

KADİR HAS UNIVERSITY  
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING  
PROGRAM OF MSc IN ELECTRONICS ENGINEERING

**SEMI PERSISTENT RADIO RESOURCE ALLOCATION  
FOR MACHINE TYPE COMMUNICATIONS IN 5G AND  
BEYOND CELLULAR NETWORKS**

Zaid HAJ HUSSIEN

MASTER'S THESIS

İSTANBUL, April 2018

Zaid HAJ HUSSEIN

M.S. Thesis

2018



**SEMI PERSISTENT RADIO RESOURCE ALLOCATION  
FOR MACHINE TYPE COMMUNICATIONS IN 5G AND  
BEYOND CELLULAR NETWORKS**

Zaid HAJ HUSSIEN

MASTER'S THESIS

Submitted to the Graduate School of Science and Engineering of Kadir Has University  
in partial fulfillment of the requirements for the degree of Master's in the Program of  
ELECTRONICS ENGINEERING

İSTANBUL, April, 2018

DECLARATION OF RESEARCH ETHICS /  
METHODS OF DISSEMINATION

I, Zaid HAJ HUSSIEN, hereby declare that;

- this Master's Thesis is my own original work and that due references have been appropriately provided on all supporting literature and resources;
- this Master's Thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- I have followed "Kadir Has University Academic Ethics Principles" prepared in accordance with the "The Council of Higher Education's Ethical Conduct Principles"

In addition, I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations.

Furthermore, both printed and electronic copies of my work will be kept in Kadir Has Information Center under the following condition as indicated below:

- The full content of my thesis/project will be accessible from everywhere by all means.

Zaid HAJ HUSSIEN

---

04/04/2018

KADIR HAS UNIVERSITY  
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

**ACCEPTANCE AND APPROVAL**

This work entitled **SEMI PERSISTENT RADIO RESOURCE ALLOCATION FOR MACHINE TYPE COMMUNICATIONS IN 5G AND BEYOND CELLULAR NETWORKS** prepared by **Zaid HAJ HUSSIEN** has been judged to be successful at the defense exam held on **04/04/2018** and accepted by our jury as **MASTER'S THESIS**.

APPROVED BY:

(Asst. Prof. Yalçın ŞADİ) (Advisor) (Kadir Has University)



(Asst. Prof. Arif Selçuk Öğrenci) (Kadir Has University)



(Prof. Hakan Ali ÇIRPAN) (Istanbul Technical University)



I certify that the above signatures belong to the faculty members named above.



(Assoc. Prof. Dr. Ebru Demet AKDOĞAN)

Dean of Graduate School of Science and Engineering

DATE OF APPROVAL: (04/04/2018)

## TABLE OF CONTENTS

<b>TABLE OF CONTENTS</b> .....	
<b>ABSTRACT</b> .....	<b>i</b>
<b>ÖZET</b> .....	<b>ii</b>
<b>ACKNOWLEDGMENTS</b> .....	<b>iii</b>
<b>LIST OF TABLES</b> .....	<b>v</b>
<b>LIST OF FIGURES</b> .....	<b>vi</b>
<b>LIST OF SYMBOLS AND ABBREVIATIONS</b> .....	<b>vii</b>
<b>1. INTRODUCTION</b> .....	<b>1</b>
1.1 M2M Communications in Cellular Networks .....	1
1.2 Related Work.....	4
1.3 Original Contributions.....	9
1.4 Organization .....	9
<b>2. SCHEDULING PERIODIC TASKS</b> .....	<b>10</b>
2.1 Periodic Tasks Model .....	10
2.2 Scheduling Algorithms .....	11
2.2.1 Earliest deadline first (EDF) .....	12
2.2.2 Rate monotonic scheduling (RM) .....	12
2.3 Partitioned Multiprocessor Tasks Scheduling Algorithms.....	14
<b>3. FLEXIBLE PHYSICAL LAYER ARCHITECTURE</b> .....	<b>17</b>
3.1 Flexibility & Waveforms.....	18
3.2 Flexibility & Numerology Design.....	21
3.3 Beneficial Use for M2M Communications .....	23
3.4 System Model and Assumptions .....	25
<b>4. MINIMUM BANDWIDTH RESOURCE ALLOCATION PROBLEM</b> .....	<b>29</b>
4.1 Problem Description .....	29
4.2 NP-Hardness .....	32
4.3 Optimization Problem .....	33
<b>5. FAST MINIMUM-BAND MAXIMUM-UTILIZATION ALGORITHM</b> <b>(SINGLE SUBCARRIER CASE)</b> .....	<b>34</b>
5.1 Algorithm Description.....	34

