,i}$	state in Semi-Markov process
T	average state holding time for all states
T_C	duration that starts after the time instant when the UE sends a preamble and ends at the time instant when the contention resolution timer expires
T_E	duration that starts after the time instant when the UE sends a preamble and ends at the time instant which is the end of the random access response window
T_j	channel access latency if the aggregated packet is successfully delivered at jth try
T'_j	duration for an unsuccessful random access at the jth try
T_S	duration that starts after the time instant when the UE sends a preamble and ends at the time when a UE receives the contention resolution message from eNB
W	backoff window size
α	packet loss rate limit
β	power saving factor to measure energy saving with packet aggregation
λ	packet arrival rate
$\pi_{j,i}$	stationary probability for state $S_{j,i}$
τ	probability that a UE is sending a preamble in one subframe

It has to be noticed that a packet may be generated during random access. In this case, this packet generation does not trigger the state transition. Instead, a UE remains at the current state and this new generated packet will be sent along with the existing packet when the random access is successful.

Denoting p_j, $j \in [1, M-1]$, as the unsuccessful probability for the jth preamble transmission, the state transition probability from $S_{j,0}, j \in [1, M-1]$ to $S_{j+1,i}, i \in [0, W-1]$, is p_j/W.

An unsuccessful random access can be caused by wireless channel error or collision, therefore we have

$$p_j = p_c + p'_j - p'_j p_c \tag{3.6}$$

where p_c is the preamble collision probability and p'_j is the error probability caused by wireless channel for the jth preamble transmission.

Denoting $\pi_{j,i}$ as the stationary probability of state $S_{j,i}$, we have

$$\begin{cases} \pi_{0,n} = \pi_{0,0}, n \in [1, N], \\ \pi_{1,i} = \pi_{0,N} 1/W + \pi_{1,i+1}, i \in [0, W-2], \\ \pi_{j,i} = \pi_{j-1,0} p_{j-1}/W + \pi_{j,i+1}, j \in [2, M], i \in [0, W-2], \\ \pi_{1,W-1} = \pi_{0,N} 1/W, \\ \pi_{j,W-1} = \pi_{j-1,0} p_{j-1}/W, j \in [2, M], \end{cases} \tag{3.7}$$

With the first, second and fourth equation in equation system (3.7), we have

$$\pi_{1,0} = \pi_{0,N} \tag{3.8}$$

$$\pi_{1,i} = \frac{W-i}{W}\pi_{0,N} = \frac{W-i}{W}\pi_{1,0}, i \in [1, W-1]. \tag{3.9}$$

By the use of the third and fifth equation in equation system (3.7), we get

$$\pi_{j,0} = p_{j-1}\pi_{j-1,0}, j \in [2, M] \tag{3.10}$$

$$\pi_{j,i} = \frac{W-i}{W} p_{j-1}\pi_{j-1,0} = \frac{W-i}{W}\pi_{j,0}, j \in [2, M], i \in [1, W-1]. \tag{3.11}$$

As the sum of all state stationary probabilities is one, we have

$$\begin{aligned} 1 &= \sum_{n=0}^{N} \pi_{0,n} + \sum_{j=1}^{M}\sum_{i=0}^{W-1} \pi_{j,i} \\ &= \pi_{0,0}(N+1) + \sum_{j=1}^{M} \pi_{j,0} \sum_{i=0}^{W-1} \frac{W-i}{W} \\ &= \pi_{0,0}(N+1) + \sum_{j=1}^{M} \pi_{j,0} \frac{W+1}{2} \\ &= \pi_{0,0}(N+1) + \frac{W+1}{2} \sum_{j=1}^{M}\prod_{i=0}^{j-1} p_i \pi_{0,N} \\ &= \pi_{0,0}[(N+1) + \frac{W+1}{2} \sum_{j=1}^{M}\prod_{i=0}^{j-1} p_i] \end{aligned} \tag{3.12}$$

where p_0=1.

Therefore,

$$\pi_{0,0} = 1/[N+1+\frac{W+1}{2}\sum_{j=1}^{M}\prod_{i=0}^{j-1}p_i] \tag{3.13}$$

which is a function of p_c.

With equations (3.8) and (3.10), the stationary probability $\pi_{j,0}, j \in [1, M]$ is given by

$$\pi_{j,0} = \prod_{i=0}^{j-1}p_i\pi_{0,N} = \prod_{i=0}^{j-1}p_i\pi_{0,0} \tag{3.14}$$

which is also a function of p_c.

Now let us calculate the state holding time for this Semi-Markov process model.

It is obvious that the state holding time for $S_{0,N}$ and $S_{j,i}$, $j \in [1, M], i \in [1, W-1]$ is 1 ms. Moreover, the average state holding time for state $S_{0,n}, n \in [0, N-1]$ is $1/\lambda$.

The calculation for the state holding time $S_{j,0}, j \in [1, M]$, is less obvious.

There are four different results for a random access:

1. the preamble is delivered without collision but with wireless channel error. The probability for this case is $p'_j(1-p_c)$, where p_c is the collision probability and p'_j is the error probability for the jth preamble transmission.

2. the preamble is transmitted with collision but the transmitted preamble is not correctly detected by eNB due to wireless channel error. The probability for this case is $p_c p_{E,j}$, where $p_{E,j}$ is the probability that a preamble cannot be detected by eNB when collision happens for the jth random access. Assuming amount of preamble allocated for contention based random access is N_C, and the amount of MTC device in the cell is N_M, we can calculate:

$$p_{E,j} = \sum_{n=1}^{N_M-1}\binom{N_M-1}{n}\tau^n(1-\tau)^{N_M-1-n}\sum_{i=1}^{n}\binom{n}{i}(\frac{1}{N_C})^i(1-\frac{1}{N_C})^{n-i}E_{j,i+1}. \tag{3.15}$$

   where $E_{j,i+1}$ is the probability that a preamble cannot be decoded by eNB when it is sent by $i+1$ UEs for the jth random access, and τ is the probability that a UE is sending a preamble in one subframe.

3. the preamble is transmitted with collision and the transmitted preamble is correctly detected by eNB. The probability for this case is $p_c(1-p_{E,j})$. This case is quite typical in LTE. For example, if a preamble is sent by two UEs, since the location of these two UEs are different, two peaks for the same preamble may appear at the eNB side. The probability that all these two peaks cannot be detected by eNB is relative low. Indeed, that is why contention resolution mechanism is used in random access. Contention resolution is used to handle the problem where multiple UEs send same the preamble and then send the L2/L3 message on the same resource.

4. the preamble is successfully transmitted and without collision. The probability for this case $(1 - p'_j)(1 - p_c)$.

The state holding time for these four events can be calculated respectively as:

1. For the first case, if the transmitted preamble is not correctly received by eNB, then no random access response (RAR) is sent by eNB to that UE. Therefore, a UE re-starts a random access when the random access response window ends. The state holding time for this case is denoted as T_E which is the duration that starts after the time instant when the UE sends a preamble and ends at the time instant which is the end of the random access response window as described in the last section.

2. For the second case, a UE cannot receive the RAR packet. Therefore the state holding time is also T_E, which is the same as the first case.

3. For the third case, multiple UEs send packets (L2/L3 message) on the same resource. Assuming that none of collided packets can be decoded by eNB, a collided UE cannot receive the contention resolution message in this case. Hence, a UE will restart a random access when the contention resolution timer expires. The state holding time for this case is denoted as T_C which is the duration that starts after the time instant when the UE sends a preamble and ends at the time instant when the contention resolution timer expires.

4. For the fourth case, state holding time is denoted as T_S which is the duration that starts after the time instant when the UE sends a preamble and ends at the time when a UE receives the contention resolution message from eNB.

Hence the expected state holding time for state $S_{j,0}, j \in [1, M]$, is

$$h_j = p'_j(1 - p_c)T_E + p_c p_{E,j} T_E + p_c(1 - p_{E,j})T_C + (1 - p'_j)(1 - p_c)T_S. \tag{3.16}$$

With the above results, the proportion of time that a UE is at $S_{j,0}, j \in [1, M]$ is

$$Q_j = \frac{\pi_{j,0} h_j}{T}. \tag{3.17}$$

where

$$T = \pi_{0,N} + \sum_{n=0}^{N-1} \pi_{0,n} \frac{1}{\lambda} + \sum_{j=1}^{M} \sum_{i=1}^{W-1} \pi_{j,i} + \sum_{j=1}^{M} \pi_{j,0} h_j. \tag{3.18}$$

is the average state holding time for all states.

A UE triggers a random access in state $S_{j,0}, j \in [1, M]$, and the time used to transmit a preamble is 1ms. Therefore the probability that a UE is sending a preamble is

$$\tau = \sum_{j=1}^{M} \frac{1}{h_j} Q_j. \tag{3.19}$$

which is a function of p_c as Q_j and h_j are the functions of p_c.

For a given UE, its preamble is collided if this preamble is also sent by any other UEs. Therefore, then collision probability p_c is calculated by

$$p_c = \sum_{i=1}^{N_M-1} \binom{N_M - 1}{i} \tau^i (1-\tau)^{N_M-1-i} (1 - (\frac{N_C - 1}{N_C})^i). \tag{3.20}$$

which is a function of τ.

It can be seen that equations (3.19) and (3.20) comprise a nonlinear equation system, which could be solved by numerical methods. The pseudo code to solve this nonlinear equation system is shown in Algorithm 1 . Therefore, we can get the collision probability p_c and τ.

Algorithm 1 Numerical method to solve the nonlinear equation system (3.19) and (3.20)

```
Input: preamble collision rate $p_c$ is initialized as 0.0002
Output: preamble collision rate $p_c$ and transmission probability $\tau$
while $p_c < 1$ do
    use (3.19) to compute transmission probability $\tau$
    use (3.20) to compute new preamble collision rate $p_{c1}$
    if $|p_c - p_{c1}| < 0.002$ then
        Break;
    else
        $p_c = p_c + 0.0002$
    end if
end while
```

An unsuccessful a random access can be caused by wireless channel error or collision as described above, hence the duration for an unsuccessful random access at the jth try is

$$T'_j = \frac{p'_j(1-p_c) + p_c p_{E,j}}{p'_j(1-p_c) + p_c} T_E + \frac{p_c(1-p_{E,j})}{p'_j(1-p_c) + p_c} T_C. \tag{3.21}$$

the average state holding time for all states.

If a random access is successful at the first try, the channel latency T_1 includes the backoff time and time to perform a random access. Hence

$$T_1 = T_S + W/2. \tag{3.22}$$

If a random access is successful at the jth try ($j > 2$), the channel access latency T_j includes the latency caused by the precedent unsuccessful random access and the latency of the current random access. Therefore, we have

$$T_j = \sum_{n=1}^{j-1} [T'_n + W/2] + T_S + W/2, j > 2. \tag{3.23}$$

Then the expected time used to deliver an aggregated packet is

$$d' = \frac{1 - p_1}{1 - \prod_{j=1}^{M} p_j} T_1 + \sum_{j=2}^{M} \frac{(1-p_j) \prod_{n=1}^{j-1} p_n}{1 - \prod_{j=1}^{M} p_j} T_j \tag{3.24}$$

where $p_j = p'_j + (1 - p'_j)p_c$ is the unsuccessful probability of a random access at the jth try.

With the above results, in the term of the first aggregated packet, the overall latency: time to accumulate packets plus time to deliver the aggregated packet is

$$d = \frac{N-1}{\lambda} + d' \tag{3.25}$$

where the $\frac{N-1}{\lambda}$ is the waiting time needed to aggregate N packets.

One usage for packet aggregation is to reduce packet loss rate. It is obvious that the preamble collision rate decreases as the packet aggregation number increases, i.e., the preamble collision rate can be very small if we aggregate large number of packets. However, the latency increase with the packet aggregation number, which is not desirable for real time MTC applications. Therefore, the optimal packet aggregation number should be set such that the packet loss rate is less than the constraint while the latency is kept as small as possible:

$$\begin{aligned} \arg\min_{N} \quad & d \\ \text{subject to} \quad & \prod_{i=1}^{M} p(i) < \alpha, \\ & N < N_{max}, \end{aligned} \tag{3.26}$$

where α is the packet loss rate limit, and N_{max} is the the maximum allowed amount of aggregated packets which is determined by buffer size. As we do not have a closed form of d, therefore this optimization cannot be solved by any specific optimization method. Instead, we use exhaustive search to solve this problem.

The another usage for packet aggregation is energy saving. It is obvious that, by the use of packet aggregation, a MTC device reduces the number of random access, which hence saves the energy. To measure this effect, here we define the energy saving factor β as

$$\beta = 1 - \frac{E_t + P_a T_a}{N E_t + N P_a T_u} \tag{3.27}$$

where E_t is the energy used for preamble and L2/L3 message transmission in one random access, P_a is the average power when a UE is at active state, T_a is resulted latency when using packet aggregation, and T_u is the latency when packet aggregation is not used. A UE usually spends much more power for transmission than active state [82], i.e., $E_t \gg P_a T_a$ and $N E_t \gg N P_a T_u$. Therefore, for simplicity, we can approximate the above formula as [2] :

$$\beta = 1 - \frac{1}{N}. \tag{3.28}$$

For a power constrained MTC device, packet aggregation can be used for energy saving. However, as we discussed above, packet aggregation increases packet latency. Therefore,

[2] This is a very roughly approximation. When packet aggregation number N increases, T_a also increases, therefore the power spent at active state $P_a T_a$ and $N P_a T_u$ may not be omitted.

the number of aggregated should be carefully selected to comply with delay constraint:

$$\begin{aligned} \underset{N}{\text{arg max}} \quad & \beta \\ \text{subject to} \quad & d < d_{lim}, \\ & N < N_{max}, \end{aligned} \tag{3.29}$$

where d_{lim} is the delay limit.

3.3.2 Packet Aggregation for Random Access with Multiple Types of Traffic

In the last subsection, we consider single type of traffic in our model. In that method, we reduces the packet collision rate at the expense of increasing the channel access latency. Here we consider two types of traffic in our model: real time traffic with small latency requirement and non-real time traffic with large latency requirement. Here, we reduce collisions through packet aggregation for the real time as well as non-real time traffic. Moreover, for the non-real time traffic with larger latency requirement, more packets are aggregated in order to avoid large latency increase for the real time traffic.

Here we assume that the number of MTC device with real time traffic is N_r and the number of MTC device with non-real time traffic is N_n. The packet aggregation number for real time and non-real time MTC device is denoted as a_r and a_n, respectively. With the Semi-Markov process model proposed in the last subsection, we can easily derive the probability that a MTC with real time traffic or non-real time traffic is sending a preamble: τ_r and τ_n, respectively. Similar to that derived in the last subsection, τ_r and τ_n are the functions of the preamble collision probability. The preamble collision probability for real time traffic is:

$$p_r' = \sum_{i=0}^{N_r-1} \binom{N_r-1}{i} \tau_r^i (1-\tau_r)^{N_r-1-i} \sum_{j=0}^{N_n} \binom{N_n}{j} \tau_n^j (1-\tau_n)^{N_n-1-j} (1-(\frac{N_C-1}{N_C})^{i+j}). \tag{3.30}$$

and the preamble collision rate for non-real time traffic is:

$$p_n' = \sum_{i=0}^{N_n-1} \binom{N_n-1}{i} \tau_n^i (1-\tau_n)^{N_n-1-i} \sum_{j=0}^{N_r} \binom{N_r}{j} \tau_r^j (1-\tau_r)^{N_r-1-j} (1-(\frac{N_C-1}{N_C})^{i+j}). \tag{3.31}$$

which are functions of τ_r and τ_n. By the use of numerical calculation method, we can get the preamble collision probability p_r' and p_n', and preamble transmission probabilities τ_r and τ_n. We also notice that when τ_r and τ_n are small, $p_r' \approx p_n'$, which could simplify the numerical calculation.

With the above formulas, similar to the method used in the last subsection, the latency for the real time traffic d_r and non-real time traffic d_n, and the packet loss rate for the real time p_r and non-real time traffic p_n can be calculated.

As introduced above, the transmission for the non-real time can be postponed in order to reduce latency for the real time traffic (objective function). However, though the delay requirement for the non-real time is usually large, the packet aggregation number for the non-real time should not be too large to violate its delay requirement (the first constraint). Moreover, similar to that in the last subsection, with packet aggregation we also want to keep

Table 3.2: Symbols used in Section 3.3.2

Notation	Definition
N_n	number of MTC device with real time traffic
N_r	number of MTC device with non-real time traffic
a_n	packet aggregation number for real time
a_r	packet aggregation number for non-real time
τ_n	the probability a MTC with real time traffic is sending a preamble
τ_r	the probability a MTC with non-real time traffic is sending a preamble

the packet drop rate less than the threshold (the second and third constraint). Therefore, the optimal packet aggregation number is:

$$\begin{aligned} \underset{a_r, a_n}{\text{arg min}} \quad & d_r \\ \text{subject to} \quad & d_n < D_n, \\ & p_r < P_r, \\ & p_n < P_n, \\ & a_r < A_r, \\ & a_n < A_n, \end{aligned} \tag{3.32}$$

where D_n is the delay constraint for the non-real time traffic, and P_r and P_n are the packet loss rate threshold for the real time and non-real time traffic, and A_r and A_n are the maximum packet aggregation number for the real time and non-real time traffic which are determined by the size of buffer.

3.3.3 Results

Parameters

We provide some results in this section. The parameters for random access are shown in Table 3.3. In Table 3.3, we use the value specified in [12] for contention response window T_{RARW} and contention resolution timer T_{timer}, and the value specified in [60] for backoff window W. The value for parameter T_{RAR} and T_{CR} is obtained by assuming that the time used to decode preamble, SR message, and contention resolution message is 3ms, respectively. Here the number of preamble is set to 20, since other preambles is allocated for other traffics.

In case of no collision, the preamble detection rate is assume to be $1 - \frac{1}{e^j}$ as that used in [3], where $j \in [1, M]$ indicates the jth preamble transmission. When a preamble are sent by multiple UEs, it can always be correctly decoded, i.e., $E_{j,i}$=0 when $i \geq 2$. This assumption is quite typical in LTE. If a preamble is sent by two UEs, two peaks appear at the eNB side. The probability that neither peak can be decoded by eNB is relatively low.

Table 3.3: Parameters

Parameter	Description	Value
N_C	Number of preamble	20
M	Transmission limit	5
T_C	State holding time if a preamble is transmitted with collision and it is correctly detected by eNB ($T_C = T_{RAR} + T_D + T_{timer}$)	32 ms
T_{CR}	Duration which starts at time instant when a UE sends the SR message and ends at the time instant when a UE decodes the contention resolution message	8 ms
T_D	Time used to decode a RAR message	3 ms
T_E	State holding time if no RAR message is received by UE ($T_E = T_{RAR} + T_{RARW}$).	15 ms
T_{RAR}	Duration that starts at the end of a preamble's transmission and ends at the time instant when the RAR message can be received	5 ms
T_{RARW}	Random access response window	10 ms
T_S	State holding time if a preamble is corrected received and without collision ($T_S = T_{RAR}+T_D+T_{CR}$)	16 ms
T_{timer}	Contention resolution timer	24 ms
W	Backoff window size	30

Model validation

We develop the random access mechanism with MATLAB following 3GPP 36.321 [60]. Here we set the packet aggregation number N to 1 and 2 to represent the regular random access and random access with packet aggregation. We compare the simulated results with the analytical results obtained using our proposed method in Fig.3.6 and Fig.3.7 [3]. We can see that the analytical results match the simulation results, when hence validate our method.

[3] Here we compare preamble collision rate instead of packet loss rate. The reason is that sometimes the packet loss rate is very small, e.g., it is very close to 0 when $\lambda = 1/300$, N=2 and number of UE is less than 1500. Therefore, the difference between simulated results and analytical results cannot be clearly seen in the figure when comparing.

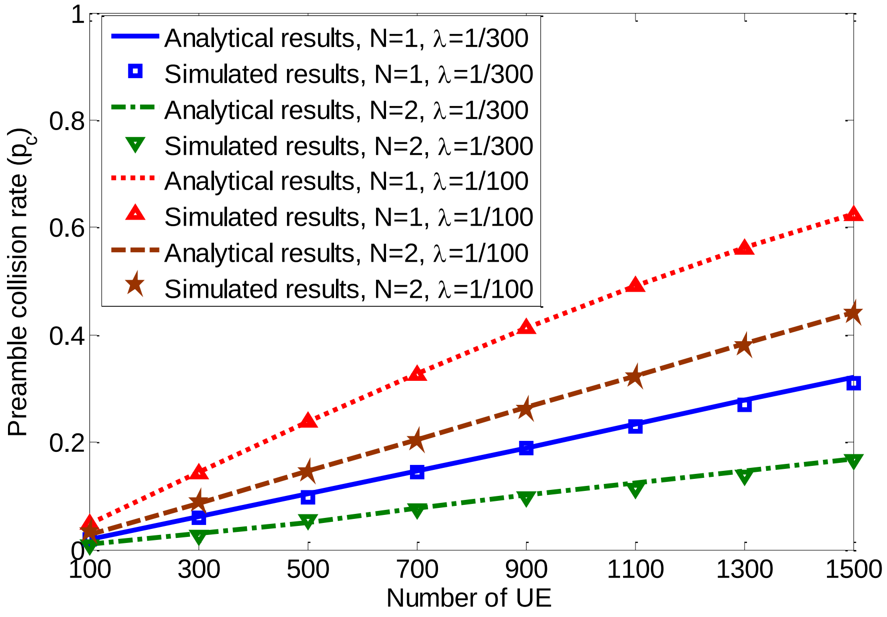


Figure 3.6: Preamble collision comparison

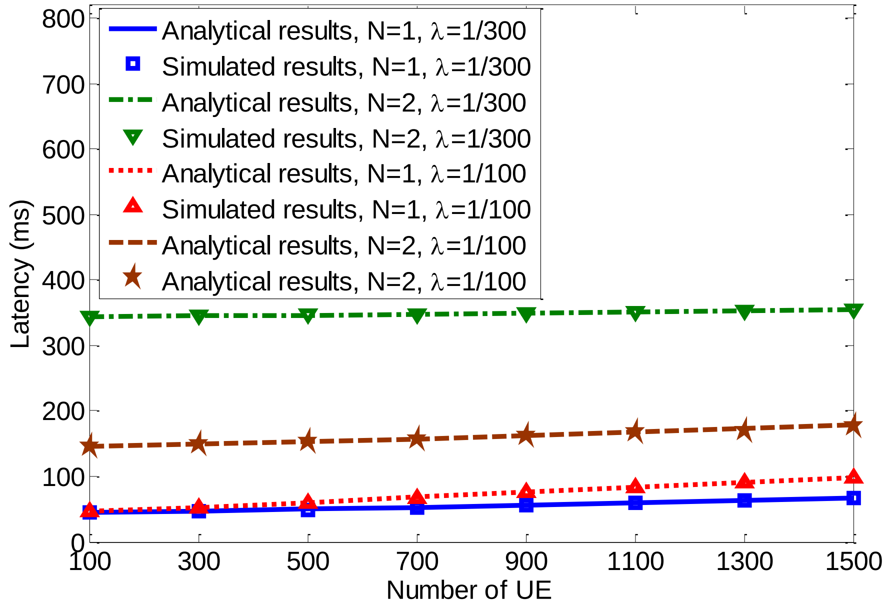


Figure 3.7: Latency comparison

Packet Aggregation for Packet Loss Rate Reduce with Single Type of Traffic

We vary the number of UE and packet arrival rate to see the performance of the proposed method: number of aggregated packet, packet loss rate, latency, and power saving factor.

Here the packet loss rate threshold α is set as 0.1 and the maximum number of aggregated packets N_{max} is 50. The average packet interval is 100 ms, 200 ms, and 500 ms, which is much larger than the backoff window size.

Fig. 3.8 shows the amount of aggregated packet under different number of UEs and packet arrival rate λ (packets/ms) when using our proposed method. It can be seen that the amount of aggregated packet non-decreases with the increase of packet arrival rate or number of UE . This is reasonable since the collision rate increases with packet arrival rate or number of UE. If the collision rate is larger than the given threshold, aggregating more packets is needed to lower the collision. Otherwise, the amount of aggregated packet may not be increased. For instance, when λ=1/100 and the number of UE is 2000, the amount of aggregated packet is 2 and the packet drop rate is 0.093 which is very close to the threshold 0.1. Therefore, when the number of UE increases to 2500, the amount of aggregated packets increase to 3, which reduces the packet drop rate to 0.07 (see Fig. 3.9). We also notice that the packet aggregation number is always 1 when $\lambda = 1/500$. The reason is that the packet loss rate is always less than threshold (0.1) when number of UE increases. For example, the loss rate is 0.05 when $\lambda = 1/500$ and number of UE is 3500. **We also find that the packet aggregation number are different even if the number of average new transmission are the same**. For example, the packet aggregation number is 1 when $\lambda = 1/100$ and number of UE is 1000 (the average new transmissions in one subframe is 10). In contrast, the packet aggregation number is 2 when $\lambda = 1/300$ and number of UE is 3000 (the average new transmissions in one subframe is also 10). The reason is explained as following. In the above example, the latency is 80 ms when $\lambda = 1/100$ and number of UE is 1000. Hence 55 % packets are generated during random access ($1 - e^{(-80/100)} = 0.5507$). As we explained, these new generated packets during a random access are delivered with the existing packets by one transmission [4]. Therefore, only 45 % new generated packet trigger random access attempts. In contrast, the latency is 100 ms when $\lambda = 1/300$ and number of UE is 3000. And the percentage of packets which are generated during random access is 28% ($1 - e^{(-100/300)} = 0.28$). Hence, 72% packet new generated packets trigger random access attempts, which is much higher than case where $\lambda = 1/100$ and number of UE is 1000. Therefore, its collision rate is increased. As a result, the packet aggregation number is set to 2 to lower the packet loss rate.

Fig. 3.9 demonstrates the packet loss rate when using the packet aggregation results shown in Fig. 3.8. It can be seen that with our method the packet loss rate is lower than the packet loss rate threshold (0.1), which validates our method. In contrast, without packet aggregation the packet loss rate is high especially when $\lambda = 1/100$ and the number of UE is larger than 2000.

As discussed above, we lower the packet loss rate at the expense of latency increase. Fig. 3.10 compares the channel access latency with or without packet aggregation. We can see that the latency is increased when using packet aggregation. For example, the latency is increased from 110 ms to 193 ms when $\lambda = 1/100$, number of UE is 2000, and the packet aggregation number is 2. We can imagine that if the delay constraint for this MTC application is 150 ms, then aggregating 2 packets is not feasible. Therefore, for real time MTC application, if the resulted latency is larger than delay constraint after using packet aggregation, more preambles and/or PRACH resource should be allocated by eNB. While for a MTC application with

[4] The Aloha protocol does not has this mechanism. Therefore, the performance of random access and Aloha are different under the same packet arrival rate and number of UE.

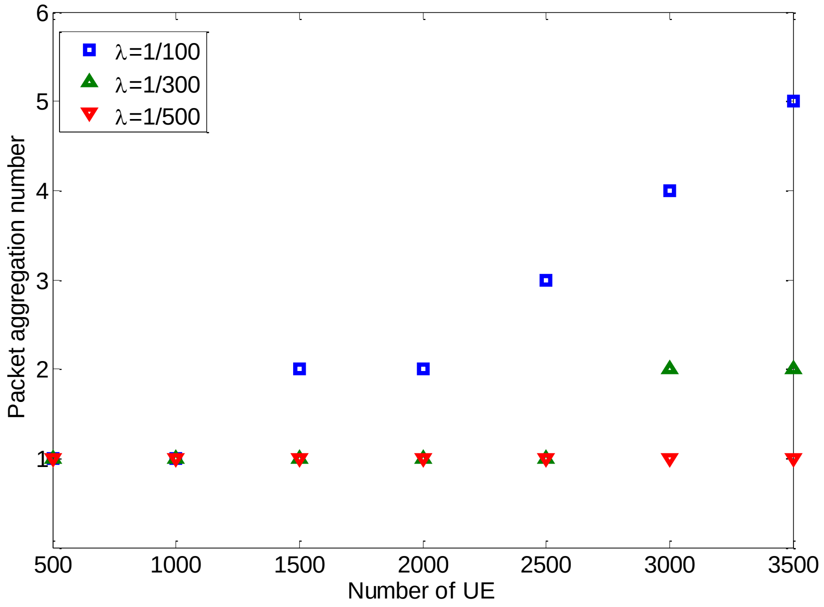


Figure 3.8: Amount of aggregated packet

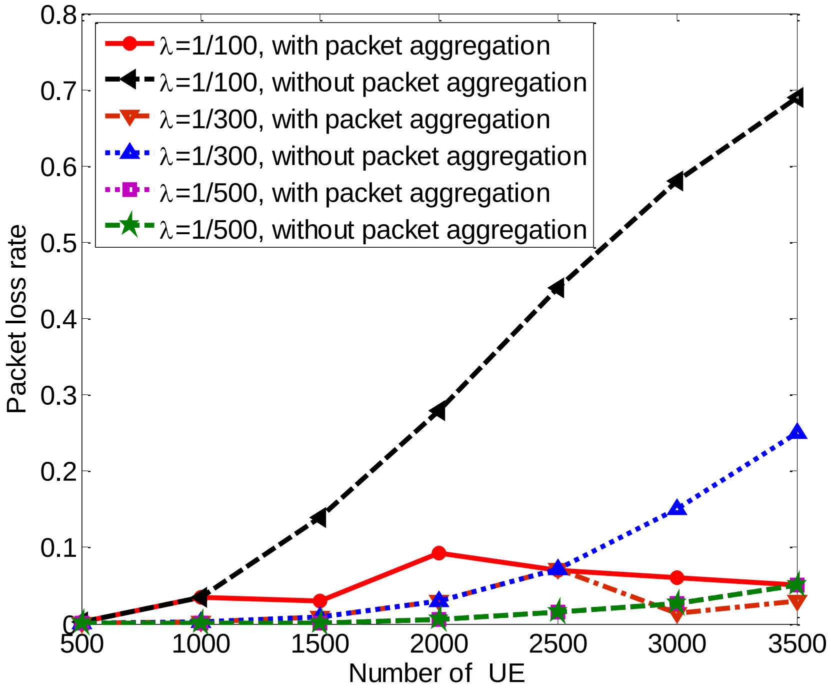


Figure 3.9: Packet loss rate

elastic delay constraint, by the use of packet aggregation the packet loss rate can be reduced to a very small value.

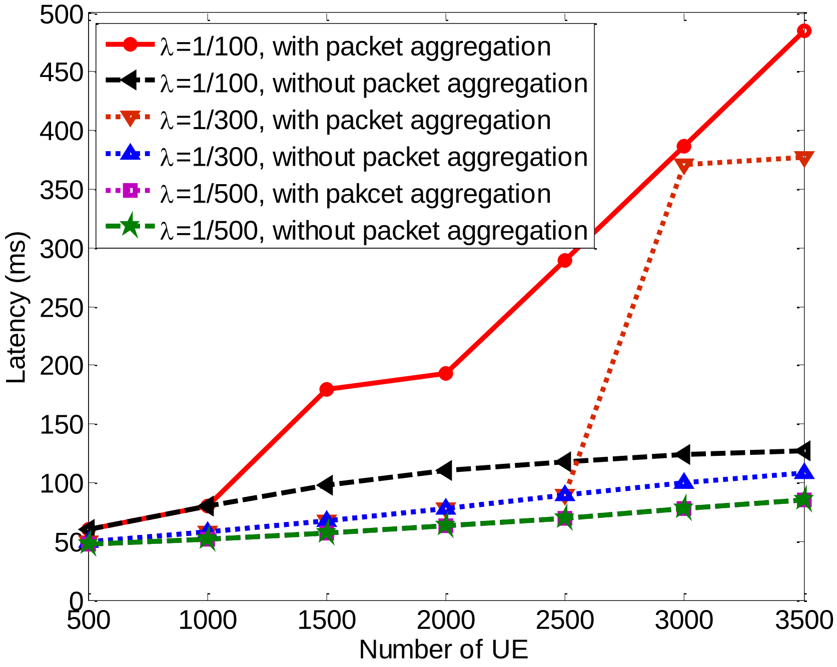


Figure 3.10: Latency

Packet Aggregation for Energy Saving with Single Type of Traffic

In this part, we show the results of packet aggregation for energy saving. Here we set the delay limit to 300 ms, which is related to one of the delay requirements for non- Guaranteed Bit Rate (GBR) bears in LTE [9][5]. Since the delay requirement is 300ms, packet aggregation is not feasible for the traffic where λ=1/300 (packet/ms) or 1/500. Therefore, in this subsection, we only consider the case where λ=1/100.

Fig.3.11 shows the packet aggregation number for different number of UE. We can see that the packet aggregation number is always 3 when number of UE is no larger than 3000, and it decreases to 2 when the number of UE is 3500. Therefore, the power saving factor is 0.67 when the number of UE is no larger than 3000, and it is 0.5 when the number of UE is 3500.

Fig.3.12 demonstrates the delay when using the results shown in Fig.3.11. It can be found that the delay is constantly less than the limit (300 ms). We also notice the when the number of UE is 3000, the resulted latency is very close to 300 ms (it is 299 ms). Therefore, when the number of UE further increases to 3500, the packet aggregation number is decreased to 2 in order to maintain the delay constraint.

Fig.3.13 shows the packet loss rate when using the results shown in Fig.3.11. We find that the packet loss rate is less than 0.1 when number of UE is no larger than 2500. However,

[5] The packet delay limit specified for non-GBR bear includes the delay of air interface as well as of core network. Since the delay of air interface is usually much larger than that of core network, we use this value as the delay for air interface for approximation.

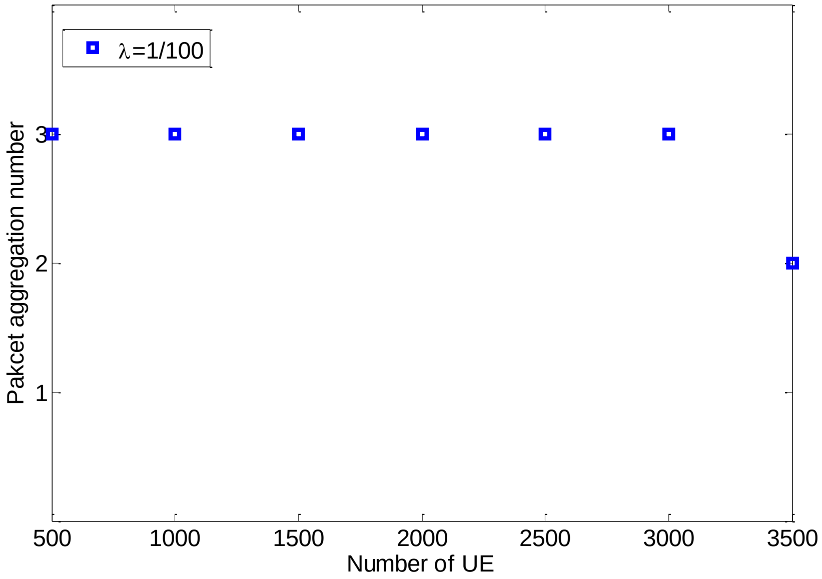


Figure 3.11: Amount of aggregated packet

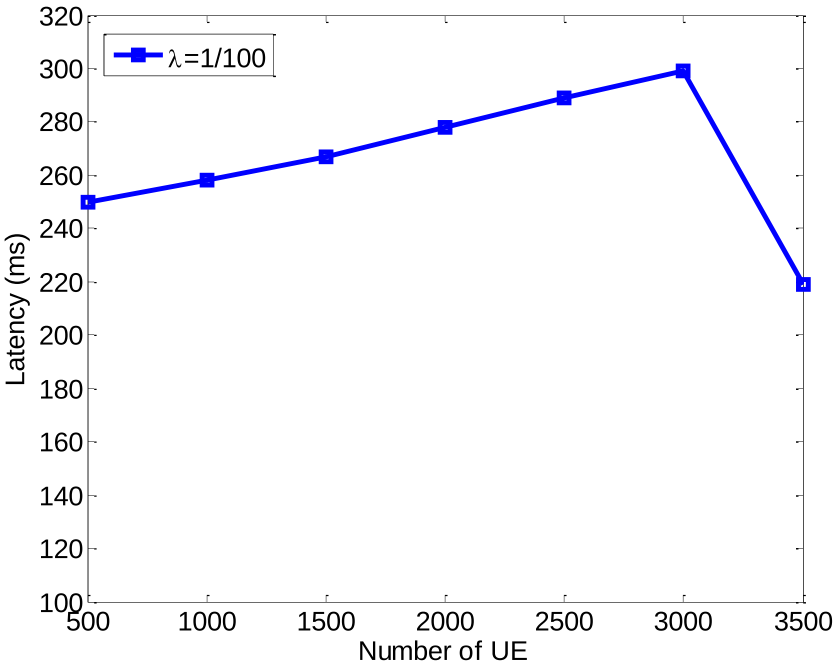


Figure 3.12: Latency

it increases to 0.14 when number of UE is 3000 and it is 0.45 when the number of UE is 3500, which violate the packet loss requirement if the packet loss requirement equals 0.1. In this case, where the delay requirement and packet loss requirement cannot be satisfied at the same time, more preamble and/or PRACH resource should be allocated. With more preamble or PRACH resource, the collision rate can be reduced. Hence, a UE do not need to aggregate large number of packets. As a result, the delay constraint can be satisfied.

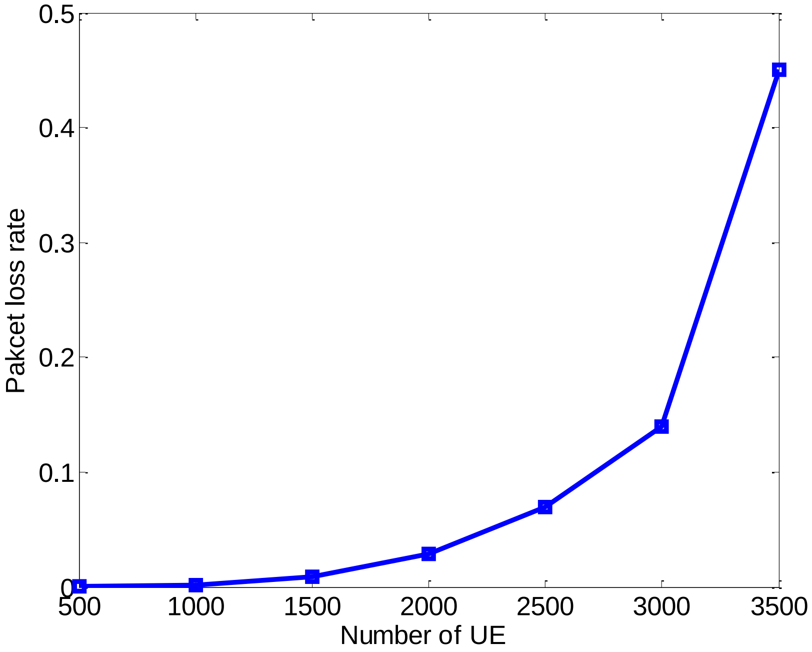


Figure 3.13: Packet loss rate

Packet Aggregation for Packet Loss Rate Reduce with Multiple Types of Traffic

Here we apply the packet arrival rate of 1/100 for real time and 1/300 for non-real time traffic to highlight the packet aggregation trade-off between realtime and non-realtime traffics for smaller number of terminal (note that for a smaller packet arrival rate, e.g. 1/500, the trade-off occurs for larger number of terminals). The number of UE with non-real time traffic is set to 1000, and its delay requirement is 5000ms. The packet loss rate for real time and non-real time traffic is 0.1 and the packet aggregation number limit for real time and non-real time traffic is 20 and 50, respectively.

Fig. 3.14 presents the number of aggregated packets for UEs with real time and non-real time traffic. We can see that the packet aggregation number for UEs with non-real time traffic increases when the number of UE with real time traffic changes from 500 to 1500. The reason is that the packet loss rate increases drastically with number of UE, which requires that more packets should be aggregated to reduce packet loss rate. Moreover, we also find that the packet aggregation number for the real time traffic remains the value of 1 regardless of number of UE varies from 500 to 1500. This is also reasonable as our objective function is to minimize the latency for the UE with real time traffic. To meet this object, the packet aggregation number for UEs with real traffic should be kept as small as possible.

Fig. 3.15 shows the packet loss rate when using the packet aggregation results presented in Fig. 3.14. For comparison, we also plot the packet loss rate without packet aggregation. We find that without packet aggregation, the packet loss rate increases with number of UE with real time traffic. For example, the packet loss rate increases from 0.05 to 0.29 when number of UE with real time traffic increases from 500 to 1500, which violates the loss rate constraint. To

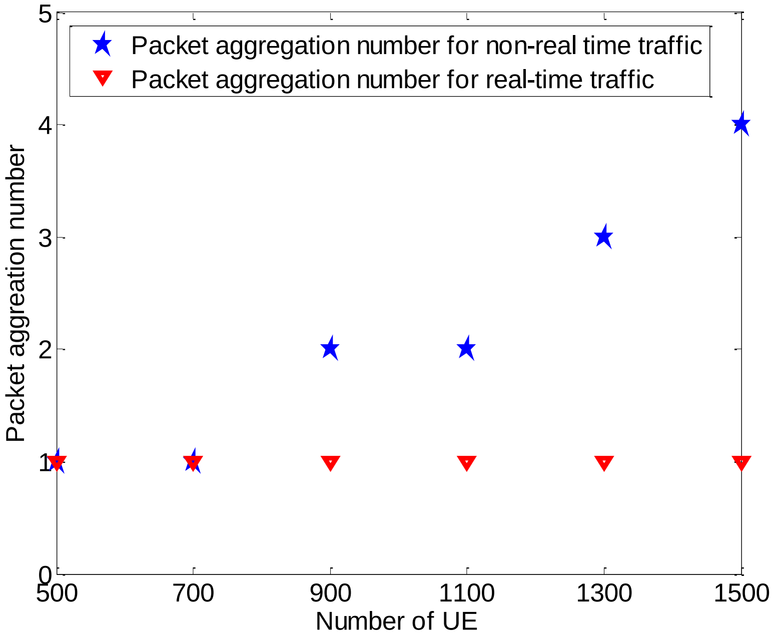


Figure 3.14: packet aggregation number for UEs with real time or non-real time traffic

comply with packet loss requirement, as shown in Fig. 3.14, the packet aggregation number is set to 4 for the UE with non-real time, which reduces the packet loss rate to 0.09.

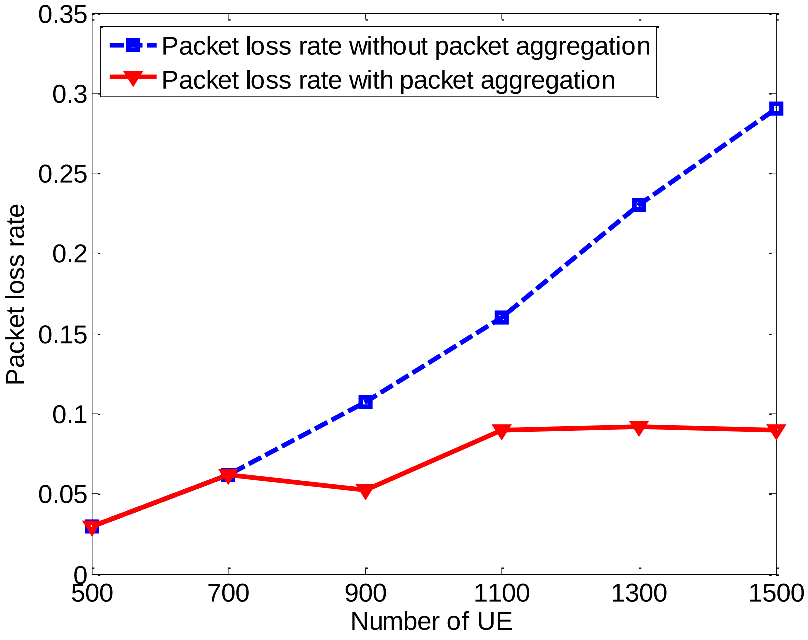


Figure 3.15: packet loss rate comparison with and without packet aggregation

Fig. 3.16 shows the latency when using the packet aggregation results presented in Fig. 3.14. The latency for the UE with non-real time traffic increases greatly from 79 ms to 991 ms when the number of UE increases from 500 to 1500. This phenomenon is reasonable as the UE with non-real time traffic has to aggregated 4 packets when the number of UE increases is 1500, which hence increases the latency. In contrast, the latency for UE with real time traffic remains almost constant (less than 100ms) as the number of UE increases, which is desirable for real time applications. The reason for this phenomenon is the packet aggregation number for UE with real time traffic is always 1.

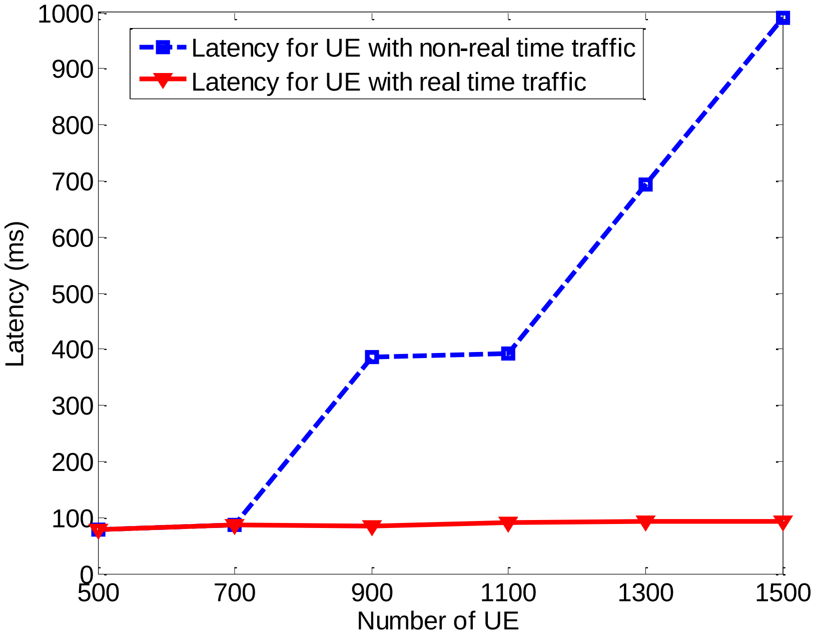


Figure 3.16: Latency

3.3.4 Discussion

In this section, we propose a packet aggregation method. With this method, we can reduce packet loss rate or save power consumption. However, there are some limitations for this method. Firstly, this method requires larger UE memory to store multiple packets, which may increase the cost of UE. Secondly, the packet aggregation method increases the packet size. As a result, the packet loss rate for the aggregated packet is also increased, which may in turn increase the latency. Thirdly, we find that the channel access latency is very large when there are large number of UE. For example, when the packet arrival rate is 1/100 (packet/ms), the number of UE is 1500, and packet aggregation number is 2, the resulted latency is 180 ms (Fig. 3.10). In contrast, if Semi-persistent scheduling is applied and the resource period is set to 100ms, the average latency would be 50 ms for the best case, which is much smaller than random access. However, there are also some problems with Semi-persistent scheduling: (1) there is signaling overhead to configure the parameters for Semi-persistent scheduling; (2) the MCS used for uplink transmission is fixed during the uplink scheduling interval ranging from 10ms to 640ms, which is not efficient for the time varying channel; (3) the resource allocation is also fixed during the uplink scheduling interval, which is neither efficient for transmission of packets with different size nor robust for time varying channel. Therefore, some investigations are needed to evaluate the performance of Semi-persistent scheduling for MTC uplink channel access. Maybe we can find some break points to determine the usage of Semi-persistent scheduling or random access: in some cases, random access with packet aggregation should be used; while in other cases, Semi-persistent scheduling should be applied.

The packet aggregation can also be applied when the packet size becomes so small, e.g. less than 20 bytes, where the signaling overhead to schedule the packet becomes dominant with respect to the payload. Moreover, the minimum resource block in LTE contains 12 subcarriers with a duration of 1 ms (14 OFDM symbols), which might be too large for transmission of small size packet. By the use of packet aggregation, packet size becomes large, which alleviates the overhead.

3.4 TTI Bundling for Machine Type Communications with Random Access

3.4.1 General Idea for TTI Bundling

In the last section, we introduce the packet aggregation method to improve the performance of random access for machine type communication. In this section, we propose a TTI bundling scheme. As introduced in the last section, an unsuccessful random access can be caused by two cases: (1) erroneous transmission of the preamble in the first stage. In this case a UE cannot receive the RAR message from eNB and it will re-start a random access when the RAR window ends, which introduces latency. (2) Unsuccessful transmission for the L2/L3 message in the third stage, which is mainly caused by preamble collision. In this case, a UE re-triggers a random access when the contention resolution timer expires, which also introduces latency. Moreover, to alleviate collision in random access, a UE usually has to backoff certain amount of time before sending a preamble, which increases latency. Therefore, we can see that *a UE has to wait for certain amount of time before starting a new random access if the previous random access fails*, which introduces large channel access latency (might be unacceptable for some real-time MTC application) and also increases power consumption as a UE has to spend more time in the active state.

In this section, we also consider that a UE uses random access to send SR to apply for resource from eNB. To reduce the channel access latency in random access, we propose a TTI bundling scheme as shown in Fig.3.17. With the proposed scheme, a UE sends several randomly selected preambles in consecutive subframes to perform multiple random attempts, which is referred to as TTI bundling for random access. Then, for every correctly received preamble, the eNB sends a RAR message[6]. Thirdly, a UE sends the L2/L3 message with the allocated resource (if a UE is allocated with multiple amount of resource, it just use one

[6] It is possible that multiple preambles sent in several consecutive subframes by one UE are correctly received by eNB. We notice that these preambles have the same timing advance value. Therefore, to save resource, only one RAR is sent by eNB. However, this method also has a problem: if two UEs, which have the same distance with eNB, successfully send two preambles in two consecutive subframes, only one RAR is sent by eNB. The

of them). Finally, the eNB acknowledges the correctly received L2/L3 message through the contention resolution message. Here we consider the multiple random access in consecutive subframes as a random access round. It is obvious that if one of these preambles in a random access round is correctly received by eNB and without collision, the random access round is successful, which *eliminates the time that a UE has to wait for to start a new random access after an unsuccessful random access*.

It seems that increasing the number of bundling TTIs yields higher successful probability for a random access round and thus reduces latency. However, this is not always true. Actually, the preamble collision rate increases with the number of bundling TTIs since each UE has to trigger more transmissions, which may in turn reduce the successful probability for a random access round. Therefore, the number of bundling TTI should be carefully selected such that the successful probability for a random accesses round can be maximized and hence the access latency is minimized.

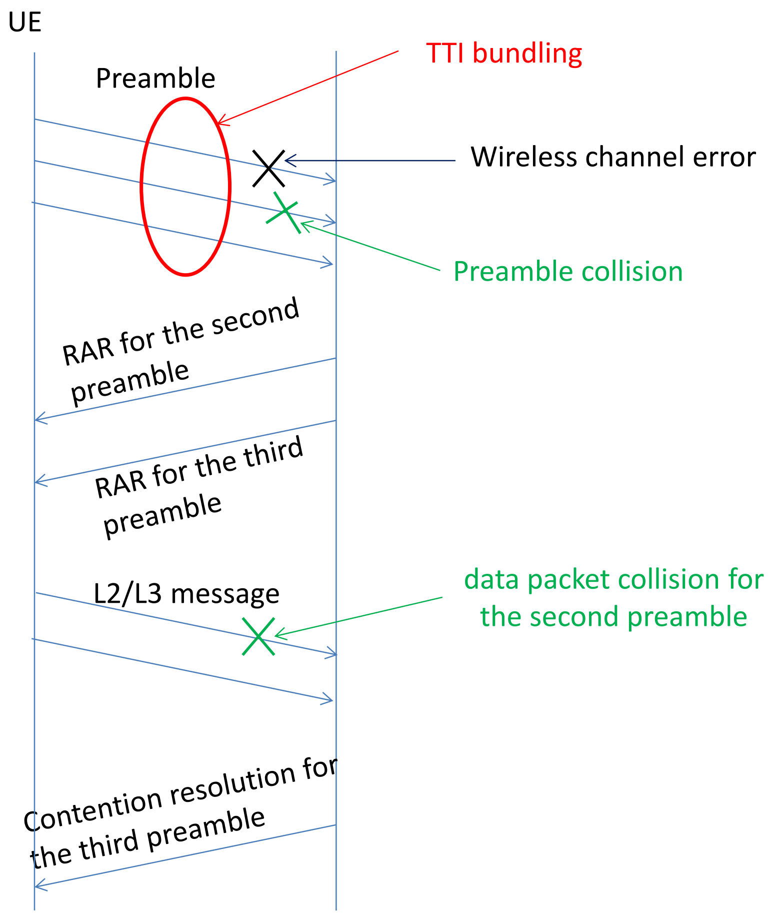

UE
Preamble
TTI bundling
Wireless channel error
Preamble collision
RAR for the second preamble
RAR for the third preamble
L2/L3 message
data packet collision for the second preamble
Contention resolution for the third preamble

Figure 3.17: contention based random access with TTI bundling

probability that problem happens is not very high. Therefore, for the ease of analysis, this problem is not considered here.

It can be seen that with TTI bundling a UE has to send several preambles to reduce the random access latency, which may increase the power consumption as sending more preambles spends more power. Therefore, for a UE with small battery capacity, the TTI bundling number should be adjusted, i.e., the TTI bundling number is not necessarily the optimal one which minimizes the latency, to save some power.

3.4.2 Optimal TTI Bundling for Random Access

To find the optimal TTI bundling number which minimizes the access latency, a Semi-Markov process model is proposed. Based on this model, we derive the channel access latency as a function of the number of bundling TTIs. Therefore, the optimal value which minimizes the channel access latency can easily be selected.

To use the Semi-Markov process model, we have to make some assumptions:

1. Similar to the assumption used in Bianchi's paper [69], here we also assume that each packet collides with constant and independent probability. The assumption is feasible when the backoff window and number of UE are large.

2. Regardless of the packet size, all the packets in a UE's buffer can be sent by one uplink transmission. This assumption handle the following problem: during the random access, it is possible that new packets are generated. With this assumption, these new packets are delivered with precedent packet. Therefore, when a UE re-starts at the initial state, there is no packet in its buffer. Moreover, due to the memoryless characteristic, the probability that a packet arrives in one subframe is not changed.

3. The packet arrival is Poisson distributed.

4. The random access channel is available in every subframe, which is related to random access resource configuration index 14 [13]. This assumption is reasonable for the scenario where there are massive MTC devices, for example: thousands of sensors for MTC applications in smart city [65].

5. The probability τ that a station will attempt transmission in one subframe is constant across all subframes [69].

6. Failures caused by wireless channel error between different random accesses in one random access round are independent.

Fig.3.18 shows the proposed Semi-Markov process model for random access with TTI bundling, where there are three types of state: idle, backoff, and random access. Specifically,

- idle state $S_{0,E}$ means there is no packet in the UE's buffer.

- backoff state $S_{i,j}, i \in [1, M], j \in [0, W_i - 1]$, means that the UE in the ith backoff stage and the backoff counter is j, where M is the transmission limit and $W_i - 1$ is the maximum backoff counter size for ith backoff stage.

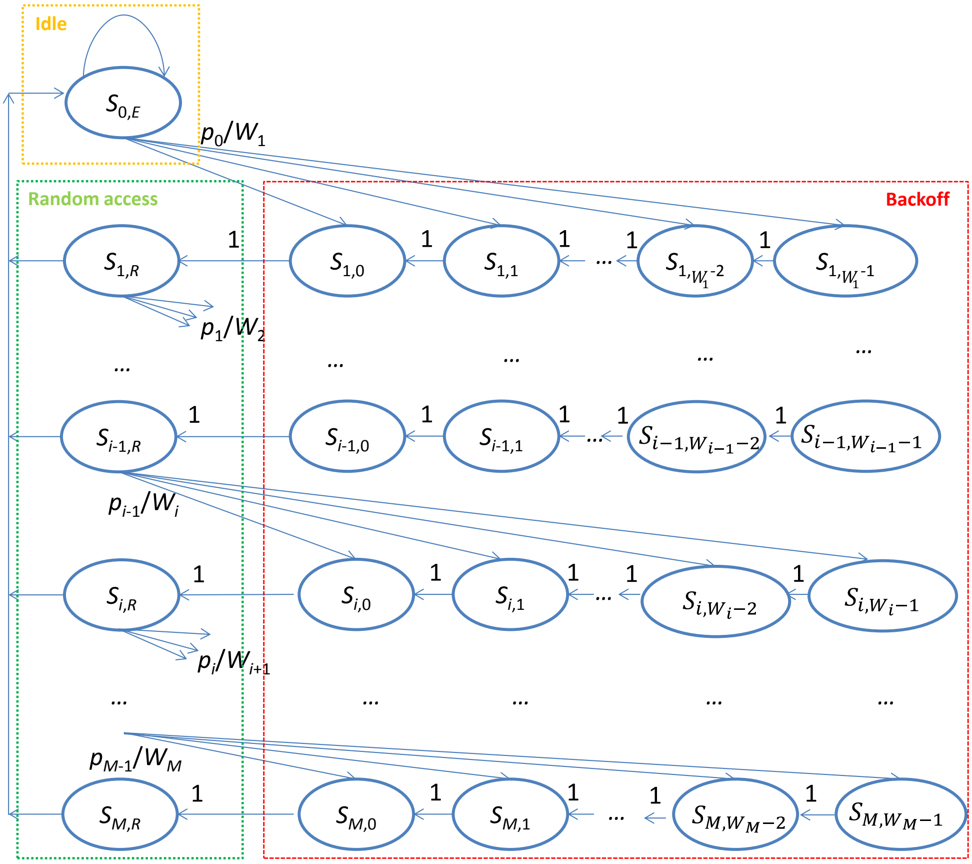


Figure 3.18: Semi-Markov process model for random access with TTI bundling

- transmission state $S_{i,R}$, $i \in [1, M]$, means that a UE is performing multiple random access attempts, i.e., sending preambles or L2/L3 messages, and waiting for the response from eNB, such as RAR or contention resolution message.

A UE transfers between states as following:

- When a UE is at state $S_{0,E}$, if a packet arrives in one subframe then UE selects a random backoff counter j over $[0, W_1 - 1]$ and transfers to state $S_{1,j}$ to start the first backoff. Otherwise it remains at state $S_{0,E}$.
- When a UE is at state $S_{i,j}, i \in [1, M], j \in [0, W_i - 1]$, it transfers to $S_{i,j-1}$ after 1ms.
- When a UE is at state $S_{i,R}$, $i \in [1, M - 1]$, it transfers to state $S_{0,E}$ if a contention resolution indicating a successful random access is received. In contrast, it transfers to state $S_{i+1,j}$ (j is randomly selected over $[0, W_{i+1} - 1]$) to start another random access round if the $n, n \in [1, N]$, random accesses in one random access round all fail, where n is the number of preambles that are sent for one random access round and N is the limit of the number of preambles that are sent for one random access round.
- When a UE is at state $S_{M,R}$, whether the random accesses in this round are successful or not, it transfers to state $S_{0,E}$ when the random access round ends.

It has to be noticed that a packet may be generated during random access. In this case, this packet generation does not trigger the state transition. Instead, a UE remains at the current state and this new generated packet will be sent along with the existing packet when the random access is successful.

Denoting the probability that a packet arrives during one subframe (1ms) is p_0, the state transition probability from $S_{0,E}$ to $S_{1,j}$, $j \in [0, W_1 - 1]$, is p_0/W_1. Similarly, we denote p_i, $i \in [1, M-1]$, as the unsuccessful probability for the ith random access round, therefore the state transition probability from $S_{i,R}$, $i \in [1, M-1]$ to $S_{i+1,j}, j \in [0, W_{i+1} - 1]$ is p_i/W_{i+1}.

Denoting $\pi_{i,j}$ as the stationary probability for state $S_{i,j}$, it can be calculated as:

$$\begin{cases} \pi_{1,W_1-1} = \pi_{0,E}\dfrac{p_0}{W_1} \\ \pi_{1,j} = \pi_{0,E}\dfrac{p_0}{W_1} + \pi_{1,j+1}, j \in [0, W_1 - 2]. \\ \pi_{i,W_i-1} = \pi_{i-1,R}\dfrac{p_{i-1}}{W_i}, i \in [2, M] \\ \pi_{i,j} = \pi_{i-1,R}\dfrac{p_{i-1}}{W_i} + \pi_{i,j+1}, i \in [2, M], j \in [0, W_i - 2]. \end{cases} \tag{3.33}$$

With the first and second equations in equation system (3.33), we have:

$$\pi_{1,j} = (W_1 - j)\frac{p_0}{W_1}\pi_{0,E}, j \in [0, W_1 - 1]. \tag{3.34}$$

$$\pi_{1,R} = p_0\pi_{0,E}. \tag{3.35}$$

By the use of the third and fourth equations in equation system (3.33), we get

$$\pi_{i,j} = (W_i - j)\frac{p_{i-1}}{W_i}\pi_{i-1,R}, i \in [2, M], j \in [0, W_i - 1]. \tag{3.36}$$

$$\pi_{i,R} = p_{i-1}\pi_{i-1,R}, i \in [2, M]. \tag{3.37}$$

The sum of the all state's stationary probabilities is 1, which yields:

$$\begin{aligned} 1 &= \pi_{0,E} + \sum_{i=1}^{M}\pi_{i,R} + \sum_{i=1}^{M}\sum_{j=0}^{W_i-1}\pi_{i,j} \\ &= \pi_{0,E} + \sum_{i=1}^{M}\prod_{j=0}^{i-1}p_j\pi_{0,E} + \sum_{i=1}^{M}\sum_{j=0}^{W_i-1}\frac{W_i - j}{W_i}\pi_{i,R} \\ &= \pi_{0,E} + \sum_{i=1}^{M}\prod_{j=0}^{i-1}p_j\pi_{0,E} + \sum_{i=1}^{M}\frac{W_i+1}{2}\prod_{j=0}^{i-1}p_j\pi_{0,E}. \end{aligned} \tag{3.38}$$

Therefore, we have

$$\pi_{0,E} = \frac{1}{1 + \sum_{i=1}^{M}\prod_{j=0}^{i-1}p_j + \sum_{i=1}^{M}\frac{W_i+1}{2}\prod_{j=0}^{i-1}p_j}. \tag{3.39}$$

Table 3.4: Symbols used in Section 3.4

Symbol	Definition
d	average channel access latency
d'_i	latency caused by the ith unsuccessful random access round
M	transmission limit for random access
n	number of preambles that are sent for one random access round
N	limit of number of preambles that are sent for one random access round
N_{HARQ}	maximum number of HARQ transmissions
N_p	number of available preambles for random access
N_u	total amount of UE
p_0	probability that a packet arrives during one subframe
p_c	preamble collision probability
$p_{E,i}$	error probability caused by wireless channel for a preamble in the ith random access round
p_{ES}	error rate to send the SR information
$p_{F,i}$	unsuccessful probability for one random access in the ith random access round
p_i	unsuccessful probability for the ith random access round
$p^N_{i,k}$	probability if the state holding time is determined by the kth preamble in the ith random access round
$P_{NRAR,i}$	probability that no RAR is received in one random access of the ith random access round
$P_{RAR,i}$	probability that the collision happens for a random access in the ith random access round and UE receives the RAR message
$r_{i,j+1}$	detection rate for the preamble in the ith random access round when $j+1$ UEs (one UE plus j contending UEs) send the same preamble
$S_{j,i}$	state in Semi-Markov process
t_n	time instant when a UE sends the last preamble
t_j	time instant when a UE sends the jth preamble
T^R	state holding time if none of the random access in the ith random access round is successful, and the UE receives the RAR message for the last transmitted preamble
T_{CR}	duration which starts at time instant when a UE sends SR and ends at the time instant when a UE decodes the contention resolution message
T_D	time used to decode a RAR message
$T^S_{i,j}$	state holding time if the jth random access in the ith random access round is successful
$T^N_{i,k}$	state holding time if the state holding time is determined by the kth preamble in the ith random access round
$T_{i,R}$	average holding time for state $S_{i,R}$
$T_{i,TX}$	average duration used to send preamble in the ith random access round
T'_i	latency caused by the precedent unsuccessful random access if a random access is successful at the ith random access round
T_{RAR}	duration that starts at the end of a preamble's transmission and ends at the time instant when the RAR message can be received
T_W	duration which starts at the time instant when a UE sends a preamble and ends at the last subframe of RAR window

Table 3.5: Symbols used in Section 3.4-continued

Symbol	Definition
T_{timer}	duration for contention resolution timer
W_i	backoff window size for ith random access
α	threshold used to measure the latency increase when using a smaller TTI bundling number
λ	packet arrival rate
$\pi_{j,i}$	stationary probability for state $S_{j,i}$
τ	probability that a UE sends a preamble in one subframe

Now let us calculate the state transition probability. Assuming the packet arrives following Poisson distribution with arrival rate λ, the probability that a packet arrives in one subframe is

$$p_0 = 1 - e^{-\lambda}. \tag{3.40}$$

One random access round is unsuccessful if all the random access in that round is unsuccessful, therefore

$$p_i = p_{F,i}^n \tag{3.41}$$

where $p_{F,i}$ is the unsuccessful probability for one random access in the ith random access round.

As mentioned in Sect. 3.2, an unsuccessful random access is caused by erroneous transmission for the preamble or the unsuccessful delivery of the L2/L3 message (the L2/L3 message considered here is the SR information). More specifically, the unsuccessful delivery of the SR information is also caused by two sub-cases[7]: (1) collision of the preamble. In this case multiple UEs send the SR information on the same resource which leads to the failure for the SR information delivery. (2) the SR information message is corrupted due to wireless channel error. In this case, the preamble is correctly received by eNB and free of collision. However, the SR information cannot be successfully decoded by eNB due to the wireless channel error. With the above analysis, we have

$$p_{F,i} = p_{E,i} + (1 - p_{E,i})p_c + (1 - p_c)(1 - p_{E,i})p_{ES}^{N_{HARQ}} \tag{3.42}$$

In the above equation p_c is the collision rate for a preamble; $p_{E,i}$ is the error probability caused by wireless channel for a preamble in the ith random access round; p_{ES} is the error rate to send the SR information and N_{HARQ} is the maximum number of HARQ transmissions. Since the SR information is of very small size, therefore its error rate p_{ES} is very small (less than 0.1) and hence $p_{ES}^{N_{HARQ}} \approx 0$ considering that N_{HARQ} is usually larger than 2. With this result, we have

$$p_{F,i} \approx p_c + p_{E,i} - p_c p_{E,i}, \tag{3.43}$$

i.e., an unsuccessful random access is only caused by preamble collision which causes the unsuccessful delivery of the L2/L3 message or the erroneous transmission for the preamble.

[7] The case we consider here is more realistic than that in the last section.

From the perspective of one UE, collision happens when there are other UEs selecting the same preamble, therefore

$$p_c = \sum_{i=1}^{N_u-1} \binom{N_u-1}{i} \tau^i (1-\tau)^{N_u-1-i} (1-(1-\frac{1}{N_p})^i). \tag{3.44}$$

In the above equation N_u is the total amount of UE; τ is the probability that a UE sends a preamble in one subframe; N_p is the number of available preambles for random access.

Now let us calculate the state holding time for this Semi-Markov process model. It is obvious that the state holding time for $S_{0,E}$ and $S_{i,j}, i \in [1, M], j \in [0, W-1]$ is 1ms.

For the UE at $S_{i,R}, i \in [1, M]$, the calculation for state holding time is less obvious. We denote the duration that starts at the end of a preamble's transmission and ends at the time instant when receiving the RAR message for that preamble as T_{RAR} and the time used to decode the RAR message as T_D. Therefore, the SR message is sent $T_{RAR} + T_D$ ms after the preamble's transmission if the RAR message is received (no wireless error for the transmitted preamble). As stated above, when the UE is at $S_{i,R}, i \in [1, M]$, the state transition happens when one random access is successful or all the random access in one random access round fail. Hence, we calculate the state holding time by three cases [8]:

1. The jth, $j \in [1, n]$, random access in the ith, $i \in [1, M]$, random access round is successful.

   The probability for the first case $p^S_{i,j}$ is

   $$\prod_{k=1}^{i-1} p_k p_{F,i}^{j-1}(1-p_{F,i}), i > 1 \tag{3.45}$$

   or

   $$p_{F,1}^{j-1}(1-p_{F,1}), i = 1. \tag{3.46}$$

   When a random access succeeds, the UE transfers to the initial state $S_{0,E}$ after decoding the contention resolution message. Denoting T_{CR} as the average duration which starts at time instant when a UE sends the SR message and ends at the time instant when a UE decodes the contention resolution message, the state holding time for the first case is

   $$T^S_{i,j} = j + T_{RAR} + T_D + T_{CR} \tag{3.47}$$

2. None of the random access in the ith, $i \in [1, M]$, random access round is successful, and the UE receives the RAR message from eNB for the last transmitted preamble.

   In this case, a UE sends the L2/L3 message. However, as the random access is unsuccessful, it cannot receive the contention resolution message. This UE will transfer to the initial state $S_{0,E}$ when the contention resolution timer expires.

[8] For the ease of analysis, here consider three cases to calculate the state holding time, which is different from Section 3.3.1

Therefore, the state holding time for the second case is

$$T^R = n + T_{RAR} + T_D + T_{timer} \tag{3.48}$$

where T_{timer} is the duration for contention resolution timer. The probability for this second case p_i^R is

$$\prod_{k=1}^{i-1} p_k p_{F,i}^{n-1} P_{RAR,i}, i > 1 \tag{3.49}$$

or

$$p_{F,1}^{n-1} P_{RAR,1}, i = 1 \tag{3.50}$$

where $P_{RAR,i}, i \in [1, M]$, is the probability that the collision happens for a random access in the ith random access round and UE receives the RAR message. The $P_{RAR,i}$ is calculated by

$$P_{RAR,i} = \sum_{n=1}^{N_u-1} \binom{N_u - 1}{n} \tau^n (1-\tau)^{N_u-1-n} \sum_{j=1}^{n} \binom{n}{j} (\frac{1}{N_p})^j (1 - \frac{1}{N_p})^{n-j} r_{i,j+1}. \tag{3.51}$$

In the above equation $r_{i,j+1}$ is the detection rate for the preamble in the ith random access round when $j+1$ UEs (one UE plus j contending UEs) send the same preamble. If $r_{i,j+1} \approx 1$ for $j \geq 1$, i.e., a preamble can mostly be detected when it is sent by multiple UEs, then $P_{RAR,i} \approx p_c$.

3. None of the random access in the ith, $i \in [1, M]$, random access round is successful, and the UE does not receive RAR for the last random access.

   In this case, after sending the last preamble (the nth preamble) the minimum time that a UE will stay at state $S_{i,R}, i \in [1, M]$, is T_W, where T_W is the duration which starts at the time instant when a UE sends a preamble and ends at the last subframe of the RAR window. However, if the UE receives its RAR for a precedent random access (not the last one) in this round, it cannot transfer to the initial state $S_{0,E}$ before the contention resolution timer ends, which may affect the state holding time. Denoting the time instant when a UE sends the jth, $j \in [1, n]$, preamble as t_j and assuming the RAR is received by UE for this preamble, the time instant when the contention resolution timer ends is $t_j + T_{RAR} + T_D + T_{timer}$. Therefore, for a preamble transmission which can affect the UE's state holding time at state $S_{i,R}, i \in [1, M]$, its transmission time instant should satisfy the following the condition

   $$t_j + T_{RAR} + T_D + T_{timer} > t_n + T_W \tag{3.52}$$

   where t_n is the time instant when a UE sends the last preamble.

   Since $t_n = t_j + (n - j)$, the above calculation is rewritten as

   $$j + T_{RAR} + T_D + T_{timer} > n + T_W \tag{3.53}$$

   The minimum index satisfying equation (3.53) is denoted as z.

   Therefore, the state holding time is determined by the kth, $k \in [z, n-1]$, preamble if

(a) the RAR is received for this preamble,

(b) the random accesses are unsuccessful for the preambles sent before it,

(c) no RAR messages are received for the preambles sent after it.

Accordingly, the state holding time for the third case is

$$T_{i,k}^{N} = k + T_{RAR} + T_{D} + T_{timer} \tag{3.54}$$

and its probability $p_{i,k}^{N}$ is

$$\prod_{j=1}^{i-1} p_j p_{F,i}^{k-1} P_{RAR,i} P_{NRAR,i}^{n-k}, i > 1 \tag{3.55}$$

or

$$p_{F,1}^{k-1} P_{RAR,1} P_{NRAR,1}^{n-k}, i = 1 \tag{3.56}$$

where $P_{NRAR,i}, i \in [1, M]$, is the probability that no RAR is received in one random access of the ith random access round. The RAR is not sent to a UE if the transmitted preamble sent by one UE (or multiple UEs) is not correctly detected by eNB, therefore we have

$$P_{NRAR,i} = \sum_{n=0}^{N_u-1} \binom{N_u-1}{n} \tau^n (1-\tau)^{N_u-1-n} \sum_{j=0}^{n} \binom{n}{j} (\frac{1}{N_p})^j (1-\frac{1}{N_p})^{n-j} (1 - r_{i,j+1}). \tag{3.57}$$

It is also possible that no RAR is received for all the random accesses whose index are larger than z. Then the state holding time is $T_{i,n}^{N} = n + T_W$ and its probability $p_{i,n}^{N}$ is

$$\prod_{j=1}^{i-1} p_j p_{F,i}^{z-1} P_{NRAR,i}^{n-z+1}, i > 1 \tag{3.58}$$

or

$$p_{F,i}^{z-1} P_{NRAR,i}^{n-z+1}, i = 1 \tag{3.59}$$

If $r_{i,j+1} \approx 1$ for $j \geq 1$, i.e., a preamble can mostly be detected when it is sent by multiple UEs, we have $P_{NRAR,i} \approx (1 - p_c) p_{E,i}$.

With the above results, the average holding time $T_{i,R}, i \in [1, M]$, for state $S_{i,R}$ is

$$T_{i,R} = \frac{\sum_{j=1}^{n} p_{i,j}^{S} T_{i,j}^{S} + p_i^R T^R + \sum_{j=z}^{n} p_{i,j}^{N} T_{i,j}^{N}}{\sum_{j=1}^{n} p_{i,j}^{S} + p_i^R + \sum_{j=z}^{n} p_{i,j}^{N}}. \tag{3.60}$$

When a UE is at state $S_{i,R}, i \in [1, M]$, the average duration which is used for sending preambles is

$$T_{i,TX} = \frac{\sum_{j=1}^{n} p_{i,j}^{S} j + p_i^R n + \sum_{j=z}^{n} p_{i,j}^{N} n}{\sum_{j=1}^{n} p_{i,j}^{S} + p_i^R + \sum_{j=z}^{n} p_{i,j}^{N}}. \tag{3.61}$$

Therefore, the proportion of time that a UE is sending a preamble, i.e., the probability that a UE sends a preamble in one subframe, is

$$\tau = \sum_{i=1}^{M} \frac{\pi_{i,R} T_{i,TX}}{T} \tag{3.62}$$

where

$$T = \pi_{0,E} + \sum_{i=1}^{M} \sum_{j=1}^{W-1} \pi_{i,j} + \sum_{i=1}^{M} \pi_{i,R} T_{i,R}. \tag{3.63}$$

is the average holding time for all the states.

It can be seen that equations (3.62) and (3.44) comprise a equation system with two unknowns p_c and τ, which can be solved by the use of numerical method.

Provided that the $i \in [1, M]$ random access round is unsuccessful, then the duration that UE stays at state $S_{i,R}$ is the latency introduced by this random access round. Denoting the latency in this case as d_i', it is calculated by

$$d_i' = \frac{p_i^R T^R + \sum_{j=z}^{n} p_{i,j}^N T_{i,j}^N}{p_i^R + \sum_{j=z}^{n} p_{i,j}^N}. \tag{3.64}$$

If a random access if successful at the first random access round, no latency is caused by precedent random access. Therefore, we have

$$T_1' = 0. \tag{3.65}$$

However, if a random access is successful at the ith random access round, the latency caused by the precedent unsuccessful random access is

$$T_i' = \sum_{j=1}^{i-1} d_j', i \in [2, M]. \tag{3.66}$$

No let us calculate the access latency for random access which is defined as the duration that starts at the time instant when a UE wants to trigger a random access and ends at the time when that UE receives a contention resolution message indicating the random access is successful. Assuming the random access is succeed in the jth transmission of the ith random access round, the access latency includes (1) the duration of $\sum_{j=1}^{i} \frac{W_j}{2}$ which is used for backoff, (2) the duration of $T_{i,j}^S$ which is the time used for the successful random access in the current round, and (3) the latency T_i' which is used for the precedent unsuccessful random access rounds.

With above analysis, the average channel access latency caused by random access is calculated by

$$d = \sum_{j=1}^{n} \frac{p_{F,1}^{(j-1)}(1-p_{F,1})}{1-\prod_{i=1}^{M} p_i}(\frac{W_1}{2} + T_{1,j}^S) + \sum_{i=2}^{M} \sum_{j=1}^{n} \frac{\prod_{k=1}^{i-1} p_k p_{F,i}^{(j-1)}(1-p_{F,i})}{1-\prod_{i=1}^{M} p_i}(\sum_{k=1}^{i} \frac{W_k}{2} + T_{i,j}^S + T_i'). \tag{3.67}$$

The optimal TTI bundling number which minimizes the access latency is

$$\begin{aligned} \arg\min_{n} \quad & d \\ \text{subject to} \quad & n < N \end{aligned} \tag{3.68}$$

where N is the limit of number of preambles that are sent for one random access round. The L2/L3 message is sent $T_{RAR} + T_D$ ms after the first preamble's transmission if the preamble is correctly received by eNB. As a UE cannot send a L2/L3 message as well as a preamble at the same time, the maximum bundling TTI number N should be no larger than $T_{RAR} + T_D$. Since we do not have a closed form of d in the term of n, therefore the above optimization problem can only be solved by exhaustive search.

As mentioned in Sect.3.2, for some power constrained MTC device, the TTI bundling number is not necessarily to be the optimal one which minimizes the latency. Instead, the TTI bundling number can be smaller than optimal one such that some power is saved. To this end, we propose a method as shown in **Algorithm** 2 to set the proper TTI bundling number for power constrained MTC device. In **Algorithm** 2, α is the threshold used to measure the latency increase when using a smaller TTI bundling number: if the latency increase is smaller the threshold, using a smaller TTI bundling number is feasible; otherwise, a smaller TTI bundling number should not be used to avoid a larger latency increase than desired.

Algorithm 2 TTI bundling number selection for power constrained MTC device

```
Input: $n$:solution for (3.68); the minimum latency $d$; $\alpha$
Output: the optimal TTI bundling number $n^*$
while $n > 1$ do
    use (3.67) to compute the latency $d_1$ when the TTI bundling number is $n - 1$;
    if $\frac{d_1-d}{d} < \alpha$ then
        $d = d_1$;
        $n = n - 1$;
    else
        Break;
    end if
end while
```

3.4.3 Results

The parameters are shown in Table 3.6.

In case of no collision, the preamble detection rate is assume to be $1 - \frac{1}{e^i}$ as that used in [3], where $i \in [1, M]$ indicates the ith preamble transmission. When a preamble are sent by multiple UEs, it can always be correctly decoded, $r_{i,j+1} \approx 1$ for $j \geq 1$. This assumption is quite typical in LTE. If a preamble is sent by two UEs, two peaks appear at the eNB side. The probability that neither peak can be decoded by eNB is relatively low.

Table 3.6: Parameters

Parameter	Description	Value
N	limit of number of preambles that are sent for one random access round	8
N_p	Number of preamble	20
M	Transmission limit	5
T_{CR}	Duration which starts at time instant when a UE sends the SR message and ends at the time instant when a UE decodes the contention resolution message	8 ms
T_D	Time used to decode a RAR message	3 ms
T_{RAR}	Duration that starts at the end of a preamble's transmission and ends at the time instant when the RAR message can be received	5 ms
T_{RARW}	Random access response window	10 ms
T_{timer}	Duration for contention resolution timer	24 ms
T_W	Duration which starts at the time instant when a UE sends a preamble and ends at the last subframe of RAR window ($T_W = T_{RARW} + T_{RAR}$)	15 ms
W_i, $i \in [1, 5]$	Backoff window size	30

Model Validation

Firstly, to validate the method which is used to calculate latency (equation 3.67), we compare the simulation results with the analytical results obtained using equation (3.67) in Fig. 3.19. The TTI bundling number n is set to 1 and 2 to demonstrate that our method is applicable to the regular random access (n=1) as well as the random access with TTI bundling (n=2). The packet arrival rate is set to be $\lambda = 1/100$ and $1/50$ packet/ms. As it is shown in Fig. 3.19, the analytical results match the simulation results when $n = 1$ or 2, which validates our method.

TTI bundling for Regular MTC Device

Fig. 3.20 demonstrates the optimal number of bundling TTI under different number of UE and packet arrival rate. We can see that the optimal TTI bundling number non-increases as the number of UE and packet arrival rate increases. The reason for this phenomenon is: the preamble collision rate increases with larger number of UE and packet arrival rate. Therefore, when the number UE or packet arrival rate becomes large, a UE should bundle smaller (or same) number of TTI to avoid the collision rate increase. Moreover, we also find that the TTI bundling number of lower packet arrival rate is larger than that of higher one. This is reasonable since the packet collision rate increases with packet arrival rate. Therefore, smaller TTI bundling should be used to avoid serve collision when packet arrival rate increases.

Fig. 3.21 compares the latency obtained using the results shown in Fig. 3.20 to the latency without using TTI bundling. We find that with TTI bundling the latency is greatly reduced when the number of UE is less than 900 for $\lambda = 1/100$ or the number of UE is less than 600

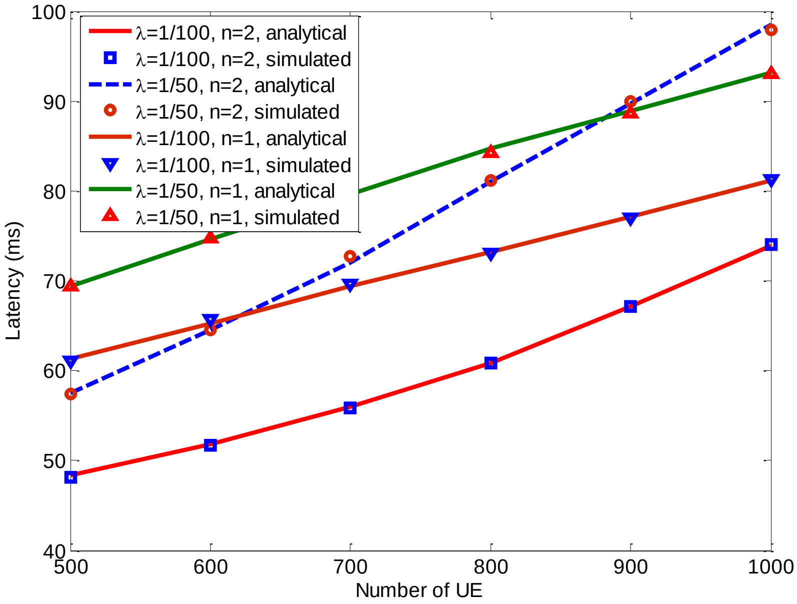


Figure 3.19: Comparison of simulation and analytical results

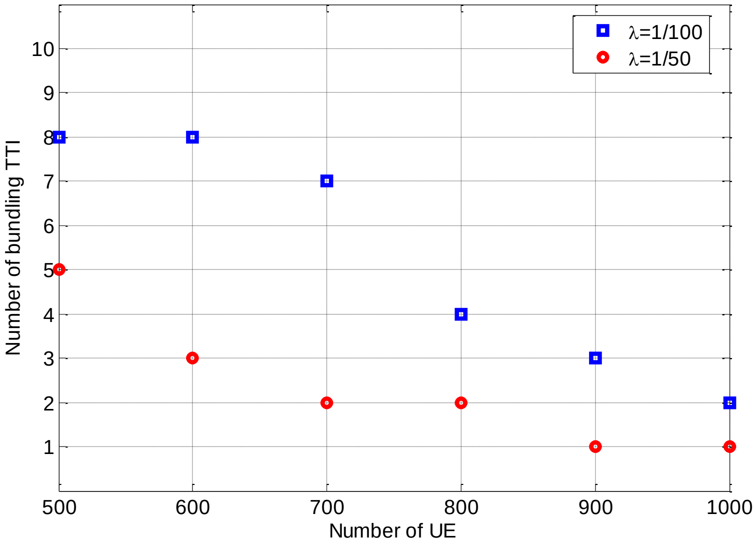


Figure 3.20: Optimal number of bundling TTI

for $\lambda = 1/50$. Concretely, The latency is reduced by 46% (61ms to 33 ms) at maximum when $\lambda = 1/100$ and the number of UE is 500. In contrast, the gain for using TTI bundling becomes smaller for other cases. The reason for this phenomena has two aspects: (1) when preamble collision rate is not very high, bundling multiple TTIs greatly increases the successful rate for random access. For example, the collision rate is 0.22, the first random access round successful rate is 0.50 and the latency is 61ms when $\lambda = 1/100$, $n = 1$, and $N_u = 500$. When the TTI bundling number n increases to 8, though the collision rate increases to 0.36, the first random access round successful approximately equals 1 which greatly reduces the latency to 33ms as demonstrated in Fig.3.21. (2) When the preamble collision rate is high, bundling multiple TTIs increases the preamble collision rate to a very high level. As a result, the random access round successful rate is not significantly improved and hence the latency is not greatly reduced. For example, the collision rate is 0.48; the first random access round successful rate is 0.33 and the latency is 85ms when $\lambda = 1/50$, $n = 1$, and $N_u = 800$. When the TTI bundling number n increases to 2, the preamble collision rate jumps to 0.65 and the successful rate for the first random access round only increases to 0.45 which slightly reduces the latency to 81ms as shown in Fig.3.21.

Therefore, to reduce the channel access latency for a network where preamble collision rate is high, we should firstly lower the preamble collision rate (it can be achieved by allocating more preambles). Secondly, we apply the TTI bundling scheme.

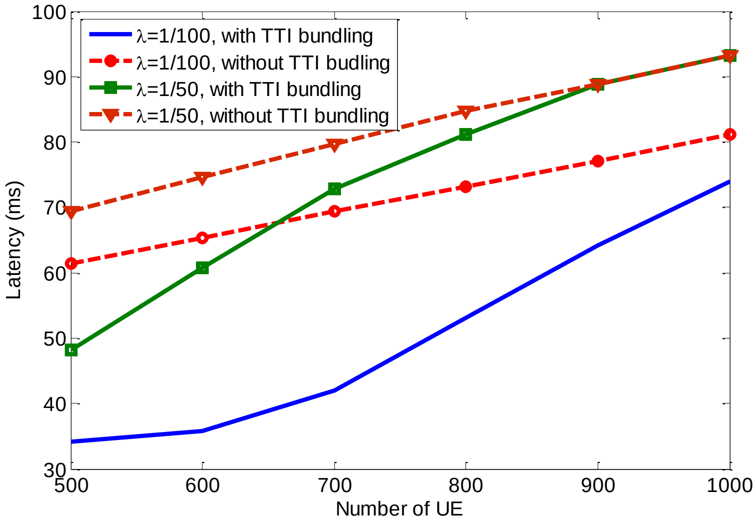


Figure 3.21: Latency comparison with and without TTI bundling

TTI Bundling for Power Constrained MTC Device

Fig.3.22 shows the channel access latency under different TTI bundling number when $\lambda = 1/100$ and $1/50$, and the number of UE $N_u = 500$. We find that the latency decreases very slowly when the TTI bundling number becomes larger than 5 for $\lambda = 1/100$. The reason for this phenomenon is that the first round random access successful rate is very high (0.93) when the TTI bundling number is 5 and $\lambda = 1/100$, and it increases slightly to 0.99 as the TTI bundling number varies from 5 to 8. Therefore, the latency only decreases from 36 ms to 34 ms. We notice that the power consumption increases with TTI bundling number as UE has to send more preambles. Therefore, for some power constrained MTC device, setting the TTI bundling number to 5 might also be a reasonable choice. Similarly, for the case of $\lambda = 1/50$, setting TTI bundling number to 4 might also be feasible for a power constrained MTC device as the latency decreases insignificantly from 49 ms to 48 ms as TTI bundling varies from 4 to 5 (the optimal TTI bundling number). To this end, a TTI bundling method for power constrained MTC device (**Algorithm** 2) is proposed.

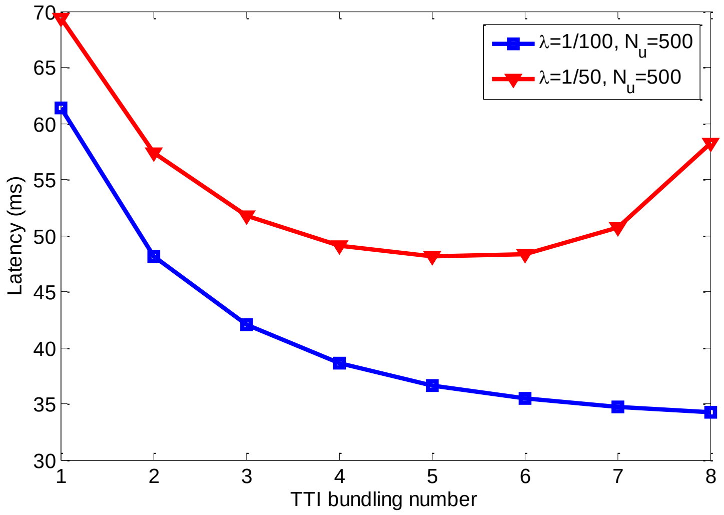


Figure 3.22: Latency under different TTI bundling number

Fig. 3.23 and Fig. 3.24 compare the TTI bundling number calculated from (3.68) to those obtained using **Algorithm** 2, where α is set to 5% and 10%, and $\lambda = 1/100$ and $1/50$, respectively. We can find that the TTI bundling number is reduced when using **Algorithm** 2 and larger α yields smaller TTI bundling number.

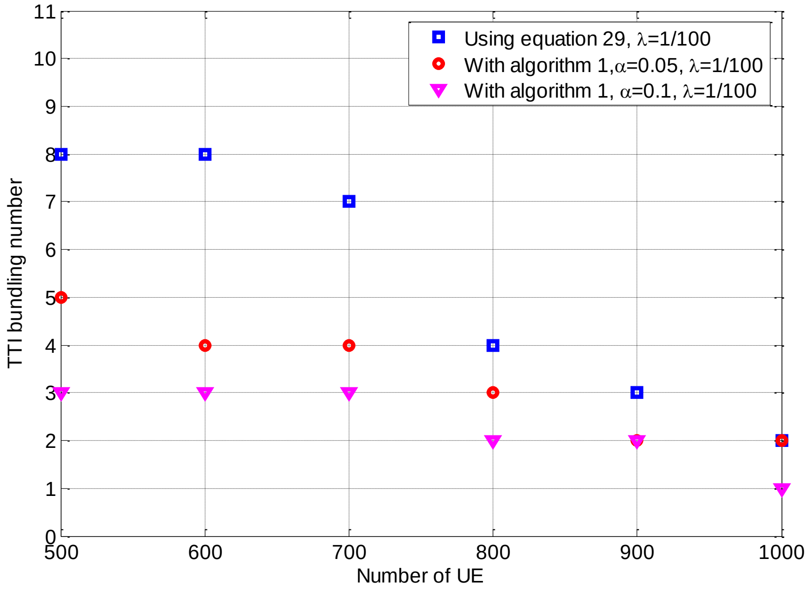


Figure 3.23: TTI bundling number comparison: regular TTI bundling to Algorithm 1

Since using less TTI bundling number saves power, in order to measure the power reduction, here we define the power saving factor β as the ratio of the saved TTI bundling number when using **Algorithm** 2 to the TTI bundling number obtained by the use of (3.68). Fig. 3.25 and Fig. 3.26 shows the power saving factor used to measure the saved TTI bundling number shown in Fig. 3.23 and Fig. 3.24, respectively. From these two figures, it is can be easily

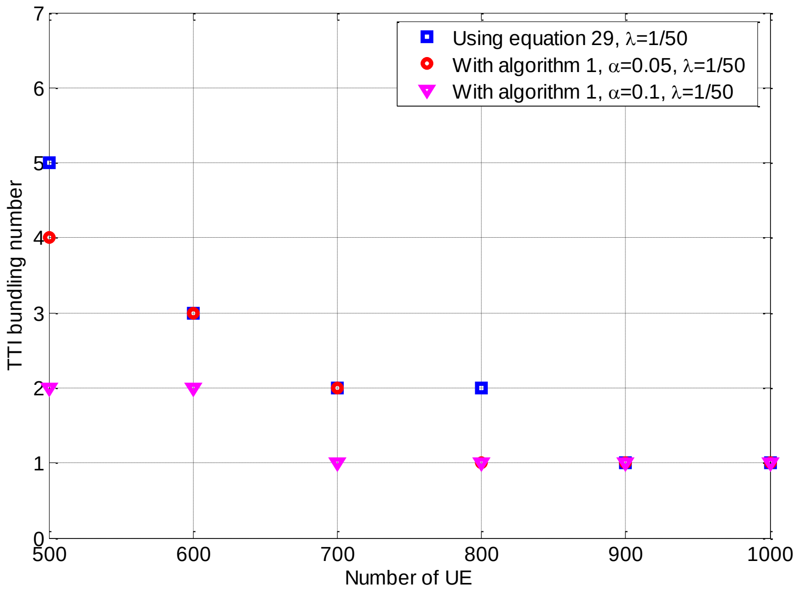


Figure 3.24: TTI bundling number comparison: regular TTI bundling to Algorithm 1

found that using larger α yields higher (or same) power saving factor as UE tends to use smaller TTI bundling number when α becomes larger.