5.2 Approximation Ratio Performance.....	37
<b>6. FAST MINIMUM-BAND MAXIMUM-UTILIZATION ALGORITHM</b>	
<b>(MULTIPLE SUBCARRIER CASE).....</b>	<b>39</b>
6.1 Multi-Subcarrier Effect Analysis .....	39
6.2 Minimum Bandwidth Optimal Subcarrier Spacing Algorithm (OSC).....	40
6.3 Multi-Subcarrier Fast Minimum-Band Maximum-Utilization Algorithm (FMM-OSC).....	42
<b>7. PERFORMANCE EVALUATION .....</b>	<b>44</b>
7.1 Fast Minimum-Band Maximum-Utilization Algorithm (FMM).....	44
7.2 Minimum Bandwidth Optimal Subcarrier Spacing Algorithm (OSC).....	46
7.3 Multi-Subcarrier Fast Minimum-Band Maximum-Utilization Algorithm.....	52
<b>8. CONCLUSION.....</b>	<b>54</b>
<b>REFERENCES.....</b>	<b>55</b>

# SEMI PERSISTENT RADIO RESOURCE ALLOCATION FOR MACHINE TYPE COMMUNICATIONS IN 5G AND BEYOND CELLULAR NETWORKS

## ABSTRACT

The fast growth of machine-to-machine (M2M) communications in cellular networks brings the challenge of satisfying diverse Quality-of-Service (QoS) requirements of massive number of machine type communications (MTC) devices with limited radio resources. In this study, we first introduce the minimum bandwidth resource allocation problem for M2M communications in 5G and beyond cellular networks. NP-hardness of the problem is proven. Then, we propose a fast and efficient polynomial-time algorithm exploiting the periodicity of the MTC traffic based on persistent resource allocation. We prove a mathematical performance result for this algorithm considering a special case of the problem. We elaborate on the expected flexible physical layer structure and study its possible effects on our algorithm. Simulations show that the proposed algorithm outperforms the previously proposed clustering-based radio resource algorithms significantly and performs very close to optimal.

**Keywords:** 5G Cellular Networks, M2M Communications, Radio Resource Allocation, Flexible Physical Layer Architecture.

# 5G VE ÖTESİ HÜCRESEL AĞLARDA MAKİNE TİPİ İLETİŞİM İÇİN YARI-KALICI RADYO KAYNAK DAĞITIMI

## ÖZET

Hücresele ağlarda makineler arası iletişimin hızlı büyümesi, çok büyük sayıda makine tipi iletişim aracının servis kalitesi gerekliliklerinin kısıtlı radyo kaynaklarıyla karşılanması zorluğunu da beraberinde getirmektedir. Bu çalışmada, ilk olarak 5G ve ötesi hücresele ağlarda makineler arası iletişim için minimum band genişliğinde kaynak dağıtımını problemini sunmaktayız. Problemin NP-zor olduğu kanıtlanmaktadır. Sonrasında, kalıcı kaynak dağıtımına dayanan ve makine tipi iletişim trafiğinin periyodikliğinden yararlanan hızlı ve etkin bir polinom-zamanlı algoritma önermekteyiz. Problemin özel bir durumunu ele alarak, bu algoritma için matematiksel bir performans sonucu kanıtlamaktayız. Simülasyonlar, önerilen algoritmanın daha önce önerilmiş gruplama tabanlı radyo kaynak dağıtım algoritmasına belirgin şekilde üstün geldiğini ve optimale çok yakın performans gösterdiğini göstermektedir.

**Anahtar Sözcükler:** 5G Hücresele Ağlar, Makineler Arası İletişim, Radyo Kaynak Dağıtım, Esnek Fiziksel Katman Yapısı.

## ACKNOWLEDGMENTS

The success and final outcome of this thesis required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my thesis. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I do respect and feel thankful of Assist. Prof. Dr. Yalçın ŞADI, for providing me an opportunity to do the thesis work with him and giving me all support and guidance, which made me complete the thesis. I am extremely thankful to him for providing such a nice support and guidance, although he had busy schedule managing his academic work.

I would not forget to remember Assist. Prof. Dr. Selçuk Öğrenci, Prof. Dr. Hakan ÇIRPAN, and Assoc. Prof. Dr. Serhat ERKÜÇÜK for their encouragement and moreover for their timely support and guidance till the completion of my thesis work.