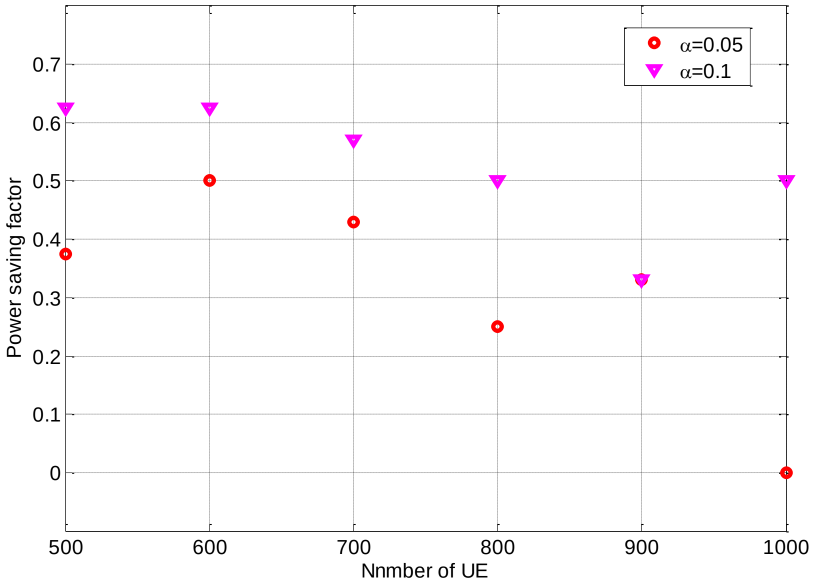


Figure 3.25: Power saving factor for different α with Algorithm 1, $\lambda = 1/100$

However, power saving is achieved at the cost of latency increase. Fig. 3.27 and Fig. 3.28 show the latency when using the TTI bundling number shown in Fig. 3.23 and Fig. 3.24, as well as the latency without TTI bundling. We can see that larger α yields higher channel access latency, which is the result of using smaller TTI bundling. Therefore, the configuration

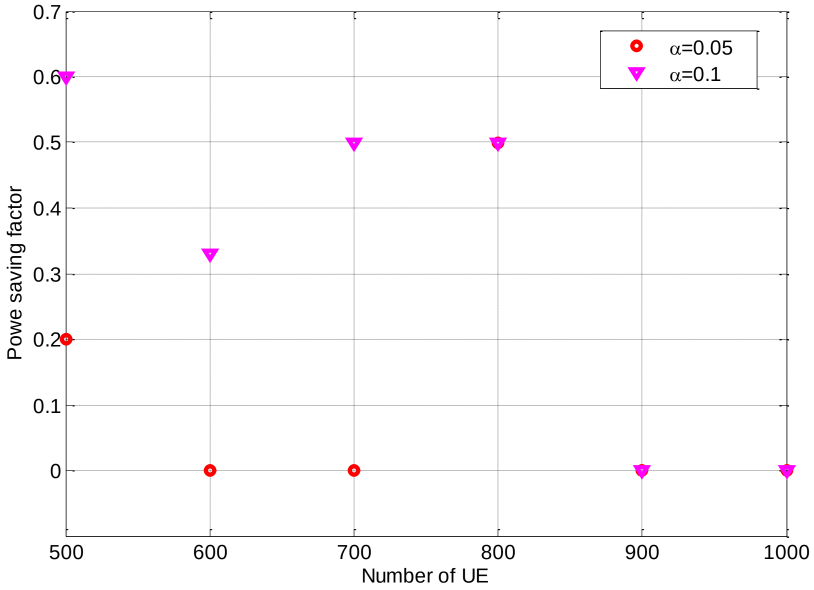


Figure 3.26: Power saving factor for different α with Algorithm 1, $\lambda = 1/50$

for α should consider power capacity of the device as well as the latency requirement. For a MTC device requires very short latency or has large power capacity α should be configured as a smaller value or even zero, while it can be configured as a larger value if the MTC device has small power capacity or does not require very short latency.

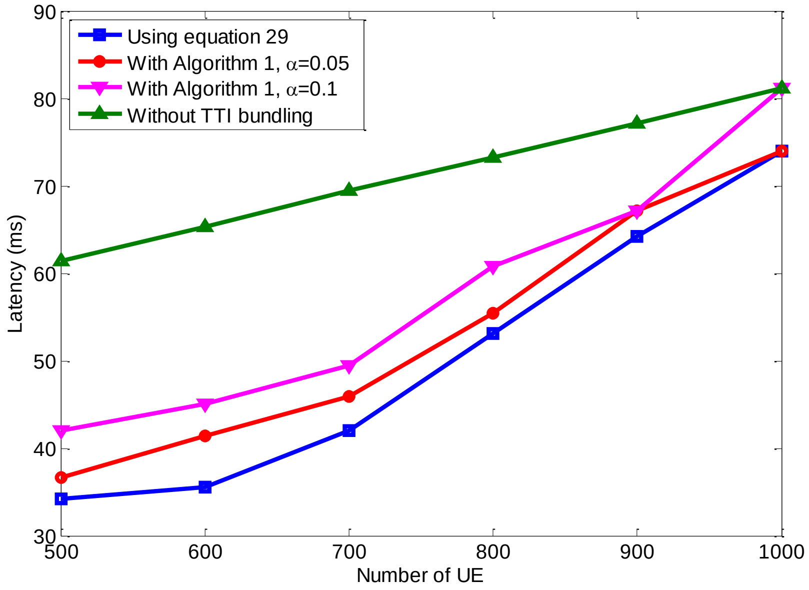


Figure 3.27: Latency comparison, $\lambda = 1/100$

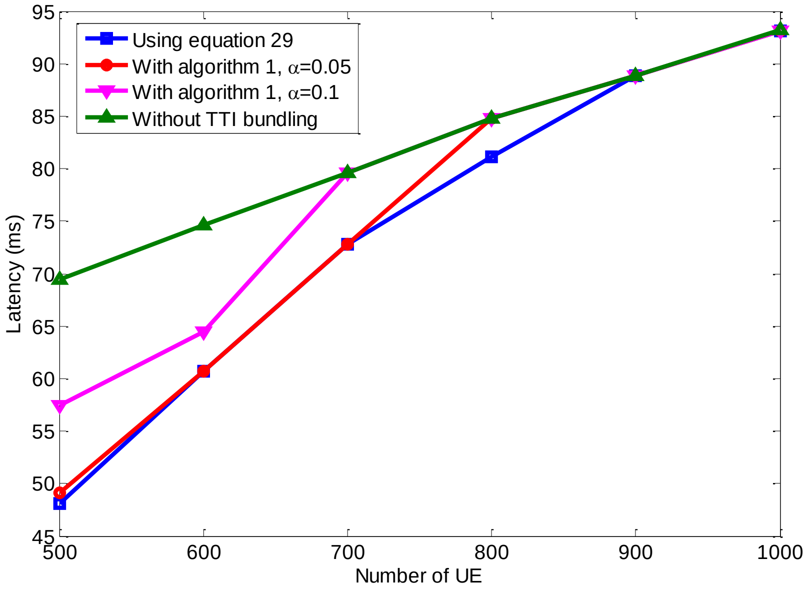


Figure 3.28: Latency comparison, $\lambda = 1/50$

3.5 Usage of the Proposed Methods

The packet aggregation method is simple yet efficient techniques that control the collision rate. It can be easily implemented in LTE with minimum modifications (no modification to PHY/MAC layers). For single type of traffic with elastic delay constraint, the packet loss rate can be reduced to a small value the until the memory limit is reached. While for single type of traffic with strict delay constraint, the packet loss rate can only be reduced to a certain extent. In this case, if the resulted packet loss rate is higher than threshold, more preambles/or PRACH resource should be allocated by eNB. For power constrained MTC device with elastic delay constraint, its energy consumption can greatly be reduced. While for power constrained MTC device with strict delay constraint, its energy consumption can be reached until reaching the delay constraint. For multiple types of traffic, non-real time traffic can be aggregated to reduce the packet loss rate until reaching the memory limit. Moreover, if the resulted the latency is larger than threshold, packets for real-time traffic should also be aggregated to reduce the packet loss rate. Finally, if the latency is still larger than threshold, more preambles/or PRACH resource should be allocated by eNB.

The TTI bundling method is also applicable to LTE, but requires some small modifications to the protocol stack. It reduces latency at the expense of energy (sending multiple preambles). Therefore, the TTI bundling method is applicable to MTC applications which requires very low latency and with adequate power budget.

3.6 Conclusion

Random access is essential for machine type communication uplink channel access in LTE. However, it may suffer from high collision rate due to the large number of simultaneous transmission, which causes large latency. To address this problem, we introduce a packet aggregation method. With the proposed method, a UE does not start a random access until the aggregated packets in the buffer reaches the given threshold. However, this method introduces extra channel access latency which is used to accumulate certain amount of packets. We propose a Semi-Markov process method to analyze the random access procedure with packet aggregation and derive the packet loss rate and channel access latency as functions of amount of aggregated packets. Therefore, the optimal amount of aggregated packet which satisfies the packet loss requirement while keeping the latency as small as possible can be found. The simulation result shows that proposed method set the amount of aggregated packet properly and the packet drop rate is greatly reduced. We also propose an energy saving method for power constraint MTC device: a device with elastic delay constraint can greatly save power through packet aggregation until reaching the delay limit.

In the second method, we propose a TTI bundling method to reduce the latency caused by unsuccessful random access. With TTI bundling, a UE sends one or multiple preambles in one random access round to increase the random access successful rate. To find the optimal TTI bundling number which minimizes the channel access latency, we propose a Semi-Markov process model and formulate the access latency as a function of TTI bundling number. The simulation result shows that by the use of TTI bundling the channel access latency is greatly reduced when the preamble collision rate is low while the gain for using TTI bundling becomes smaller if the preamble collision rate is high. Moreover, we also propose a TTI bundling number scheme to save the extra power caused by multiple preamble transmissions, which is crucial to power constrained MTC device. By setting the threshold α, we can find a good tradeoff between latency decrease and power consumption increase.

Chapter 4
Contention Based Access

4.1 Introduction

As explained in the last chapter, random access is playing an important role in machine type communications in LTE, for example sending scheduling request to enable data packet transmission, RRC connection setup, uplink synchronization. In the last chapter, we introduce the packet aggregation and TTI bundling methods to improve the performance for random access in LTE. However, we find that random access has some limits in latency reduction. Figure 4.1 shows the data packet transmission procedure with random access. Firstly, by the use of random access, a UE sends scheduling request (SR) to apply resource from eNB. Then, the eNB sends schedule grant (SG) to allocated resource for that UE. Thirdly, the UE sends the buffer statue report (BSR). Fourthly, the eNB allocates resource for the UE according to BSR. Finally, the UE sends the data packet. Assuming the time used to decode a packet is 3ms, the total latency for a data packet delivery using random access is 27ms for the best case. However, the latency greatly increases when preamble collision or wireless transmission error happens. We can see that there are considerable signaling overhead to send a data packet, which motives us to design a new uplink channel access method to enable fast data packet transmission for machine type communication in LTE. It has to be mentioned that for RRC connection establishment and uplink synchronization, the standard random access method will be used.

In the proposed method, UEs are not allocated with specific resources, but rather with a pool of common resources where they randomly select for the data transmission. Collision may happen if more than two UEs use the same resource. In this case, dedicated resources are allocated for data retransmissions provided that C-RNTI of the collided UEs can be correctly decoded based on the MU-MIMO detection technique. Thus significant latency gain and overhead reduction is achieved by bypassing the SR, BSR, and preamble procedures used in regular scheduling and random access methods.

The reminder of the chapter is organized as following. Section 4.2 introduces the general idea for the proposed contention based access (CBA) method. In Section 4.3 , detailed low layer signaling enhancement to enable CBA technique in current LTE specification (Rel. 11)

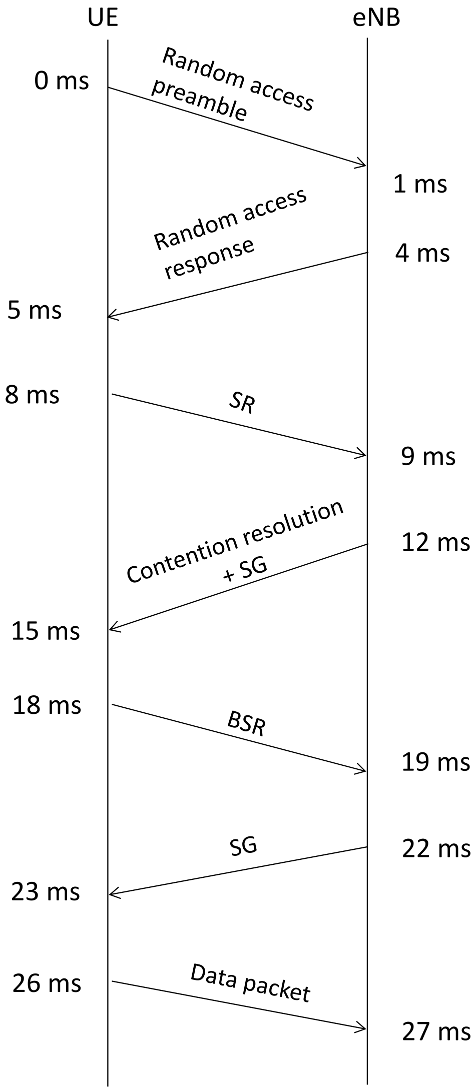


Figure 4.1: Data packet delivery with random access

is presented. Section 4.4 presents a resource allocation scheme for CBA. Section 4.5 provides the simulation results. Section 4.6 provides two examples to use the proposed CBA method and Section 4.7 concludes this chapter and presents the future work.

4.2 Contention Based Access

4.2.1 General Idea for Contention Based Access

To provide a low latency uplink channel access for MTC over LTE, a new resource allocation method, called contention based access (CBA), is proposed. The main feature of CBA is that the eNB does not allocate resources for a specific UE. Instead, the resource allocated by the eNB is applicable to all or a group of UEs and any UE which has data to transmit randomly selects resource blocks among the available resource (see Fig.4.2). The procedure for contention based uplink access is shown in Fig.4.3 (Here we assume that the UE is uplink synchronized[1]). Firstly, the UE receives the resource allocation information which indicates the resource allocated for CBA. Assuming the CBA resource is available in each subframe, a UE wait for 0.5 ms to receive the scheduling grant (SG) information for CBA. Then, after decoding the resource allocation information which costs 3 ms, the UE sends the frame on the randomly selected resources. The latency for this whole procedure is 7.5 ms in the best case, which is much smaller than that of the random access (27 ms). In some sense, the contention based access is similar to the standard random access method: for both methods the resource allocation is common. However, the contention based access method has some specific characteristics.

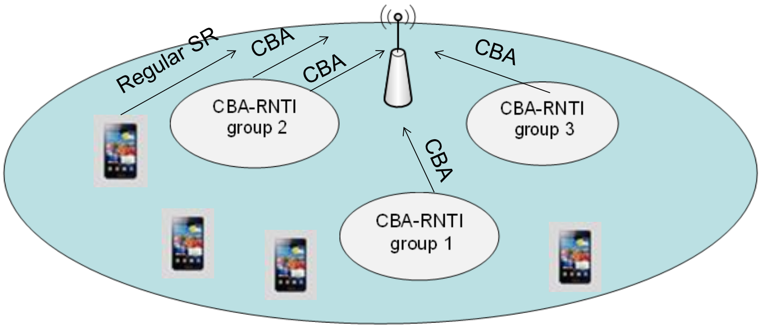


Figure 4.2: Uplink channel access with contention based access

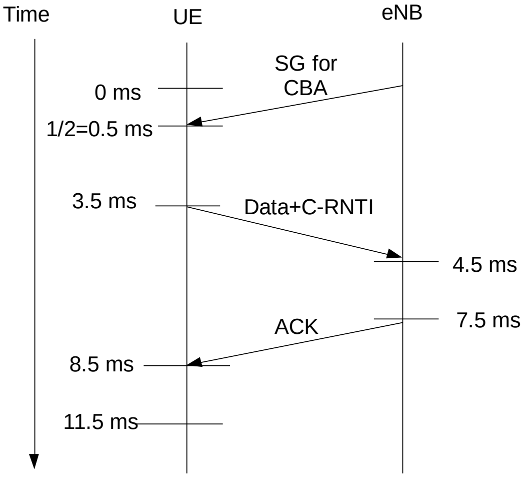


Figure 4.3: Contention based access

[1] This UE is not necessarily to be RRC connected, which will be explained in Section 4.6.

As the CBA resources are not UE specific but rather allocated for all or a group of UEs, collisions may happen when multiple UEs select the same resource. In a network with sporadic traffic, the collision probability is very low, which means most transmissions are free of collision and therefore CBA method outperforms the regular scheduling method in view of latency. However, in a dense network the collision probability is very high, which means lots of retransmission are needed and hence the latency is increased. For example supposing the total available resource block in a subframe is 50, the collision probability is 0.06 if 3 UEs transmit in the subframe, while the collision probability increases to 0.99 if 20 UEs transmit in the subframe.

To solve the above problem, the following method is used. Each UE sends its identifier, C-radio network temporary identifier (C-RNTI), along with the data on the randomly selected resource. Since the C-RNTI is of very small size, therefore it can be transmitted with the most robust modulation and channel coding scheme (MCS) without introducing huge overhead. By the use of MU-MIMO detection, these highly protected CRNTIs might be successfully decoded even if they are sent on the same time-frequency resource. Upon the successfully decoding for the collided C-RNTIs, the eNB triggers regular scheduling for the corresponding UEs as shown in Fig.4.4. Therefore, a UE can retransmit the packet using the regular HARQ procedure. The overall latency for this whole scheduling procedure is still not larger than that of the regular scheduling.

For the collided UEs whose C-RNTIs are not decoded, neither dedicated resource (SG) nor ACK information is received; those UEs have to retransmit the packets as shown in Fig.4.5. It has to noted that the retransmissions is still based on CBA, which is referred as HARQ for CBA as it is different from the regular HARQ procedure (In regular HARQ, dedicated resource is allocated for a UE with retransmission).

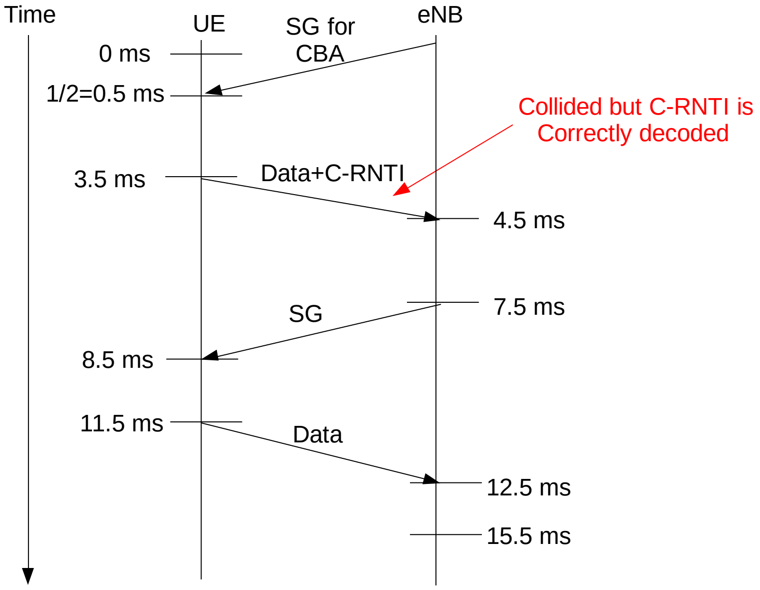


Figure 4.4: Contention based access with collision detection

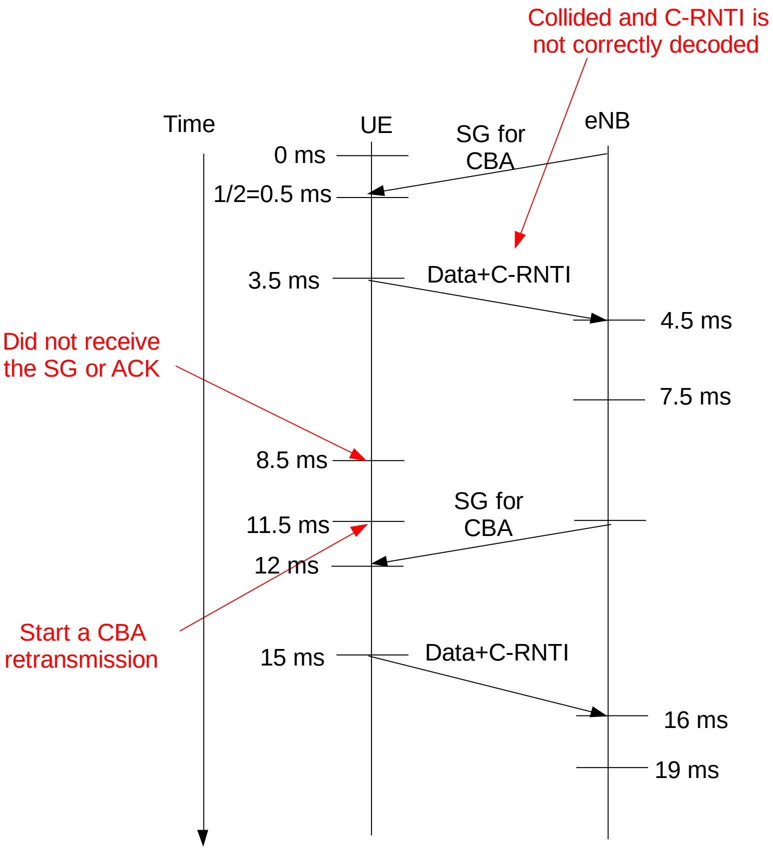


Figure 4.5: Contention based access with retransmission

4.3 Implementation of Contention Based Access in LTE

In order to implement the proposed contention based access method in LTE, some modifications are needed. Before presenting the detailed modifications to LTE, we first introduce the regular uplink scheduling method in LTE. Fig.4.6 demonstrates a regular uplink scheduling procedure in LTE:

1. a UE sends the SR information on the physical uplink control channel (PUCCH) to request resource from eNB.
2. a eNB decodes the SR packet and allocates resource for that UE.
3. a UE sends its buffer state report (BSR) on the allocated resource.
4. a eNB allocates suitable amount of resource for the UE according to its BSR information. The resource allocation information is sent with DCI 0 [2]. The MCS index used for uplink transmission and the cyclic shift used for reference signal are also specified in

[2] DCI format 0 is used for uplink resource allocation, while DCI format 1/1A/1B/1C/1D and 2/2A is used for downlink [76].

DCI 0. This DCI 0 information is attached by a 16-bit CRC, where the CRC parity bits are scrambled with the C-RNTI such that UEs can identify the UE-specific resource.

5. A UE uses its C-RNTI to identify its DCI 0 information. With this DCI 0 information, a UE finds the resource allocation information, the uplink MCS index and cyclic shift. Then it sends the packet on the allocated resource with the specified MCS and cyclic shift.

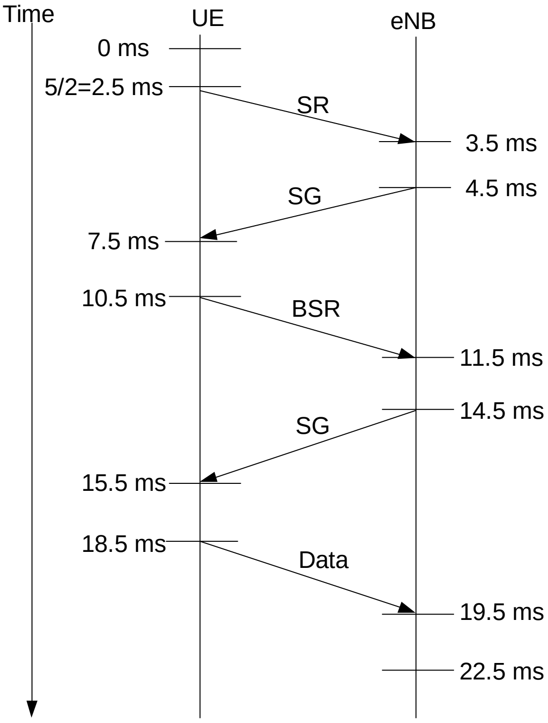


Figure 4.6: Uplink packet scheduling in LTE

4.3.1 Enhancements to the RRC Signaling Related to 3GPP TS 36.331

Signaling to Inform UEs about the CBA-RNTI

The CBA-RNTI, which is used by a UE to decode the resource allocated information for CBA, is allocated by eNB during the RRC connection reconfiguration procedure for the data radio bearer (DRB) establishment. To implement this procedure, the CBA-RNTI has be added to the RadioResourceConfigDedicated information element as defined in [12]. It should be mentioned that the CBA-RNTI is not UE specific. Instead, all UEs or a group of UEs have a common CBA-RNTI configured by RRC signaling.

4.3.2 Enhancement to the PHY Signaling Related to 3GPP TS 36.212

Signaling to Inform UEs about the CBA Resource Allocation

To adapt to the resource allocation for CBA, a new DCI format, DCI format 0A, is defined. The DCI format 0A is used to inform UEs about the resource allocated to CBA. The content of DCI format 0A is shown in Tab.4.1, where N_{RB}^{UL} is number of resource block in the uplink. The CRC for DCI format 0A is scrambled with a new defined radio network temporary identifier CBA-RNTI. With the CBA-RNTI, the UE decodes the DCI format 0A to locate the resource allocated for CBA. As the resource allocation is not UE specific, multiple UEs may select the same resource, which causes collisions.

Table 4.1: Field of DCI format 0A

Information	Type Number of Bits	Purpose
Hopping flag	1	Indicates whether PUSCH frequency hopping is performed
Resource block assignment	$\log_2 N_{RB}^{UL}(N_{RB}^{UL}+1)$	Indicates assigned resource blocks

Signaling to Inform eNB about the Selected MCS

In CBA, the MCS used for uplink transmission is not informed by eNB. Instead, UEs determine the MCS independently [3]. Therefore, the UE should inform the eNB about the selected MCS index so that the uplink frame can be properly decoded. To inform the eNB about the selected MCS index, the following method is proposed.

A new type of control element is defined as shown in Fig. 4.7, which includes the uplink MCS index as well as C-RNTI. This new control element and data form a MAC PDU as shown in Fig. 4.8, which is sent on the physical uplink shared channel (PUSCH). Fig. 4.9 shows the subheader of this new control element. The Logical Channel ID (LCID) identifies the logical channel instance of the corresponding MAC SDU or the type of the corresponding MAC control element or padding. For this new defined control element, its LCID is set to be 01011. The definitions for the other fields in the subheader can be found in [60] .

It has to be noted that the MCS for CBA data and control information are usually different. To achieve that, the CBA data and control information are treated independently: high MCS is used for CBA data while low MCS is employed for CBA control information. The resulted bit streams (one for CBA data and other for CBA control information)are then assembled for further processing, e.g., resource mapping. This is common in LTE. For example: the uplink channel quality indicator (CQI) can be multiplexed with data, and sent on PUSCH.

[3] In CBA, cyclic shift for uplink reference signal is used for channel estimation. The cyclic shift is random selected by UEs. The eNB identify a UE's cyclic shift by trying all the possible values. The one which attains the highest peak value is considered as the used cyclic shift.

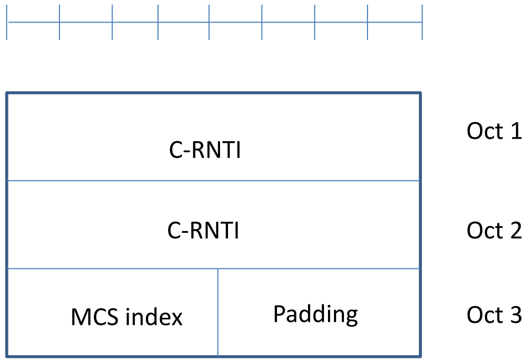


Figure 4.7: Control element for MCS and C-RNTI

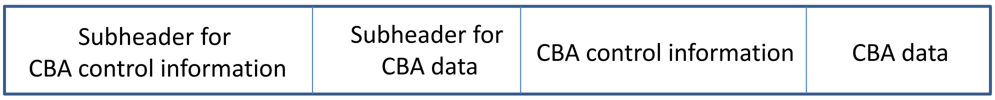


Figure 4.8: MAC PDU for CBA transmission

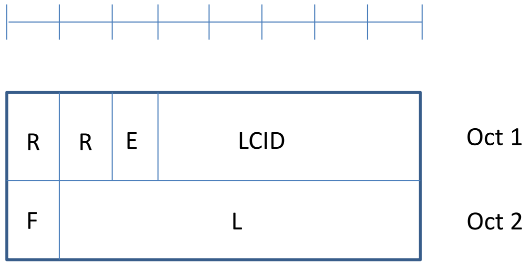


Figure 4.9: Subheader for CBA control element

4.3.3 Enhancement to the PHY Procedures related to 3GPP TS 36.213

UE Procedure to Locate the Resource Allocated for CBA

With the allocated CBA-RNTI which is obtained during the RRC connection reconfiguration procedure procedures, a UE can locate the resource allocated for CBA by decoding the DCI format 0A information.

UE Hybrid ARQ (HARQ) Procedure for CBA

UEs should find the ACK/NACK information after sending the data frame such that a new transmission or a retransmission can be properly performed. To adapt to CBA, the method to locate the ACK/NACK information is described as follows.

The ACK information is sent by eNB for a correct received CBA packet, which is the same as the one specified in [76]. The physical HARQ indicator channel (PHICH) index is implicitly associated with the index of the lowest uplink resource block and the cyclic shift used for corresponding contention-based access. Therefore, UEs which successfully send frames can find the corresponding ACK information without extra signaling. The details for this method can found in [76]. If two UEs select the same resource and use the same cyclic shift, we assume that none of the two packets can be decoded by eNB as the eNB cannot correctly estimate channel information for these two UEs. As a result, no ACK is sent [4]. On the other hand, if two UEs select the same resource and use different cyclic shifts, it is possible that one of the two collided packet is correctly decoded. In this case, an ACK is sent. The UE who sends the correctly received packet can locate the ACK; the other UE cannot find the ACK since it uses a different cyclic shift.

For UEs whose C-RNTI is correct detected but data is corrupted, here we also introduce a new method to represent the NACK information. As introduced in Section 4.2, for these unsuccessful UEs with successfully decoded C-RNTI, the eNB triggers regular scheduling by sending the DCI 0 information (not DCI 0A). In the DCI 0 information, the new data indicator (NDI) field is set to 0 to represent the NACK information. Hence once a UE receives the DCI 0 information with NDI 0, it infers that the last transmissions is unsuccessful. And then, this UE starts a retransmission on the dedicated resource indicated by DCI 0 information. For the FDD system, the UE starts the retransmission three subframes later after receiving the DCI with format 0 as shown in Fig.4.4.

There are also some UEs whose C-RNTIs cannot be successfully decoded. For these UEs, they cannot receive any information at the expected time. As a result, new retransmissions with CBA are performed as shown in Fig.4.5.

CBA Data Reception

As the CBA data format is different from the legacy format as shown in Fig.4.8, a eNB should decode the CBA packet in a different way. Specifically, for the data sent on the CBA resource, the eNB should be aware of that part of payload is the UL-SCH data while the other part is MCS and C-RNTI. With this method, the eNB can try to decode the C-RNTIs for the collided UEs.

Resource Allocation for Correctly Received C-RNTI

For an erroneous packet with correctly decode C-RNTI, the eNB allocates dedicated resource for the corresponding UE and informs the UE through the DCI 0 information. With this allocated resource, a UE can send the data packet without collision.

With the above modifications to the LTE standard, we can implement the contention based access method using the following proposed architecture.

[4] In some case, even if two UEs use the same cyclic shift, one of packets can be correctly decoded by eNB. As a result, an ACK is received by both UEs. How to handle this problem would be considered in our future work.

4.3.4 CBA Architecture

The architecture for CBA at the UE side is shown in Fig.4.10.

1. The UE uses the received RRC message to configure its MAC and PHY layers. With RNTI allocated for CBA transmission, the UE decodes the DCI 0A information to locate the resource allocated for CBA.
2. With the resource allocation information for CBA, the resource selection module randomly selects resource and passes the result to physical configuration module.
3. The physical configuration module sets the proper parameters for transmission using the randomly selected cyclic shift and the results from the resource selection module.
4. Finally the transmission module sends the data and the CBA control information following the instructed configuration.
5. A UE uses the regular HARQ procedure to retransmit the packet if SG is received or it retransmit the packet by the use of CBA (CBA-HARQ) if nothing is received.

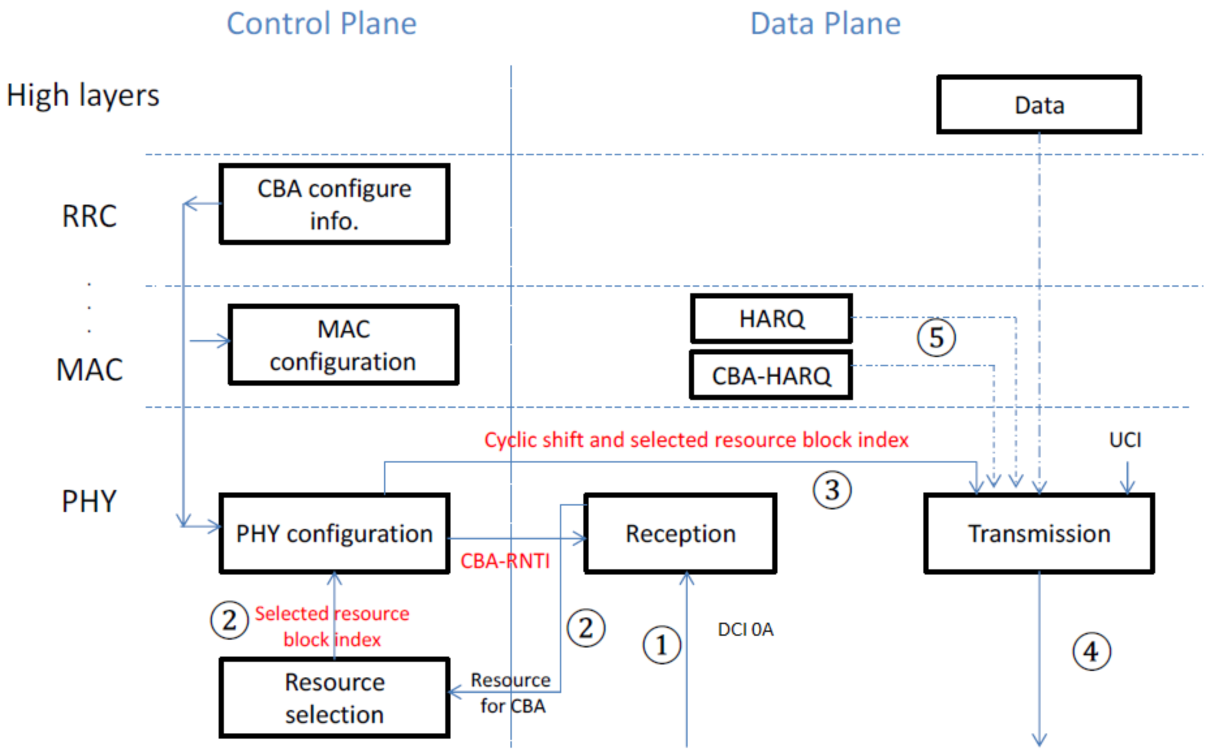


Figure 4.10: CBA architecture at UE side

The architecture for CBA at the eNB side is shown in Fig.4.11

1. The eNB sends the CBA configuration information to UEs through the RRC message. In addition to that, the eNB also performs resource allocation for CBA, for which the related resource allocation information is sent with DCI 0A and the CRC parity bits of DCI 0A is scrambled by RNTIs allocated for CBA transmission.
2. As for the reception, the ACK information is sent to the transmission module for the correct received CBA packets such that it can be sent on the PHICH channel.
3. The successfully decoded C-RNTIs of the collided UEs are sent to regular scheduling module and hence the specific resource can be allocated for those UEs and informed through the DCI 0 information

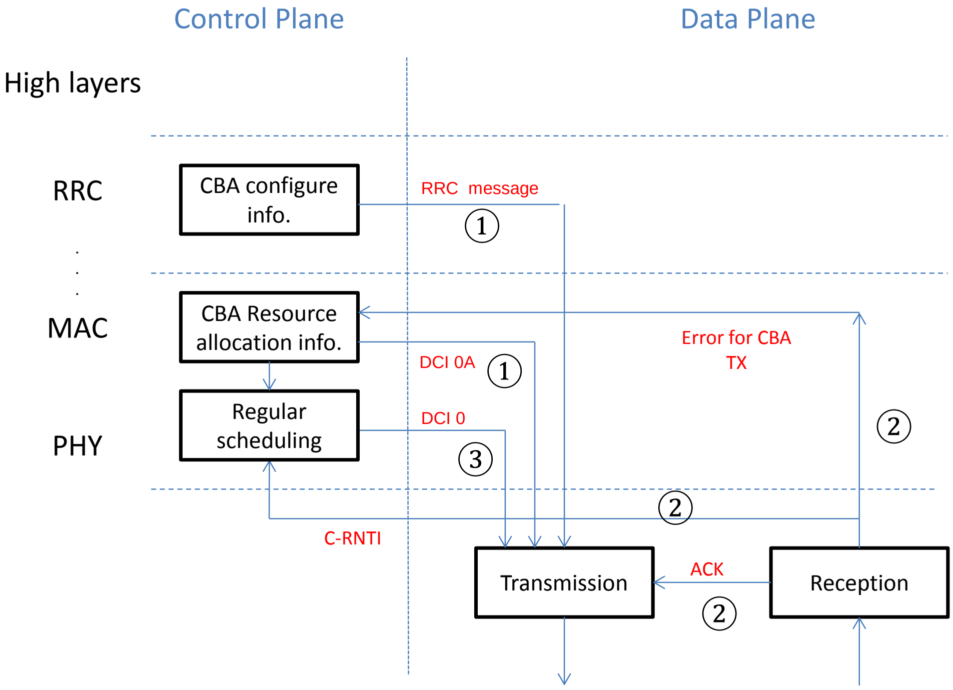


Figure 4.11: CBA architecture at eNB side

4.4 Resource Allocation Scheme for Contention Based Access

The main target for resource allocation is to assign the proper amount of resource such that the latency constraints are satisfied and the allocated resources are efficiently used. Accurate resource allocation for CBA is very important as it is directly connected to latency experienced by the application traffic.

Let us denote the total number of resource elements allocated for one CBA transmission as N_{RACH}. This contains the amount of resource elements used for control information transmission, denoted as N_{ctrl} in addition to those reserved for data N_{data}, i.e.,

$$N_{RACH} = N_{ctrl} + N_{data}. \tag{4.1}$$

Therefore, The spectral efficiency of the control information is

$$R_c = 24/N_{ctrl}(\text{bits/RE}) \tag{4.2}$$

under the assumption that the control information comprises 24 bits (16 bits for C-RNTI, 4 bits for MCS and 4 padding bits). Similarly, the spectral efficiency of the data is

$$R_d = M_{data}/N_{data}(\text{bits/RE}) \tag{4.3}$$

where M_{data} is the bit of data payload.

For each contention based access transmission, we have the following events:

1. neither the control information nor the data are detected, which is denoted as E_1;
2. the control information is not detected but the data is detected, which is denoted as E_2;
3. the control information is detected but the data is not detected, which is denoted as E_3;
4. both the control information and data are detected, which is denoted as E_4.

In order determine the probability of each event we take a an approach based on instantaneous mutual information. This asymptotic measure yields a lower bound on the above probabilities for perfect channel state information at the receiver. To this end, the received signal of *m*th antenna at resource element *k* is

$$y_m[k] = \sum_{u=0}^{N_u-1} H_{m,u}[k]x_u[k] + Z_m[k], m = 0, \cdots, N_{\text{RX}} - 1 \tag{4.4}$$

where $H_{m,n}[k]$ is the channel gain for user u at antenna m, $x_u[k]$ is the transmitted signal, $Z_m[k]$ is the noise, and N_u is the random number of active users transmitted on this resource block.

The normalized sum-rate for N_u contending users based on mutual information for both data and control portions is computed as

$$I_{\text{X}} = \frac{1}{N_u N_{\text{X}}} \sum_{k=0}^{N_{\text{X}}-1} \log_2 \det \left(\mathbf{I} + \sum_{u=0}^{N_u-1} \gamma_u \mathbf{H}_u[k] \mathbf{H}_u^*[k] \right) \tag{4.5}$$

where X is represents either control or data, $\gamma_n, n = 0, \cdots, N_u - 1$, is the received signal-to-noise ratio (SNR) and $\mathbf{H}_i[k] = \begin{pmatrix} H_{0,n}[k] & H_{1,n}[k] & \cdots & H_{N_{\text{RX}}-1,n}[k] \end{pmatrix}^T$. The use of this expression requires the two following assumptions. Firstly, all channels can be estimated at the receiver irrespective of the number of contending users. This has to make proper use of the cyclic shifts to guarantee contention-free access for channel estimation. In practice, for loaded cells with only CBA access, this will require association of UEs to orthogonal CBA resources (in time/frequency) and on a particular CBA resource a maximum of 12 contenting UEs can be accommodated. Secondly, the expression assumes Gaussian signals and that the eNB receiver uses an optimal multi-user receiver (i.e. it performs complete joint detection.) These expressions can be found in [77].

Assuming there are i active UEs contending on the same CBA resource unit, for a given UE the probabilities of the four events caused by one CBA transmission are:

$$P_{E_1,i} = P_{S,i} + (1 - P_{S,i})p(I_{ctrl} < R_c, I_{data} < R_d), \tag{4.6}$$

$$P_{E_2,i} = (1 - P_{S,i})p(I_{ctrl} < R_c, I_{data} > R_d), \tag{4.7}$$

$$P_{E_3,i} = (1 - P_{S,i})p(I_{ctrl} > R_c, I_{data} < R_d), \tag{4.8}$$

$$P_{E_4,i} = (1 - P_{S,i})p(I_{ctrl} > R_c, I_{data} > R_d), \tag{4.9}$$

where $P_{S,i}$ is the probability that other UEs use the same cyclic shift on one CBA resource unit provided that there are i contending UEs (we assume that if the multiple UEs select the same cyclic neither control information nor data can be decoded by eNB as the eNB cannot correctly estimate the channel). In general, the control information is more protected than the data, i.e., $R_c < R_d$, so $P_{E_2,i} \approx 0$.

Then the expected value for the probabilities of the four events are:

$$P_1 = \sum_{i=1}^{N} P_{A,i} P_{E_1,i} \tag{4.10}$$

$$P_2 = \sum_{i=1}^{N} P_{A,i} P_{E_2,i} \approx 0, \tag{4.11}$$

$$P_3 = \sum_{i=1}^{N} P_{A,i} P_{E_3,i} \tag{4.12}$$

$$P_4 = \sum_{i=1}^{N} P_{A,i} P_{E_4,i} \tag{4.13}$$

where $P_{A,i}$ is the probability that there are i active UEs contending on one CBA resource unit; N is the total amount of UEs in a cell.

To minimize the latency for the MTC traffic, the CBA resource should be available in each subframe. The resource allocation can be performed in the following steps:

1. Set the CBA resource unit
2. Initialize the amount of CBA resource unit to 1
3. Calculate the probabilities of the four events caused by a CBA transmission.

4. Calculate the latency based on the measured amount of CBA resource unit
5. If the estimated latency is larger than the latency constraint, increase the amount of resource unit by one and go back to step 3. Else end

It can be seen that as the latency decreases with amount of CBA resource unit, therefore with the above method we always find the minimum amount of CBA resource. It has to noted that here we assume that there is always enough resource. For a system which has a constraint on CBA resource, more intelligent scheduler can used to address this problem, for example a scheduler which consider the priorities between real time and non-real time traffics.

4.4.1 Estimation of the Probabilities of Events in Step 3

To estimate the probabilities of the four events caused by a CBA transmission, we drive a Semi-Markov chain model as shown in Fig. 4.12, where

- S_0 means that there is no packet in the UE's buffer,
- S_{2i-1}, $i \in [1, M]$, means the ith CBA transmission of the UE, where M is the transmission limit,
- S_{2i}, $i \in [1, M-1]$, means that the UE is waiting for the ACK or SG information.

The UE transfers between states as:

- When the UE is at state S_0, if a packet arrives, it transfers to states S_1 to start the first transmission; otherwise it remains at state S_0
- When the UE is at state S_{2i-1}, $i \in [1, M-1]$, it sends the packet and transfers to S_{2i}
- When the UE is at state S_{2M-1}, it sends the packet and transfers to S_0.
- When the UE is at state S_{2i} $i \in [1, M-1]$: (1) if ACK is received it transfers to state S_0; (2) if SG is received it sends the packet as shown in Fig. 4.4 and then transfers to state S_0; (3) if neither ACK nor SG is received at the expected time instant, it transfers to state S_{2i+1} to retransmit the packet as shown in Fig. 4.5.

Denoting $p_{i,j}$ as the state transition probability from state S_i to state S_j, $i, j \in [1, 2M-1]$, the state stationary probability of state i can be calculated as:

$$\begin{cases} \pi_0 = \pi_0 p_{0,0} + \sum_{i=1}^{M-1} \pi_{2i} p_{(2i),0} + \pi_{(2M-1)} p_{(2M-1),0} \\ \pi_{2i-1} = \pi_{2i-2} p_{(2i-2),(2i-1)}, i \in [1, M] \\ \pi_{2i} = \pi_{2i-1} p_{(2i-1),(2i)}, i \in [1, M-1]. \end{cases} \tag{4.14}$$

With the above equations, we can get

$$\pi_i = \prod_{j=1}^{i} p_{(j-1),j} \pi_0, i \in [1, 2M-1]. \tag{4.15}$$

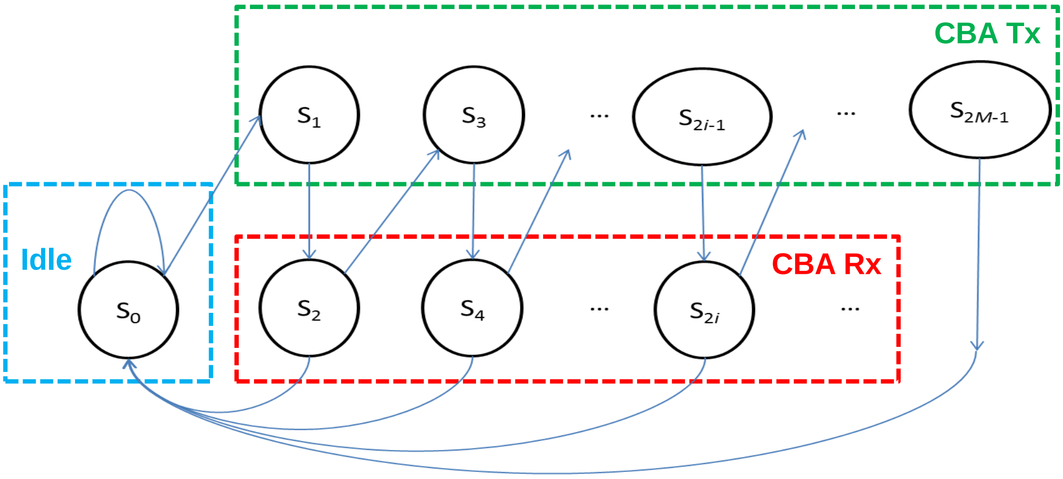


Figure 4.12: Markov chain model for contention based access

Substituting (4.15) into the following equation

$$\sum_{i=0}^{2M-1} \pi_i = 1, \tag{4.16}$$

we can get

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^{2M-1} \prod_{j=1}^{i} p_{(j-1),j}}. \tag{4.17}$$

The state transition probability can be calculated as following. In each subframe (1ms) if a packet arrives, the UE transfers from state S_0 to state S_1. Supposing the packet arrives following a Poisson distribution with the arrival rate λ, we have $p_{0,1} = 1 - e^{-\lambda}$. When the UE is at state S_{2i-1}, after transmission it transfers to state S_{2i}, therefore $p_{(2i-1),2i} = 1$, $i \in [1, M-1]$.