*for their silent never-ending support which started on the day I opened my eyes to this  
world until this moment and for their hidden persistent prayers,*

*my parents*

*for every single courageous man or women who keep struggling against oppression,*

*my nation*

## LIST OF TABLES

<b>Table 2.1</b> - Worst-Case Performance Ratio For Task Assignment Heuristics .....	16
<b>Table 7.1</b> - FMM Algorithm Performance over CBA .....	45
<b>Table 7.2</b> - Optimality Performance of FMM Algorithm .....	46

## LIST OF FIGURES

<b>Figure 1.1</b> - M2M Communications Architecture Proposed by ETSI .....	2
<b>Figure 1.2</b> - General Scheduling Process in LTE/LTE-A .....	5
<b>Figure 2.1</b> - Summary of Multiprocessor Task Scheduling Algorithms .....	16
<b>Figure 3.1</b> - Different Waveform Parameters Providing Flexibility .....	23
<b>Figure 3.2</b> - Flexible Subcarrier Spacing for Heterogeneous Service Requirements.....	24
<b>Figure 3.3</b> - Jitter Requirement Definition .....	26
<b>Figure 3.4</b> - Resource Block Structure .....	27
<b>Figure 3.5</b> - Multi-subcarrier spacing physical structure .....	28
<b>Figure 4.1</b> - A Unit Frequency Band (UFB) .....	30
<b>Figure 4.2</b> - UFB Utilization Definition.....	31
<b>Figure 5.1</b> - FMM vs. CBA Algorithms.....	36
<b>Figure 7.1</b> – Bandwidth Reduction by Multi-Subcarrier Spacing Values for Devices of a Single Cluster with Period= 9 ms.....	48
<b>Figure 7.2</b> - Bandwidth Reduction by Multi-Subcarrier Spacing Values for Devices of a Single Cluster with Period= 26 ms .....	49
<b>Figure 7.3</b> - Bandwidth Reduction by Multi-Subcarrier Spacing Values for Devices of a Single Cluster with Period= 53 ms .....	50
<b>Figure 7.4</b> - Bandwidth Reduction by Multi-Subcarrier Spacing Values for Devices of a Single Cluster with Period= 100 ms .....	51
<b>Figure 7.5</b> - FMM Algorithm/FMM-MC Algorithm Performance Comparison .....	52
<b>Figure 7.6</b> - FMM Algorithm using 6 different scaled subcarrier spacing values .....	53

## LIST OF SYMBOLS AND ABBREVIATIONS

3GPP	The 3rd Generation Partnership Project
5G	Fifth Generation
AGTI	Access Grant Time Interval
ACB	Access Class Barring
AMAM	Adaptive Massive Access Management
CAT-NB1	Category Narrow Band 1
CP	Cyclic Prefix
CP-OFDM	Cyclic Prefix - Orthogonal Frequency Division Multiplexing
DSL	Digital Subscriber Line
DL	Downlink
EDF	Earliest Deadline First
eMBB	Enhanced Mobile BroadBand
ETSI	European Telecommunications Standards Institute
eNB	Evolved Node B
FBMC	Filter Bank Multi-Carrier
FDD	Frequency Division Duplex
FC-OFDM	Flexibly Configured OFDM
FMM	Fast Minimum-Band Maximum-Utilization
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
GFDM	Generalized Frequency Division Multiplexing
H2H	Human to Human
IoT	Internet of Things
IP	Internet Protocol
IEEE	Institute of Electrical and Electronics Engineers
ICI	Inter-Carrier Interference
ISI	Inter-Symbol-Interference
LTE	Long Term Evolution
LTE-A	Long Term Evolution - Advanced
LAN	Local Area Network

MTC	Machine Type Communications
mMTC	Massive Machine Type Communications
M2M	Machine to Machine
MBB	Mobile BroadBand
MIMO	Massive Multiple-Input Multiple-Output
NB-IoT	Narrow Band Internet of Things
NR	New Radio
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OSC	Optimal Subcarrier Spacing
OQAM	Offset Quadrature amplitude modulation
OBE	Out of Band Emissions
PARP	Peak-to-Average Power Ratio
PRB	Physical Resource Blocks
QoS	Quality of Service
Rel.13	3GPP Release 13
RB	Resource Blocks
RM	Rate-Monotonic
RA	Random Access
RAT	Radio Access Technology
TTI	Transmission Time Interval
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
UB	Utilization Bound
URLLC	Ultra-Reliable Low Latency Communication
UFB	Unit Frequency Band
UFMC	Universal Filtered Multi-Carrier
UE	User Equipment
UL	Uplink
V2X	Vehicle-to-Anything
VoIP	Voice over Internet Protocol
W3C	World Wide Web Consortium