When the UE is at state S_{2i}, it transfers to state S_{2i+1} if neither ACK nor SG is received, i.e., it transfers state S_{2i+1} if neither the control information nor the data are detected, therefore

$$p_{2i,(2i+1)} = P_1. \tag{4.18}$$

With derived transition probability, π_0 can be calculated as

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^{M-1} 2(1-e^{-\lambda})P_1^{i-1} + (1-e^{-\lambda})P_1^{M-1}} \tag{4.19}$$

and π_i can be calculated using (4.15). We can see that π_i, $i \in [1, 2M-1]$, is a function of P_1.

Now let us calculate the state holding time D_i (in ms) for state S_i, $i \in [1, 2M-1]$. In state S_0 for every subframe the UE checks if a packet arrives. If so, it transfers to state S_1, therefore $D_0 = 1$.

In state S_{2i-1} as shown in Fig. 4.3 the UE first waits for resource allocation information for CBA and then sends the packet; finally it transfers to state S_{2i}, therefore $D_{2i-1} = 3.5$, $i \in [1, M]$.

When the US is in state S_{2i}: (1) if ACK is received it transfers to state S_0, the state holding time for this case is $11.5 - 3.5 = 8$ms as shown in Fig. 4.3; (2) if SG is received it sends the packet on the allocated resource as shown in Fig. 4.4 and then transfers to state S_0, the state holding time for this case is $11.5 - 3.5 = 8$ms; (3) if neither ACK or SG is received at the expected time instant, the UE transfers to state S_{2i+1} to start a retransmission as shown in Fig. 4.5, the state holding time for this case is also $11.5 - 3.5 = 8$ms. Hence, $D_{2i} = 8$, $i \in [1, M - 1]$.

Denoting Q_i, $i \in [1, 2M - 1]$, as the proportion of time that the UE is in state i, it can be calculated as

$$Q_i = \frac{\pi_i D_i}{\sum_{i=0}^{2M-1} \pi_i D_i}, \tag{4.20}$$

which is a function of P_1.

A UE trigger a CBA transmission in state S_{2i-1} and the time used for a CBA transmission is 1ms. Therefore the probability that a UE is performing a CBA transmission is

$$\tau = \sum_{i=1}^{M} Q_{2i-1} \frac{1}{D_{2i-1}}. \tag{4.21}$$

which is also a function of P_1.

For a UE which is performing a CBA transmission, the probability that there are i another UEs contending on the same CBA resource is

$$P_{C,i} = \sum_{j=i}^{N-1} \binom{N-1}{j} \tau^j (1-\tau)^{N-1-j} \binom{j}{i} (\frac{1}{N_{RE}})^i (1 - \frac{1}{N_{RE}})^{j-i} \tag{4.22}$$

where $i \in [0, N - 1]$, N is the total amount of UEs in a cell and N_{RE} is the amount of CBA resource unit.

Therefore, the probability that there are i contending UEs use the same CBA resource unit is

$$P_{A,i} = P_{C,(i-1)}, i \in [1, N]. \tag{4.23}$$

which is a function of τ.

Moreover assuming the amount of UE which contends on the same CBA resource is i, for a given active UE the probability that other UE selects the same cyclic shift is

$$P_{S,i} = 1 - (\frac{11}{12})^{i-1}. \tag{4.24}$$

It has to be mentioned that above equation holds since the maximum available cyclic shifts in one CBA resource unit is 12. Hence, with equations (4.24) and (4.6) we can calculate $P_{E_1,i}$ for i contending UEs.

With the above results, the probability for the first event is

$$P_1 = \sum_{i=1}^{N} P_{A,i} P_{E_1,i} \tag{4.25}$$

which is a function of τ.

We can see that equations (4.21) and (4.25) comprise a system of equations with two unknown P_1 and τ, which could be solved by numerical methods. Hence, we can calculate P_3 and P_4 using (4.12)-(4.13), respectively.

4.4.2 Estimation of the Latency in Step 4

With the results derived in last subsection, we can estimate the latency for given amount of CBA resource.

As stated at the beginning of this section, for each CBA transmission we have four events. Here we denote the packet transmission latency for the four events as T_1, T_2, T_3, and T_4 (in ms), respectively. Hence the average latency can be calculated as:

$$T = P_1T_1 + P_2T_2 + P_3T_3 + P_4T_4. \tag{4.26}$$

As $P_2 \approx 0$, so the above equation can be simplified as:

$$T = P_1T_1 + P_3T_3 + P_4T_4. \tag{4.27}$$

For an unsuccessful CBA transmission where both data and control information cannot be decoded retransmission happens 11.5ms after the initial transmission as shown in Fig. 4.5, therefore T_1 can rewritten as $T_1 = T_5 + 11.5$, where T_5 is packet delivery latency for a new CBA transmission. Moreover, as shown in Fig. 4.3 and 4.4, we have T_3=15.5, and T_4=7.5. With the above results, we have $T = (T_5 + 11.5)P_1 + 15.5P_3 + 7.5P_4$.

Since $E(T) = E(T_5)$, the expected channel access latency is

$$E(T) = \frac{11.5P_1 + 15.5P_3 + 7.5P_4}{1 - P_1}. \tag{4.28}$$

where P_1, P_3 and P_4 are calculated in the third step.

4.5 Simulation Results

To evaluate the performance of proposed contention based access method and its resource allocation scheme, simulations are performed using MATLAB. We assume that the system is operating FDD-LTE; the SNR is set to 5 dB; transmission limit M is set to 5 and the number of receiving antennas is 2; the coding rate for the control information $R_C = 0.2$. For simplicity we have assumed a line-of-sight dominant channel model with randomized angle-of-arrival at the eNB receiver in order to model the $\mathbf{H}_i[k]$. The CBA resource unit is set to be 6 resource blocks, i.e. $6 \times 12 = 72$ subcarriers, which is same as the resource of the PRACH channel. Moreover, the packet size is assumed to be of small size, following an exponential distribution with average packet size of 100 bits. The packet arrival rate λ is 1/100 packet/ms.

4.5.1 Performance Evaluation

Firstly, to validate the proposed contention based access (CBA) method, we compare the channel access delay of CBA with that of random access (referred as PRACH method). We compare these performance of two methods with the same amount of resources. Concretely, for the CBA method, we allocate one CBA resource unit containing 6 resource blocks in every subframe. While for the PRACH method, the preamble is set as 64 and the PRACH resource configuration index is set to 14, which occupies the same resource as CBA (6 resource blocks in every subframe) and it is the maximum allowed resource for PRACH in LTE. The transmission limit for random access is 5; the random access response window size is 10; and the contention resolution timer is 24 ms.

Fig.4.13 shows the simulation results. We can see that the latency of CBA is much smaller than that of the PRACH method. It shows that the latency gain to use CBA is around 30 ms, which validates that CBA outperforms the PRACH method in the term of latency.

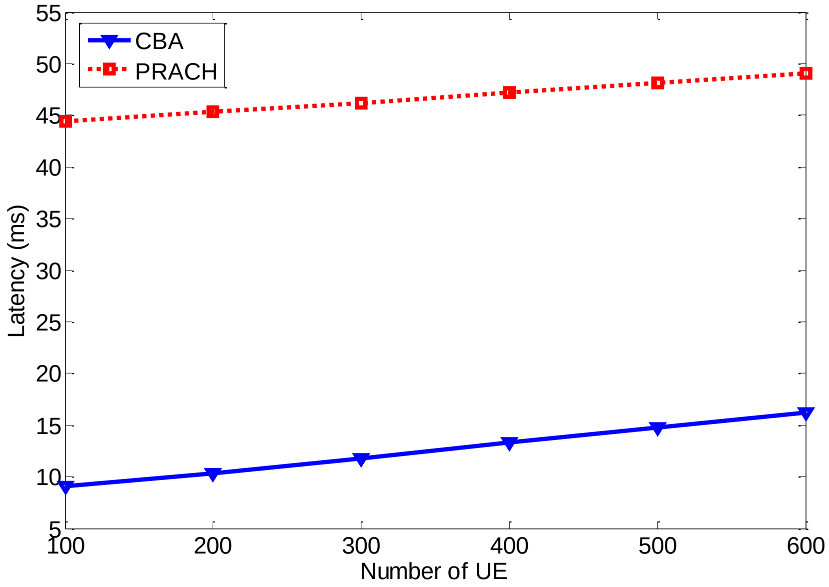


Figure 4.13: Latency comparison

Then, we analyze the effect of different parameters (coding rate of control information, number of receiving antenna, CBA resource unit size) on the CBA performance.

The coding rate R_c determines the amount of resource used for control information. More resource is used for control information when the coding rate decreases, which makes the control information more robust to the wireless channel error. However, since the resource for CBA transmission is fixed (the resource is shared by information and data), the resource used for data transmission is reduced when the coding rate decreases, which indicates that the data becomes more sensitive to wireless channel error. A robust control information is more likely received by eNB which reduces latency, while a sensitive data is more easily corrupted which in turn increases latency. Therefore, the coding rate R_c has strong effect on the

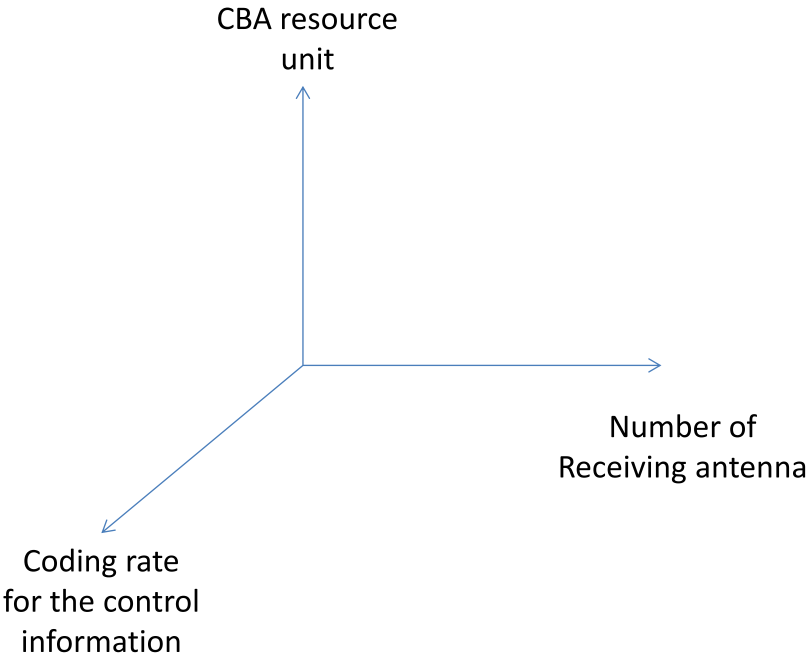

CBA resource unit
Number of Receiving antenna
Coding rate for the control information

Figure 4.14: Effect of parameters on the CBA performance

CBA performance. Fig. 4.15 shows the effect of R_c(coding rate for the control information) on the CBA performance. We can see that the latency when $R_c = 0.3$ is less than other two cases. For example, the latency is 15.6ms when $R_c = 0.3$ and number of user is 600, while it is 15.9 or 17.4 when R_c equals 0.2 or 0.1, respectively. Therefore, to achieve the best performance of CBA, the code rate for the control information should be carefully selected.

Using more receiving antenna increases the successful rate to receive the control information as well as data. Fig. 4.16 presents the latency under different number of receiving antennas. It can found that the latency decreases when the number of receiving antennas increases. For example, the latency decreases from 15.9 ms to 12.2ms when the number of UE is 600 and number of receiving antenna increases from 2 to 3, and it further reduces to 10.9 ms when the umber of receiving antenna increases to 4.

Assuming the total amount resource allocated for CBA is fixed, the size of the CBA resource unit also effect the performance. Larger CBA unit size is beneficial to the transmission of data and control information. However, larger CBA unit size yields smaller amount of CBA resource unit, which increases the collision rate for cyclic shift and hence the latency. Fig. 4.17 demonstrates the latency under different CBA resource unit. It is shown that by setting the CBA resource unit size to 6 resource blocks, the minimum latency can be achieved. Therefore, to attain the best performance of CBA, the CBA resource unit size should be carefully tuned.

We implemented CBA on the OpenAirInterface platform (http://www.openairinterface.org/). The performance of the CBA is also validated by

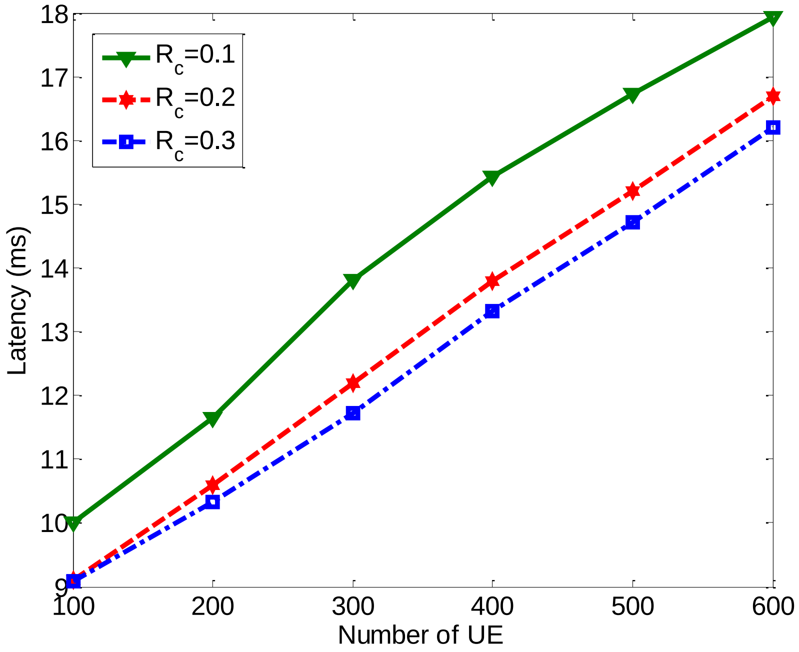


Figure 4.15: Effect of R_c on the CBA performance

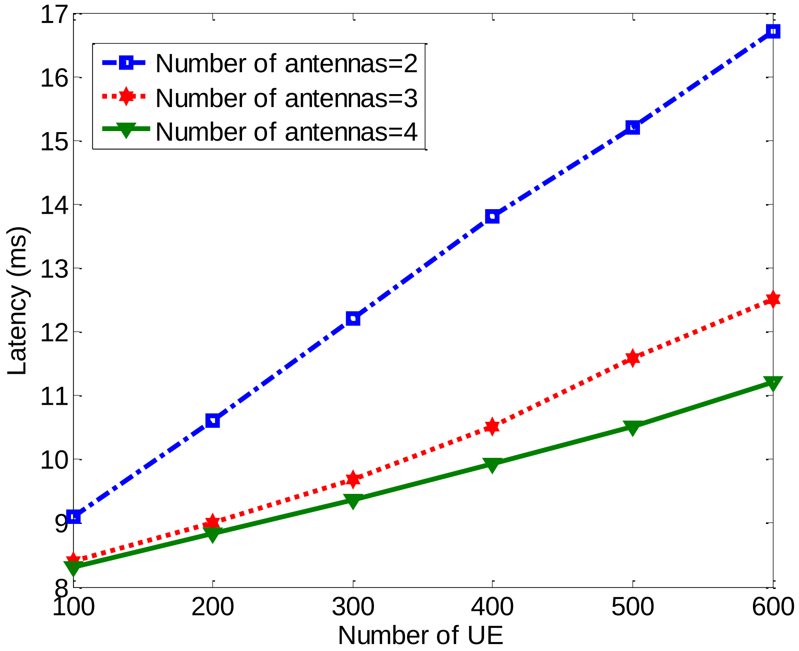


Figure 4.16: Effect of number of receiving antennas on the CBA performance

the use of OpenAirInterface, which can be found at Deliverable 5.3 of the LOLA project (www.ict-lola.eu/deliverables/wp5-integration-and-validation).

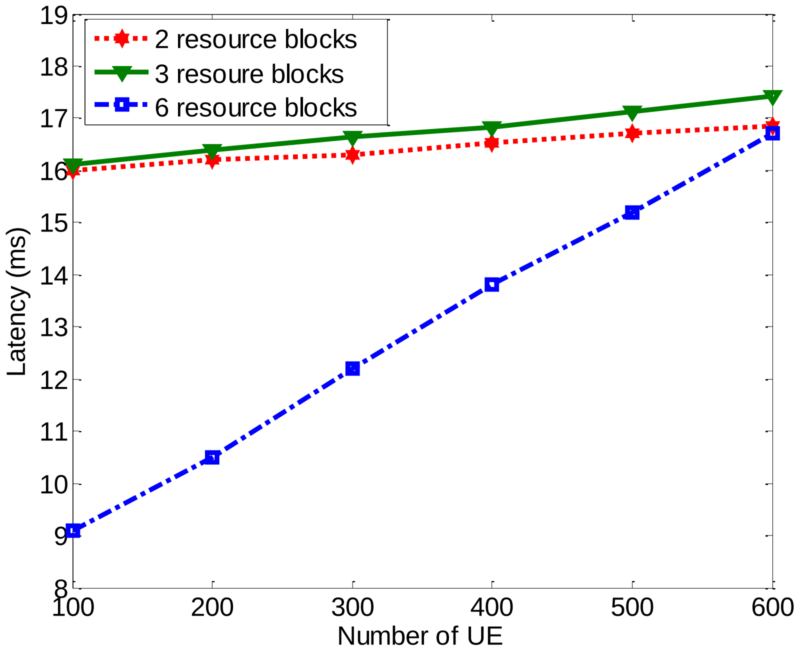


Figure 4.17: Effect of CBA resource unit size on the CBA performance

4.5.2 Performance of the Resource Allocation Scheme

Fig. 4.18 shows the resource allocation results using our proposed method with different packet arrival rates λ (packets/ms) and number of UEs when the delay constraint is 30ms. We can see that the allocated resource units non-decrease with the increase number of UEs and/or packet arrival rate. This is because the packet collision rate increases with number of UEs and packet arrival rate, which hence increases latency. To satisfy the delay constraint, more resource should be allocated. For instance, when $\lambda = 1/30$ and the number of UE is 300, the CBA resource unit is one and the latency is 28.9 ms which is very close to the threshold 30ms. Therefore, when the number of UE increases to 400, two CBA resource units are allocated which reduces the latency to 18ms. Similarly, when the number of UEs reaches 600, the CBA resource unit is increased to three, and the latency decreases to 18ms. Fig. 4.19 demonstrates the delay when using the allocated amount of resource shown in Fig. 4.18. It can be seen that the delay is smaller than the delay constraint 30ms, which validates the proposed resource allocation method.

4.6 Application Scenario for Contention Based Access

Contention based access method can be used for UEs which are uplink synchronized. Here, we provide two examples where CBA can be used.

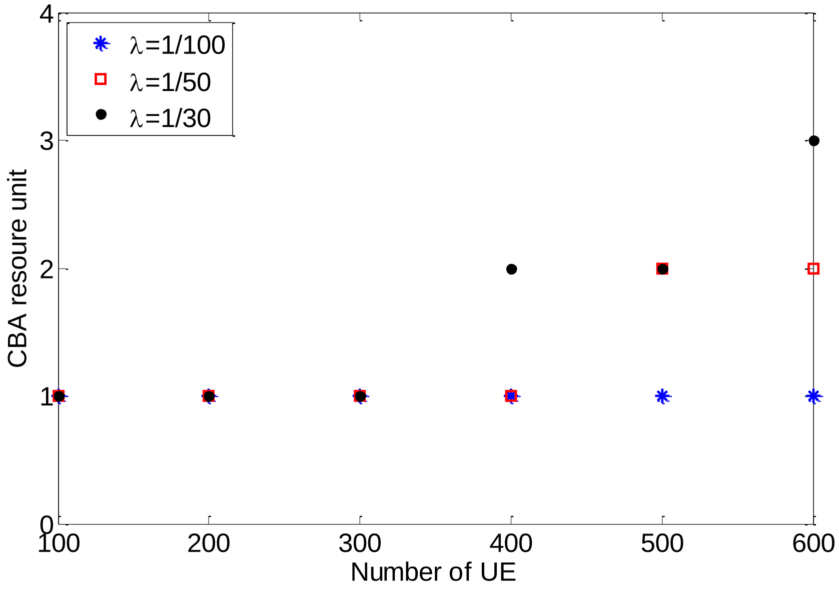


Figure 4.18: CBA resource allocation for different number of UEs

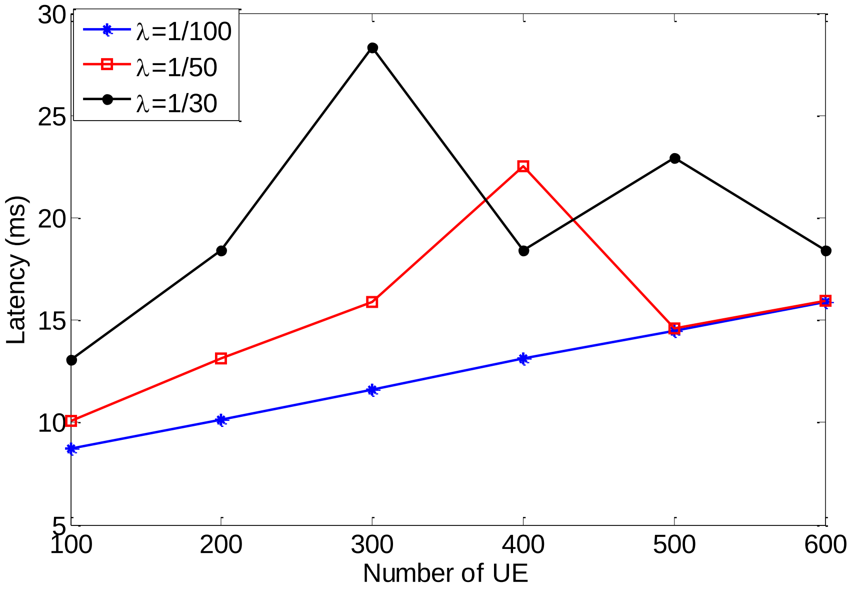


Figure 4.19: Latency of CBA resource allocation method

On-line Gaming

Though CBA is designed for MTC, it is also applicable to on-line gaming (Team Fortress 2, Open Arena, Dirt2). In online gaming, there are frequent data transmissions between eNB and UE to exchange and update the gaming information, which makes the UE always uplink synchronized with eNB.

Fixed Sensor Network

The CBA method also be used to fixed sensor network. For a MTC device, after using the random access to get RRC connected to the network, it may switch off the embeded system to save power, which makes the device transfer to the RRC_IDLE state. With the regular method, when this device needs to trigger a uplink transmission, it should first uses random access to get connected to the network and the apply for resource from eNB, which takes a time period around 30 ms (best case) and consumes much power. However, this problem can also be addressed using CBA if this device is stationary.

If a MTC device is stationary, then timing advance message obtained in the initial random access procedure is always effective. For this type of MTC device, we introduce a new RRC state: RRC-pre-connected state. With the RRC-pre-connected state, a UE's context can be kept at eNB even if its RRC connection is released. Hence, when a new data comes, a stationary MTC device can send it directly through CBA because the timing advance information is still effective and its context is also available at eNB.

4.7 Conclusion

To eliminate the signaling overhead of the random access for data transmission, a contention based access (CBA) method is proposed in this chapter. With CBA, UEs select resource randomly without indications from eNB, which saves signaling overhead and hence the latency can be reduced. To address the problem of collision, a control header (C-RNTI) with higher protection combined with MU-MIMO detection at the eNB allows for the identification the collided UEs and their transport format so that it could allocate specific contention-free resources in subsequent subframes. To enable CBA in LTE, we present the modifications to the LTE standard as well as a resource allocation method. The proposed resource allocation method finds the minimum needed resource for CBA by increasing the amount of resource until the estimated latency less than the latency requirement.

We carried simulations to validate CBA and the proposed resource allocation scheme. The simulation results show that: (1) CBA outperforms the random access method in the term of latency; (2) using the proposed resource allocation method for CBA, the latency constraint can be satisfied; (3) the coding rate for the control information, the CBA resource unit size, and the number of receiving antenna have strong effect on the performance of CBA.

CHAPTER 5

Discontinuous Reception Modeling and Optimization

5.1 Introduction

In the last two chapters we propose several schemes in order to reduce the uplink channel access latency for machine type communications in LTE. In this chapter, we introduce methods to improve the performance for MTC downlink reception. It is well known that most MTC devices are battery powered, for example sensors in oil pipeline, smart water meter, etc. Therefore, lowering the power consumption, which prolongs the MTC device's life time and hence reduce the deployment cost, is among the primary requirements. To achieve this, discontinuous reception (DRX) is employed in LTE/LTE-A network. With DRX, a UE only turns on the receiver at some pre-defined time points while sleeps at others. If a packet arrives at eNB but the target UE is sleeping, the packet is buffered at eNB and will be delivered to that UE when it wakes up. Therefore, it can be seen that the DRX mechanism attains power savings at the expense of an extra delay. It is preferred that the DRX parameters are selected such that the power saving is maximized while the application delay constraint is satisfied. However, the optimal trade-off between the power saving factor and wakeup delay is unknown.

Authors in [78,79] present analytical methods to model the DRX mechanism in UMTS. However, LTE introduces two types of DRX cycles which is different form the single DRX cycle in UMTS. Hence, the models used in UMTS are not applicable to the LTE case. References [80, 81] provide methods to model the LTE DRX mechanism in the presence of bursty and Poisson traffic, respectively. However, they do not take into account the ON duration, which is part of every short and long DRX cycle. They assume that a packet (always) arrives during the sleep period and has to be delayed and buffered. In practice, a packet may arrive during the ON part of a cycle and be sent by the eNB (base station) right away. This is not accounted for in the aforementioned models, leading to inaccurate estimates for the power-saving factor and average latency. Reference [82] proposes a DRX power consumption model for MTC service with deterministic packet interval. Reference [83] provides a single threshold adaptive configuration DRX mechanism to save more power while maintaining the throughput. Reference [84] presents a scheduling method for delay sensitive service in LTE. With this method, the packet loss rate caused by sleeping process during DRX can be reduced. Authors in [85] investigate the energy-saving provided DRX and its impact on the QoS performance of VoIP traffic by simulations. Reference [86] introduces the light sleeping mode to improve the performance of DRX. The idea of light sleeping mode is to turn off the power amplifier but leave the other components on such that power is saving while fast wakeup is enabled.

In this chapter, we present two methods to analyze the detailed DRX mechanism in LTE/LTE-A. In the first method, we assume that the traffic is Poisson distributed. With this assumption, a semi-Markov chain model is proposed to analyze the DRX mechanism. We do model the On duration parameter, which in LTE/LTE-A takes values between 1 and 200ms [12], by using two type of states to differentiate the On duration from the sleep period of short or long DRX cycle and show that it has a significant impact on the DRX performance. With this model, one can calculate the power saving factor and latency for a given DRX parameter set, which can be used to select the suitable DRX parameter. We use simulation to validate our results.

Different from the first method which requires the traffic is Poisson distributed, the second method is applicable to all kinds of ***sporadic*** traffic (Poisson distributed, uniform distributed, etc.). With the second method, we also provide a simple method to find the optimal DRX parameter which maximizes the power saving factor while maintaining the latency requirement.

The reminder of this chapter organized as follows. Section 5.2 gives a brief introduction to the DRX mechanism in LTE. Section 5.3 presents the a DRX modeling method for Poisson distributed traffic. Section 5.4 provides a method to analyze the DRX mechanism for sporadic traffic and Section 5.5 concludes this chapter.

5.2 DRX Mechanism in LTE

The mechanism of DRX is specified in [60] and shown in Fig. 5.1 [9]. When DRX is enabled, the UE wakes up and checks for the downlink scheduling information during the subframes referred to as the on duration (the period of on duration is denoted as T_{ON}), which is located at the beginning of short/long DRX cycles. If not scheduled, the UE goes back to sleep for the purpose of power saving. Otherwise, it starts an inactivity timer T_0 and enters the continuous reception mode to check the scheduling information at every subframe. The Inactivity Timer will be restarted if the UE is scheduled before the expiry of the timer. Otherwise, the UE starts a short DRX Cycle T_S. If the UE is not scheduled after several short DRX, which is specified by the DRX short cycle timer N, the UE starts the Long DRX cycle T_L to save more power (T_L is a multiple of T_S as specified in [12]).

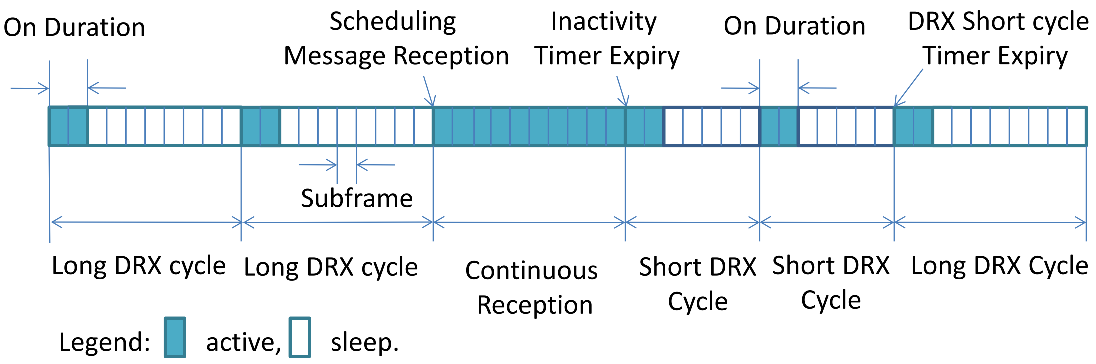


Figure 5.1: DRX procedure in LTE [9]

5.3 DRX Modeling for Poisson Distributed Traffic

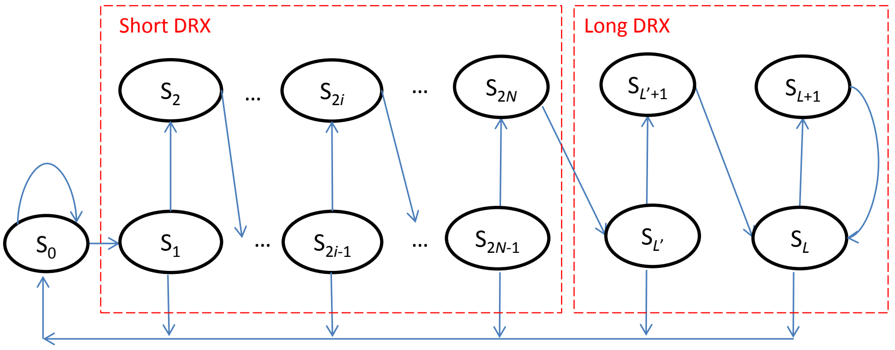
5.3.1 Semi-Markov Chain Model for DRX in LTE

Here we assume that the traffic is Poisson distributed. The DRX mechanism can be regarded as a Semi-Markov chain model as shown in Fig. 5.2. The states of this Semi-Markov process model are defined as following:

- State S_0 means that the UE is in continuous reception,
- State S_{2i-1} means that the UE is in the active period of the ith short DRX $i \in [1, N]$,
- State S_{2i} means that the UE is in the sleep period of the ith short DRX $i \in [1, N]$,
- State $S_{L'}$ means that the UE is in the active period of the first long DRX,
- State $S_{L'+1}$ means the UE is in the sleep period of the first long DRX,
- State S_L means that the UE is in the active period of other long DRX,
- State S_{L+1} means the UE is in the sleep period of other long DRX.

The transitions between states are:

1. When the UE is at state S_0, if it is not scheduled before the expiry of the *Inactivity Timer*, the UE transfers to state S_1; otherwise it restarts the *Inactivity Timer* and remains at state S_0.
2. When the UE is at state S_{2i-1}, $i \in [1, N]$, if it is not scheduled before the expiry of the *On Duration*, the UE transfers to state S_{2i} and starts sleep; otherwise, the UE transfers to S_0.
3. When the UE is at state S_{2i}, $i \in [1, N-1]$, after sleeping for a period of $T_S - T_{ON}$ it wakes up and transfers to state S_{2i+1}.



State	Description
S_0	Continuous reception mode
S_{2i-1}	Active period of the ith short DRX $i \in [1, N]$
S_{2i}	Sleep period of the ith short DRX $i \in [1, N]$
$S_{L'}$	Active period of the first long DRX
$S_{L'+1}$	Sleep period of the first long DRX
S_L	Active period of other long DRX
S_{L+1}	Sleep period of other long DRX

Figure 5.2: Semi-Markov Chain model for DRX

4. When the UE is at state S_{2N}, after sleeping for a period of $T_S - T_{ON}$ it transfers to state $S_{L'}$ to start the first long DRX cycle.

5. When the UE is at state $S_{L'}$ if it is not scheduled before the expiry of the *On Duration*, it transfers to state $S_{L'+1}$ and starts sleep; otherwise, the UE transfers to S_0.

6. When the UE is at state $S_{L'+1}$, after sleeping period of $T_L - T_{ON}$ it wakes up and transfers to state S_L.

7. When the UE is at state S_L if it is not scheduled before the expiry of the On *Duration*, it transfers to state S_{L+1} and starts sleep; otherwise, the UE transfers to S_0.

8. When the UE is at state S_{L+1}, after sleeping period of $T_L - T_{ON}$ it wakes up and transfers to state S_L.

The sleeping period is $T_S - T_{ON}$ before entering into state $S_{L'}$ while it is $T_L - T_{ON}$ when entering into state S_L, which is the reason why we use two states ($S_{L'}$ and S_L) to differentiate the first long DRX cycle from other DRX cycles.

For this Semi-Markov chain model, we start with the calculation for the stationary probability for each state and then derive the states' holding times. Denoting $p_{i,j}$ as the transition probability from state S_i to state S_j, the stationary probability of state i, π_i can be calculated as:

$$\begin{cases} \pi_i = \pi_{i-1} p_{i-1,i}, \; i \in [1, 2N] \\ \pi_{L'} = \pi_{2N} p_{2N,L'}, \\ \pi_{L'+1} = \pi_{L'} p_{L',L'+1}, \\ \pi_L = \pi_{L'+1} p_{L'+1,L} + \pi_{L+1} p_{L+1,L}, \\ \pi_{L+1} = \pi_L p_{L,L+1}. \end{cases} \tag{5.1}$$

With the above equations, we can get:

$$\pi_i = \pi_0 \prod_{j=1}^{i} p_{j-1,j}, i \in [1, 2N] \tag{5.2}$$

$$\pi_{L'} = \pi_0 p_{2N,L'} \prod_{j=1}^{2N} p_{j-1,j}, \tag{5.3}$$

$$\pi_{L'+1} = \pi_0 p_{L',L'+1} p_{2N,L'} \prod_{j=1}^{2N} p_{j-1,j}, \tag{5.4}$$

$$\pi_L = \pi_0 \frac{p_{L'+1,L} p_{L',L'+1} p_{2N,L'}}{1 - p_{L,L+1} p_{L+1,L}} \prod_{j=1}^{2N} p_{j-1,j}, \tag{5.5}$$

$$\pi_{L+1} = \pi_0 \frac{p_{L,L+1} p_{L'+1,L} p_{L',L'+1} p_{2N,L'}}{1 - p_{L,L+1} p_{L+1,L}} \prod_{j=1}^{2N} p_{j-1,j}. \tag{5.6}$$

The state transition probability for this model is calculated as following. Here we assume that packet arrival rate of the Poisson distributed traffic is λ, therefore the packet interval time T' follows an exponential distribution with expected value $1/\lambda$. There are totally eight types of state transition as described above. We start with the calculation for the first case. Recall the Markov chain model described above we can see that after receiving a packet the UE is at state S_0 for the period of T_0 at most. Assuming the buffered packets in a short/long DRX cycle for one UE can be delivered by one transmission in the subsequent on duration[1], the transition from state S_0 to S_1 is only triggered by the event that another packet does not arrive before the expiry of the Inactivity Timer. Hence the state transition probability $p_{0,1} = p(T' > T_0) = e^{-\lambda T_0}$.

Similarly,

$$p_{1,2} = e^{-\lambda T_{ON}}. \tag{5.7}$$

[1] this assumption is realistic for MTC applications as most MTC traffic is *uplink dominated*

When the UE is at state S_{2i}, $i \in [1, N-1]$, it transfers to state S_{2i+1} with probability 1, i.e. $p_{2i,2i+1} = 1$.

Similarly,

$$p_{2N,L'} = 1. \tag{5.8}$$

$$p_{L'+1,L} = 1. \tag{5.9}$$

$$p_{L+1,L} = 1. \tag{5.10}$$

When the UE is at state S_{2i+1}, $i \in [1, N-1]$, if it receives a packet which arrived at eNB during the state S_{2i} and On duration, it transfers to state S_0; otherwise it transfers to stats S_{2i+2}. Therefore,

$$p_{2i+1,2i+2} = p(T' > T_S) = e^{-\lambda T_s}, i \in [1, N-1]. \tag{5.11}$$

Similarly,

$$p_{L^prime,L'+1} = e^{-\lambda T_s}. \tag{5.12}$$

$$p_{L,L+1} = e^{-\lambda T_L}. \tag{5.13}$$

Now let us calculate the holding time H_i of the semi Markov process at state S_i $(i = 0, 1, ..., 2N, L', L'+1, L, L+1)$.

$E(H_0)$. When UE is at state S_0, the packet arrives after the expiry of the Inactivity timer with probability p_{01} or it arrives at the ith subframe of the Inactivity timer with probability p_i. Therefore,

$$H_0 = p_{0,1}T_0 + \sum_{i=1}^{T_0} T_i p_i, \tag{5.14}$$

where T_i is the state holding time when the packet arrives at the ith subframe.

When the UE is at state S_0, the probability that the packet arrives at the ith subframe of the Inactivity timer is

$$p_i = p(i-1 < T' < i) = e^{-(i-1)\lambda} - e^{-i\lambda}, i \in [1, T_0]. \tag{5.15}$$

If a packet arrives at the ith subframe of the Inactivity timer, a new continuous reception is started. Hence

$$T_i = i + H_0. \tag{5.16}$$

Substituting (5.16) in (5.14), we can get

$$E(H_0) = T_0 + \sum_{i=1}^{T_0} ip_i/p_{0,1} = \frac{1 - e^{-\lambda T_0}}{(1 - e^{-\lambda})e^{-\lambda T_0}}. \tag{5.17}$$

$E(H_{2i-1})$, $i \in [2, N]$. When UE is at state S_{2i-1}, there are three cases for packet arrival: (i) the packet arrives after the expiry of the On duration with probability $p_{2i-1,2i}$, (ii) the packet arrives at the jth subframe of the On duration with probability p_j^{on}, (iii) the packet arrived during the last sleep period (sleep period of the $(i-1)$th short DRX cycle) with probability p_s.

Hence,

$$H_{2i-1} = p_{2i-1,2i}T_{ON} + \sum_{j=1}^{T_{ON}} T_j^{ON} p_j^{ON} + T_s p_s. \tag{5.18}$$

where T_j^{ON} is the state holding time when the packet arrives at the jth subframe of the On duration and T_s is the state holding time when the packet arrived during the last sleep period.

When the UE is at state S_{2i-1}, the probability that the packet arrived during the sleep period of the $(i-1)$th short DRX cycle is

$$\begin{aligned} p_s &= p(T' < T_s - T_{ON}) \\ &= 1 - e^{-(T_s - T_{ON})\lambda} \end{aligned} \tag{5.19}$$

and the probability that the packet arrives at the jth subframe of the On duration

$$\begin{aligned} p_j^{ON} &= p(T_s - T_{ON} + j - 1 < T\prime < T_s - T_{ON} + j) \\ &= e^{-(T_s - T_{ON} + j - 1)\lambda} - e^{-(T_s - T_{ON} + j)\lambda}, j \in [1, T_{ON}]. \end{aligned} \tag{5.20}$$

If a packet arrived during the sleeping period of the $(i-1)$th short DRX, it is delivered at the first subframe of the next On duration. Hence, the state holding state for this case is

$$T_s = 1. \tag{5.21}$$

Moreover, when a packet arrives at jth subframe of the On duration the state holding time is

$$T_j^{ON} = j, j \in [1, T_{ON}]. \tag{5.22}$$

Substituting equations (5.19)-(5.22) in equ(5.18), we can get

$$\begin{aligned} H_{2i-1} &= p_{2i-1,2i}T_{ON} + \sum_{j=1}^{T_{ON}} jp_j^{ON} + (1 - e^{-(T_s - T_{ON})\lambda}) \\ &= \frac{e^{-\lambda(T_s - T_{ON})} - e^{-\lambda T_s}}{1 - e^{-\lambda}} + 1 - e^{-\lambda(T_s - T_{ON})}. \end{aligned} \tag{5.23}$$

$E(H_1)$. When the UE is at state S_1, there are two cases for packet arrival: (i) the packet arrives after the expiry of the On duration with probability $p_{1,2}$, (ii) the packet arrives at the jth subframe of the On duration with probability p_j^1. Therefore

$$H_1 = p_{1,2}T_{ON} + \sum_{j=1}^{T_{ON}} T_j^1 p_j^1, \tag{5.24}$$

where $T_j^1 = j$ is the state holding time when the packet arrives at the jth subframe of the On duration.

When the UE is at state S_1, the probability that the packet arrives at the ith subframe of the On duration is

$$\begin{aligned} p_j^1 &= p(i-1 < T' < i) \\ &= e^{-(i-1)\lambda} - e^{-i\lambda}, i \in [1, T_{ON}]. \end{aligned} \tag{5.25}$$

Therefore, (5.24) can be rewritten as $H_1 = \frac{1-e^{-\lambda T_{ON}}}{1-e^{-\lambda}}$.

$E(H_{L'})$. When UE is at state $H_{L'}$, there are three cases for packet arrival: (i) the packet arrives after the expiry of On Duration with probability $p_{L'(L'+1)}$; (ii) the packet arrives at the jth subframe of On duration with probability $p_j^{L'\_ON}$; (iii) the packet arrived during the last sleep period (sleep period of the Nth short DRX cycle) with probability $p_{L'\_S}$, $i \in [2, N]$. Therefore,

$$H_{L'} = p_{L'(L'+1)}T_{ON} + \sum_{j=1}^{T_{ON}} T_j^{L'\_ON} p_j^{L'\_ON} + T_{L'\_S}p_{L'\_S}, \tag{5.26}$$

where $T_j^{L'\_ON}$ is the state holding time when the packet arrives at the jth subframe of the On duration and $T_{L'\_S}$ is the state holding time when the packet arrived at the last sleep period.

When UE is at state H'_L, the probability that the packet arrived during sleep period of the last short DRX cycle is

$$\begin{aligned} p_{L'\_S} &= p(T' < T_S - T_{ON}) \\ &= 1 - e^{-(T_S - T_{ON})\lambda} \end{aligned} \tag{5.27}$$

and the probability that the packet arrives at the jth subframe of the On duration

$$\begin{aligned} p_j^{L'\_ON} &= p(T_s - T_{ON} + j - 1 < T' < T_s - T_{ON} + j) \\ &= e^{-(T_s - T_{ON} + j - 1)\lambda} - e^{-(T_s - T_{ON} + j)\lambda}, j \in [1, T_{ON}]. \end{aligned} \tag{5.28}$$

Similar to the case of $E(H_{2i-1})$, $T_{L'\_S} = 1$ and $T_j^{L'\_ON} = j$. With the above results, we have

$$H_{L'} = \frac{e^{-\lambda(T_S - T_{ON})} - e^{-\lambda T_S}}{1 - e^{-\lambda}} + 1 - e^{-\lambda(T_S - T_{ON})}. \tag{5.29}$$

$E(H_L)$. When UE is at state S_L, there are also three cases for packet arrival: (i) the packet arrives after the expiry of On duration with probability $p_{L,L+1}$; (ii) the packet arrives at the jth subframe of On duration with probability $p_j^{L\_ON}$; (iii) the packet arrived during the last sleep period with probability p_{L_S}. Hence,

$$H_L = p_{L,L+1}T_{ON} + \sum_{j=1}^{T_{ON}} T_j^{L\_ON} p_j^{L\_ON} + T_{L\_S}p_{L\_S}, \tag{5.30}$$

where $T_j^{L\_ON}$ is the state holding time when the packet arrives at the jth subframe of the On duration and T_{L_S} is the state holding time when the packet arrived at the last sleep period.

When UE is at state H_L, the probability that the packet arrived during sleep period of the last long DRX cycle is

$$\begin{aligned} p_{L\_S} &= p(T' < T_L - T_{ON}) \\ &= 1 - e^{-(T_L - T_{ON})\lambda} \end{aligned} \tag{5.31}$$

and the probability that the packet arrives at the jth subframe of the On duration is

$$\begin{aligned} p_j^{L\_ON} &= p(T_L - T_{ON} + j - 1 < T' < T_L - T_{ON} + j) \\ &= e^{-(T_L - T_{ON} + j - 1)\lambda} - e^{-(T_L - T_{ON} + j)\lambda}, j \in [1, T_{ON}] \end{aligned} \tag{5.32}$$

Similar to the case of $E(H_{2i-1})$, $T_{L\_S} = 1$ and $T_j^{L\_ON} = j$. Therefore (5.30) can be rewritten as

$$\begin{aligned} H_L &= p_{L,L+1}T_{ON} + \sum_{j=1}^{T_{ON}} j p_j^{L\_ON} + (1 - e^{-(T_L - T_{ON})\lambda}) \\ &= \frac{e^{-\lambda(T_L - T_{ON})} - e^{-\lambda T_L}}{1 - e^{-\lambda}} + 1 - e^{-\lambda(T_L - T_{ON})} \end{aligned} \tag{5.33}$$

The state holding time for other state is easy to get: $H_{2i} = T_s - T_{ON}, i \in [1, N]$, $H_{L'+1} = T_L - T_{ON}$, and $H_{L+1} = T_L - T_{ON}$.

5.3.2 Power Saving Factor and Wake Up Delay

Based on the results derived in the last section, the proportion of time that the UE in the sleep period of short DRX can be calculated as

$$\begin{aligned} p_{sd} &= \frac{\sum_{i=1}^{N} \pi_{2i} H_{2i}}{\sum_{i=0}^{2N} \pi_i H_i + \sum_{i=0}^{1} \pi_{L'+i} H_{L'+i} + \sum_{i=0}^{1} \pi_{L+i} H_{L+i}} \\ &= \frac{e^{-\lambda(T_0 + T_{ON})} \frac{1 - e^{-\lambda N T_s}}{1 - e^{-\lambda T_s}} (T_s - T_{ON})}{T} \end{aligned} \tag{5.34}$$

where

$$
\begin{aligned}
T = {} & \frac{1-e^{-\lambda T_0}}{(1-e^{-\lambda})e^{-\lambda T_0}} + e^{-\lambda T_0}\frac{1-e^{-\lambda T_{ON}}}{1-e^{-\lambda}} + e^{-\lambda(T_0+T_{ON})}\frac{1-e^{-\lambda N T_S}}{1-e^{-\lambda T_S}}(T_S - T_{ON}) \\
& + (e^{-\lambda(T_0+T_{ON})}\frac{1-e^{-\lambda(N-1)T_S}}{1-e^{-\lambda T_S}} + e^{-\lambda(T_0+T_{ON}+(N-1)T_S)})(\frac{e^{-\lambda(T_s-T_{ON})}-e^{-\lambda T_s}}{1-e^{-\lambda}} + 1 - e^{-\lambda(T_s-T_{ON})}) \\
& + (e^{-\lambda(T_0+T_{ON}+NT_S)} + \frac{e^{-\lambda(T_0+T_{ON}+NT_S+T_L)}}{1-e^{-\lambda T_L}})(T_L - T_{ON}) \\
& + \frac{e^{-\lambda(T_0+T_{ON}+NT_S)}}{1-e^{-\lambda T_L}}(\frac{e^{-\lambda(T_L-T_{ON})}-e^{-\lambda T_L}}{1-e^{-\lambda}} + 1 - e^{-\lambda(T_L-T_{ON})}).
\end{aligned}
\tag{5.35}
$$

Similarly, the proportion of time that the UE in the sleep period of long DRX is

$$
\begin{aligned}
p_{Ld} &= \frac{\pi_{L'+1}H_{L'+1} + \pi_{L+1}H_{L+1}}{\sum_{i=0}^{2N}\pi_i H_i + \sum_{i=0}^{1}\pi_{L'+i}H_{L'+i} + \sum_{i=0}^{1}\pi_{L+i}H_{L+i}}. \\
&= \frac{(e^{-\lambda(T_0+T_{ON}+NT_s)} + \frac{e^{-\lambda(T_0+T_{ON}+NT_s+T_L)}}{1-e^{-\lambda T_L}})(T_L - T_{ON})}{T}.
\end{aligned}
\tag{5.36}
$$

Therefore the power saving factor, which is defined as ratio of time that UE is at the power saving states to the time that the UE is at all states, is

$$\alpha = p_{sd} + p_{Ld}. \tag{5.37}$$

It has to be noted that: here we use a simple method to measure effect of power saving. A more practically way would be: (1) we should measure the realistic power consumption value in the power saving state and active state; (2) we should consider the energy consumed for transition between active state and power saving state; (3) we should also consider the proportional of power consumed by modem compared to the whole embeded system (if it is a small proportion, there is no need for us to optimize the DRX procedure).

According to the property of Poisson distribution, the packet arriving time over short or long DRX follows uniform distribution. Hence, the wake up delay, which is the interval between the time when a packet arrives at the eNB and the time when the packet is delivered by eNB, caused by short DRX and long DRX is $d_s = (T_s - T_{ON})/2$ and $d_L = (T_L - T_{ON})/2$ respectively.

Finally, the overall wake up latency is

$$d = p_{sd}d_s + p_{Ld}d_L. \tag{5.38}$$

As $d_S \leq d_L$, therefore $\alpha d_S \leq d \leq \alpha d_L$, i.e. the wake up latency d is upper bounded by αd_L and lower bounded by αd_S. Moreover, from (5.37) and (5.38) we can see that the power saving factor and wake up latency tradeoff is affected by T_{ON}, which is different from the results in [80], [81].

5.3.3 Simulation Validation

To validate the proposed method, simulations are carried out. The DRX mechanism is developed following the protocols in [60] with a MATLAB based system level simulator and the

DRX parameters used in section is realistic as this specified in [12]. Firstly, we compare the simulated results with analytical results obtained using our method under different packet arrival rate λ. The DRX parameters are listed in Table 5.1. From Fig.5.3, we can see that the analytical results are very close to the simulated ones, which validates our method.

Then, we carried simulations to see the effect of different DRX parameter on the DRX performance. It should be noted that the packet arrival rate is $\lambda = 1/30$ or $1/3$ packet/second to represent the sporadic (average packet interval $1/\lambda \gg T_S$, T_L) and non-sporadic traffic, respectively.

Table 5.1: Simulation parameters for results in Fig.5.3-5.9

Figure index	Parameters
5.3	$T_0 = 20ms$, $T_{ON} = 80ms$, $T_S = 160ms$, $N = 16$, $T_L = 320ms$
5.4	$T_{ON} = 80ms, T_S = 160ms, N = 16, T_L = 320ms$
5.5	$T_0 = 20ms, T_S = 160ms, N = 16, T_L = 320ms$
5.6	$T_0 = 20ms, T_{ON} = 80ms, N = 16, T_L = 2T_Sms$
5.7	$T_0 = 20ms, T_{ON} = 30ms, N = 16, T_L = 320ms$
5.8	$T_0 = 20ms, T_{ON} = 80ms, T_S = 160ms, T_L = 320ms$
5.9	$T_0 = 20ms, T_{ON} = 80ms, T_S = 160ms, N = 16$

1. *Effect of inactivity timer*. Fig.5.4 demonstrates the DRX performance under different inactivity timer. We find that the simulated results match the analytical results, which also validates our model. Moreover, for the sporadic traffic we notice that the latency decreases slightly from 84ms to 77ms and power saving factor changes from 0.73 to when 0.67 when the inactivity timer increases from 20 to 2560. The reason for this phenomenon is that for the sporadic traffic a UE is at the sleep state most of the time. Therefore, the effect of inactivity time is weak. In contrast to that, for the non-sporadic traffic the latency and power saving factor changes heavily with the inactivity timer. The reason is that a UE stays more time at the active state when inactivity timer increases. As a result, the latency and power saving factor is greatly reduced. From the simulation results, we can conclude that the effect of inactivity timer on the DRX performance is weak for sporadic traffic, while it is strong for non-sporadic traffic.

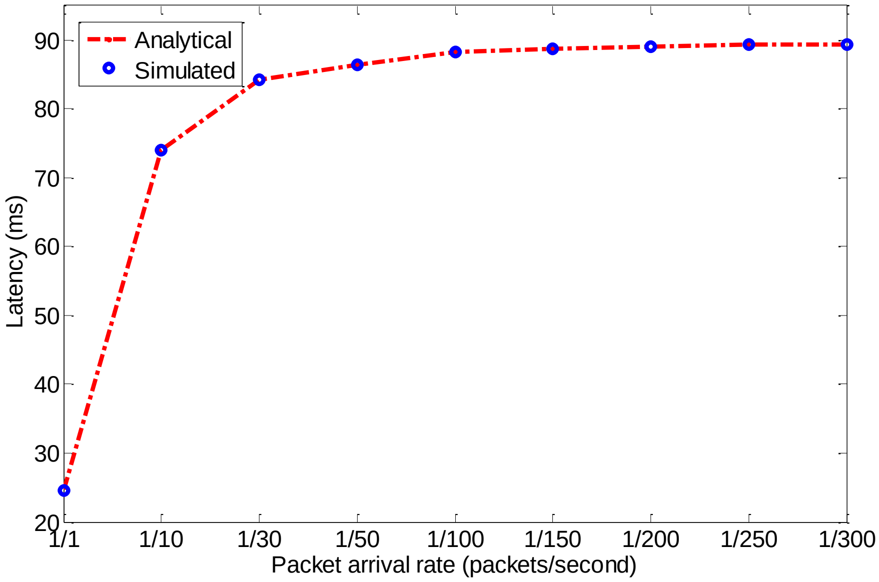
2. *Effect of on duration*. Fig. 5.5 shows the DRX performance under different on duration. We can see that the simulated results match the analytical results, which validates our model. Furthermore, for both the sporadic and non-sporadic traffic we find that the latency decreases greatly. Therefore, it is be conclude that on duration has strong effect on DRX performance regardless of the type of traffic. The reason for this phenomenon is that on duration determines the proportion of time that a UE is at active state. A UE spends more time at active state when the on duration increases, which decreases the power saving factor and latency.

3. *Effect of short DRX cycle*. As specified in [12], the long DRX cycle is a multiple of short DRX cycle. In order to comply with the constraint, here we set T_L as $T_L = 2T_S$. Fig. 5.6 shows that the latency increases and the power saving changes drastically when the short DRX cycle increases from 128 to 640 for both sporadic and non-sporadic traffic. Therefore, short DRX cycle seems to have a strong effect on the DRX performance.

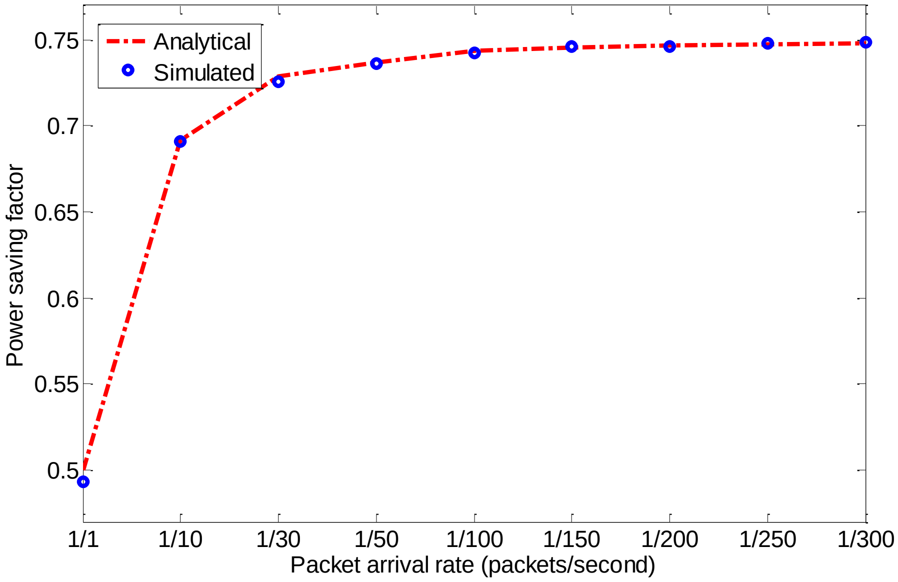
However, the long DRX cycle also varies in this simulation since $T_L = 2T_S$, which also affects the DRX performance. Therefore, the effect of short DRX cycle on the DRX performance is unclear. To clarify this issue, we carried another simulation. In this simulation, T_L is fixed as 320 and T_S varies from 40 to 320, which still complies with the constraint that the long DRX cycle is a multiple of short DRX cycle. In Fig. 5.7, we can find that for the sporadic traffic the latency decreases and power saving factor slightly changes when the short DRX cycle increases from 10 ms to 160 ms, while for the non-sporadic traffic the DRX performance varies significantly. This reason is that for the sporadic traffic a UE is at the long sleep state most of the time. Therefore, the effect of short DRX cycle is weak in this case. While for the non-sporadic traffic, a UE spends considerable amount of time in the short sleep state. As a result, the short DRX cycle has a strong effect on the DRX performance for the non-sporadic traffic. In addition to that, we also find that the latency does not always decrease when the short DRX cycle increases. The reason for this phenomenon is: (1) when $T_S < T_L$, increasing the short DRX cycle makes the UE spends more time at the short sleep state and hence reduce the latency; (2) when $T_S = T_L$, there is only long sleep period for a UE. Therefore, the latency is obviously increased.

4. *Effect of DRX short cycle timer*. Fig.5.8 demonstrates the DRX performance with different DRX short cycle timer. We can find that the simulated results match the analytical results, which also verifies our method. Furthermore, for the sporadic traffic the results show that the latency decreases from 89.5 ms to 84 ms and the power saving factor decreases from 0.748 to 0.729 when the DRX short cycle timer increases from 1 to 16, which indicates that DRX short timer has a weak effect on the DRX performance for the sporadic traffic. The reason is that for the sporadic traffic the time of overall short DRX cycle ($N * T_S$) is much smaller than the average packet interval, which makes a UE is at long sleep state most of the time. Therefore the effect of DRX short cycle timer is weak. Different with that, for the non-sporadic traffic the DRX performance greatly changes with the DRX short cycle timer. This is because of when the DRX short cycle timer increases, a UE stays more time in the short sleep state. Hence, the power saving factor and latency is reduced as the short DRX cycle is smaller than long DRX cycle.

5. *Effect of long DRX cycle*. Fig.5.9 shows the DRX performance under different long DRX cycle. It is shown that the stimulated results match the analytical results, which validates our method. We also notice that the latency and power saving factor varies dramatically when the long DRX cycle changes for both sporadic and non-sporadic traffic. Take the result of sporadic traffic as an example, the latency increases from 19 ms to 1106 ms when the long DRX cycle varies from 160 to 2560, while the power saving factor increases from 0.49 to 0.93. Therefore, it is evident that the long DRX cycle has a strong effect on the DRX performance. The reason for this phenomenon is that a UE usually spends large proportion of time at the long sleep state. As a result, when the long DRX cycle increases, a UE spends more time at sleep state, which increases the power saving factor and latency.

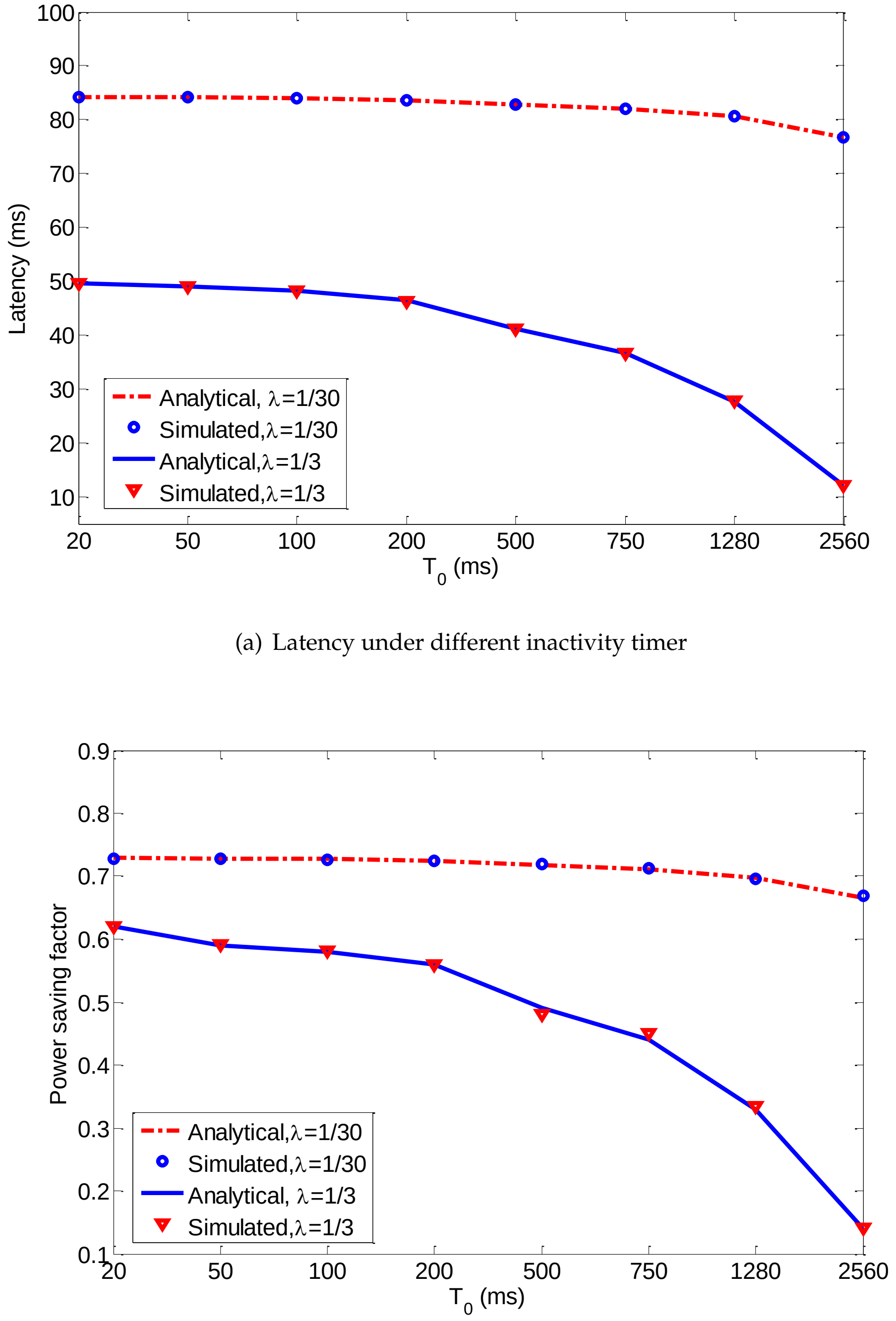


(a) Latency under different packet arrival rate



(b) Power saving factor under different Inactivity timer

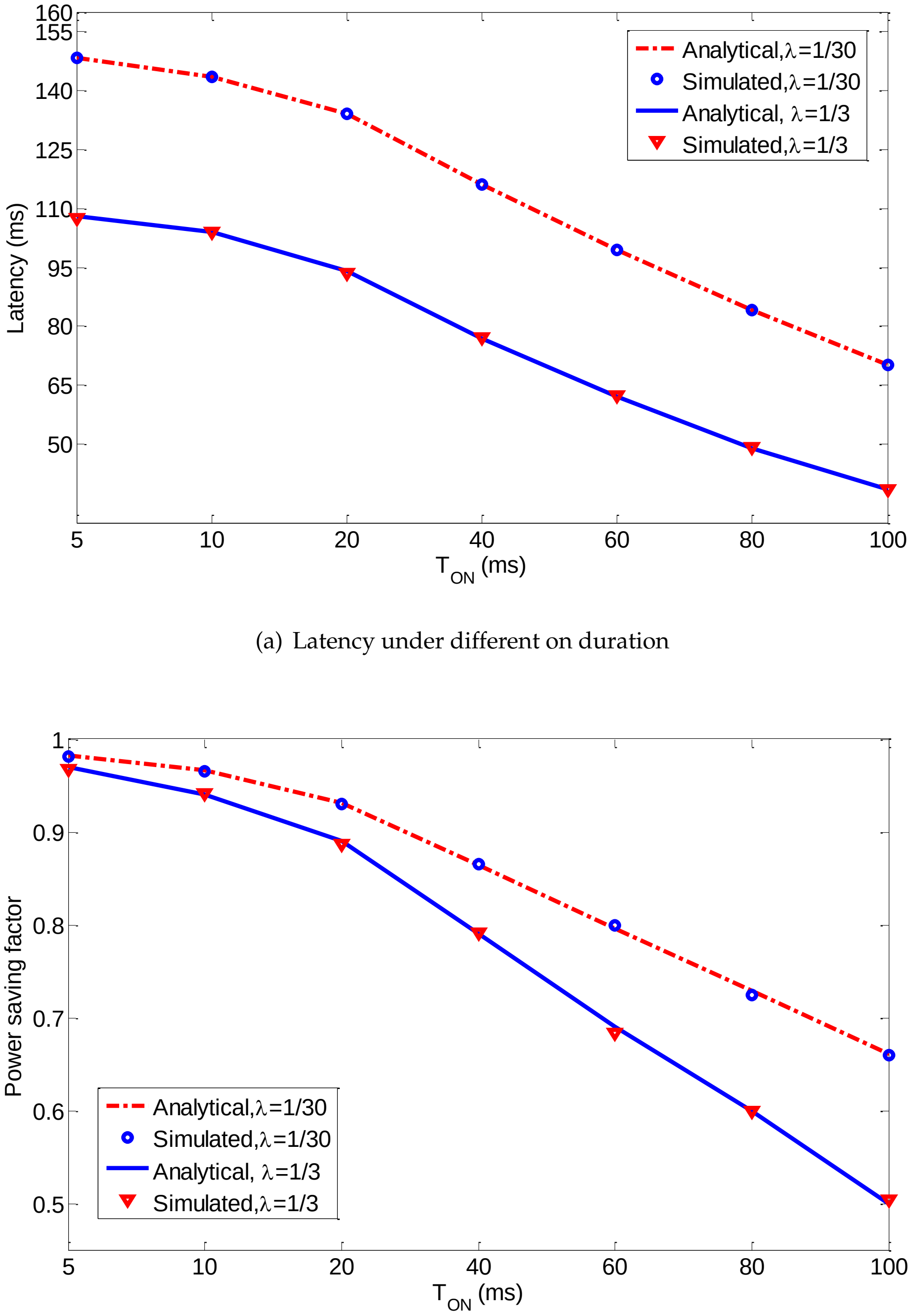
Figure 5.3: DRX performance under different packet arrival rate



(a) Latency under different inactivity timer

(b) Power saving factor under different Inactivity timer

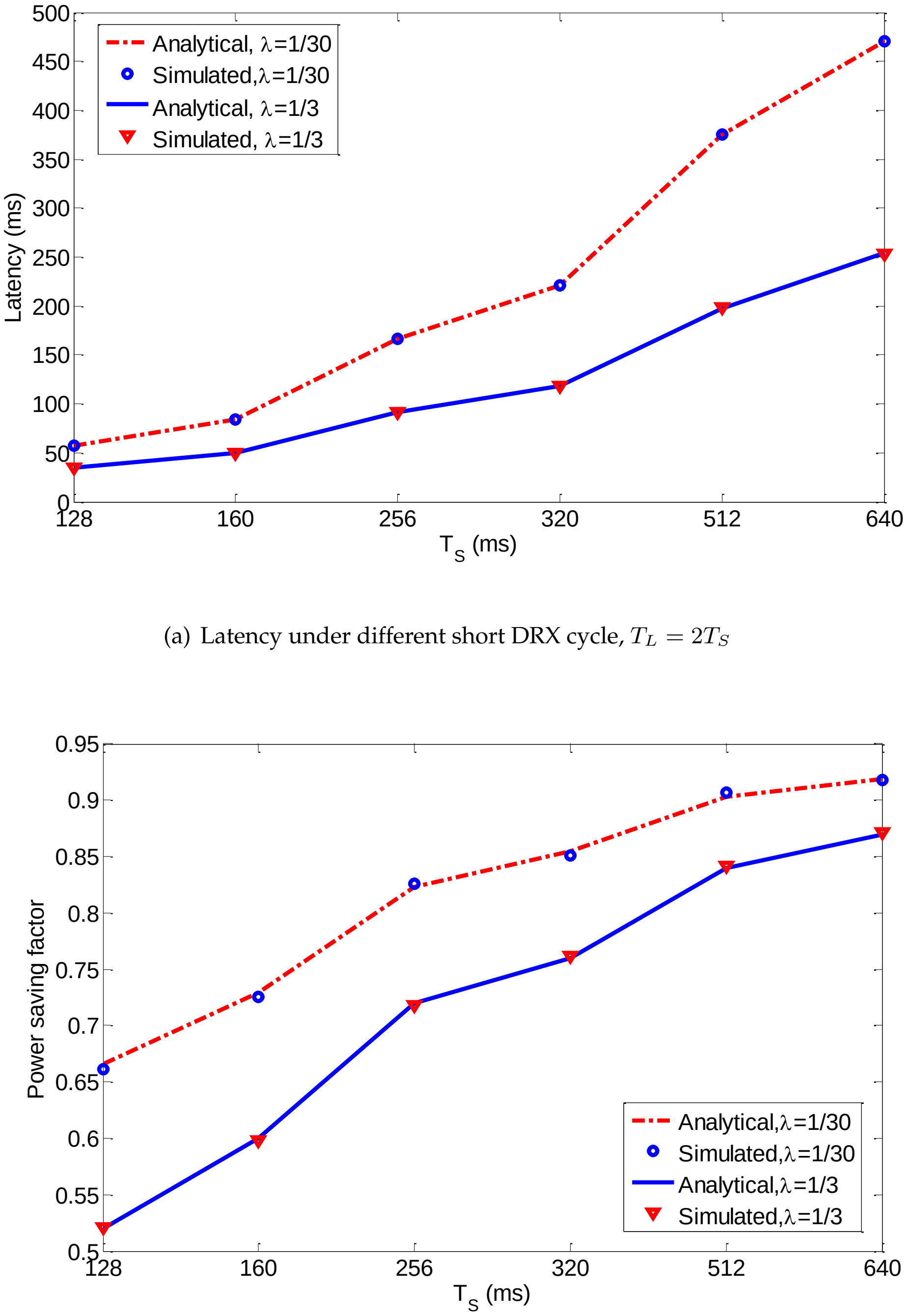
Figure 5.4: DRX performance under different inactivity timer



(a) Latency under different on duration

(b) Power saving factor under different on duration

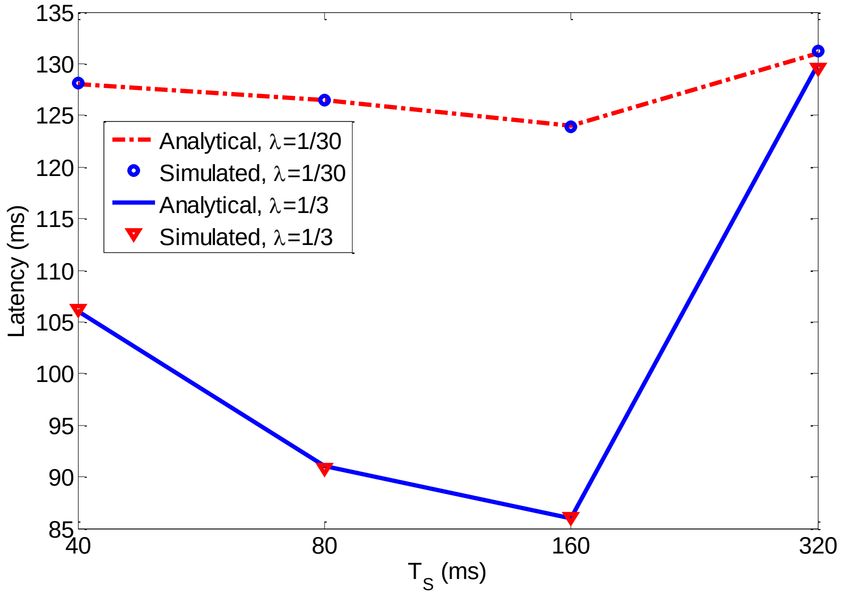
Figure 5.5: DRX performance under different on duration



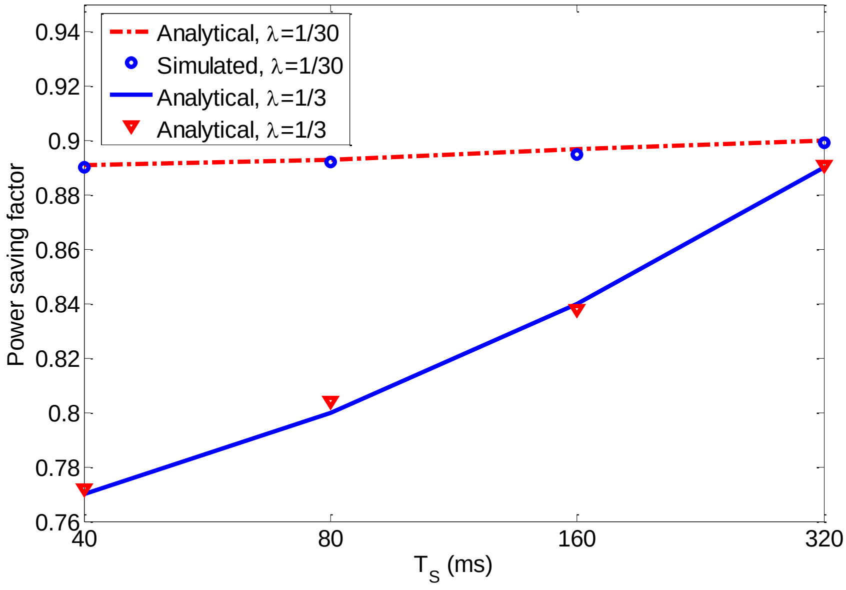
(a) Latency under different short DRX cycle, $T_L = 2T_S$

(b) Power saving factor under different short DRX cycle, $T_L = 2T_S$

Figure 5.6: DRX performance under different short DRX cycle, $T_L = 2T_S$

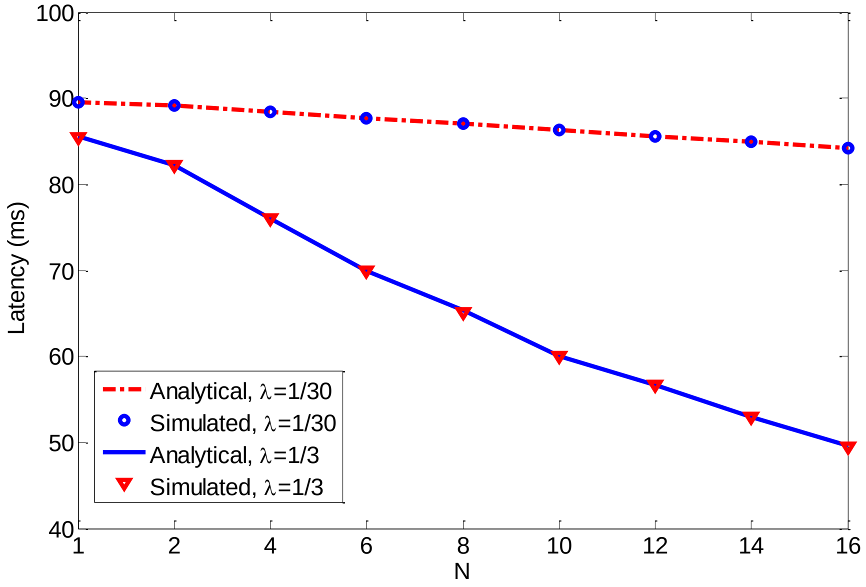


(a) Latency under different short DRX cycle

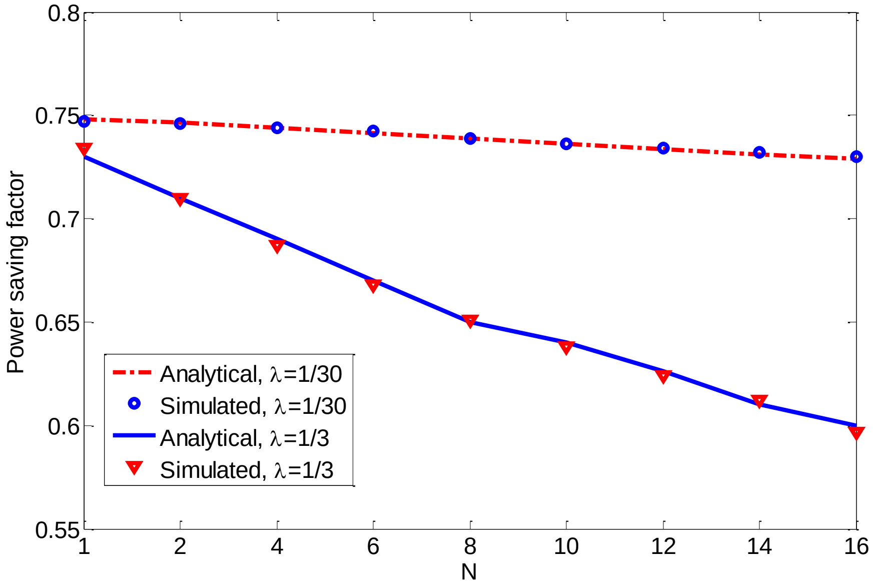


(b) Power saving factor under different short DRX cycle

Figure 5.7: DRX performance under different short DRX cycle

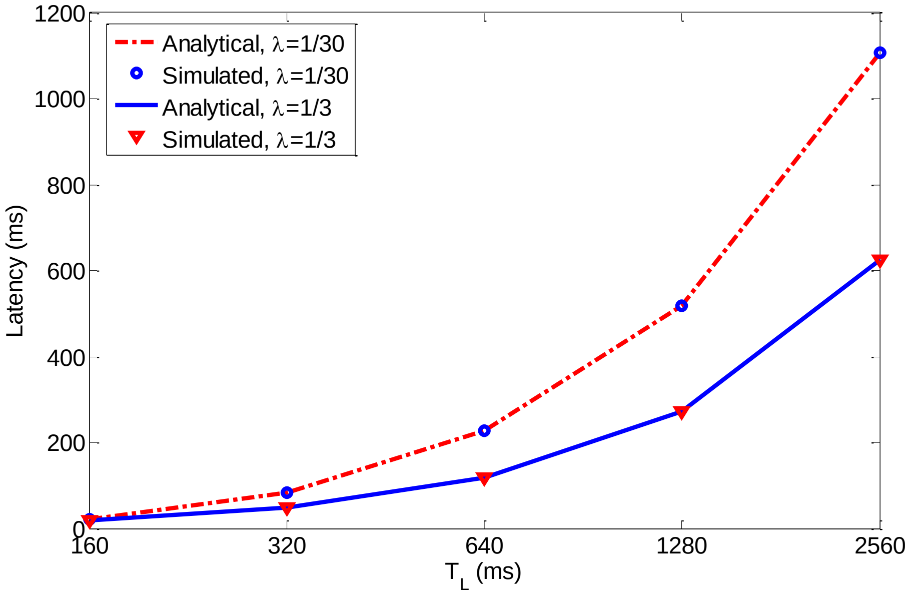


(a) Latency under different DRX short cycle timer

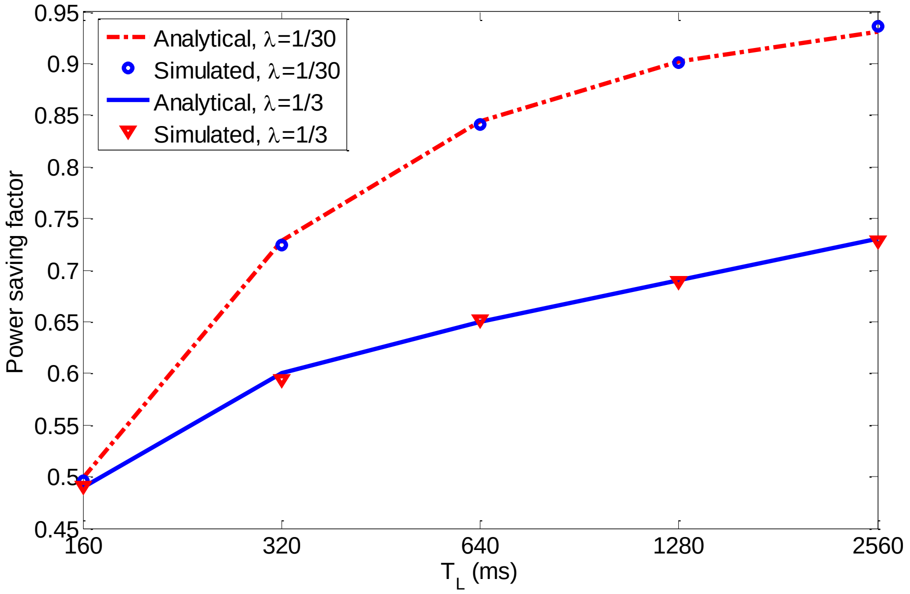


(b) Power saving factor under different short DRX cycle timer

Figure 5.8: DRX performance under different DRX short cycle timer



(a) Latency under different long DRX cycle



(b) Power saving factor under different long DRX cycle

Figure 5.9: DRX performance under different long DRX cycle

Through the above results, our proposed model is evidently validated. Moreover, we find that for the sporadic traffic the on duration and long DRX cycle have a strong effect on the DRX performance than inactivity timer, short DRX cycle, and DRX short cycle timer, which should be taken in account when selecting the optimal DRX parameters.

5.4 DRX Modeling and Optimization for Sporadic Traffic

5.4.1 Introduction

In the last section, we proposed a Semi-Markov model to analyze the DRX mechanism for MTC application. In that model, we assume the traffic is Poisson. However, as there are various types of MTC applications, this assumption may not always be realistic for all types of MTC traffic. For example, in [3] the traffic is assumed to be uniformly distributed. For the non-Poisson traffic, the model proposed in the last section is not applicable. Therefore, in this section we propose another method which is applicable to analyze the DRX mechanism with generic sporadic MTC traffic model.

5.4.2 DRX modeling for MTC Application with Generic Sporadic Traffic

Fig. 5.10 shows the packet delivery procedure in LTE with DRX. Here we consider two packets arrived at the eNB subsequently. As the target UE is at sleep mode, these two packets have to wait for certain time before being delivered. Here we assume that the first packet is delivered with latency d_1, while the second packet is delivered with latency d_2. The interval between the first and second packet arrival is denoted as t, which follows certain distribution (Poisson, Pareto, uniform, etc.). Here we also assume that the traffic is sporadic, i.e., the probability that two packets arrived in one short or long DRX cycle is negligible, which is quite reasonable for MTC applications. With this assumption, it can be easily infer that t is always larger than d_1. From Fig. 5.10, we can see that the second packet arrives at the eNB $t - d_1$ ms after the time instant when the first packet is delivered. Therefore, for the second packet, whether the UE is at the active mode or sleep mode is determined by the amount of time $t - d_1$. For example, after the delivery of the first packet, the inactivity timer is running for a period of T_0. If $t - d_1 < T_0$, then the second packet arrives at the active period, where no latency is introduced. Similarly, if $T_0 + T_{ON} < t - d_1 < T_0 + T_S$, the second packet arrives at the sleep period of the first short DRX cycle, which causes a latency of $T_0 + T_S - (t - d_1)$. Furthermore, if $T_0 + NT_S + T_{ON} < t - d_1 < T_0 + NT_S + T_L$, the second packet arrives at the sleep period of the long DRX cycle, which incurs a latency of $T_0 + NT_S + T_L - (t - d_1)$. To generalize, the latency if a packet arrives during the sleep period of the ith short DRX is calculated by

$$d'_i = \int_{T_0+(i-1)T_S+T_{ON}}^{T_0+iT_S} [T_0 + iT_S - (t - d_1)]f(t)dt \tag{5.39}$$

where $f(t)$ is the probability density function (PDF) of the packet interval t. Similarly, the latency if a packet arrives during the sleep period of a long DRX cycle is calculated as

$$d^*_j = \int_{T_0+NT_S+(j-1)T_L+T_{ON}}^{T_0+NT_S+jT_L} [T_0 + NT_S + jT_L - (t - d_1)]f(t)dt \tag{5.40}$$

Table 5.2: DRX parameters

Notation	DRX parameter
N	short DRX cycle timer
M	number of long DRX cycle
T_0	Inactivity timer
T_L	long DRX cycle
T_{ON}	On duration
T_S	short DRX cycle

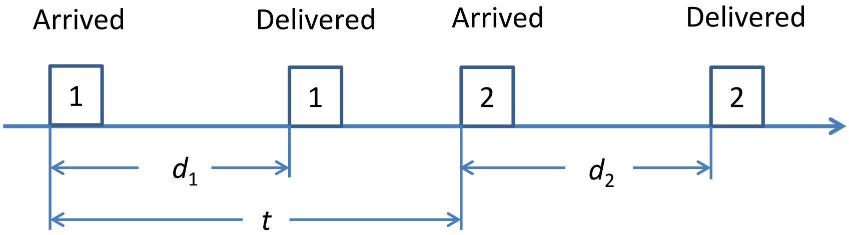


Figure 5.10: packet delivery procedure with DRX

With the above results, the overall latency is

$$d_2 = \sum_{i=1}^{N} d_i' + \sum_{j=1}^{M} d_j^*. \tag{5.41}$$

where M is the number of long DRX cycle. The value of M depends on the traffic pattern. For uniform traffic, the packet interval is finite, i.e., a UE will be at the continuous reception state after certain time. Therefore, M is also finite. While for Pareto and Poisson traffic, the packet interval is infinite, i.e., a UE can stay at the long DRX state as long as possible. As a result, the value of M is also infinite.

In this equation, we have two unknowns d_1 and d_2, which cannot be solved directly. To solve this equation, here we assume the average downlink packet interval per UE is much larger than the short or long DRX cycle, i.e., $E(t) \gg T_S, T_L$. This assumption is realistic and comes from the observation that most MTC downlink traffic is *sporadic*. For example, as proposed in [87]- [89] the packet interval for MTC traffic is 30 and 300s, while the maximum short and long DRX cycle is 640ms and 2.56s respectively (much smaller than packet interval). Moreover, for delay sensitive MTC applications the long DRX cycle is usually set to be the order of several hundred milliseconds to comply with latency requirement, which also makes the long DRX cycle much smaller than the packet interval. Since the packet latency d_1 is less than T_L, i.e., $d_1 < T_L$, it is obvious that $E(t) \gg d_1$. Therefore, $t - d_1 \approx t$. With this approximation equations (5.39) and (5.40) are simplified as

$$d_i' = \int_{T_0+(i-1)T_S+T_{ON}}^{T_0+iT_S} (T_0 + iT_S - t) f(t) dt \tag{5.42}$$

$$d_j^* = \int_{T_0+NT_S+(j-1)T_L+T_{ON}}^{T_0+NT_S+jT_L} (T_0 + NT_S + jT_L - t)f(t)dt \tag{5.43}$$

Hence, latency can be calculated with equation (5.41).

The power saving factor, defined as ratio of time that UE is at the power saving mode to the time that the UE is at all modes (sleep and active), is calculated as

$$\begin{aligned}\alpha = &\sum_{i=1}^{N} F(T_0 + iT_S) - F(T_0 + (i-1)T_S + T_{ON}) \\ &+ \sum_{j=1}^{M} F(T_0 + NT_S + jT_L) - F(T_0 + NT_S + (j-1)T_L + T_{ON})\end{aligned} \tag{5.44}$$

where the function $F(t)$ is the cumulative distribution function (CDF) of the packet interval time $f(t)$.

It can be seen that the derivation of latency and power saving factor is quite simple if the traffic sporadic.

5.4.3 Examples

To demonstrate the method proposed in the last section, we provide two examples in this section.

Uniform Distributed Traffic

Firstly, we assume the packet interval is uniform distributed as that proposed in [3]. Specifically, we assume the packet interval is uniform distributed over $[a, b]$. With (5.42), we have

$$d_i' = \frac{(T_S - T_{ON})^2}{2(b-a)}. \tag{5.45}$$

By the use of (5.43), we have

$$d_j^* = \frac{(T_L - T_{ON})^2}{2(b-a)}. \tag{5.46}$$

Since the packet interval is uniformly distributed over $[a, b]$, the maximum packet interval is $b - a$. Moreover, it is known that time consumed by continuous reception and short DRX cycle is $T_0 + NT_S$. Therefore, the number of long DRX cycle M is calculated as

$$M = \frac{b - a - T_0 - NT_S}{T_L}. \tag{5.47}$$

Therefore, the latency is

$$d = N\frac{T_S - T_{ON}}{b-a}\frac{T_S - T_{ON}}{2} + M\frac{T_L - T_{ON}}{b-a}\frac{T_L - T_{ON}}{2}. \tag{5.48}$$

In the above equation, $\frac{T_S - T_{ON}}{b-a}$ is the ratio that a short DRX cycle to the packet interval and $\frac{T_S - T_{ON}}{2}$ is the average latency when a packet arrives at the short DRX cycle. Therefore, it is easy to understand that the first term of the above equation is the expected latency caused by short DRX cycle. Similarly, it is obvious that the second term of the above equation is the average latency caused by long DRX cycle.

The power saving factor is

$$\alpha = N\frac{T_S - T_{ON}}{b-a} + M\frac{T_L - T_{ON}}{b-a}. \tag{5.49}$$

Pareto Distributed Traffic

Here, we assume the packet interval follows Pareto distributed, i.e.,

$$f(t|\beta, T_m)\begin{cases} \frac{\beta T_m^\beta}{t^{\beta+1}}, & \text{for } T_m \le t \le \infty; \beta, T_m > 0 \\ 0, & \text{for } t < T_m \end{cases} \tag{5.50}$$

where T_m is the scale parameter and β is the shape parameter.

The the cumulative distribution function (CDF) of a Pareto random variable with parameters scale parameter T_m and shape parameter β is:

$$F(t|\beta, T_m)\begin{cases} 1-(\frac{T_m}{t})^\beta, & \text{for } T_m \le t \le \infty; \beta, T_m > 0 \\ 0, & \text{for } t < T_m \end{cases} \tag{5.51}$$

Depending on the value of T_m, here are several cases to calculate latency and power saving factor:

1. $T_m < T_0 + T_{ON}$. With (5.42), we have

$$\begin{aligned} d_i' = (T_0 + iT_S)[(\frac{T_m}{T_0 + (i-1)T_S + T_{ON}})^{-\beta} - (\frac{T_m}{T_0 + iT_S})^{-\beta}] \\ - T_m^\beta \frac{\beta}{1-\beta}[(T_0 + iT_S)^{1-\beta} - (T_0 + (i-1)T_S + T_{ON})^{1-\beta}] \end{aligned} \tag{5.52}$$

With (5.43), we get

$$\begin{aligned} d_j^* = (T_0 + NT_S + jT_L)[(\frac{T_m}{T_0 + NT_S + (j-1)T_L + T_{ON}})^{-\beta} - (\frac{T_m}{T_0 + NT_S + jT_L})^{-\beta}] \\ - T_m^\beta \frac{\beta}{1-\beta}[(T_0 + NT_S + jT_L)^{1-\beta} - (T_0 + NT_S + (j-1)T_L + T_{ON})^{1-\beta}] \end{aligned} \tag{5.53}$$

where M is the number of long DRX cycle. The packet interval t follows Pareto distribution, therefore its value can be as large as infinite. As a result, the number of long DRX cycle M is also infinite, which makes it difficult to calculate the second item in

5.41. For the reason of simplicity, here we use the quantile function for Pareto distributed variable [90]:

$$F^{-1}(p) = \frac{T_m}{(1-p)^{1/\beta}}. \tag{5.54}$$

By setting p in the last formula, we can calculate T_1: the probability that packet interval t is smaller than T_1 equals p. If p is large enough, we can assume that packet interval t is always smaller than T_1. Noticing that the time consumed by continuous reception and short DRX cycle is $T_0 + NT_S$, the number of long DRX cycle M is

$$M = \frac{T_1 - (T_0 + NT_S)}{T_L}. \tag{5.55}$$

Hence, we can calculate the latency as

$$d_2 = \sum_{i=1}^{N} d'_i + \sum_{j=1}^{M} d^*_j. \tag{5.56}$$

The power saving introduced by the ith short DRX cycle is

$$\alpha'_i = (\frac{T_m}{T_0 + (i-1)T_S + T_{ON}})^\beta - (\frac{T_m}{T_0 + iT_S})^\beta \tag{5.57}$$

and by the jth long DRX cycle is

$$\alpha^*_j = (\frac{T_m}{T_0 + NT_S + (j-1)T_L + T_{ON}})^\beta - (\frac{T_m}{T_0 + NT_S + jT_L})^\beta. \tag{5.58}$$

The power saving factor is

$$\alpha = \sum_{i=1}^{N} \alpha'_i + \sum_{j=1}^{M} \alpha^*_j. \tag{5.59}$$

2. $T_0 + n_1 T_S < T_m < T_0 + n_1 T_S + T_{ON}$, $n_1 \in [0, N-1]$. Then latency is

$$d_2 = \sum_{i=n_1+1}^{N} d'_i + \sum_{j=1}^{M} d^*_j \tag{5.60}$$

and the power saving factor is

$$\alpha = \sum_{i=n_1+1}^{N} \alpha'_i + \sum_{j=1}^{M} \alpha^*_j. \tag{5.61}$$

3. $T_0 + n_2 T_S + T_{ON} < T_m < T_0 + (n_2+1)T_S$, $n_2 \in [0, N-1]$. The latency caused by the (n_2+1)th short DRX cycle is

$$\begin{aligned} d^{\circ}_{n_2+1} &= (T_0 + (n_2+1)T_S)[1 - (\frac{T_m}{T_0 + (n_2+1)T_S})^{-\beta}] \\ &- T_m^{\beta} \frac{\beta}{1-\beta}[(T_0 + (n_2+1)T_S)^{1-\beta} - (T_m)^{1-\beta}] \end{aligned} \tag{5.62}$$

and the power saving factor is

$$\alpha^{\circ}_{n_2+1} = 1 - (\frac{T_m}{T_0 + (n_2+1)T_S})^{\beta}. \tag{5.63}$$

The overall latency and power saving factor is

$$d_2 = d^{\circ}_{n_2+1} + \sum_{i=n_2+2}^{N} d'_i + \sum_{j=1}^{M} d^*_j \tag{5.64}$$

and

$$\alpha = \alpha^{\circ}_{n_2+1} + \sum_{i=n_2+2}^{N} \alpha'_i + \sum_{j=1}^{M} \alpha^*_j. \tag{5.65}$$

4. $T_0 + NT_S + n_3 T_L < T_m < T_0 + NT_S + n_3 T_L + T_{ON}$, $n_3 \in [0, M-1]$. In this case, the latency and power saving factor is

$$d_2 = \sum_{j=n_3+1}^{M} d^*_j \tag{5.66}$$

and

$$\alpha = \sum_{j=n_3+1}^{M} \alpha^*_j. \tag{5.67}$$

5. $T_0 + NT_S + n_4 T_L + T_{ON} < T_m < T_0 + NT_S + (n_4+1)T_L$, $n_4 \in [0, M-1]$. The latency caused by the (n_4+1)th long DRX cycle is

$$\begin{aligned} d^{\dagger}_{n_4+1} &= (T_0 + NT_S + (n_4+1)T_L)[1 - (\frac{k}{T_0 + NT_S + (n_4+1)T_L})^{-\beta}] \\ &- k^{\beta} \frac{\beta}{1-\beta}[(T_0 + NT_S + (n_4+1)T_L)^{1-\beta} - (T_m)^{1-\beta}] \end{aligned} \tag{5.68}$$

and the corresponding power saving factor is

$$\alpha^{\dagger}_{n_4+1} = 1 - (\frac{T_m}{T_0 + NT_S + (n_4+1)T_L})^{\beta}. \tag{5.69}$$

The overall latency and power saving factor is

$$d_2 = d^{\dagger}_{n_4+1} + \sum_{j=n_4+2}^{M} d^*_j \tag{5.70}$$

and

$$\alpha = \alpha^{\dagger}_{n_4+1} + \sum_{j=n_4+2}^{M} \alpha^*_j. \tag{5.71}$$

5.4.4 Model Validation

To validate the proposed model, we carried out simulations with a MATLAB based simulator. This simulator is developed in compliance with the protocols in [60] and the parameters are listed in Table 5.3. For the Pareto distribution, the expected value is $\frac{\beta T_m}{\beta-1}$. Here we set $\beta = 2$, therefore we have different scale parameter T_m for different packet intervals, which are listed in Table 5.4. The parameter p in formula (5.54) is set to 0.99.