WLAN	Wireless Local Area Network
WiMAX	Worldwide Interoperability for Microwave Access
ZT-OFDM	Zero-Tail OFDM
ZP	Zero Prefix

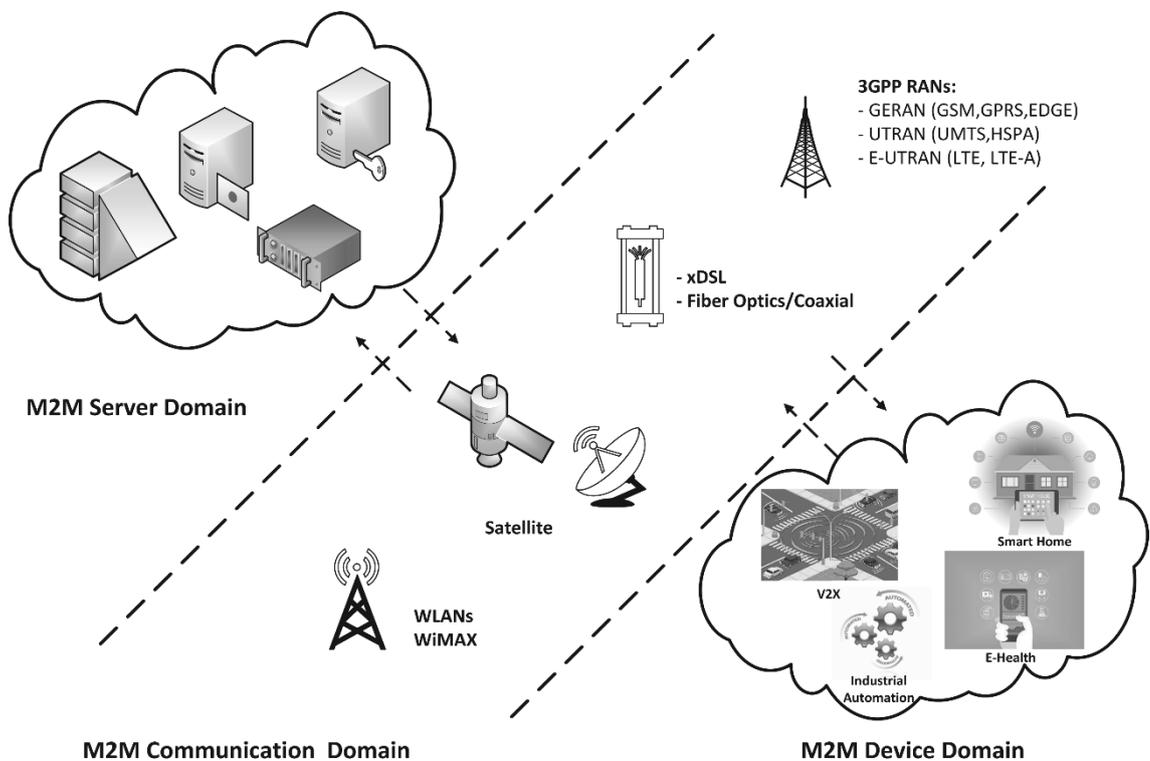
# 1. INTRODUCTION

## 1.1 M2M Communications in Cellular Networks

By 2019, more than % 40 percent of all connected devices are projected to be machine-type communications (MTC) devices (Pepper, 2015). MTC or what sometimes is called Machine to machine communications (M2M) has been heavily discussed in academia and industry in the past few years inspecting their traffic characteristics and QoS requirements, trying to forecast and predict their potential effect in technology and our lives. However, this topic looks important enough to catch big and deep-rooted companies and standardization bodies' interests and push them to conduct studies about it. Companies like Intel, standardization bodies like IEEE, ETSI, 3GPP, and W3C have started different projects about the next mobile communication generation (5G) and M2M communication attended strongly in these projects (Mehmood et al., 2017).

MTC can be defined as a communication between a set of devices such as sensors/actuators and a cloud-based server through a wired or a wireless access network far from any human supervision or intervention (Mehmood et al., 2017) (Ghavimi & Chen, 2015) (Wu et al., 2011). This type of communications covers a wide range of applications, services, and use cases. Knowing these promising applications and services will demonstrate the wide opportunities waiting for the market. The smart grid is one of these applications (Fan et al., 2014) where smart meters can be integrated with electric power, gas or water supplying networks to collect information and send it to a server for automated control and monitoring functions. This can help to get considerable savings in consumption. Vehicle to everything (V2X) is another exciting application of M2M communication. Starting from traffic congestion control, safe automated driving, vehicle accidents reporting and emergency calling and not ending with maintenance notifications reported by distributed sensors inside the vehicle, M2M communications will play a

significant role in these applications (Wu et al., 2011). E-Health is a new concept to improve the quality of patient care remotely and is one of the important M2M use cases (Chen, Kwang-Cheng, 2012). Planted or wearable sensors can be used to send different health reports like blood pressure, temperature, and heart rates to related doctors and medical centers. There are many other applications for M2M communications such as industrial automation, tracking and tracing, smart homes and environmental monitoring...etc. Figure 1.1 illustrates the communication architecture standardized by ETSI for M2M communications.



**Figure 1.1 - M2M Communications Architecture Proposed by ETSI**

M2M communications' main traffic is in the uplink direction and traffic model can be categorized mainly into two categories. One is event-driven traffic which is sent only when a specific event occurs. The other is time-controlled traffic which is sent periodically between wake and sleep modes. Both categories share a common set of QoS requirements and features which are listed below (Wu et al., 2011):

- A massive number of connections: this is one of the most obvious characteristics of M2M communications. A huge number of distributed devices which try to get access and send data to the network simultaneously.
- High reliable connections: there are many M2M applications which carry sensitive data such as security, disaster management or health care which need highly reliable and secure connections.
- Small/Burst data transmission: most of the M2M applications are involved in surveillance, sensing, and control functions (e.g.; Temperature measuring, traffic counting...). These kinds of tasks generate small packets and send them to servers.
- A wide range of delay requirements: latency constraints of M2M data transmissions can vary from a very stringent requirement such as V2X connectivity or health issues where traffic safety and lifesaving depend heavily on quick response to delay-tolerant traffic.
- Extreme low power consumptions due to very limited access to power sources.

M2M devices may be stationary (e.g. home/factory sensors) or mobile (e.g. vehicular devices) connected directly to the access medium or through an aggregator in case of these devices have power and cost constraints. These aggregators are smart devices which collect data from simpler devices and process it, then send it to relate servers. As illustrated in Figure 1.1, the access network may be either wired (xDSL, fiber optics...) or wireless (Mobile network, WLAN, WiMAX...). Wired networks may have good advantages for high reliability, high data rates, and security but it is not an effective choice to support M2M applications because of high costs and lack of mobility support. On the other hand, although short-range wireless networks (e.g. LAN) are cheaper and provide mobility, they have a non-global coverage which affects mobility limits. This limitation besides low rates and weak security make mobile networks which have a ready-to-use infrastructure with global coverage, high data rates, and good security a strong candidate to carry M2M communications. Therefore, an extensive research has been conducted to reach this goal through the past few years (Ghavimi & Chen, 2015).