Firstly, we compare the simulated results with the analytical results under different average packet intervals for uniform and Pareto traffic. The results are shown in Fig. 5.11. We find that the simulated results match the analytical results very well when the average packet interval is larger than 5 seconds. In contrast to that, there are some estimation error when the average packet interval is 0.25, 0.5 and 1 second. For example, for Pareto traffic when the average packet interval is 0.25 second the estimated (analytical) latency is 13 ms while the simulated latency is 17.47 ms. The reason is: when the average packet interval is 0.25 second (250ms), it is not much larger than T_S (128ms) and T_L (256ms), which is not consistent with our assumption. Therefore, the analytical results do not match the analytical results. However, when the average packet interval increases to 5 seconds (5000 ms), the packet interval becomes much larger than T_S and T_L, by which the analytical results match the analytical results with small errors. Hence it can be concluded that by the use of our model the performance of DRX with sporadic traffic can be correctly calculated.

In addition, we carried out more simulations to validate our model and to see that effect of different DRX parameters on its performance. The parameters are listed in Table 5.3. It should be noted that here the traffic used in simulation is sporadic (the average packet interval T=30 seconds).

Table 5.3: Simulation parameters for results in Fig. 5.11-5.17

Figure index	Parameters
5.11	$T_0 = 20ms, T_{ON} = 80ms, T_S = 128ms, N = 2, T_L = 256ms$
5.12	$T = 30s, T_{ON} = 80ms, T_S = 128ms, N = 2, T_L = 256ms$
5.13	$T = 30s, T_0 = 20ms, T_S = 128ms, N = 2, T_L = 256ms$
5.14	$T = 30s, T_0 = 20ms, T_{ON} = 80ms, N = 2, T_L = 2T_S$
5.15	$T = 30s, T_0 = 20ms, T_{ON} = 80ms, N = 2, T_L = 1024ms$
5.16	$T = 30s, T_0 = 20ms, T_{ON} = 80ms, T_S = 128ms, T_L = 256ms$
5.17	$T = 30s, T_0 = 20ms, T_{ON} = 80ms, T_S = 128ms, N = 2$

Table 5.4: Scale parameter for different packet interval

Average packet interval (second)	Value of T_m (ms)
0.25	125
0.5	250
1	500
5	2500
10	5000
20	10000
30	150000
40	200000

1. *Effect of inactivity timer*. Fig.5.12 demonstrates the DRX performance under different inactivity timer for uniform and Pareto traffic. It is shown that the simulated results match the analytical results and DRX performance of uniform and Pareto traffic are almost the same. Moreover, we find that the latency decreases slightly when inactivity timer changes. For example, regarding the uniform traffic the latency changes from 59ms to 57ms and power saving factor changes from 0.68 to when 0.67 when the inactivity timer increases from 20 to 2560. Therefore, it is reasonable to argue that the inactivity timer has weak effect on the DRX performance.

2. *Effect of on duration*. Fig. 5.13 shows the DRX performance under different on duration. The simulated results matches the analytical results and DRX performance of uniform and Pareto traffic are almost the same. Furthermore, it is shown that the latency decreases greatly from 124 ms to 47ms and the power saving factor is reduced drastically from 0.98 to 0.60 when the on duration increases from 5 to 100. The reason for this phenomenon is that on duration determines the proportion of time that a UE is at active state. A UE spends more time at active state when the on duration increases. Therefore, on duration has strong effect on DRX performance.

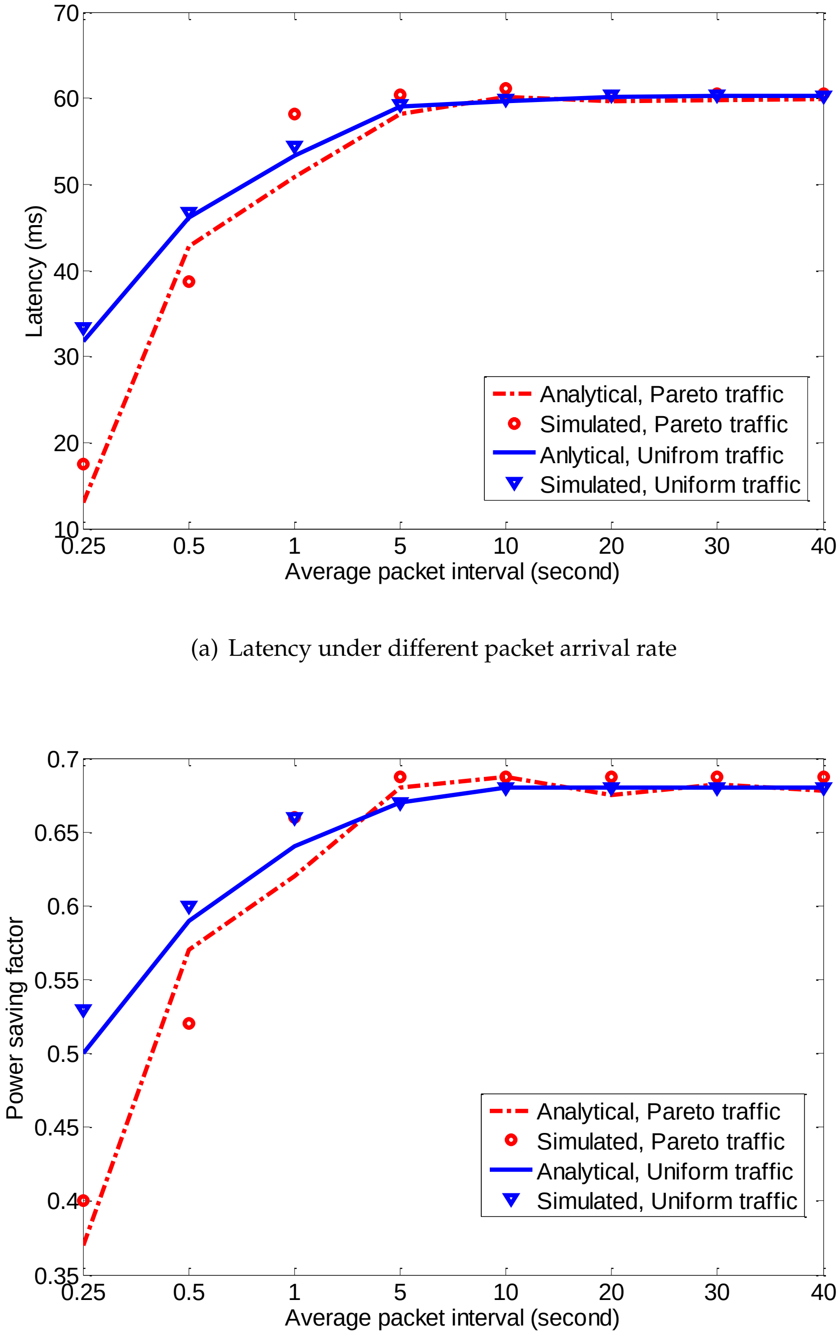
3. *Effect of short DRX cycle*. As specified in [12], the long DRX cycle is a multiple of short DRX cycle. To comply with that, here firstly we set T_L as $T_L = 2T_S$. From Fig. 5.14, we find that the latency increases from 60 to 555 and the power saving changes from 0.68 to 0.93 when the short DRX cycle increases from 128 to 640. Therefore, it seems that short DRX cycle has a strong effect on the DRX performance. However, it is has to be

noted that the long DRX cycle also varies in this simulation since $T_L = 2T_S$, which also affects the DRX performance. Therefore, whether short DRX cycle has a strong effect on the DRX performance is unclear. To clarify this issue, we carried another simulation. In this simulation, $T_L = 1024$ and T_S varies from 128 to 1024 while complying with the constraint that the long DRX cycle is a multiple of short DRX cycle. The results are shown in Fig. 5.15. We can see that the latency and power saving changes slightly when the short DRX cycle increases from 128 ms to 1024 ms. Hence, it can concluded that short DRX has a weak effect on the DRX performance.

4. *Effect of DRX short cycle timer*. Fig.5.16 demonstrates the DRX performance with different DRX short cycle timer. We can find that the simulated results match the analytical results and difference of the DRX performance between uniform and Pareto traffic are small. Furthermore, the results show that the latency and power saving factor are almost constant for Pareto traffic when the DRX short cycle timer changes. For the uniform traffic, the latency decreases from 60 ms to 59 ms and the power saving factor decreases from 0.68 to 0.67 when the DRX short cycle timer increases from 1 to 16. Therefore, it is reasonable to conclude that DRX short timer has a weak effect on the DRX performance.

5. *Effect of long DRX cycle*. Fig.5.17 shows the DRX performance under different long DRX cycle. Firstly, we can see that the stimulated results match the analytical results and DRX performance of uniform and Pareto traffic are almost the same. Secondly, it is shown that the latency and power saving factor varies dramatically when the long DRX cycle changes. Concretely, the latency increases from 9 ms to 944 ms when the long DRX cycle varies from 128 to 2048, while at the same time the power saving factor increases from 0.37 to 0.95. Therefore, it is evident that the long DRX cycle has a strong effect on the DRX performance.

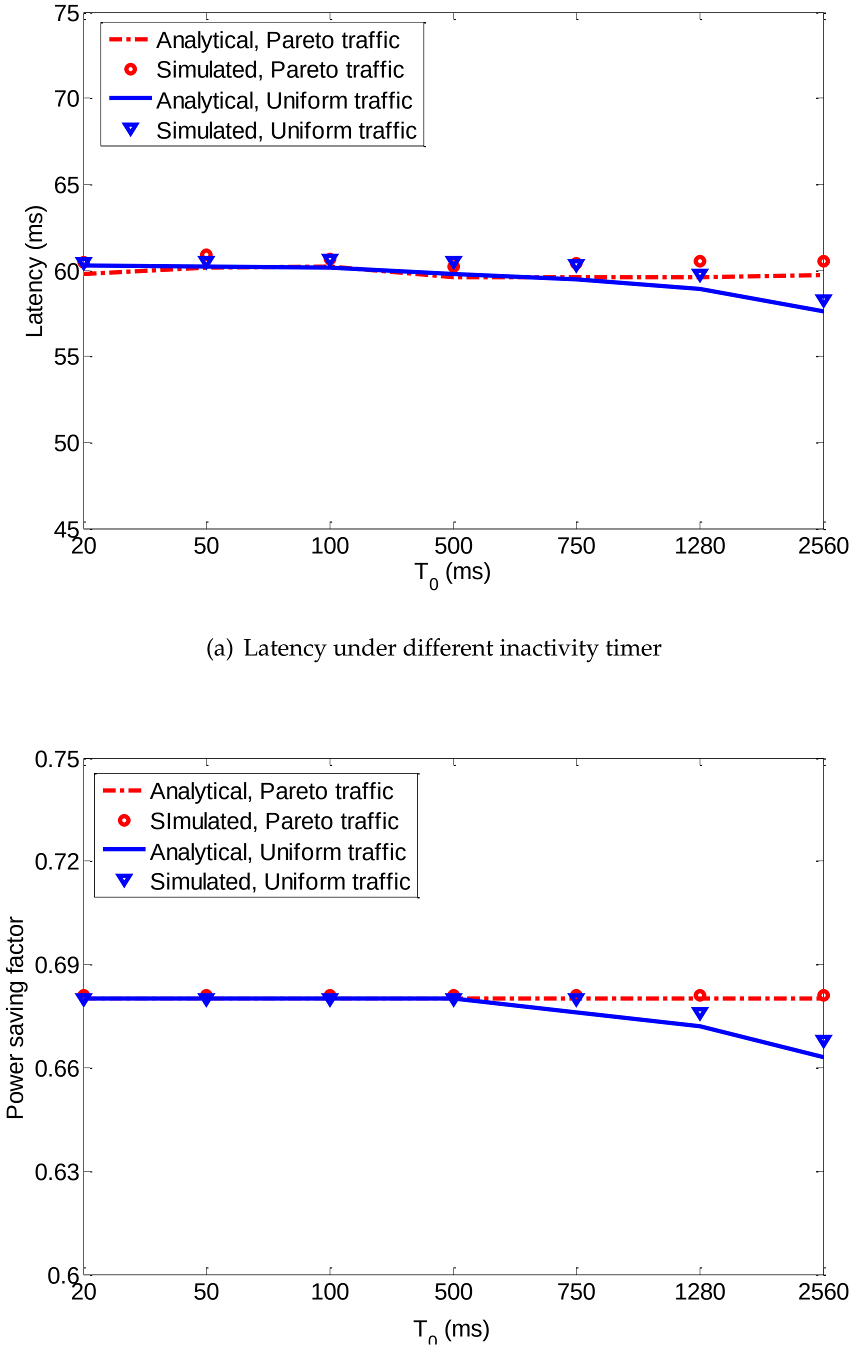
Through simulations, we observe that the on duration and long DRX cycle has strong effect on the DRX performance, while the inactivity timer, short DRX cycle, and DRX short cycle timer has weak effect on the DRX performance. The reason for this phenomenon is that: as the traffic is sporadic, the packet interval is much larger than the inactivity, and short sleep period (short DRX cycle × short DRX cycle timer). As a result, a UE is mostly at the long sleep state, and therefore the DRX performance is determined by on duration and long DRX cycle. With this observation, in order to find the optimal DRX parameter with low complexity, it is feasible that the on duration and long DRX cycle is carefully selected while the other DRX parameters can be flexibly determined. Moreover, through simulations we also notice that the DRX performance of uniform traffic and Pareto traffic are almost the same.



(a) Latency under different packet arrival rate

(b) Power saving factor under different Inactivity timer

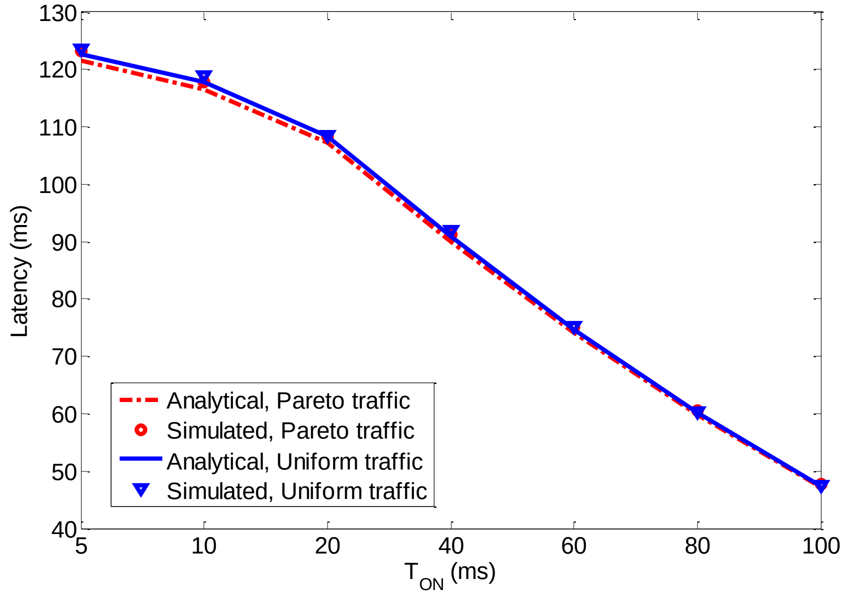
Figure 5.11: DRX performance under different packet arrival rate



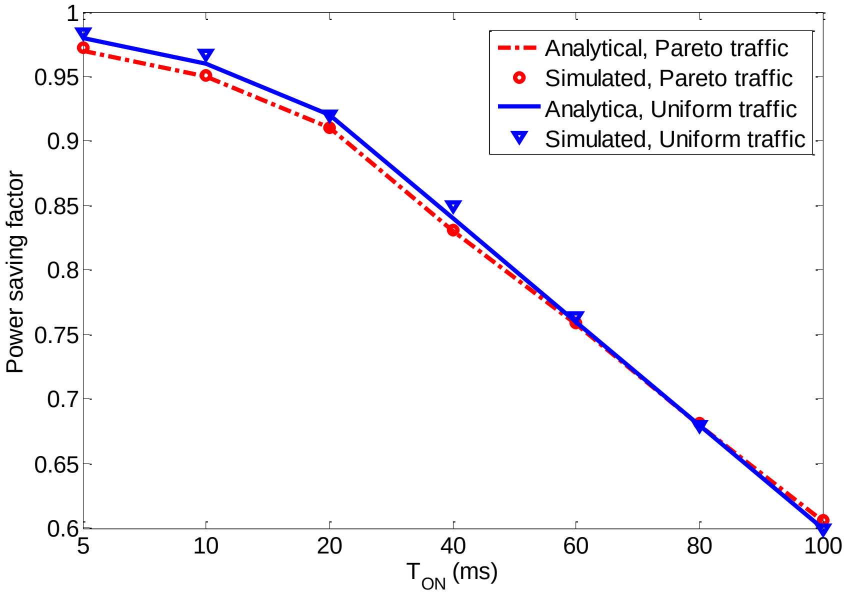
(a) Latency under different inactivity timer

(b) Power saving factor under different Inactivity timer

Figure 5.12: DRX performance under different inactivity timer

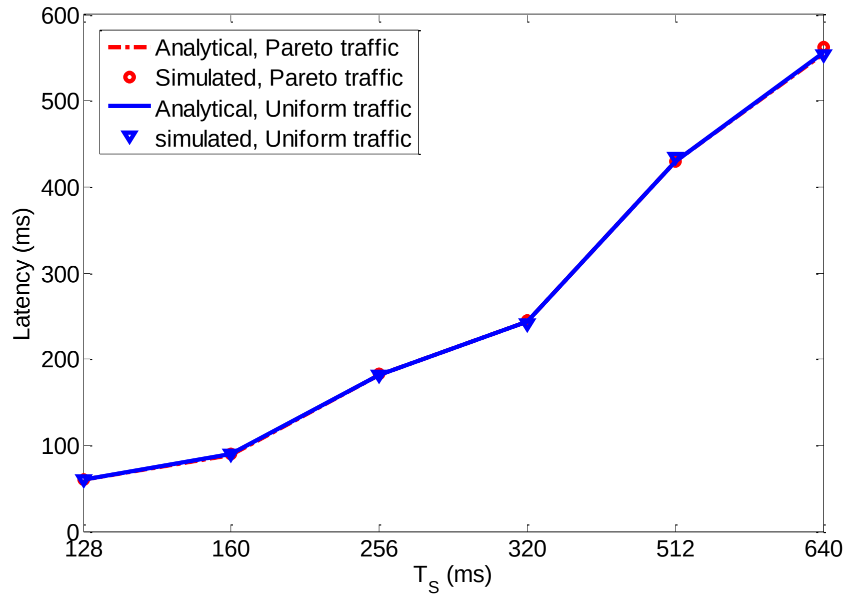


(a) Latency under different on duration

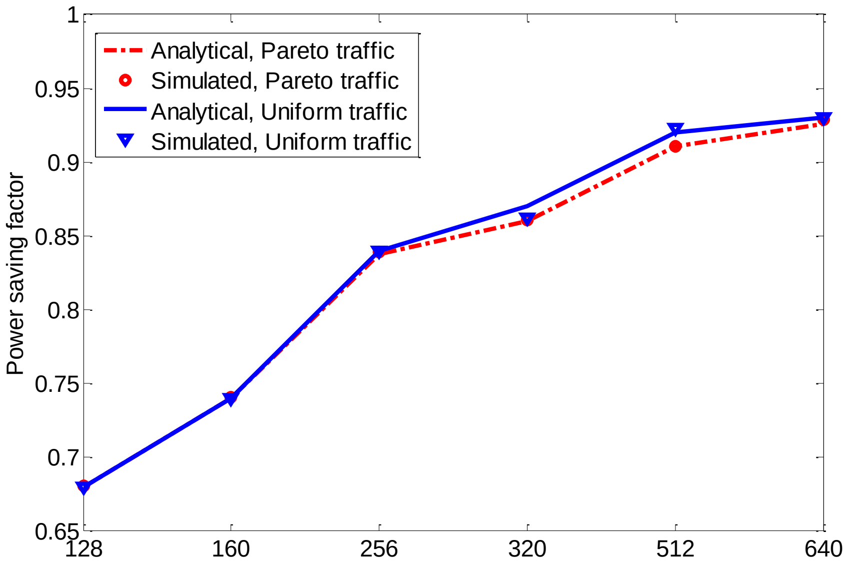


(b) Power saving factor under different on duration

Figure 5.13: DRX performance under different on duration

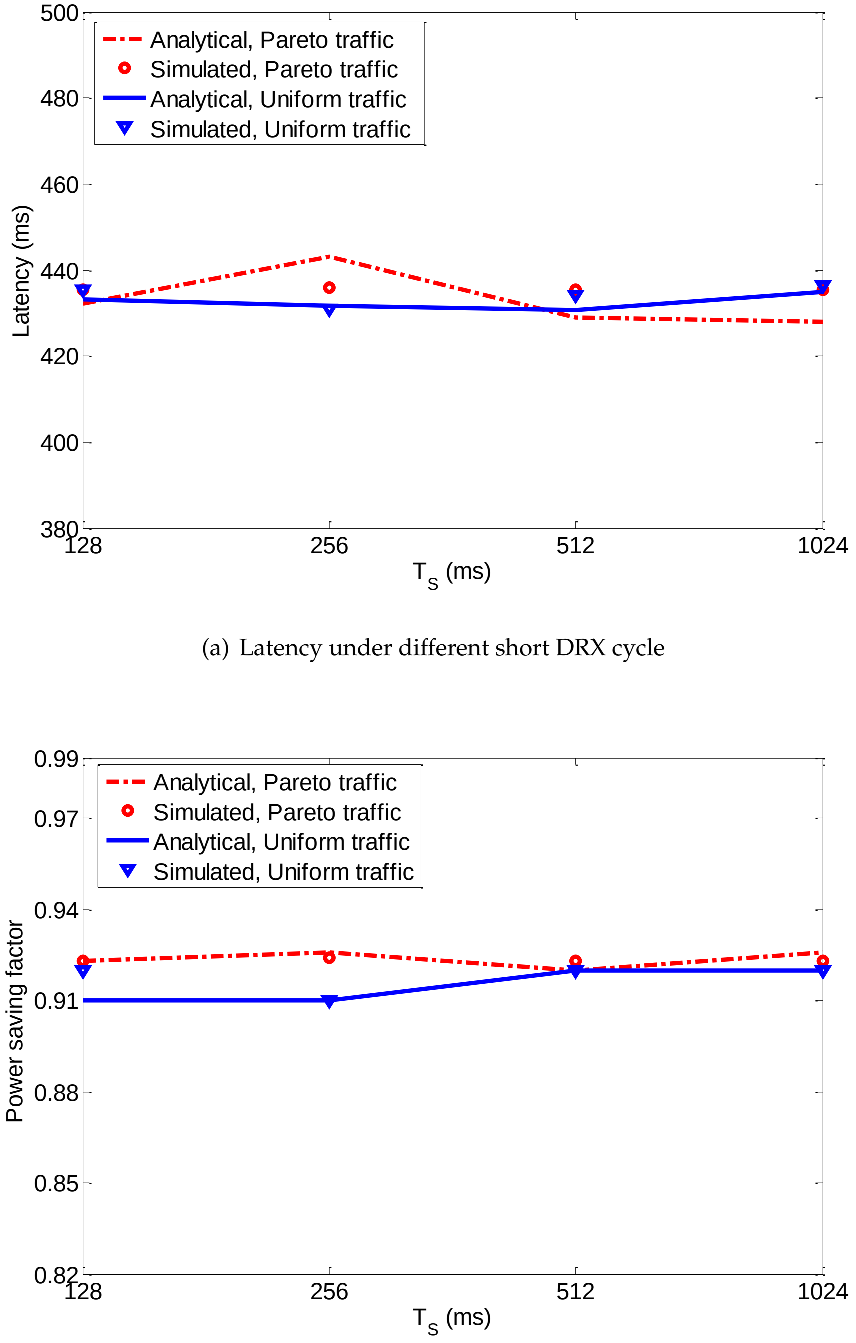


(a) Latency under different short DRX cycle, $T_L = 2T_S$



(b) Power saving factor under different short DRX cycle, $T_L = 2T_S$

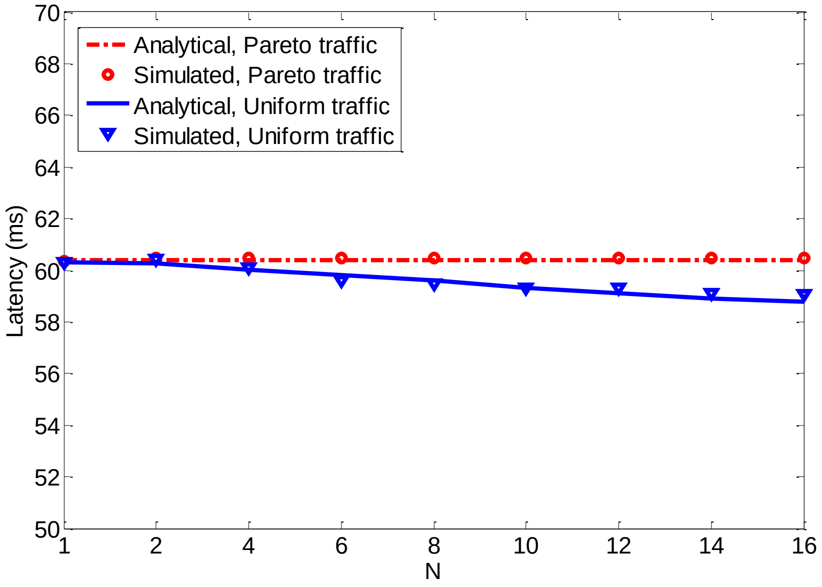
Figure 5.14: DRX performance under different short DRX cycle, $T_L = 2T_S$



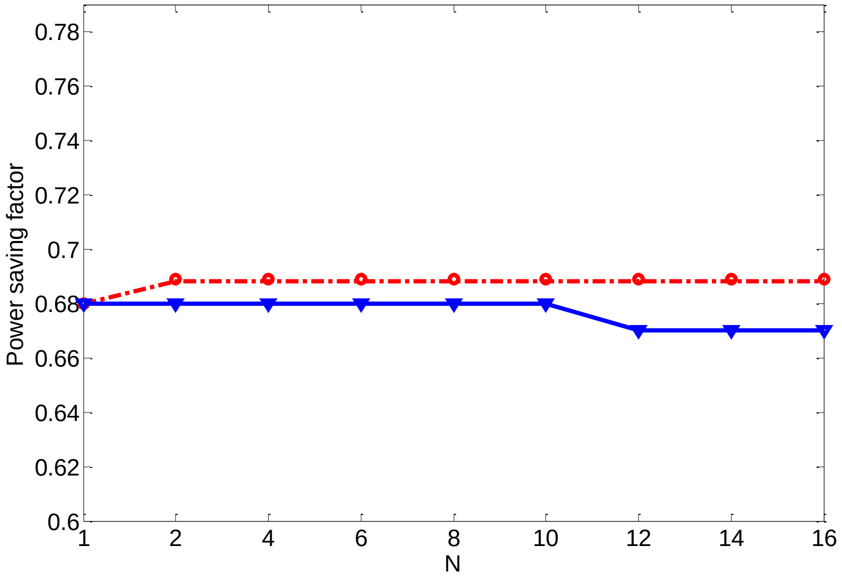
(a) Latency under different short DRX cycle

(b) Power saving factor under different short DRX cycle

Figure 5.15: DRX performance under different short DRX cycle

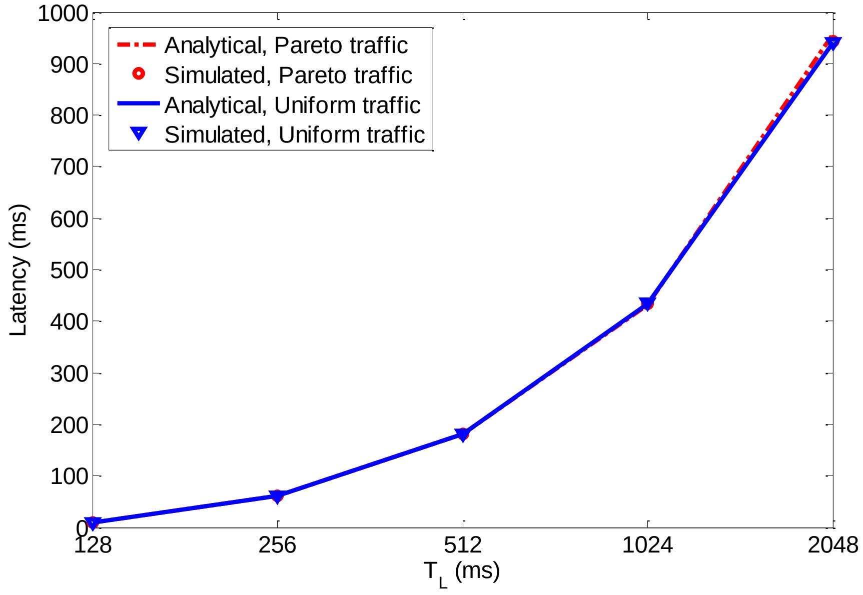


(a) Latency under different DRX short cycle timer

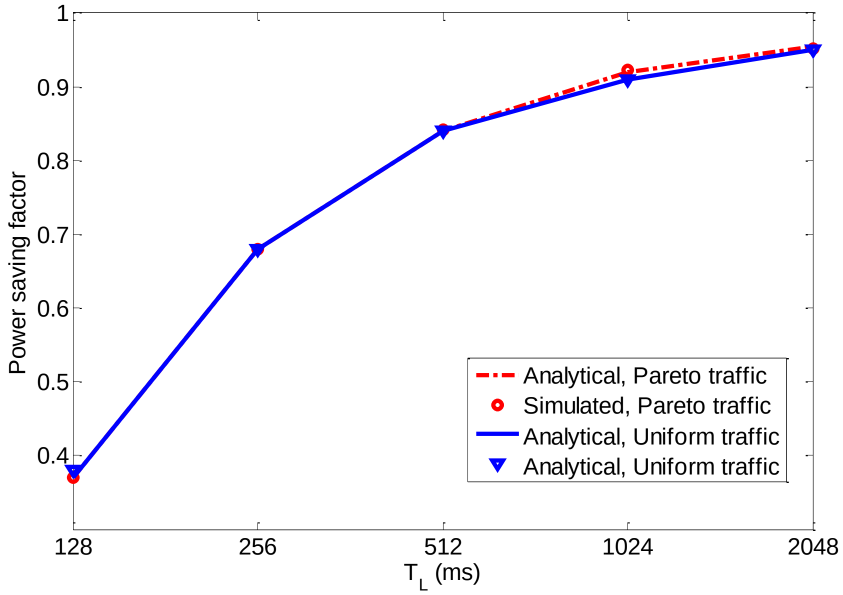


(b) Power saving factor under different short DRX cycle timer

Figure 5.16: DRX performance under different DRX short cycle timer



(a) Latency under different long DRX cycle



(b) Power saving factor under different long DRX cycle

Figure 5.17: DRX performance under different long DRX cycle

5.4.5 Optimal DRX Parameters Selection

In the last section, we propose a method to analyze the DRX performance with sporadic traffic and validate the proposed method through simulations. Through simulations, we notice that the on duration and long DRX cycle have strong effect on the DRX performance while the inactivity timer, short DRX cycle, and DRX short cycle timer have weak effect on the DRX performance. Hence, for the sake of simplicity, we only consider on duration and long DRX cycle when selecting the optimal DRX parameters. The optimal DRX parameter set would be the one which maximizes the power saving factor while maintaining the latency constraint, i.e., it is calculated as

$$\begin{aligned} &\underset{(T_{ON},T_L)}{\text{arg max}} && \alpha \\ &\text{subject to} && d \leq D \end{aligned} \tag{5.72}$$

where D is the latency constraint. As the power saving factor increases with latency, therefore the maximum power saving factor is achieved when the latency reaches the constraint. Accordingly, the above optimization problem can also be written as

$$\begin{aligned} &\underset{(T_{ON},T_L)}{\text{arg max}} && \alpha \\ &\text{subject to} && d = D \end{aligned} \tag{5.73}$$

which can easily be solved by Lagrange multiplier method.

Here we provide an example to demonstrate how the optimal DRX parameter is selected. We assume that the packet interval is uniformly distributed over [0, 30000], i.e., $T = 15000$. The inactivity timer (T_0), short DRX cycle (T_S), DRX short cycle timer (N) is set to the medium value as: 500, 512, and 8 respectively. Therefore, $M = \frac{25404}{T_L}$, the power saving factor is

$$\begin{aligned} \alpha &= N\frac{T_S - T_{ON}}{b-a} + M\frac{T_L - T_{ON}}{b-a} \\ &= 8\frac{512 - T_{ON}}{30000} + \frac{25404}{T_L}\frac{T_L - T_{ON}}{30000} \\ &= \frac{512 - T_{ON}}{3750} + \frac{0.85(T_L - T_{ON})}{T_L} \end{aligned} \tag{5.74}$$

and the latency is

$$\begin{aligned} d &= N\frac{T_S - T_{ON}}{b-a}\frac{T_S - T_{ON}}{2} + M\frac{T_L - T_{ON}}{b-a}\frac{T_L - T_{ON}}{2} \\ &= 8\frac{512 - T_{ON}}{30000}\frac{512 - T_{ON}}{2} + \frac{25404}{T_L}\frac{T_L - T_{ON}}{30000}\frac{T_L - T_{ON}}{2} \\ &= \frac{(512 - T_{ON})^2}{7500} + \frac{0.42(T_L - T_{ON})^2}{T_L} \end{aligned} \tag{5.75}$$

Assuming the delay constraint is 300 ms, by the use of the Lagrange multiplier method, the optimal $T_L = 628$, $T_{ON} = 1$, and the achieved power saving factor is $\alpha = 0.982$. It has to be noted that as specified in [12] the value 628 is not the standardized value for T_L. In order to be consistent with the LTE standard, T_L can be adjust as $T_L = 512$ (T_{ON} is still set to 1), which slightly degrades the power saving factor to 0.981.

As we notice that the DRX performance of Pareto and uniform traffic are almost the same for sporadic traffic. Therefore, we can set the DRX parameter for the Pareto traffic same as those for the uniform traffic. For example, here we set T_0=500, T_S=512, N=8, T_L=512, T_{ON}=1 for a Pareto traffic (average packet interval 15000 ms) with delay constraint 300 ms. The resulted power saving factor is 0.99 and latency is 252 ms which is less than the constraint (300ms).

5.5 Conclusion

In this chapter, we propose two methods to analyze the DRX mechanism for Poisson traffic and sporadic traffic, respectively.

In the first method, we introduce a Semi-Markov chain model to analyze the DRX mechanism for MTC over LTE. The model accurately derives the wake up latency and power saving factor by calculating the stationary probabilities and holding times for the active and sleeping states. Concretely, in our method we firstly derive the stationary probabilities of active and sleeping states; then calculate the holding time for each state; finally derive the wake up latency and power saving factor which are the two metrics of DRX mechanism. The proposed method is validated through simulations. Moreover, for the sporadic traffic we find that on duration and long DRX cycle have a strong effect on the DRX performance than inactivity timer, short DRX cycle, and DRX short cycle timer.

In the second method, we propose a method to analyze the DRX mechanism with sporadic traffic. Different from the method proposed in the last section which requires that the traffic is Poisson distributed, this method is applicable to all kinds of traffic (Poisson distributed, uniform distributed, etc.). The proposed method is validated through simulations for Uniform and Pareto distributed traffic. We find that when the traffic is sporadic (packet interval is larger than 5 seconds), the simulated results match the analytical ones, which validates our method. In addition to that, we also find that on duration and long DRX cycle have strong effect on the DRX performance while the inactivity timer, short DRX cycle, and DRX short cycle timer has weak effect on its performance. Based on this observation, we present a method to select the optimal DRX parameter (on duration and long DRX cycle), which maximizes the power saving factor while maintaining the delay constraint.

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

Machine type communications is one of the most promising applications provided by LTE/LTE-A networks. However, it is challenging to accommodate MTC in LTE due to its special characteristics and requirements. This thesis addresses the challenges incurred from machine type communications in LTE/LTE-A. Specifically, we propose several methods to improve the performance of MTC uplink access and downlink reception.

In chapter 3, a packet aggregation method is proposed. With the proposed packet aggregation method, a UE triggers a random access when the aggregated packets in the buffer reaches the given threshold. This method reduces the packet collision rate or power consumption at the expense of an extra latency which is used to accumulate certain amount of packets. Therefore, the tradeoff is carefully selected between packet loss rate/power consumption and access latency. We derive the packet loss rate and channel access latency as functions of amount of aggregated packets using a Semi-Markov chain model. With the derived results, the optimal amount of aggregated packets which satisfies the packet loss requirement and keeps the latency as small as possible can be found. With packet aggregation, a UE reduces number of random access, which hence saves energy. Therefore, we also propose a energy saving method with packet aggregation: a UE aggregates packets as much as possible until its latency reaches delay constraint. Moreover, to reduce the access latency caused by unsuccessful random access, we also propose a TTI bundling scheme. In our method, a UE sends multiple preambles in consecutive subframes in order to increase the random access successful rate. We introduce a Semi-Markov model to analyze the random access mechanism with TTI bundling. With this model, we formulate the access latency as a function of the TTI bundling number and select the optimal TTI bundling number which minimizes the access latency. The numerical results show that the access latency is greatly reduced when the preamble collision rate is not high. A TTI bundling scheme designed for power constrained MTC device is also proposed aimed to save the extra power introduced by multiple preamble transmissions.

In chapter 4, in order to further reduce the uplink latency for MTC, we propose a contention based access method (CBA). With this method, the uplink latency can be greatly reduced by bypassing the redundant signaling in random access. We employ MU-MIMO detection at the eNB side to identify the C-RNTIs of collided UEs such that dedicated resource can be allocated for those UEs in the next round. Modifications to the LTE standard are provided in order to implement CBA in LTE. We also present a resource allocation method for CBA. With this resource allocation scheme, the minimum resource to guarantee the delay constraint is allocated.

In chapter 5, firstly we model the LTE/LTE-A discontinuous reception (DRX) mechanism for MTC applications with Poisson distributed traffic. With our model the power saving factor and wake up latency can be accurately estimated for a given choice of DRX parameters, thus allowing to select the ones presenting the best tradeoff. The proposed model is validated through simulations. We also investigate the effect of different DRX parameters on performance. Then, we also propose another method to analyze DRX mechanism with sporadic traffic which is typical for MTC application. The proposed method is validated against simulations with Uniform and Pareto traffic. We find that: (1) the DRX performance is almost the same for Uniform and Pareto traffic; (2) on duration and long DRX cycle have strong effect on the DRX performance. Based on these observations, we design a simple but effective method to choose the optimal DRX parameters (on duration and long DRX cycle), which attains the best power saving factor while maintaining the delay constraint.

6.2 Future Work

We proposed two independent methods in order to improve the performance of random access: packet aggregation and TTI bundling. Packet aggregation reduces packet loss rate or power consumption at the expense of latency, while the TTI bundling method reduces latency. Therefore, it is intuitive to combine these two methods to reduce latency as well as packet loss rate and power consumption. Moreover, in our TTI bundling method, we apply the same TTI bundling number for new transmission and retransmission. However, if an unsuccessful random access is caused by collision, bundling multiple TTIs may not increase the random access successful rate. Therefore, the TTI bundling number for retransmission should be reduced. In contrast, if an unsuccessful random access is caused by wireless channel error, more TTIs should be bundled for retransmission. To apply this idea to improve the performance of TTI bundling method is one of the future work. Finally, investigation on the performance of packet aggregation for small size packet transmission is also interesting for the future work.

Regarding the future work for CBA, firstly the grouping mechanism would be considered. The main target for grouping is to efficiently manage resource. There are different grouping strategies. For example: UEs with same traffic characteristics and QoS requirements can be allocated in the same CBA group; UEs with same channel conditions can be grouped; or UEs can be grouped based on locations. Moreover, the application of CBA to Device-to-Device (D2D) communications would be an interesting topic. D2D reduces the load of eNB and saves the power for UE. Regularly, eNB controls the D2D bearer setup between UEs and manages the resource for D2D communications. However, when the traffic is sporadic, this mechanism introduces large signaling overhead. With CBA, a UE can send packets on random selected resource, which removes the signaling overhead in D2D communications. Therefore, CBA is a feasible method to enable D2D communications in LTE. Finally, support of very small packets where the signaling and/or control overhead dominates the payload. For instance in CBA, the overhead of uplink transmission is 24 bytes, making this approach not efficient for cases where the packets are smaller than 24 bytes. One possible solution is to apply the packet aggregation technique to CBA, by which the overhead is reduced at the expense of latency.

We analyzed the DRX mechanism in LTE with Poisson and sporadic traffic. With the emergence of various new applications, more types of traffic pattern appear. Designing new methods to deal with other types of traffic would be one of the future directions.

Appendix A
Abstract for thesis in French

A.1 Introduction

Communications du type de la machine (ou la machine á communications de la machine) , par exemple : surveillance á distance , la gestion de la ville intelligente , et l'e-santé , joue un role important dans l'information et la technologie des communications (TIC). Fig. **??** montre l'évolution des communications de type de la machine (MTC) , nous pouvons voir que les communications de type de la machine évolue de l'identification par radiofréquence (RFID) pour l'Internet des objets (IdO) , oú tout ce qui peut bénéficier de la connexion réseau est connecté . Par ailleurs , le MTC est en train d'évoluer vers la société numérique , oú l'efficacité de la vie , en termes d'économie , la mobilité , l'environnement , la vie et la gouvernance , est obtenu á partir de la gestion intelligente , les TIC intégrées , et la participation active des citoyens . Bien sur, les opérateurs font pression pour la société numérique comme le marché du mobile sature et juste á la recherche de débits plus élevés ne créeront pas de nouveaux revenus , ils sont á la recherche d'un changement de paradigme pour créer de nouveaux revenus . Comme le montre la figure . Fig.A.2 [1] , le nombre d'appareils MTC est 8-9 fois plus grande que la population humaine , et dont seulement 50 millions de machines sont connectées , donc il ya un grand potentiel pour MTC . Il existe deux types de techniques pour accueillir MTC : communication filaire et communications sans fil . Par rapport á la technologie de communication filaire , la communication sans fil a des avantages pour permettre MTC : la mobilité , la facilité de déploiement et la robustesse [2] . Plus précisément , Long Term Evolution (LTE) est considérée comme une technique prometteuse pour permettre MTC en raison de sa large couverture , une faible latence et haute spectrale et l'efficacité énergétique .

Cependant , diffé rente d'une personne á (H2H) communications humaines qui les ré seaux cellulaires actuels sont principalement concus pour , MTC a des caracté ristiques et des exigences spé cifiques :

- Certaines applications MTC , telles que le controle de la pression dans olé oduc , la sé curité routiè re dans le systeme de transport intelligent (STI) , et virtuelle et de ré alité

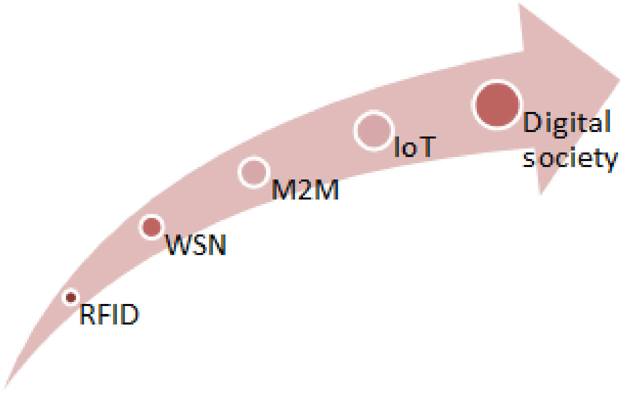
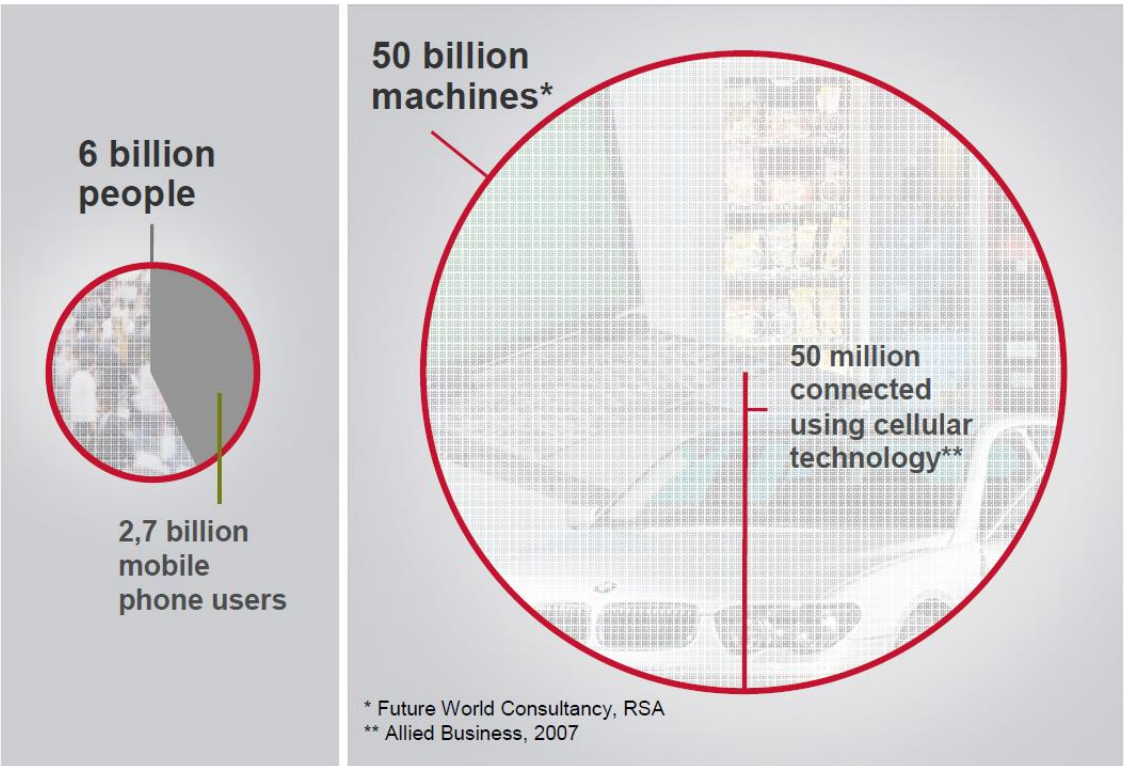


Figure A.1: Evolution des communications de type de machine


6 billion people
2,7 billion mobile phone users
50 billion machines*
50 million connected using cellular technology**

* Future World Consultancy, RSA
** Allied Business, 2007

Figure A.2: Potentiel de MTC

augmenté e , né cessite trè s peu de temps de latence , ce qui pourrait etre beaucoup plus faible que pour le trafic de voix et Internet classique .