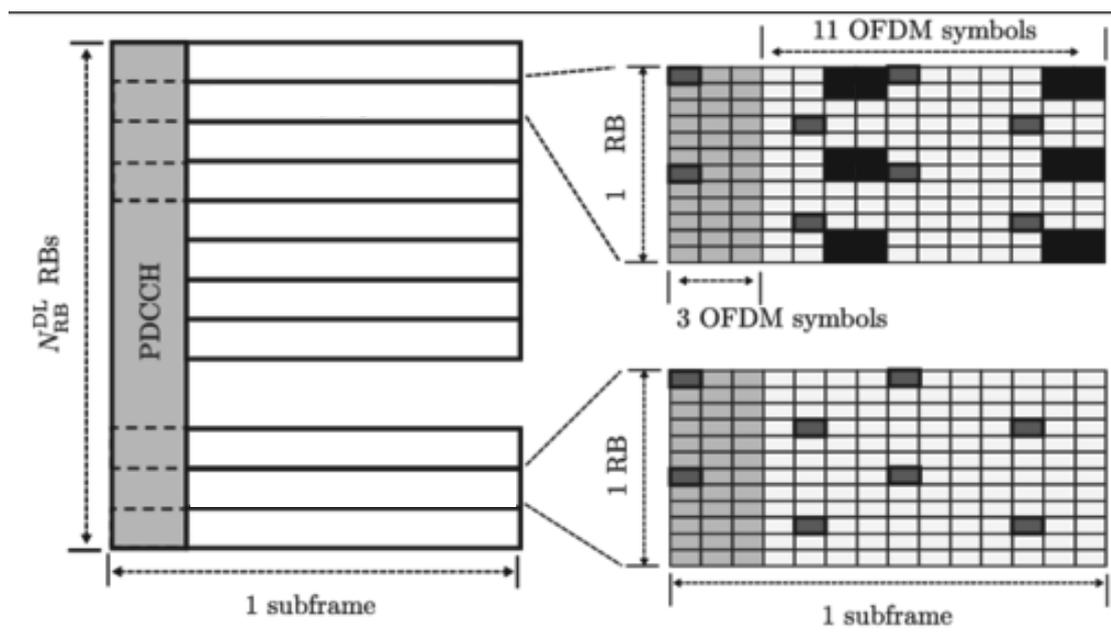
M2M communications have started to use mobile networks through a still-used second-generation technology called General Packet Radio Service (GPRS). This packet-data protocol was originally designed to support small and burst amounts of data like email browsing, one of the clearest characteristics of M2M communications. There are some other advantages for GPRS which make it a ready-to-use mobile network technology for M2M communication such as low cost, global coverage and a long experienced and tested technology by operators and vendors. These features facilitated M2M entry to the market. Despite all mentioned features, GPRS technology has limitations which hinder wide usage of it for M2M communication. The main limitation is capacity. GPRS capacity cannot exceed 150 Kbps per cell per MHz, which is very limited capacity for the M2M expected massive connections which may reach thousands of devices per cell. In addition, GPRS connection needs to be established by the device itself. These limitations made GPRS a temporary solution for M2M communications (Gotsis et al., 2012).

Starting with 3GPP Rel.13 (Schlienz & Raddino, 2016), M2M/MTC communications has been introduced to mobile systems by a new radio interface called Narrow-Band Internet-of-Things (NB-IoT) which is based on LTE. This new radio interface is standardized as simple as possible to fulfill M2M device requirements of low cost and low power consumption. Therefore, NB-IoT was designed based on some specific requirements like minimizing the signaling overhead, improve battery life, support IP and non-IP data. To fulfill these requirements, many LTE features especially advanced and sometimes basic ones were discarded from design. For instance, features like handover for connected devices, carrier aggregation, and dual connectivity are not available in NB-IoT. A new UE category was defined to tag devices support NB-IoT which is CAT-NB1.

## **1.2 Related Work**

Providing a native support in the emerging 5G cellular systems for fast-growing machine-to-machine (M2M) applications is of paramount importance. However, supporting a massive number of MTC devices is very challenging due to the problem of allocating radio resources to a large number of devices with diverse QoS requirements in the same network.

Resource allocation is a process which takes place at the base station (eNodeB) to allocate radio resources according to requests by UE or M2M devices in downlink or uplink direction. Since most of the M2M applications are expected to generate traffic mainly in the uplink direction, usually the focus is on UL scheduling. In LTE, 3GPP proposed a generic scheduling procedure based on the physical time-frequency frame structure (Figure 1.2). The minimum resource allocation unit that can be assigned to a terminal is called Physical Resource Block (PRB). PRB consists of 12 subcarriers in frequency domain each with 15 kHz (totally 180 kHz) and 7 symbols in time domain forming one time-slot of 0.5 msec. The scheduling process then can be divided into two stages: 1) a time-domain scheduling where a set of terminals are selected to be assigned PRB in the current time frame, 2) a frequency-domain scheduling where the selected terminals in the first stage are assigned PRBs. Both stages make their decisions based on different criteria like fairness, channel conditions, experienced delays and other QoS metrics.



**Figure 1.2 - General Scheduling Process in LTE/LTE-A**

Most of the resource allocation algorithms designed for cellular M2M communications are based on random access procedure to get an initial access to the network for data transmission (Dhillon H. S. et al., 2013) - (Hasan et al., 2013). However, considering the envisioned massive connectivity of MTC devices in future cellular networks, the resulting

signaling overhead introduced by these schemes is expected to put a huge burden on the network. Therefore, different solutions have been proposed to alleviate this high signaling overhead levels for M2M communications in LTE. Such solutions include backoff method which works by delaying RA attempts for UEs and M2M devices by different backoff times (Seo & Leung, 2011). Another solution is Access Class Barring (ACB) in which random access process is allowed or banned based on probability access parameter broadcasted from the network (Wang & Wong, 2015). Furthermore, in (Jang et al., 2016) a message-embedded random-access scheme is proposed to save radio resources on PUSCH by transmitting small-sized data packets during preambles (PA) transmission on the control channel. In (Lioumpas & Alexiou, 2011), authors proposed two LTE-based M2M scheduling algorithms which consider both channel conditions and the maximum delay tolerance of each device requesting a service. The first algorithm puts more weight on the channel quality for each user while the second one on the maximum delay tolerance. Authors in (Elhamy & Gadallah, 2015) proposed a technique for M2M scheduling over LTE that offers a balance between throughput and delay requirements. This technique is adaptive as its scheduling metric can be adjusted to be delay-based or channel state-based or a hybrid combination of them. In (Mostafa & Gadallah, 2017) authors deal with massive M2M connections problem by introducing a new metric called statistical priority for scheduling process. This metric can be used to evaluate the importance of information sent by M2M devices and allocate the few limited radio resources based on data uniqueness. M2M devices with unique data are given higher priority. Statistical priority metric is calculated by evaluating specific statistical attributes of the data type. Statistical attributes of three data types were handled in this paper, environmental monitoring data which represents periodic low rate data with relaxed deadlines, video surveillance data with large payloads and event-driven data with high reliability and low latency requirement.

A new flexible frame structure is proposed in (Pedersen et al., 2016) and (Pedersen et al., 2015) where an in-resource control signaling scheme is used to create an adaptable radio resource scheduler that serves each user in coherence with its service requirement, particularly latency requirement. The frame is built by resource units with minimum TTI

value for the most stringent latency requirement and each device is flexibly multiplexed on an integer number of these units according to its service need.

Another candidate strategy to overcome the problem of resulting signaling overhead introduced by these schemes is to exploit the periodic nature of most M2M traffic and use a persistent resource allocation scheme in which radio resources are allocated periodically without any additional control signaling for long durations if not for the entire lifetime of the MTC devices. In fact, this is not the first time when an application with periodic data generation is studied to be scheduled in LTE. Voice over Internet Protocol (VoIP) has characteristics in common with some M2M applications, small and periodic data transmission, and there are some proposals in literature to schedule VoIP data persistently in LTE (Jiang et al., 2007). Persistent scheduling allocates radio resources for longer periods instead of each Transmission Time Interval (TTI) which reduces signaling overhead effectively. However, due to very diverse traffic characteristics of M2M communication comparing with VoIP, we cannot use those scheduling algorithms for M2M communication.

The persistent resource allocation schemes for data transmission of MTC devices with diverse QoS requirements on same cellular network are proposed in (Lien & Chen, 2011) (Lien et al., 2011) (Gotsis et al., 2012) (Gotsis et al., 2013) where M2M devices are grouped into clusters of similar QoS characteristics and allocated a periodic access grant time intervals (AGTIs) in which all devices of the same cluster send their data within.

Authors of (Lien & Chen, 2011) and (Lien et al., 2011) proposed an LTE-based massive access management for time-controlled M2M devices which transmit a small amount of data every pre-defined period. M2M devices are grouped into clusters based on their QoS characteristics, mainly period, maximum tolerable jitter and acceptable jitter violation probability. Jitter is defined as the time difference between the time of two consecutive packet departures and the time of two consecutive packet arrivals. A sufficient but not necessary condition is introduced and proved to ensure that devices will not violate their maximum jitter tolerance during the periodic allocation process. The allocation algorithm schedules M2M devices with deterministic or statistical QoS requirement. Devices with deterministic QoS requirement has acceptable jitter violation probability equal to zero.

The base station can opportunistically utilize RBs assigned to a cluster of M2M devices with statistical QoS characteristic.

While authors (Lien & Chen, 2011) proposed a deterministic jitter bound for constant rate time-controlled M2M traffic, authors in (Gotsis et al., 2012) proposed a probabilistic delay bound for event-driven M2M devices with Poisson traffic model. An approximated analytical model for predicting the QoS performance of M2M services is introduced which relates the average traffic intensity rate with scheduling period and the specific QoS metric. Periods are calculated in terms of a statistical QoS metric, namely the delay threshold violation probability. Queue-awareness scheduling is also proposed to enhance the periodic M2M traffic.