- nombreux appareils MTC sont alimenté s par batterie . Pour ce type de dispositif MTC , faible consommation d'é nergie est extrêmement important .

- Dans certains cas , il ya des transmissions simultané es massives pour certaines applications MTC (par exemple, de la sé curité publique et la surveillance á distance), qui peuvent encourir é norme de signalisation et / ou au-dessus de la circulation pour les ré seaux cellulaires . Par exemple , les ré seaux du futur doivent prendre en charge jusqu'á 30 000 appareils MTC dans une cellule , ce qui est des ordres de grandeur plus aux exigences d'aujourd'hui [3].

- La charge utile de donné es est gé né ralement faible pour certaines applications MTC (par exemple, les compteurs intelligents et de suivi de cargaison) . La prestation efficace de ces paquets de petite taille est cruciale pour sauver des ressources du spectre pré cieux.

Ces caractéristiques et exigences imposent de grands défis pour les actuels cellulaires (UMTS , LTE , GSM) . Par conséquent , d'autres améliorations et optimisations sont nécessaires afin d'améliorer l' performances pour les applications MTC . 3GPP , ETSI , et d'autres associations (par exemple du forum WiMAX et WiFi Alliance) travaillent sur les normalisations pour MTC . 3GPP a publié plusieurs spécifications et rapports (par exemple 3GPP TS 22,368 , 37,868 et 3GPP TR 3GPP TR 36,888) pourétudier et analyser les applications MTC dans les réseaux cellulaires (GSM , UMTS , LTE) . L'Union européenne (UE) de la Commission a accordé plusieurs projets de développement de MTC , par exemple lola (http://www.ict-lola.eu/) , smartsaintander (http://www.smartsantander.eu/) , sensei (http :/ / www.sensei-project.eu/) , et exalté (http://www.ict-exalted.eu/) . Certains fournisseurs , par exemple Ericsson et Huawei , créent leurs propres visions , programmes et initiatives , tels que " 50 milliards de connectés les dispositifs " d'Ericsson , à stimuler le développement de MTC portefeuille [5]- [6]. En outre , comment accueillir MTC estégalement considéré comme un sujet important dans les futurs réseaux cellulaires (5G) .

L'objectif de cette thèse est de concevoir et d'optimiser l' accès au canal de liaison montante et la réception de liaison descendante pour les applications MTC en LTE / LTE -A . Plus précisément , pour MTC transmission en liaison montante , nous vous proposons tout d'abord un paquet méthode d'agrégation de réduire le taux de perte de paquets et la puissance ainsi que d'un système de regroupement TTI pour réduire la latence d'accès canal de liaison montante . En outre , pour les appareils MTC qui sont en liaison montante synchronisé , nous vous proposons un accès argument fondé de la méthode (CBA) qui réduisent encore le temps de latence d'accès de liaison montante . Pour la réception de liaison descendante MTC , nous proposons deux méthodes pour analyser la réception discontinue (DRX) de mécanisme en LTE . Avec notre modèle le paramètre DRX optimale qui permet d'économiser l'énergie au maximum tout en respectant la contrainte de délai peut être sélectionné.

A.2 Optimisation de l'accès alé atoire

Dans LTE, procé dure d'accès alé atoire est principalement utilisé lors de la connexion RRC mis en place, la synchronisation de liaison montante, transfert entre cellules, et demande de planification. Il existe deux types d'accès alé atoire: accès argument fondé initié es par UES et sans contention accès alé atoire coordonné par ENB. Accès alé atoire est crucial de soutenir efficacement le type de machine dans la communication LTE. La principale raison est expliqué e ci-dessous.

- accès alé atoire est principalement utilisé pour envoyer la demande d'ordonnancement (SR) de eNB que la mé thode habituelle de planification de canal de liaison montante n'est pas efficace pour l'application MTC .

  Plusieurs problèmes se posent lorsqu'on utilise un mé canisme d'ordonnancement de liaison montante ré gulière pour les communications de type de machine. Tout d'abord, pour assurer la transmission sans collision pour SR , eNB ré serve des ressources pour chaque UE à certains sous-trames . Pour le trafic MTC sporadique , comme mé canisme de ré servation de ressources devient inefficace en raison de la nature sporadique de modèle de trafic MTC . Par exemple : en supposant que la pé riode pré vue pour un trafic MTC uniformé ment ré partie est de 500 ms , si le eNB ré serve ressource SR pour que le trafic dans chaque sous-trame, seulement 1 500 de la ressource est utilisé e , ce qui entraîne un gaspillage important des ressources . D'autre part , si nous voulons sauver la ressource , l' allocation des ressources pé riodicité pourrait être ré glé que 500 ms , ce qui indique qu'une UE doit attendre 250 ms en moyenne pour accé der à la ressource . Deuxièmement, SR pé riode augmente avec le nombre d' UE (CRR relié) dans une cellule. Dans LTE , le montant maximum de ressources pour la transmission SR dans une sous-trame est de 36 . En supposant qu'il ya 1000 appareils MTC dans une cellule , le dé lai SR augmente c'est $1000 - 1036 =$ 28 ms, ce qui augmente é galement la latence . Par l'utilisation de l'accès alé atoire , une UE envoie un pré ambule sur la ressource commun de demander des ressources de eNB , ce qui é limine les problèmes de la programmation ré gulière .

- Certains appareils MTC peuvent ré sider dans l' é tat RRC IDLE la plupart du temps pour sauver la signalisation des frais gé né raux et la consommation d' é nergie . Pour ces types d'appareils MTC , ils utilisent un accès alé atoire pour envoyer une demande de connexion RRC à eNB tels que connexion RRC peut être é tabli entre les UE et eNB .

Par consé quent, on peut voir que l'acc è s alé atoire est essentiel pour les communications de type de la machine à LTE , et jouent un rôle essentiel dans la performance des applications ré alisables de MTC .

Cependant, il ya plusieurs dé fis à utiliser un acc è s alé atoire pour les communications de type machine à LTE . Il est connu que le pré ambule utilisé pour un acc è s al é atoire n'est pas UE spé cifique , par consé quent, la collision se produit lorsque plusieurs é quipements utilisateurs en utilisant le même pré ambule . Ainsi , le taux de collision devient tr è s é levé e quand il ya un nombre é norme d'UEs dans la cellule d'accé der au canal simultané ment . Par exemple, supposons que le nombre total d' UE dans une cellule est de 1000 , la probabilité de transmission de paquets pour un UE dans une sous-trame est de 0,03 et le nombre disponible de pré ambule est de 64 , alors la probabilité de collision est 0,9997 , ce qui indique que (1) le plus pré ambules ne peuvent pas être reçus par eNB et (2) é norme latence est introduit . Le deuxi è me probl è me est que un UE doit gé né ralement attendre certain laps de temps avant de commencer un autre acc è s alé atoire si l'acc è s alé atoire initiale é choue , ce qui augmente considé rablement le temps de latence . Par exemple , en supposant que la taille de la fenêtre de temporisation est de 50, la duré e de temporisation moyenne est de 25 ms avant de commencer le prochain acc è s alé atoire , qui à son tour introduit une latence supplé mentaire.

Il ya eu de nombreuses œuvres de la litté rature portant sur la performance d'acc è s alé atoire . Ré fé rence [52] pré sente un sché ma d'allocation des ressources pour l'acc è s multi-groupe spatiale alé atoire dans LTE . Auteurs dans [53] enquêter sur la probabilité de la mé thode d'acc è s alé atoire utilisé pour l'application MTC collision et d'offrir un mod è le pour calculer la probabilité de collision , la probabilité de succ è s , et les probabilité s d'inactivité des UE . Ré fé rence [54] propose un des syst è mes de contrôle d'acc è s et d'allocation des ressources communes massives pour ré duire au minimum la consommation d'é nergie totale du syst è me M2M sé lective foisé vanouissement plat et la fré quence dé coloration canal . Auteurs dans [55] proposer une mé thode de ré solution de collision pour un acc è s alé atoire sur la base de l'informationà l'avance de synchronisation fixe pour massif emplacement fixe MTC en LTE . Ré fé rence [56] introduit une nouvelle mé thode de code -é tendu pour un acc è s alé atoire en LTE . Avec la mé thode proposé e, la quantité de ressource disponible de contention esté largi , ce qui peut donc ré duire le taux de collisions en acc è s alé atoire. Ré fé rence [57] instaure un ré gime d'interdiction de classe d'acc è s coopé rative pour la stabilisation mondiale et partage l'acc è s de charge . Dans leur mé thode , chaque groupe MTC sont affecté sà la classe d'acc è s spé cifique interdisantà diffé rencier des priorité s d'acc è s . Auteurs dans [58] proposer une mé thode de gestion de l'acc è s massifà fournir une garantie de qualité de service pour les appareils MTC . Dans cette mé thode , les UE sont regroupé es en raison de leurs (QoS) exigences de qualité de service : EI avec exigence de QoS dé terministes sont ré servé s aux ressources tout en EI avec QoS douces garanties sont pré vues de façon opportuniste pour amé liorer l'efficacité des ressources d'utilisation . Ré fé rence [59] analyse le dé bit et la latence d'acc è s pour un acc è s alé atoire dans un syst è me OFMDA . Cependant, les utilisateurs de retransmission ne sont pas considé ré s dans leur mod è le , qui ne se conforme pas avec le mé canisme d'acc è s alé atoire dans LTE [60]. Auteurs dans [61] pré senter un plan d'acc è s alé atoire en priorité à fournir la qualité de service pour les diffé rentes classes de dispositifs MTC dans les ré seaux LTE - A, où les diffé rentes priorité s d'acc è s est ré alisé e en utilisant des procé dures de temporisation diffé rentes . Ré fé rence [3] examine les solutions possibles pour le contrôle de la surcharge de l'acc è s alé atoire , qui comprend : des syst è mes de classe d'acc è s interdisant , ressources RACH sé paré pour MTC , l'allocation dynamique de ressources RACH , MTC syst è me de ré duction de puissance spé cifique , fendue acc è s , et tirez ré gime fondé . Toutefois , aucune mé thode dé taillé e est fournie là .

Pour améliorer la fiabilité de l'accès aléatoire , nous proposons une méthode d'agrégation de paquets d'accès aléatoire. Avec notre méthode , un UE ne démarre pas un accès aléatoire pour chaque paquet est arrivé . Au lieu de cela , il déclenche un accès aléatoire lorsque le nombre de paquets dans la mémoire tampon atteint un certain seuil . Dans l'exemple décrit ci-dessus si l'on fixe le seuil de l'agrégation de paquets de 5 , alors la probabilité de collision est réduite à 0,21 , ce qui est bien inférieur à la valeur d'origine 0,9997 . A noter que plus le seuil de l'agrégation de paquets , plus le taux de collisions devient . Toutefois , le taux de collision préambule est réduite , au détriment du temps d'attente supplémentaire utilisé pour tamponner les paquets , ce qui peut ne pas être souhaitable pour certaines applications MTC en temps réel. Afin d'éviter latence importante pour les applications MTC en temps réel , nous tirons d'abord le taux de perte de paquets et accès canal de latence en fonction du montant de paquet agrégé en utilisant un modèle de processus semi- Markov [62] . Avec les résultats obtenus , la quantité optimale de paquets agrégés qui maintient l'exigence de taux de perte de paquet tout en minimisant le temps de latence peut être sélectionné. Il doit être mentionné que l' autre avantage d'utiliser la méthode d'agrégation de paquets est de textit économiser l'énergie en réduisant le nombre de transmissions d'accès aléatoire . Nous proposons également un procédé d'économie d'énergie , où un UE agrège les paquets aussi longtemps que la contrainte de délai est satisfaite.

Pour réduire le temps de latence causée par accès aléatoire échoue, nous vous proposons un intervalle de temps de transmission (TTI) régime regroupement . L'idée de TTI regroupement est introduit dans LTE Rel . 8 à améliorer la couverture de liaison montante pour application VoIP [63]. Dans cette méthode [63] , un UE envoie un paquet de voix sur IP à travers un faisceau de plusieurs TTI ultérieur avant de recevoir le HARQ du eNB , ce qui supprime le temps d'attente causé par des retransmissions et améliore la qualité de service pour les applications VoIP [**?**] . Inspiré par cela, nous appliquons l'idée de TTI regroupement à accès aléatoire . Avec le système de regroupement TTI proposé , un UE envoie plusieurs préambules dans plusieurs ITT / sous-trames consécutives . Par conséquent, un accès aléatoire est réussie si l'un des préambules est correctement recu par eNB et sans collision , ce qui élimine la latence due à accès aléatoire sans succès. Il est évident que la réalisation de l'accès à faible latence de canal de liaison montante par l'utilisation de la méthode de mise en faisceau TTI est bénéfique pour des applications MTC en temps réel , par exemple la surveillance d'huile / de gazoduc , d'alarme incendie à base de capteurs et détection de collision de l'automobile.

A.2.1 Agré gation des paquets de type machine Communications avec Random Access

Le principal problème de l'accès alé atoire est que la collision devient très é levé e s'il existe d'é normes UEs dans une cellule. Pour ré soudre ce problème , nous proposons une mé thode d'agré gation de paquets . Avec notre mé thode , un UE ne dé clenche pas un accès alé atoire jusqu'à ce que le paquet tampon atteint le seuil donné (Ce seuil est fixé par eNB et envoyé à chaque UE) . Comme le nombre de transmission est ré duite gr âce à l'agré gation de paquets , le taux de collision pré ambule est é galement diminué . Cependant, cette mé thode pré sente un temps de latence supplé mentaire afin d'accumuler plusieurs paquets. De plus , le taux de perte de paquets peut être augmenté en raison de la taille du paquet a augmenté , ce qui augmente é galement le temps de latence . Toutefois , pour des raisons de simplicité , nous ignorons cet effet dans notre mé thode .

Ici, nous utilisons un modèle de processus semi- Markov pour analyser notre problème . Avec cette mé thode , on peut calculer le taux de perte de paquets et la latence en fonction du nombre de paquets agré gé s. Ensuite , on sé lectionne le paquet d'agré gation nombre optimal qui minimise le taux de perte de paquet tout en maintenant l'exigence de retard . Voici des ré sultats par l'utilisation de la mé thode proposé e.

Le seuil du taux de perte de paquets est que $0,1$ et le nombre maximum de paquets agré gé s est 50 . L'intervalle de paquet moyenne est de 100 ms , 200 ms et 500 ms , ce qui est beaucoup plus grande que la taille de la fenêtre de temporisation .

Fig. A.3 montre la quantité de paquets agré gé s sous le numé ro diffé rent d'UE et le taux d'arrivé e des paquets $lambda$ (paquets / ms) lors de l'utilisation de notre mé thode proposé e . On peut voir que la quantité de paquets non- agré gé es diminue avec l'augmentation du taux d'arrivé e des paquets ou le nombre d' UE . Cela est raisonnable car les augmentations de taux de collision avec un taux d'arrivé e des paquets ou le numé ro d' UE . Si la vitesse de collision est supé rieure au seuil donné , l'agré gation de plusieurs paquets est né cessaire pour abaisser la collision . Dans le cas contraire , la quantité de paquets agré gé peut pas être augmenté e. Par exemple, lorsque λ = 1/100 et le nombre d' UE est 2000 , la quantité de paquets agré gé est 2 et le taux de perte de paquets est de 0,093 ce qui est très proche de l' seuil de 0,1 . Par consé quent, lorsque le nombre d' UE augmente à 2500 , le nombre de paquets agré gé s augmenter à trois , ce qui ré duit le taux de perte de paquet à 0,07 (Fig. A.3) . Nous remarquons é galement que le nombre d'agré gation de paquets est toujours 1 lorsque $\lambda = 1/500$. La raison en est que le taux de perte de paquet est toujours infé rieure au seuil (0,1) lorsque le nombre d'augmentations UE . Par exemple , le taux de perte est de 0,05 lorsque $\lambda = 1/500$ et le nombre d' UE est 3500 . Nous constatonsé galement que le nombre d'agré gation de paquets sont diffé rents même si le nombre de nouveau moyen de diffusion sont les mêmes .

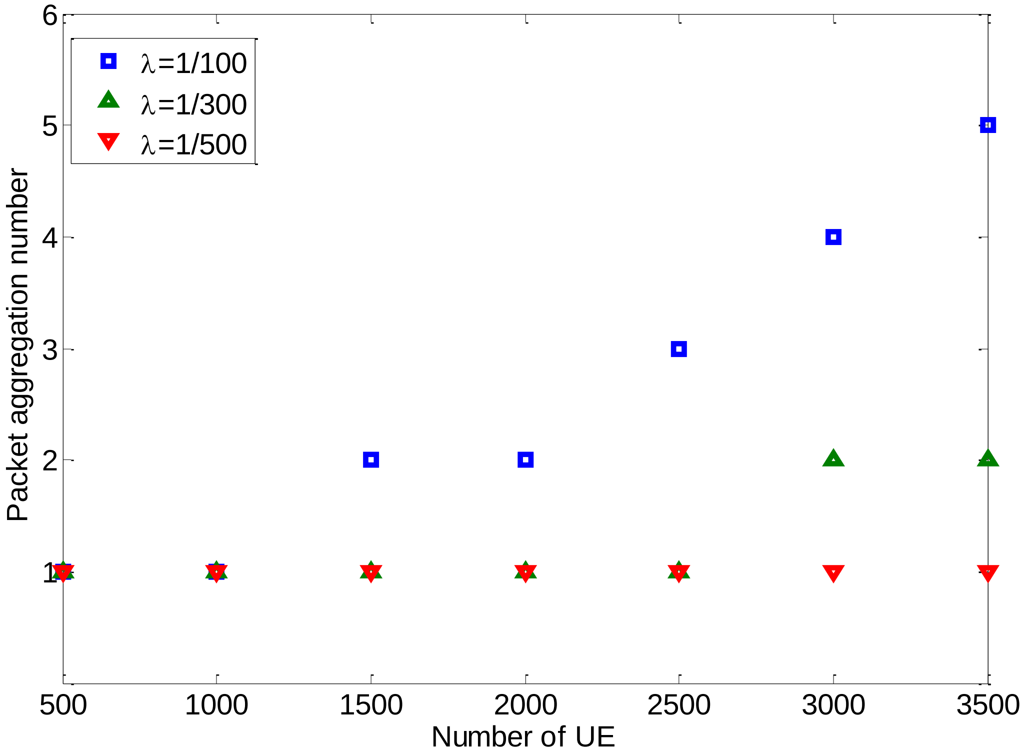


Figure A.3: Montant de paquet agrégé

Fig. A.4 démontre le taux de perte de paquets en utilisant les résultats de l'agrégation de paquets indiqués sur la Fig. A.3. Il peut être vu que par notre procédé, le taux de perte de paquets est inférieur au seuil de taux de perte de paquets (0,1), ce qui confirme notre procédé. En revanche, sans agrégation de paquets, le taux de perte de paquets est élevé surtout quand $\lambda = 1/100$ et le nombre d'UE est supérieure à 2000.

Comme discuté ci-dessus, on baisse le taux de perte de paquets au détriment de la latence augmentation. Fig. A.5 compare la latence d'accès de canal avec ou sans agrégation de paquets. Nous pouvons voir que le temps de latence est augmentée lors de l'utilisation d'agrégation de paquets. Par exemple, le temps de latence est augmenté de 110 ms à 193

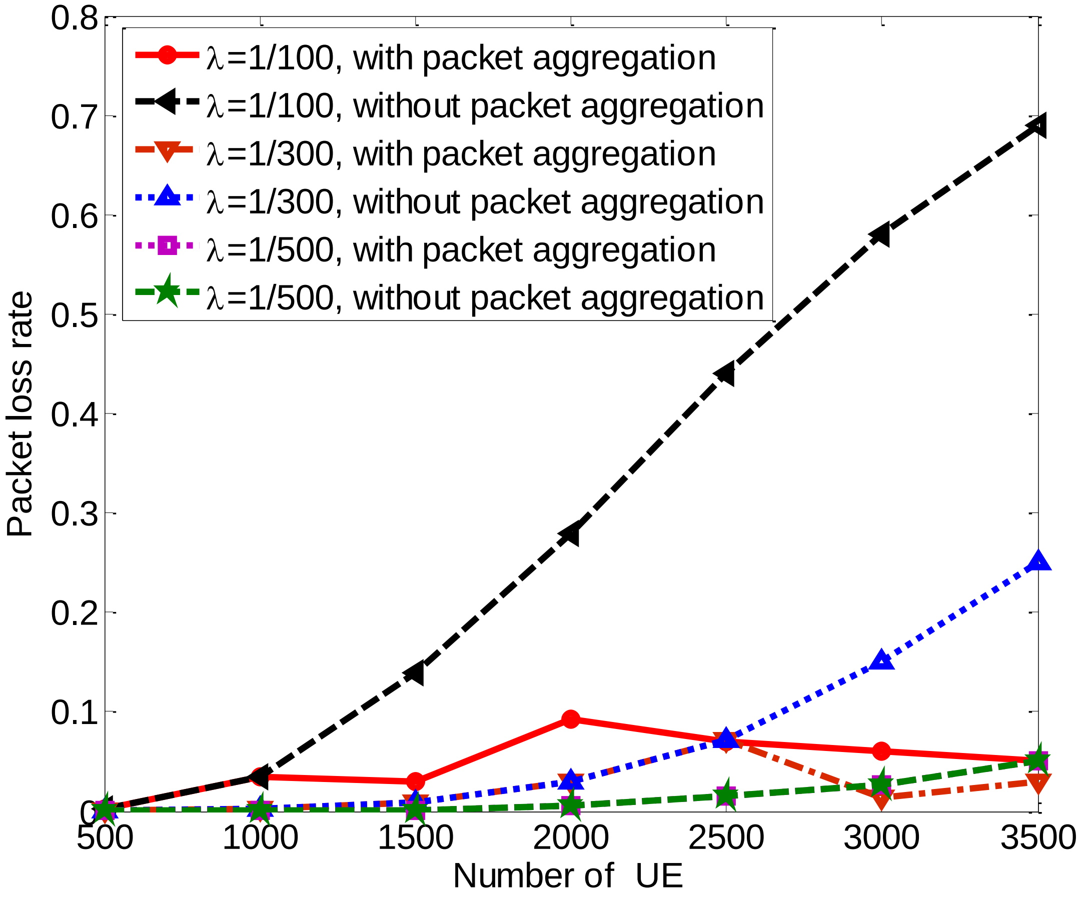


Figure A.4: Taux de perte de paquets

ms lorsque $\lambda = 1/100$, le nombre d'UE est 2000, et le nombre d'agrégation de paquets est égal à 2. On peut imaginer que si la contrainte de délai pour cette application MTC est de 150 ms, puis en agrégeant deux paquets n'est pas possible. Par conséquent, pour application en temps réel MTC, si la latence entraîné est plus grande que retarder contrainte après l'utilisation de l'agrégation de paquets, plus de préambules et / ou ressources PRACH devraient être alloués par eNB. Alors que pour une application MTC avec contrainte de délai élastique, par l'utilisation de l'agrégation de paquets, le taux de perte de paquets peut être réduite à une valeur très faible.

A.2.2 TTI Regroupement pour machine de type communications avec Random Access

Un UE doit attendre pour certain laps de temps avant de commencer un nouvel accès aléatoire si l'accès aléatoire précédente échoue, ce qui introduit une latence d'accès au canal grand (peut être inacceptable pour certains temps-réel MTC) et augmente la consommation d'énergie comme un UE a également passer plus de temps à l'état actif . Pour réduire la latence d'accès de canal d'accès aléatoire , nous proposons un schéma de regroupement TTI

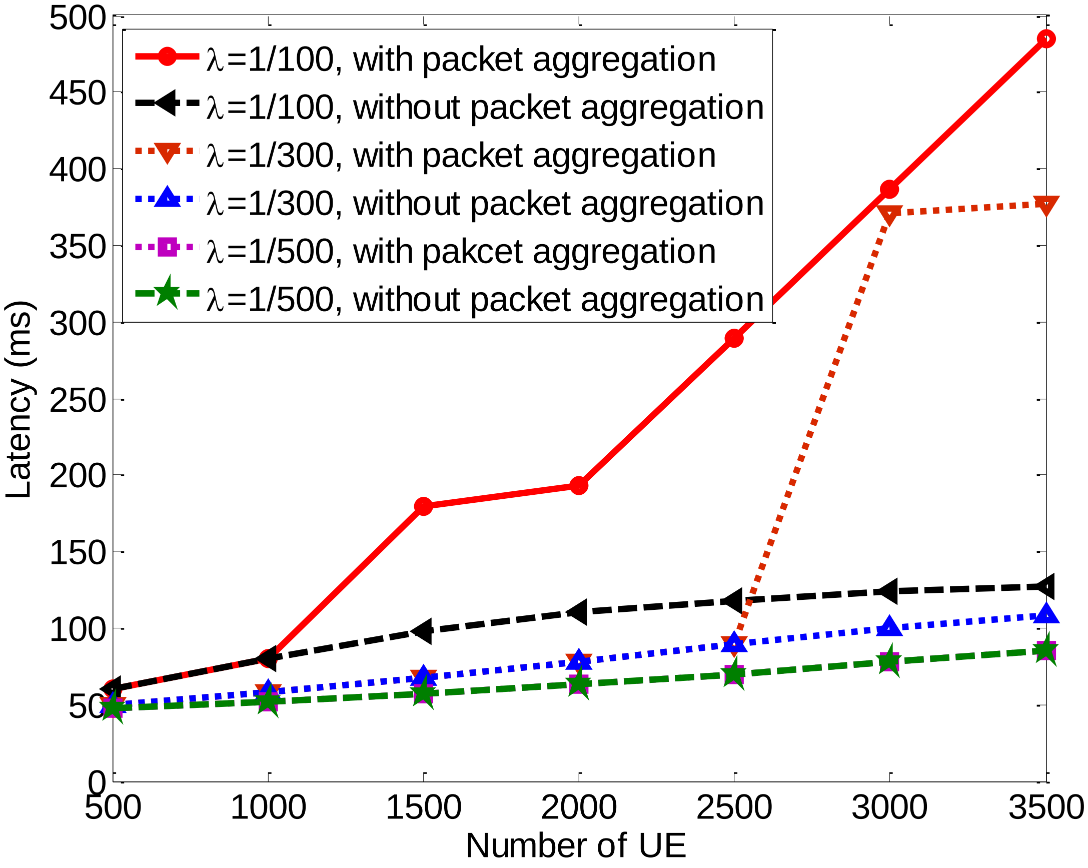


Figure A.5: latence

comme indiqué sur la Fig.A.6. Avec le système proposé , un UE envoie plusieurs préambules choisis au hasard dans les sous-trames consécutives d'effectuer plusieurs tentatives aléatoires, qui est appelée TTI regroupement pour un accès aléatoire . Alors, pour tout préambule recu correctement , le eNB envoie un message de RAR . Troisi èmement , une UE envoie le message L2/L3 avec la ressource allouée (si un UE est alloué avec une quantité multiple de ressources , il suffit d'utiliser l'un d'eux) . Enfin, le reconnait eNB le message L2/L3 correctement reç u par le message de résolution de conflit . Ici, nous considérons l' accès aléatoire multiple en sous-trames consécutives comme un tour d'accès aléatoire . Il est évident que si l'un de ces préambules dans un tour d'accès aléatoire est correctement reç u par eNB et sans collision , le long de l' accès aléatoire est réussie , qui *élimine le temps qu'une UE doit attendre pour commencer un nouvel accès aléatoire après une vive échoué* .

Il semble que l'augmentation du nombre de groupage TTI donne la probabilité de succès plus élevée pour un tour à accès aléatoire et permet ainsi de réduire la latence . Cependant, ce n'est pas toujours vrai. En fait , le préambule hausses de taux de collision avec le nombre de regroupement ITT puisque chaque UE doit déclencher plusieurs transmissions , qui peuvent à leur tour réduire la probabilité de succès d'un cycle de l'accès aléatoire . Par conséquent, le nombre de groupement TTI doit être soigneusement choisie de telle sorte que la probabilité de succès pour un accès aléatoire ronde peut être maximisée , et donc le temps de latence d'accès est minimisée.

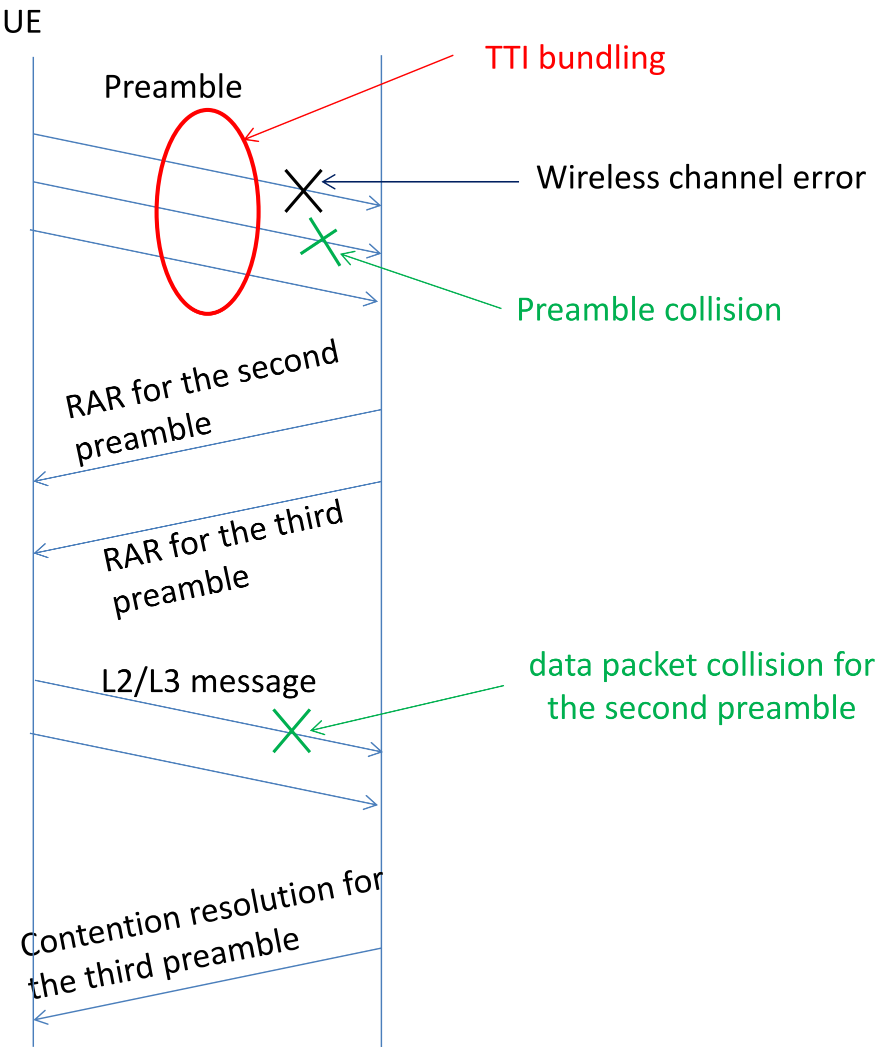

UE
Preamble
TTI bundling
Wireless channel error
Preamble collision
RAR for the second preamble
RAR for the third preamble
L2/L3 message
data packet collision for the second preamble
Contention resolution for the third preamble

Figure A.6: argument fondé accès alé atoire avec TTI regroupement

Pour trouver le numéro de regroupement TTI optimale qui minimise la latence d'accès, un modèle de processus semi-Markov est proposé. Sur la base de ce modèle, nous tirons la latence d'accès de canal en fonction du nombre de groupage ITT. Par conséquent, la valeur optimale qui minimise la latence d'accès de canal peut être facilement sélectionnée.

Voici des résultats par l'utilisation de la méthode proposée.

Fig. A.7 dé montre le nombre optimal de regroupement TTI sous le numé ro diffé rent de taux d'arrivé e UE et paquet. Nous pouvons voir que les TTI optimales numé ro de regroupement non-augmente à mesure que le nombre d'UE et paquet arrivé e des hausses de taux. La raison de ce phé nomène est la suivante: les pré ambule hausses de taux de collision avec un plus grand nombre de UE et le taux d'arrivé e des paquets. Par consé quent, lorsque le taux nombre UE ou paquet arrivé e devient grand, un UE devrait regrouper plus petit (ou même) nombre de TTI pour é viter l'augmentation du taux de collision. Par ailleurs, on constate é galement que le nombre de groupement TTI de taux d'arrivé e des paquets est infé rieure plus grande que celle de la plus é levé e. C'est raisonnable, car les paquets hausses de taux de collision avec un taux d'arrivé e des paquets. Par consé quent, plus petit regroupement TTI devrait être utilisé pour é viter servir collision lorsque les hausses de taux d'arrivé e des paquets.

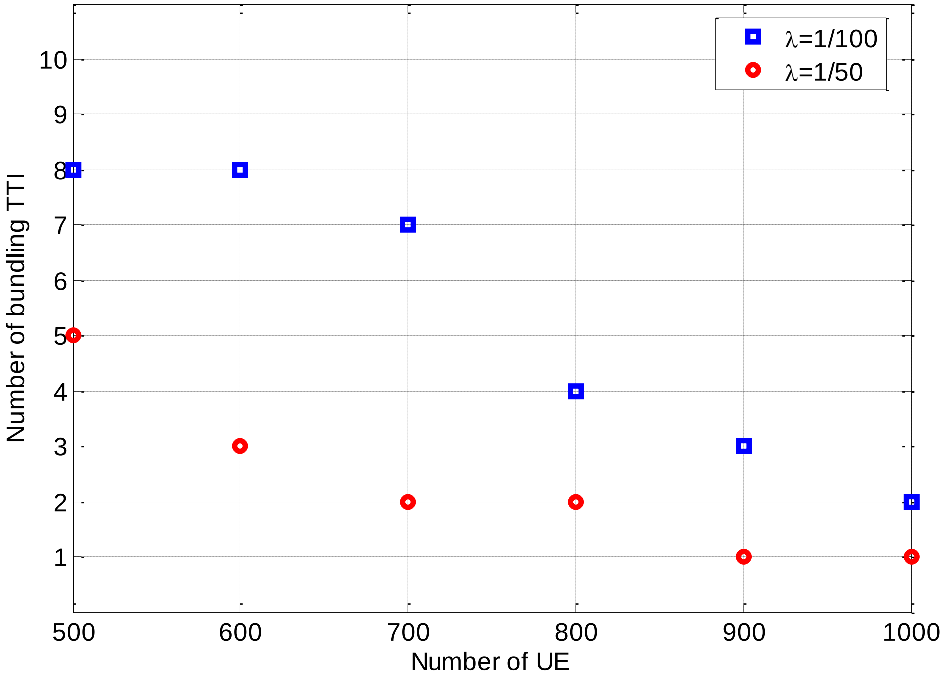


Figure A.7: Nombre optimal de regroupement TTI

Fig. A.8 compare le temps d'attente obtenue en utilisant les ré sultats montré s à la Fig. A.7 à la latence sans utiliser TTI regroupement . Nous constatons que de TTI regroupant la latence est considé rablement ré duite lorsque le nombre d' UE est moins de 900 pour $\lambda = 1/100$ ou le nombre d' UE est infé rieur à 600 pour $\lambda = 1/50$. Concrètement , la latence est ré duite de 46 % (61ms à 33 ms) au maximum quand $\lambda = 1/100$ et le nombre d' UE est de 500 . En revanche, le gain de groupage en utilisant TTI devient plus petite pour les autres cas. La raison de ce phé nomène a deux aspects : (1) lorsque le taux de collision pré ambule n'est pas très é levé , de regrouper plusieurs ITT augmente considé rablement le taux de ré ussite pour l'accès alé atoire . Par exemple , le taux de collision est $0, 22$, le premier accès alé atoire taux de ré ussite ronde est $0, 50$ et le temps de latence est de 61 ms quand $\lambda = 1/100$, $n = 1$, et $N_u = 500$. Lorsque le TTI regroupement nombre n augmente à 8 , si le taux de collision augmente à $0, 36$, le premier accès alé atoire succès ronde é quivaut à environ 1 ce qui ré duit considé rablement le temps de latence à 33 ms comme le montre la Fig.A.8. (2) Lorsque le taux de collision pré ambule est é levé , regroupant plusieurs ITT augmente le taux de collision pré ambule à un niveau très é levé . En consé quence, l' accès alé atoire rond taux de succès n'est pas sensiblement amé lioré e , et donc le temps de latence n'est pas fortement ré duite. Par exemple , le taux de collision est $0, 48$; la première à accès alé atoire taux de ré ussite ronde est $0,33$ et le temps de latence est de 85 ms quand $\lambda = 1/50$, $n = 1$, et $N_u = 800$. Lorsque le TTI regroupement nombre n augmente de 2 , le taux de collision pré ambule saute à 0,65 et le taux de ré ussite pour le premier accès alé atoire ronde augmente seulement $0,45$ qui ré duit lé gèrement la latence à 81 ms comme indiqué sur laFig.A.8.

Par consé quent, pour ré duire le temps de latence d'accès au canal pour un ré seau où le taux de collision pré ambule est é levé , nous devrions tout d'abord de ré duire le taux de collision pré ambule (il peut être atteint en allouant plus de pré ambules) . Deuxièmement , nous appliquons le système de regroupement TTI .

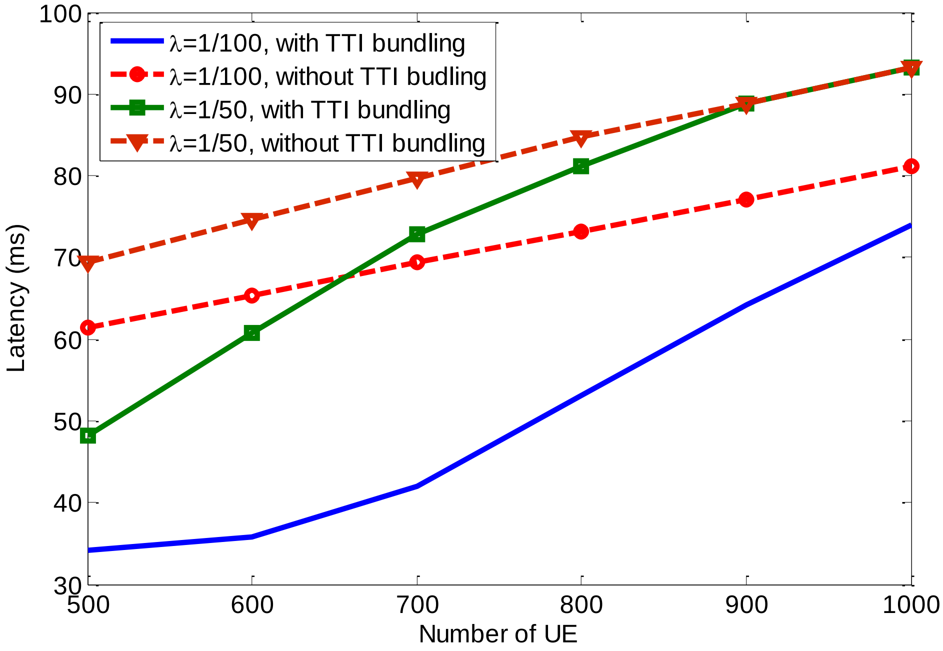


Figure A.8: Latence comparaison avec et sans regroupement TTI

A.3 Accès basé Contention

A.3.1 General Idea pour l'accès argument fondé

Pour fournir un accès au canal de liaison montante à faible latence pour MTC sur LTE , une nouvelle mé thode d'allocation des ressources , appelé affirmation accès basé (CBA) , est proposé . La principale caracté ristique de l'CBA est que le eNB n'alloue pas de ressources pour une UE spé cifique . Au lieu de cela , la ressource alloué e par l' eNB est applicable à tous les ou un groupe d'UE et de tout é quipement UE qui a des donné es à transmettre sé lectionne au hasard des blocs de ressources parmi les ressources disponibles (Fig.A.9) . La procé dure d'accès de liaison montante de contention base est repré senté sur la Fig.A.10 (Ici, nous supposons que l'UE est en liaison montante synchronisé e . Tout d'abord , l' UE reçoit l' allocation des ressources informations qui indique la ressource alloué e pour ABC . En supposant que la ressource de l'ABC est disponible en chaque sous-trame , une attente UE pour 0,5 ms pour recevoir la subvention de planification (SG) de l'information pour l'ABC . Puis, après le dé codage de la ré partition des ressources informations qui coûte 3 ms , l'UE envoie la trame sur les ressources sé lectionné es de manière alé atoire . Le temps de latence pour cet ensemble procé dure est de 7,5 ms dans le meilleur des cas , ce qui est beaucoup plus petit que celle de l' accès alé atoire (27 ms) . Dans un certain sens , l' accès argument fondé est similaire à la mé thode d'accès alé atoire norme : pour les deux mé thodes l'allocation des ressources est commune . Cependant , la mé thode d'accès de contention base a des caracté ristiques spé cifiques .

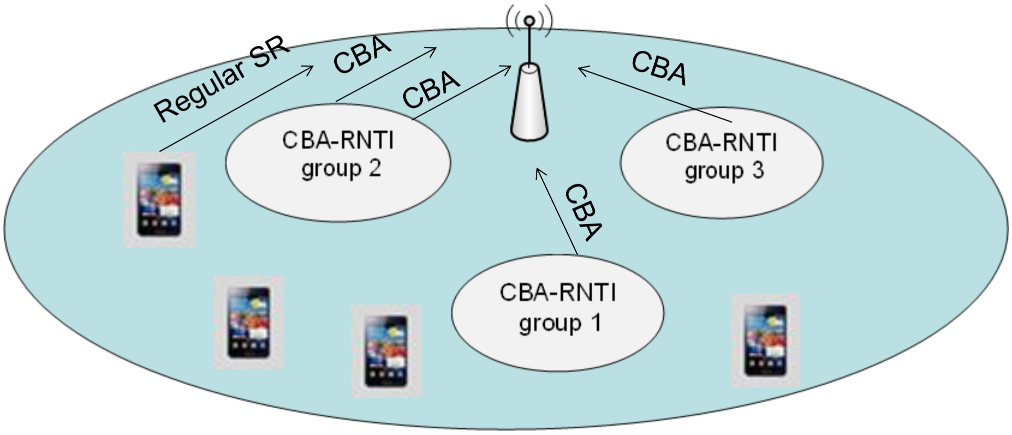


Figure A.9: Acc ès au canal de liaison montante avec acc ès argument fondé

Comme les ressources de l'ABC ne sont pas UE spé cifique , mais plutôt alloué pour la totalité ou un groupe d'UE , les collisions peuvent se produire lorsque plusieurs UEs sé lectionner la même ressource . Dans un ré seau avec trafic sporadique , la probabilité de collision est très faible, ce qui moyens la plupart des transmissions sont libres de collision et donc Mé thode ABC surpasse la mé thode d'ordonnancement ré gulier vue de latence. Cependant, dans un ré seau dense de la collision probabilité est très é levé e, ce qui signifie beaucoup de retransmission sont né cessaires et par consé quent le temps de latence est augmenté . par exemple en supposant que le bloc total des ressources disponibles dans une sous-trame est 50 , la probabilité de collision est de 0,06 , si trois UEs dans la transmettent sous-trame , et que la probabilité de collision augmente à 0,99 si 20 EI transmettent, dans le berceau .

Pour ré soudre le problème ci-dessus , le procé dé suivant est utilisé . Chaque UE envoie son identifiant , le ré seau C -radio identifiant temporaire (C -RNTI) , ainsi que les donné es sur le lieu choisi au hasard ressource . Etant donné que le C -RNTI est de très petite taille , par consé quent, il peuvent être transmis à la modulation plus robuste et le canal ré gime (MCS) de codage sans introduire de tête é norme . Par l' utilisation de la dé tection MU- MIMO , les CRNTIs hautement proté gé e pourraient être dé codé s avec succès, même si elles sont envoyé es sur la même ressource temps-fré quence. Sur le succès dé codage pour le C- RNTI collision , l' eNB dé clenche ré gulièrement planification pour les UE correspondantes comme indiqué dans la Fig.A.11 . Par consé quent , un UE peut retransmettre le paquet en utilisant la procé dure ré gulière HARQ . l' ensemble latence pour cette procé dure de planification tout n'est pas encore plus grand que celle de l' ordonnancement ré gulier .

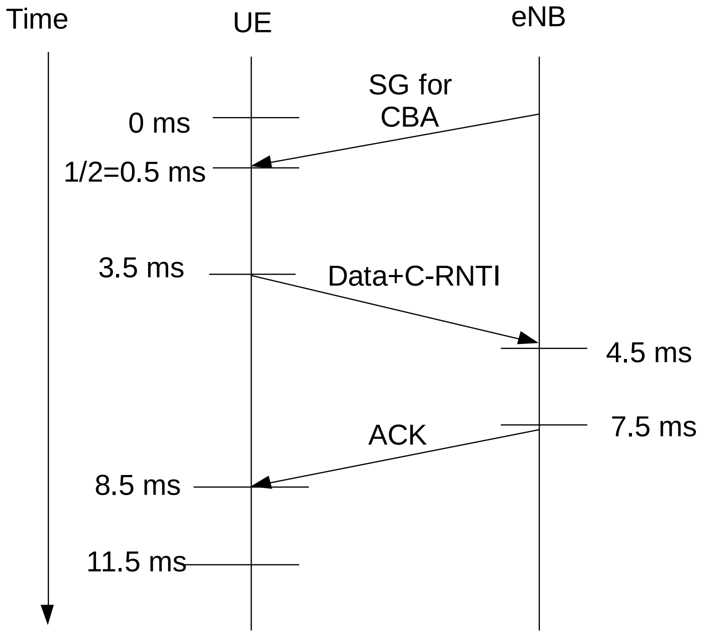


Figure A.10: L'accès de contention base

Pour les UE sont entré s en collision dont C- RNTI ne sont pas dé codé , ni ressource dé dié e (SG), ni les informations ACK est reçu ; . ces EI doivent retransmettre les paquets comme indiqué sur la Fig.A.12. Il a à noter que les retransmissions repose toujours sur ABC , qui est dé signé comme HARQ pour ABC comme elle est diffé rente de la procé dure HARQ ré gulière (Dans ré gulière HARQ , ressource dé dié e est alloué e pour une UE avec la retransmission) .

A.3.2 l'allocation des ressources pour la méthode d'accès argument fondé

L'objectif principal pour l'allocation des ressources est d'attribuer la quantité appropriée de ressources de sorte que les contraintes de latence sont remplies et que les ressources allouées sont utilisées efficacement. L'allocation des ressources précises pour l'CBA est très important car il est directement lié à la latence constatée par l'application trafic.

Pour chaque transmission d'accès argument fondé, nous avons ce qui suit événements:

1. ni les informations de contrôle, ni les données sont détectés, ce qui est noté que E_1;
2. l'information de commande n'est pas détecté, mais les données sont détectée, qui est désigné comme E_2;

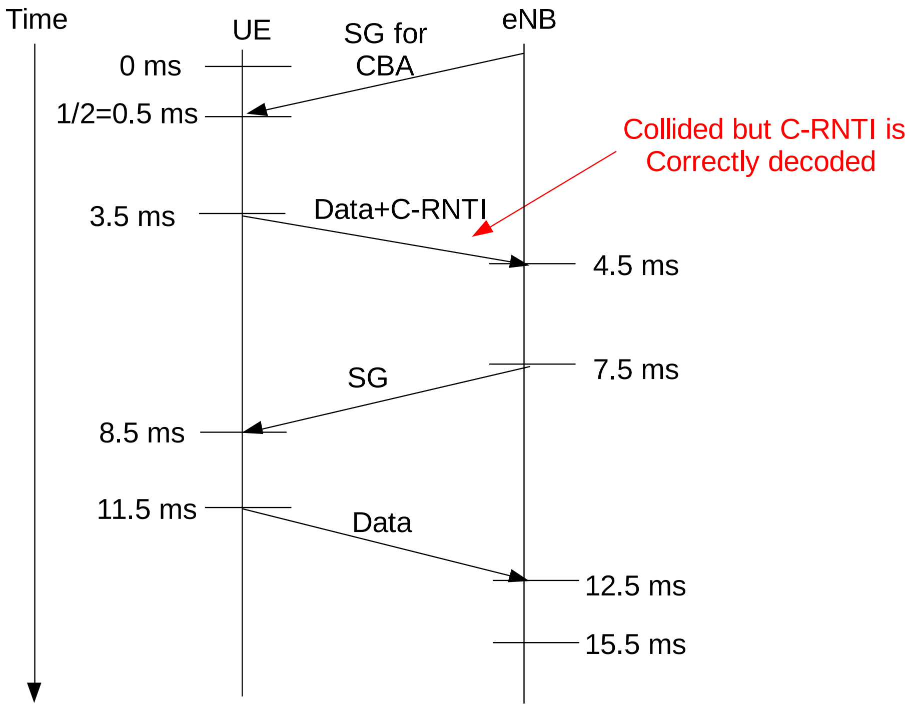


Figure A.11: L'accès de contention sur la base de la détection de collision

3. l'information de commande est détectée, mais la donnée n'est pas détecté, ce qui est notée E_3;

4. à la fois les informations de commande et de données sont détectées, qui est désigné comme E_4.

Afin de minimiser le temps de latence pour le trafic MTC, la ressource de l'CBA devrait être disponible dans chaque sous-trame. L'allocation des ressources peut être effectuée dans les étapes suivantes:

1. Définir l'unité de ressources de l'CBA

2. initialiser le montant de l'ABC unité de ressource à 1

3. Calculer les probabilités des quatre événements provoqués par une transmission de l'CBA.

4. Calculer le temps de latence en fonction de la quantité mesurée de l'ABC unité de ressource

5. Si la latence est estimée supérieure à la contrainte de latence, d'augmenter le montant de l'unité de ressource par un seul et retournez à l'étape 3. fin autre

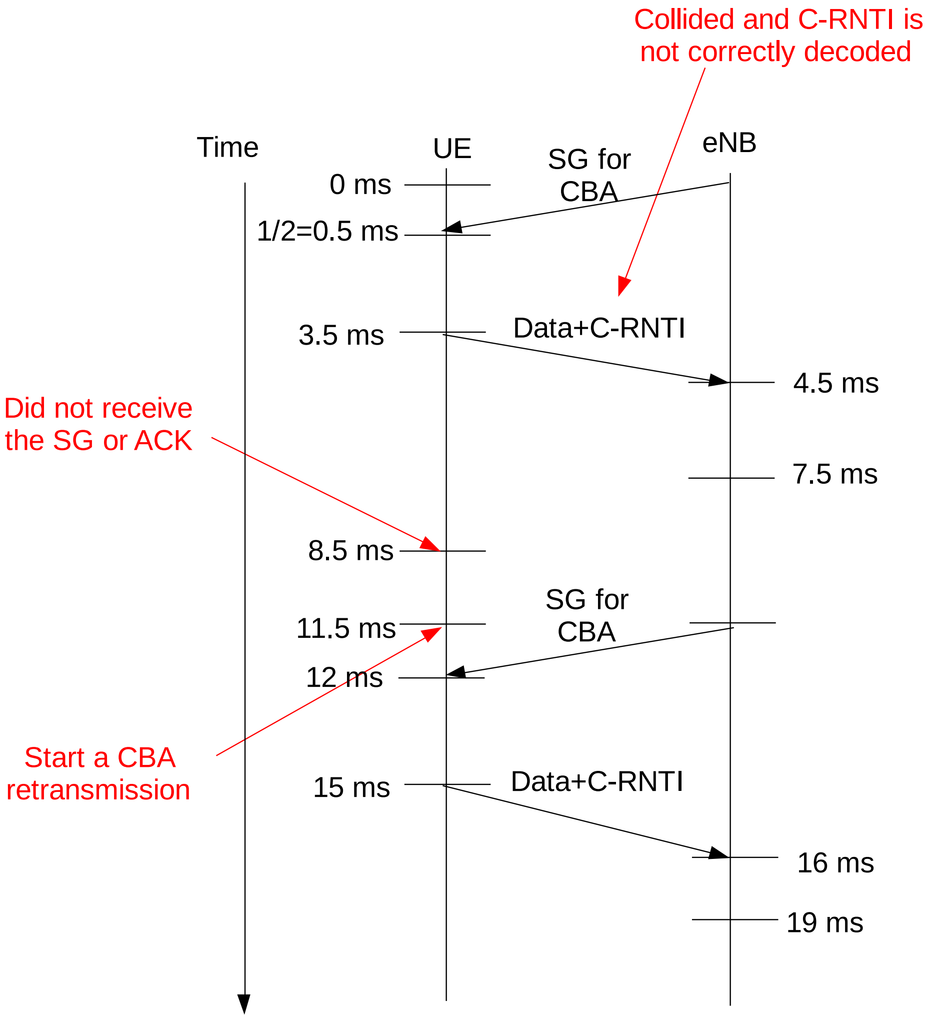


Figure A.12: L'accès de contention en fonction de retransmission

On peut voir que la latence diminue avec la quantité de CBA unité de ressource, donc avec la méthode ci-dessus, nous trouvons toujours le montant minimum de ressources de l'ABC. Il a à noter que nous supposons ici qu'il ya toujours assez de ressources. Pour un système qui a une contrainte sur les ressources de l'ABC, planificateur plus intelligent peut être utilisé pour résoudre ce problème, par exemple, un planificateur qui considèrent les priorités entre temps réel et les trafics de temps non réel.

Tout d'abord, afin de valider la thèse basée sur l'accès de la méthode (CBA) proposé, nous comparons le délai d'accès au canal de l'ABC avec celui de l'accès aléatoire (dénommé méthode PRACH). Nous comparons les performances des deux méthodes avec la même quantité de ressources. Concrètement, pour la méthode ABC, nous attribuons une unité de ressources de l'ABC contenant 6 blocs de ressources dans chaque sous-trame. Alors que pour le procédé PRACH, le préambule est défini en tant que 64 et l'indice de configuration de ressource PRACH est fixé à 14, qui occupe la même ressource que l'ABC (6 Des blocs de ressources dans chaque sous-trame) et c'est le permis ressource maximale pour PRACH en LTE. La limite de transmission pour un accès aléatoire est de 5, la taille de la fenêtre de réponse d'accès aléatoire est 10, et la résolution de contention temporisateur est de 24 ms.

Fig.A.13 montre les résultats de simulation. Nous pouvons voir que le temps de latence de l'CBA est beaucoup plus petite que celle de la méthode PRACH. Il montre que le gain de temps de latence d'utiliser l'ABC est d'environ 30 ms, ce qui confirme que l'CBA surpasse la méthode PRACH en terme de temps de latence.

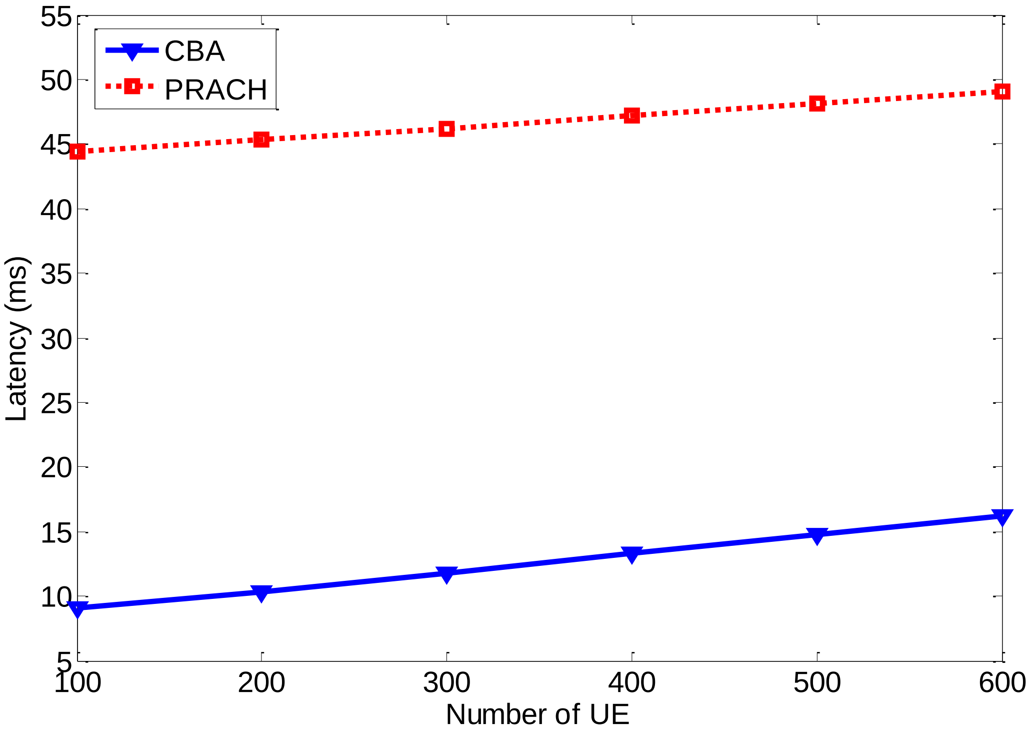


Figure A.13: Comparaison de latence

Fig. A.14 montre les résultats de l'allocation de ressources à l'aide de notre méthode proposée avec différents taux d'arrivée des paquets λ (paquets / ms) et le nombre d'UE lorsque la contrainte de délai est de 30 ms . Nous pouvons voir que les unités de ressources allouées non - diminution de l' augmentation du nombre d' EI et / ou le taux d'arrivée des paquets . C'est parce que les paquets hausses de taux de collision avec nombre d'UE et le taux d'arrivée des paquets , ce qui augmente donc la latence . Pour satisfaire la contrainte de délai , plus de ressources devrait être allouée . Par exemple, quand $\lambda = 1/30$ et le nombre d' UE de 300 , l'unité de ressources de l'ABC est un et le temps de latence est de 28,9 ms, ce qui est très proche des 30ms de seuil . Par conséquent, lorsque le nombre d' UE augmente à 400, de deux unités de ressources sont allouées ABC qui réduit le temps de latence de 18 ms . De même, lorsque le nombre d'UE atteint 600 , l'unité de ressource de l'ABC est porté à trois , et le temps de latence diminue à 18ms .

Fig. A.15 démontre le retard lors de l'utilisation du montant alloué de ressources Fig. A.14. Il peut être vu que le retard est inférieure à la contrainte de délai de 30ms, ce qui valide la méthode d'allocation des ressources proposées.

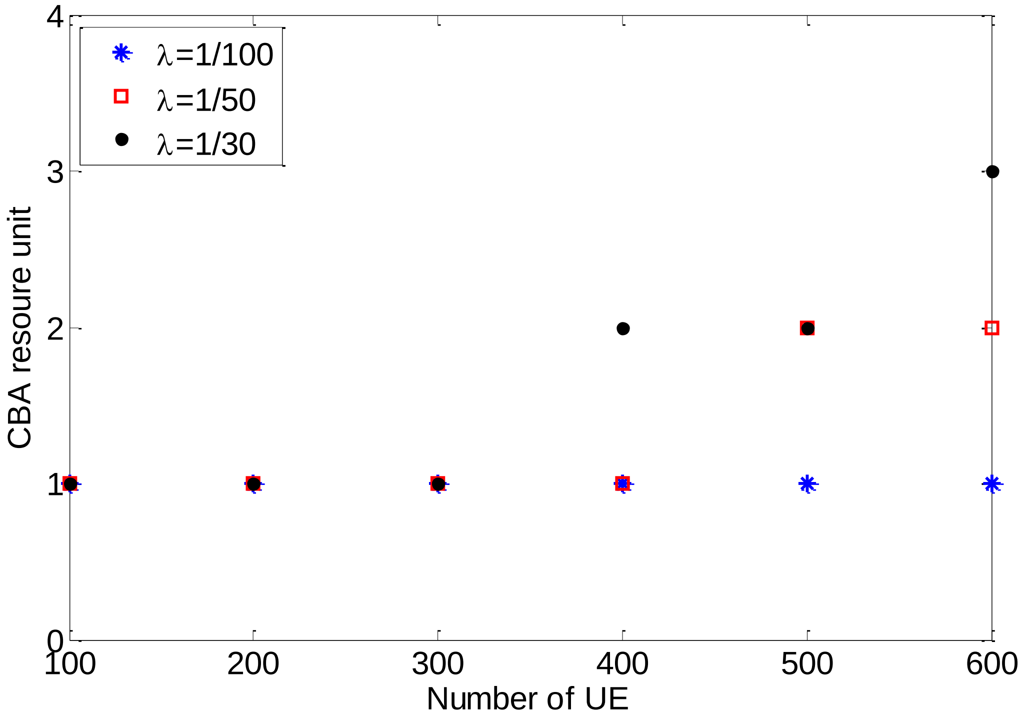


Figure A.14: CBA allocation des ressources pour un nombre différent d'UE

A.4 La modélisation de la réception discontinue et l'optimisation

Il est bien connu que la plupart des appareils MTC sont alimentés par batterie , par exemple des capteurs dans l'oléoduc , compteur d'eau intelligent , etc Par conséquent, l'abaissement de la puissance consommation, ce qui prolonge la durée de vie de l'appareil MTC et donc de réduire le coû t de déploiement , est parmi les primaires exigences . Pour ce faire, une réception discontinue (DRX) est utilisé dans la technologie LTE / LTE -A réseau. Avec DRX , un UE ne tourne que sur le récepteur à certains points de temps prédéfinis tout dort à d'autres . Si un paquet arrive à eNB mais l'UE cible dort , le paquet est tamponné à eNB et sera livré à cette UE quand il se réveille . Par conséquent, on peut voir que le mécanisme de DRX atteint économies d'énergie , au détriment d'un délai supplémentaire . il On préfè re que les paramè tres DRX sont choisis de telle sorte que l' économie de puissance est maximisée tandis que le retard d'application contrainte est satisfaite . Toutefois , le compromis optimal entre le retard de facteur d'économie d'énergie et de réactivation est inconnue .

Auteurs dans [78, 79] méthodes analytiques actuelles de modéliser le DRX mécanisme dans l'UMTS. Cependant , LTE introduit deux types de Cycles DRX qui est diffè rent du cycle DRX

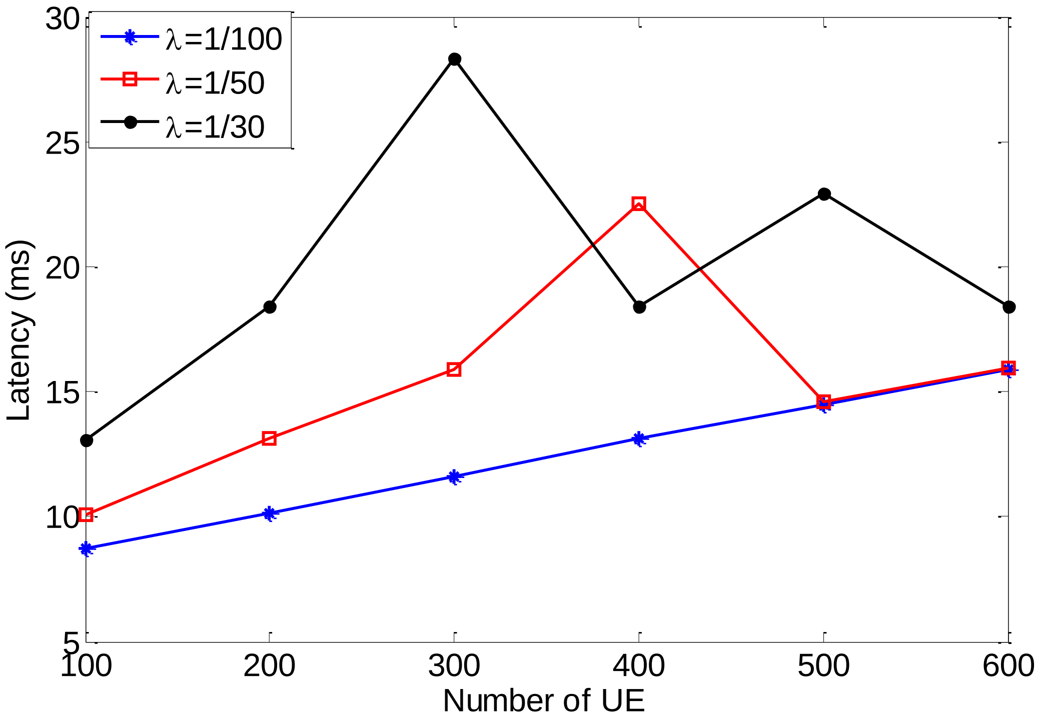


Figure A.15: Latence de l'CBA méthode d'allocation des ressources

unique UMTS . Par conséquent, les modè les utilisés dans l'UMTS ne sont pas applicables pour le cas de la technologie LTE. Références [80, 81] fournir des méthodes pour modéliser le LTE Mécanisme DRX en présence de trafic en rafales et de Poisson , respectivement . Cependant, ils ne prennent pas en compte l' ON durée , qui fait partie de chaque cycle de DRX à court et long . Ils supposent que le paquet (toujours) arrive pendant le sommeil période et doit être retardée et tamponnée . Dans la pratique , un paquet peut arriver au cours de la partie ON d'un cycle et être envoyé par la eNB (station de base) tout de suite. Ce ne sont pas comptabilisés dans le modè les mentionnés ci-dessus , menant à des estimations inexactes de la économie d'énergie facteur et la latence moyenne . Référence [82] propose un modè le de consommation d'énergie de DRX pour le service MTC avec un intervalle de paquets déterministe . Référence [83] fournit un mécanisme adaptatif configuration de DRX de seuil à sauver plus de puissance tout en maintenant le débit . Référence [84] présente une méthode de planification en cas de retard services adaptés à la technologie LTE. Avec cette méthode , le taux de perte de paquets provoquée par un procédé de couchage lors de DRX peut être réduite. Auteurs dans [85]enquêter sur la condition DRX d'économie d'énergie et son impact sur la performance QoS du trafic VoIP par simulations . Référence [86] introduit le mode de sommeil de lumiè re pour améliorer la performance de DRX . L'idée du mode de sommeil de lumiè re est d'éteindre l'amplificateur de puissance , mais laisser les autres composants tels que le pouvoir est de sauver tout réveil rapide est activée .

Nous présentons deux méthodes pour analyser le mécanisme de DRX détaillée dans LTE / LTE -A . Dans la premiè re méthode , nous supposons que le trafic est une distribution de

Poisson . Avec cette hypothè se , un modè le de chaîne semi- Markov est proposé d'analyser le mécanisme de DRX . Nous faisons modéliser le paramè tre de durée Sur , qui LTE / LTE -A prend des valeurs comprises entre 1 et 200 ms [12] , en utilisant deux types d'états de différencier la durée sur le dormir période de cycle DRX courte ou longue et montrer qu'il a un impact significatif sur la performance de DRX . Avec ce modè le , on peut calculer le facteur d'économie d'énergie et de temps d'attente pour un jeu de paramè tres DRX donné , qui peut être utilisé pour sélectionner le paramè tre de DRX approprié. Nous utilisons la simulation pour valider nos résultats .

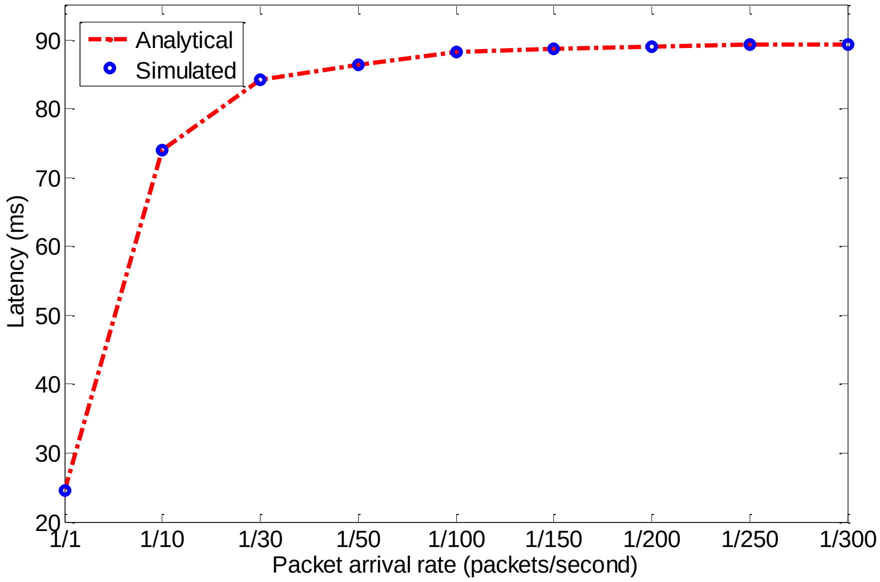
Différente de la premiè re méthode qui nécessite le trafic est une distribution de Poisson , la deuxiè me méthode est applicable à tous les types de textbf textit sporadique trafic (distribution de Poisson , uniforme distribué , etc.) Avec la deuxiè me méthode , nous fournissons également une méthode simple pour trouver le paramè tre DRX optimale qui maximise le facteur d'économie d'énergie tout en maintenant l'exigence de latence .

A.4.1 DRX modélisation pour trafic Poisson Distribué

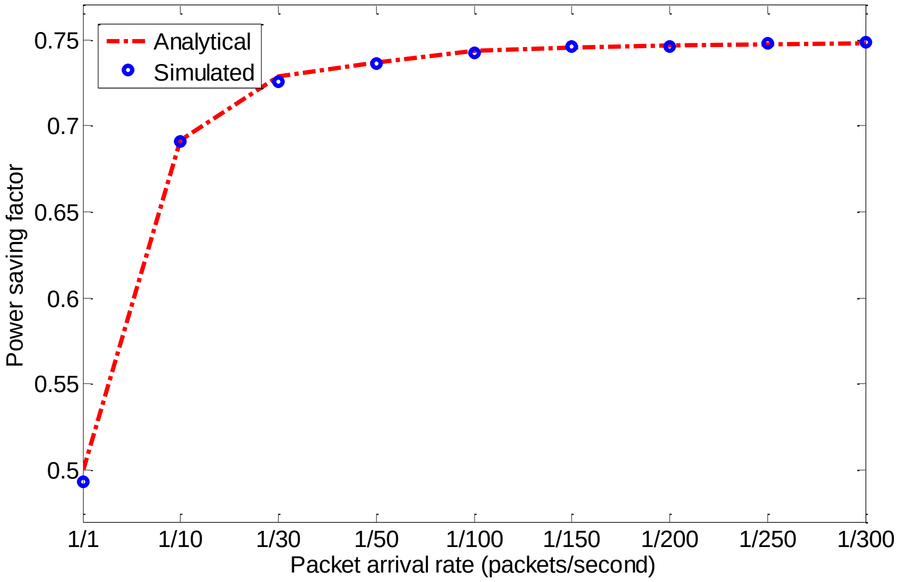
Ici, nous utilisons la méthode semi- Markov pour analyser mécanisme DRX . Nous calculons deux mesures pour DRX : facteur d'économie d'énergie et réveillons latence. Facteur de puissance d'économie est défini comme le rapport de temps que UE est à l' économie des États à la fois que l'UE est à tous les Etats le pouvoir .

Pour valider la méthode proposée , es simulations sont effectuées. nous comparons les résultats de la simulation avec les résultats d'analyse obtenus par notre méthode sous différents taux d'arrivée des paquets λ . De la Fig.A.16 , nous pouvons voir que les résultats de l'analyse sont trè s proches de celles simulées , ce qui valide notre méthode .

Ensuite , nous avons effectué des simulations pour voir l'effet du paramè tre DRX différent sur la performance de DRX . Nous constatons que pour le trafic sporadique (intervalle de paquet est beaucoup plus grand que le paramè tre DRX) , nous constatons que la durée et du cycle DRX longtemps ont une forte incidence sur la performance de DRX de temporisateur d'inactivité , cycle DRX à court et DRX à court de temps de cycle .



(a) Des temps de latence taux d'arrivée des paquets différents



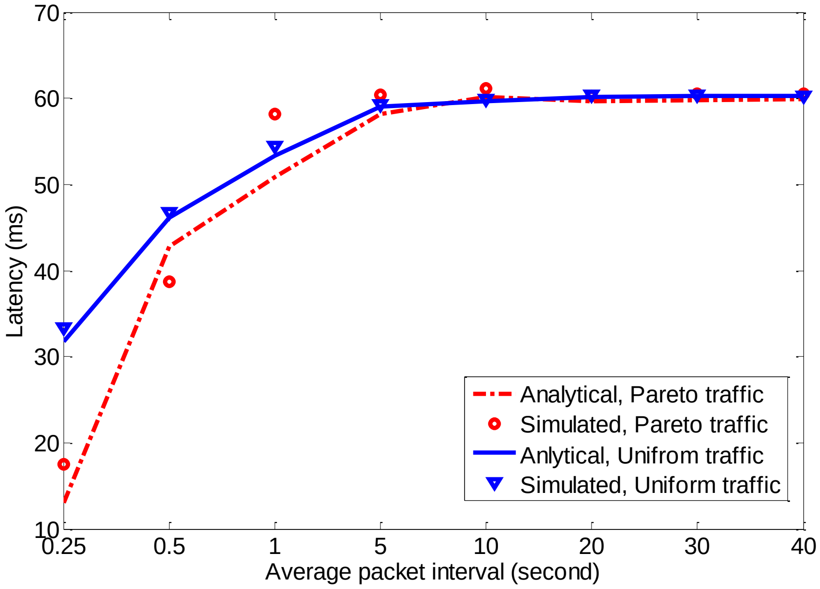
(b) Facteur d'économie d'énergie sous différentes minuterie d'inactivité

Figure A.16: Performances DRX sous le taux d'arrivée des paquets différents

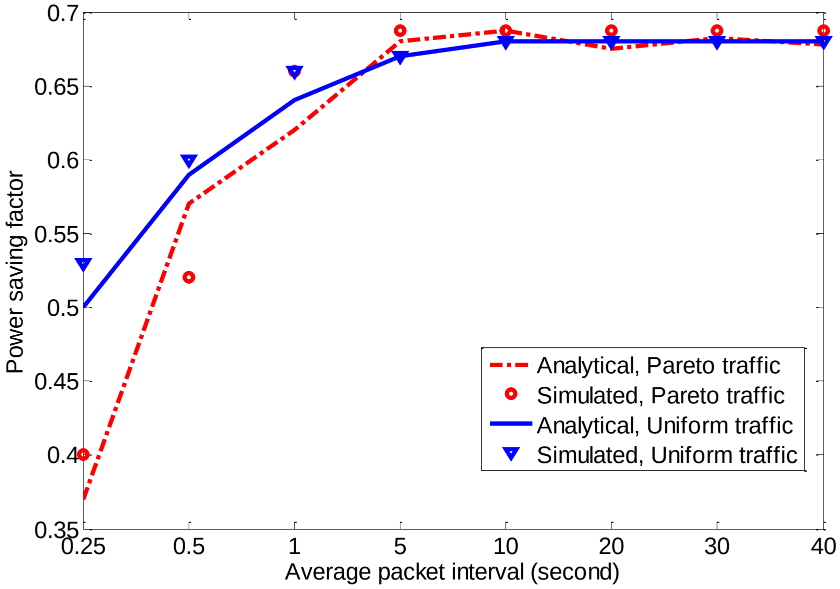
A.4.2 DRX Modélisation et optimisation de trafic sporadique

Dans la derniè re partie , nous avons proposé un modè le semi- Markov pour analyser le mécanisme de DRX pour l'application MTC . Dans ce modè le , nous supposons que le trafic est de Poisson . Cependant, comme il existe différents types d'applications MTC , cette hypothè se n'est pas toujours réaliste pour tous les types de trafic MTC . Par exemple, dans [3] le trafic est supposée être distribué uniformément . Pour le trafic non -Poisson , le modè le proposé dans la derniè re section n'est pas applicable . Par conséquent , dans cette partie, nous proposons une autre méthode qui est applicable à analyser le mécanisme de DRX avec générique sporadique modè le de trafic MTC .

Pour valider le modè le proposé , nous avons effectué des simulations avec un simulateur basé MATLAB . Nous comparons les résultats de la simulation avec les résultats d'analyse sous différents intervalles de paquets moyens pour uniforme et le trafic de Pareto . Les résultats sont montrés dans la Fig. A.17. Nous constatons que les résultats simulés correspondent aux résultats analytiques trè s bien lorsque l'intervalle entre trames est plus grande que la moyenne de 5 secondes. Contrairement à cela, il existe une erreur d'estimation , lorsque l'intervalle de paquet moyenne est de 0,25, 0,5 et 1 seconde. Par exemple , pour le trafic de Pareto lorsque l'intervalle de paquet moyenne est de 0,25 seconde, le (analytique) de latence est estimé à 13 ms alors que la latence est simulé 17,47 ms . La raison est la suivante: lorsque l'intervalle de paquet moyenne est de 0,25 seconde (250 ms) , il n'est pas beaucoup plus grande que T_S (128ms) et T_L (256ms) , ce qui n'est pas conforme à notre hypothè se . Par conséquent, les résultats d'analyse ne correspondent pas aux résultats de l'analyse . Toutefois, lorsque les paquets moyens intervalle augmente à 5 secondes (5000 ms) , l'intervalle de paquets devient beaucoup plus grande que T_S et T_L , par lequel les résultats d'analyse correspondent aux résultats d'analyse avec de petites erreurs . Par conséquent, il peut être conclu que par l'utilisation de notre modè le DRX de la performance avec un trafic sporadique peut être calculée correctement .



(a) Des temps de latence taux d'arrivée des paquets différents



(b) Facteur d'économie d'énergie sous différentes minuterie d'inactivité

Figure A.17: Performances DRX sous le taux d'arrivée des paquets différents

Bibliography

[1] H. Lenz, J. Koss, “M2M Communication – Next Revolution on Wireless Interaction,” ETSI Workshop on Machine to Machine Standardization, June 2008.

[2] Yuichi Morioka, "LTE for Mobile Consumer Devices", ETSI M2M Workshop 2011, Oct.2011.

[3] 3GPP TR 37.868, "RAN Improvements for Machine-type Communications," V. 11.0.0., Oct., 2011.

[4] Ericsson. "Ericsson Mobility Report – on the Pulse of the Networked Society", White paper, June 2013. Available at: http://www.ericsson.com/res/docs/2013/ericsson-mobilityreport-june-2013.pdf

[5] Huawei. "Internet of thins and its future", *Huawei communicate*, No. 54, Feb. 2010, pp. 23-26.

[6] Ericsson. "More than 50 billion connected devices", White paper, Feb. 2011. Available at:http://www.ericsson.com/res/docs/whitepapers/wp-50-billions.pdf

[7] Eurecom, OpenAirInterface, http://www.openairinterface.org/.

[8] Motorola, Inc. "Long Term Evolution (LTE): A Technical Overview". Technical White Paper, 2007.

[9] S. Sesia, I. Toufik, M. Baker, "LTE-The UMTS Long Term Evolution. From Theory to Practice", John Wiley and Sons, 2009.

[10] 3GPP, The Evolved Packet Core. Available at:http://www.3gpp.org/technologies/keywords-acronyms/100-the-evolved-packet-core

[11] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description," V 11.7.0 Sept., 2013.

[12] 3GPP TS 36.331, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Radio Resource Control (RRC) (Release 10)", V10.1.0, Dec.,2010.

[13] 3GPP TS 36.211, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Physical Channels and Modulation," V 11.4.0, Sept., 2013.

[14] A. Ghosh, J. Zhang, et.al., "Fundamentals of LTE", Prentice Hall, 2011.

[15] 3GPP TR 36.912, "Feasibility study for Further Advancements for E-UTRA (LTE-Advanced)," V9.2.0, March, 2010.

[16] N. Nikaein, S. Krco, "Latency for real-time machine-to-machine communication in LTE-based system architecture", *7th European Wireless Conference, Sustainable Wireless Technologies*, pp. 1-6.

[17] J. Huang, F. Qian, et.al., "A Close Examination of Performance and Power Characteristics of 4G LTE Networks", *10th international conference on Mobile systems, applications, and services*, pp. 225-238.

[18] AT&T, "Comparing LTE and 3G Energy Consumption." Available at:https://developer.att.com/developer/forward.jsp?passedItemId=11900006

[19] Anders R. Jensen, Mads Lauridsen, et.al, "LTE UE Power Consumption Model For System Level Energy and Performance Optimization", *IEEE VTC 2012-FALL*, pp.1-5.

[20] 3GPP TS 22.368, "Service Requirements for Machine-Type Communications," V10.1.0, June, 2010.

[21] Ekobus. Available at:https://mobiledevelopmentintelligence.com/products/3138-ekobus-project

[22] Y. Morioka, "LTE for Mobile Consumer Devices," ETSI Workshop on Machine to Machine Standardization, 2011.

[23] ETSI TS 102 690, "Machine-to-machine communications (M2M); functional architecture," V1.2.1, June, 2013.

[24] M2M communications, "What is M2M communications". Available at: http://www.m2mcomm.com/about/what-is-m2m/.

[25] Machina Research, "The Global M2M market in 2013". Whilte paper, 2013.

[26] Vodafone, "Machine to machine: business benefits". Available at https://m2m.vodafone.com/discover-m2m/business-benefits/.

[27] Juniper networks, "machine-to-machine (M2M)-the rise of the machines", White paper, 2011.

[28] G. Wu, S. Talwar, etal. "M2M: From mobile to embeded Internet", *IEEE Communication Maganize*, vol.49, No.4, April 2011, pp. 36-43.

[29] Taesoo Kwon, Ji-Woong Choi, "Multi-Group Random Access Resource Allocation for M2M Devices in Multicell Systems", *IEEE communication letters*, vol.16, No.6, June 2012, pp.834-837.

[30] K. Ko, M.Kim, etal., "A Novel Random Access for Fixed-Location Machine-to-Machine Communications in OFDMA Based Systems", *IEEE communication letters*, vol.16, No.9, Sept. 2012, 1428-1431.

[31] H. Thomsen, N.K Pratas, etal., "Analysis of the LTE Access Reservation Protocol for Real-Time Traffic" , *IEEE communication letters*, vol. 17, No.8, Aug.2013, pp.1616 - 1619.

[32] M. Hasan, E. Hossain, D.Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches", *IEEE Communications Magazine*, June 2013, pp.86-93.

[33] S. Lien, T. Liau, etal., "Cooperative Access Class Barring for Machine-to-Machine Communications", *IEEE Transactions on Wireless Communications*, vol. 11, No.1, Jan.2012, pp. 27-32.

[34] S. Lien, etal., "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications", *IEEE Communications Magazine*, vol.49, No. 4, April 2011, pp. 66-74.

[35] K. Zhang, etal., "Radio resource allocation in LTE-advanced cellular networks with M2M communications", *IEEE Communications Magazine*, vol 50, No.9, July 2012, pp. 184-192.

[36] A.G. Gotsis, etal., "M2M Scheduling over LTE: Challenges and New Perspectives", *IEEE Vehicular Technology Magazine*, vol. 7, No. 3, Sept. 2012, pp.34-39.

[37] C. Ho, etal., "Energy-Saving Massive Access Control and Resource Allocation Schemes for M2M Communications in OFDMA Cellular Networks", *IEEE Wireless Communications Letters*, vol. 1, No. 3, June 2012, pp.209-212.

[38] S. Chien, etal., "Power Consumption Analysis for Distributed Video Sensors in Machine-to-Machine Networks", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol.3, No. 1, March 2013, pp. 55-64.

[39] F. Gong, etal., "SER Analysis of the Mobile-Relay-Based M2M Communication over Double Nakagami-m Fading Channels", *IEEE Communications Letters*, vol. 15, No.1, Jan. 2011, pp. 34-36.

[40] H. Lee, etal., "Feasibility of cognitive machine-to-machine communication using cellular bands", *IEEE Wireless Communications*, vol. 20, No.2, April 2013, pp.97-103.

[41] A. Bartoli, etal., "Secure Lossless Aggregation Over Fading and Shadowing Channels for Smart Grid M2M Networks", *IEEE Transactions on Smart Grid*, vol.2, No.4, Dec.2011, pp. 844-864.

[42] Z.M.Fadlullah, etal., "An early warning system against malicious activities for smart grid communications", *IEEE Network*, vol. 25, No.5, Sept.2011, pp. 50-55.

[43] D.Niyato, etal., "Machine-to-machine communications for home energy management system in smart grid", *IEEE Communications Magazine*, vol.49, No.4, April 2011, pp.53-59.

[44] Z.M.Fadlullah, etal., "Toward intelligent machine-to-machine communications in smart grid", *IEEE Communications Magazine*, vol. 49, No.4, April 2011, pp. 60-65,

[45] Y. Zhang, etal., "Cognitive machine-to-machine communications: visions and potentials for the smart grid", *IEEE Network*,vol. 26. No.3, May/June 2012, pp. 6-13.

[46] Y. Zhang, etal., "Home M2M networks: Architectures, standards, and QoS improvement", *IEEE Communications Magazine*, vol.49, No.4, April 2011, pp.44-52.

[47] A. Alheraish, etal., "Design and implementation of home automation system", *IEEE Transactions on Consumer Electronics*, vol.50, No.4, July 2004, pp. 1087-1092.

[48] J.Shih, etal, "Securing M2M With Post-Quantum Public-Key Cryptography", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol.3, no.1, March 2013, pp. 106-116.

[49] D.A.Bailey, "Moving 2 Mishap: M2M's Impact on Privacy and Safety", *IEEE Security & Privacy*, vol. 10, No.1, Feb. 2012, pp. 84-87.

[50] N. Gligoric, etal., "Application-layer security mechanism for M2M communication over SMS", *20th Telecommunications Forum* (TELFOR), 2012, pp.5-8, Nov. 2012.

[51] M. Y. Cheng, G.Y. Lin, H.Y.Wei, and A.C.Hsu, "Overload control for machine-type-communications in LTE-Advanced system," *IEEE communication magazine*, vol.50, no. 6, pp. 38-45, June 2012.

[52] T.Kwon, J.W. Choi, "Multi-Group Random Access Resource Allocation for M2M Devices in Multicell Systems", *IEEE Communication Letters*, vol. 16, no. 6, pp. 834-837, June 2012.

[53] R.G.Cheng, C.H.Wei, S.L.Tsao, F.C.Ren, "RACH Collision Probability for Machine-type Communications", *Proc. of IEEE VTC*, May, 2012,pp. 1-5.

[54] C.Y.Ho, C.Y.Huang, "Energy-Saving Massive Access Control and Resource Allocation Schemes for M2M Communications in OFDMA Cellular Networks", *IEEE wirelss communication letters*, vol. 1, no. 3, pp. 209-212, June 2012.

[55] K.S. Ko, M.J.Kim, *etal.*, "A Novel Random Access for Fixed-Location Machine-to-Machine Communications in OFDMA Based Systems," *IEEE Commun. Lett.*, vol.16,No.9, pp.1428-1431, Sept. 2012.

[56] Nuno K. Pratas, *etal*, "Code-Expanded Random Access for Machine-Type Communications," *Proc. of IEEE Globecom Workshops*, Dec. 2012, pp. 1681-1686.

[57] S.Y. Lien, T.H.Liau, *etal.*, "Cooperative Access Class Barring for Machine-to-Machine Communications," *IEEE Trans. on Wirless Commun.*, vol.11, No.1, pp.27-32, Jan. 2012.

[58] S.-Y. Lien and K.-C. Chen, "Massive access management for QoS guarantees in 3GPP machine-to-machine communications," *IEEE Commun. Lett.*, vol. 15, no. 3, pp. 311-313, Mar. 2011.

[59] P. Zhou, *etal*, "An effcient random access scheme for OFDMA systems with implicit message transmission," *IEEE Trans. on Wireless Commun.*, vol.7, No.7, pp.2790-2797, Jul.2008.

[60] 3GPP TS 36.321, Evolved Universal Terrestrial Radio Access (E- UTRA); Medium Access Control (MAC)," V.11.2.0, Mar., 2013.

[61] J.P. Cheng, C.h. Lee, T.M.Lin, "Prioritized Random Access with Dynamic Access Barring for RAN Overload in 3GPP LTE-A Networks," *Proc. of IEEE Globecom Workshops*, Dec. 2011, pp. 368-372.

[62] R. Nelson, "Probability, Stochastic Processes, and Queueing Theory", Springer, 1995.

[63] 3GPP R2-072630, HARQ operation in case of UL Power Limitation, Ericsson, June 2007.

[64] R. Susitaival, M.Meyer, "LTE coverage improvement by TTI bundling," *Proc. of IEEE VTC Spring*, April, 2009, pp.1-5.

[65] 3GPP TR 36.888, "Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE," V. 2.0.0, June, 2012.

[66] LOLA Project (Achieving Low-Latency in Wireless Communications), "D3.2 Network related analysis of M2M and online-gaming traffic in HSPA" v1.0, July 2010. Available at http://www.ict-lola.eu/

[67] LOLA Project (Achieving Low-Latency in Wireless Communications), "D3.4 .M2M Traffic Model" v1.0, May 2011. Available at http://www.ict-lola.eu/

[68] Chu, D. C., "Polyphase codes with good periodic correlation properties", *IEEE Transactions on Information Theory*, vol. IT-18, pp. 531-532, July 1992.

[69] Giuseppe Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function", *IEEE Journa on Selected Areas in Communications*, Vol. 18, No. 3, pp. 535-547.

[70] F. Daneshgaran, M. Laddomada, F. Mesiti, and M. Mondin,"Unsaturated Throughput Analysis of IEEE 802.11 in Presence of Non Ideal Transmission Channel and Capture Effects", *IEEE transactions on wireless communications*, vol 7, No.4, April, 2008.

[71] Yong Shyang Liaw, Arek Dadej, Aruna Jayasuriya, "Performance Analysis of IEEE 802.11 DCF under Limited Load", *Proc. of Asia-Pacific Conference on Communications*, pp. 759-763, Oct., 2005.

[72] F. Daneshgaran, M. Laddomada, F. Mesiti, and M. Mondin, "On the Linear Behaviour of the Throughput of IEEE 802.11 DCF in Non-Saturated Conditions", *IEEE communication letters*, vol 11, No 11, 2007.

[73] Md. Siddique and J. Kamruzzaman, "Performance Analysis of m-Retry BEB Based DCF under Unsaturated Traffic Condition", *Proc. of IEEE WCNC 2010*, pp. 1-6, April, 2010.

[74] Z. Qinjuan, et al, "Bandwidth mapping model for IEEE 802.11 DCF in unsaturated condition", *IET Communications*, Vol. 6, No. 13, pp. 2007-2015.

[75] G. Cantieni, et al,"Performance analysis under finite load and improvements for multirate 802.11", *Computer communications*, vol. 28, No. 10, pp 1095-1109.

[76] 3GPP TS 36.213, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Physical layer procedure (Release 10)," V10.1.0, March, 2011.

[77] Suard, B., Xu, G., Liu, H., Kailath, T.,"Uplink Channel Capacity of Space-Division-Multiple-Access Schemes," *IEEE Transactions on Information Theory,* vol. 44, no. 4, pp. 1468-1476, 1998.

[78] S.-R. Yang, S.-Y. Yan, and H.-N. Hung, "Modeling UMTS Power Saving with Bursty Packet Data Traffic," *IEEE Transactions on Mobile Computing*, vol. 6, no. 12, pp. 1398-1408, Dec. 2007.

[79] S.-R. Yang, and Y-B. Lin,"Modeling UMTS Discontinuous Reception Mechanism," *IEEE Transactions on Wireless Communications*, vol.4, no.1, pp.312-319, Jan. 2005.

[80] L. Zhou et al. "Performance Analysis of Power Saving Mechanism with Adjustable DRX Cycles in 3GPP LTE," *Proc. of IEEE VTC*, Calgary, Canada, Spet. 2008,pp. 1-5.

[81] S. Jin and D. Qiao, "Numerical analysis of the power saving in 3GPP LTE Advanced wireless networks," *IEEE Transaction on Vehicular Technology*, vol. 61, no. 4, pp. 1779-1785, May 2012.

[82] T. Tirronen, A. Larmo, et al, "Machine-to-machine communication with long-term evolution with reduced device energy consumption", *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 4, pp. 413-426, June 2013.

[83] S. Gao, H. Tian, J. Zhu, and L. Chen, “A more power-efficient adaptive discontinuous reception mechanism in LTE,” in *Proc. of IEEE VTC*, pp. 1 –5, Sept. 2011.

[84] H. Bo, T. Hui, C. Lan, and Z. Jianchi, “DRX-aware scheduling method for delay-sensitive traffic,” *IEEE Commun. Lett.*, vol. 14, no. 12, pp.113-115, Dec. 2010.

[85] M. Polignano, D. Vinella, D. Laselva, J. Wigard, and T. Sorensens, “Power savings and QoS impact for VoIP application with DRX/DTX feature in LTE,” in *Proc. of IEEE VTC*, pp. 1 –5, May 2011.

[86] K.-C. Ting, H.-C. Wang, C.-C. Tseng, and F.-C. Kuo, “Energy-efficient DRX scheduling for QoS traffic in LTE networks,” in *Proc. of IEEE ISPA* , pp.213-218, May 2011.

[87] 3GPP R1-120056, "Analysis on traffic model and characteristics for MTC and text proposal," Huawei, HiSilicon.

[88] 3GPP R1-113656, "MTC application characteristic analysis," Huawei, HiSilicon, CMCC.

[89] 3GPP R1-114439,"Text proposal for traffic model/characteristics for MTC", Vadafone Group.

[90] The Pareto Distribution, http://www.math.uah.edu/stat/special/Pareto.html