In (Si et al., 2015), authors deal with massive MTC connections while keeping QoS requirements, mainly delay requirement, over an LTE-based network. An online-measurement-based adaptive massive access management (AMAM) is proposed which enables eNB to control all AGTI allocation periods and the number of resources allocated for each cluster based on the observed workload without any prior knowledge about the traffic statistics.

The major limitations of these works are two-fold. First, they occupy the entire bandwidth of interest while reserving a time interval to a cluster of MTC devices without considering the bandwidth efficiency and the adverse effects on human-to-human (H2H) communications. Second, meeting the stringent QoS requirements of some M2M applications served in the network becomes impossible due to the interdependence among the QoS requirements of the MTC devices allocated in the same frequency band.

In this thesis, we propose a novel resource allocation scheme that minimizes the bandwidth used by periodic M2M traffic while meeting the diverse QoS requirements of the MTC devices and allowing the admission of new MTC devices in a flexible manner.

### 1.3 Original Contributions

The contributions of this thesis can be summarized as follows;

- We describe the problem of resource allocation of M2M devices with the objective of minimizing the needed bandwidth with a constraint of meeting timing requirement of different M2M applications. We prove NP-Hardness of the problem and show that optimal solution requires an impractical exponential runtime algorithm in the size of a number of devices.
- A heuristic fixed priority-based algorithm is proposed which tries to fully utilize every single band while keep meeting timing constraints of scheduled devices. We analyze the performance of the proposed algorithm in the case of implicit deadlines.
- The need for flexible frame structure for next generations is elaborated and the compatibility of the proposed M2M resource allocation algorithm is analyzed accordingly.
- Through extensive simulations, we show the performance superiority of the proposed algorithm over existing algorithms and its adaptability with flexibility concepts of New Radio (NR).

### 1.4 Organization

The rest of this thesis is organized as follows. Chapter 2 provides a background about scheduling real tasks in distributed systems. In chapter 3 the physical structure flexibility concept is described considering provisioned 5G services with special focus on M2M communications. Then, based on these concepts system model and assumptions are given. Chapter 4 describes and formulates minimum bandwidth resource allocation problem. Chapter 5 and 6 we propose fast minimum-band maximum-utilization algorithms for single and multiple subcarrier spacing cases, respectively. Performance evaluation and results are presented in chapter 7. Finally, thesis work is concluded in chapter 8.

## 2. SCHEDULING PERIODIC TASKS

There is an analogy between scheduling periodic traffic generated by massive Machine Type Communication (mMTC) and scheduling periodic real-time tasks over multiprocessor systems. Building equivalency between them will be discussed in detail later in the chapter. 4 and 5. In this chapter, we will briefly elaborate on the well-studied problem and its proposed single and multiprocessor scheduling algorithms in the literature.

### 2.1 Periodic Tasks Model

Real-time computing systems are widely used in our life for functions like monitoring and control in many industrial and communication applications. Examples are such of engine control, robotics, traffic, time-critical packet communications, avionics systems and nuclear power plants. In such systems, almost all tasks occur infinitely, and their performance relies not only on their logical results but also on the time at which these results are produced. In other words, these tasks have deadlines must be met. If the deadline is critical and missing it causes system failure, it is said to be hard. If it is desirable to meet a task deadline but some missing is tolerable, it is said to be soft (Bertossi & Fusiello, 1997). The following discussion about hard-time tasks. Tasks model is described as follows (Zapata & Alvarez, , 2005):

We have a set of real-time tasks  $S = \{s_1, \dots, s_n\}$  where each task  $s_i \in S$  is characterized by;

- Each task is released at a specific constant rate given by period  $p_i$ .
- All instances of a task have the same worst-case execution time  $C_i$ .
- Deadline  $D_i$ ; ( $D_i = p_i$ )

- All tasks are independent, the arrival of some tasks is not affected by the arrival of any other tasks.
- The utilization factor of each task  $s_i$  is defined as  $u_i = C_i / p_i$  and the utilization factor of a set of tasks  $S$  is the sum of utilization factor of the tasks in the set  $u = \sum_{i=1}^N u_i$ . The maximum value of the total utilization is 1. If  $u > 1$ , some task will fail to meet its deadline no matter the scheduling algorithm is used. If  $u \leq 1$ , it will depend on the scheduling algorithm being used.

## 2.2 Scheduling Algorithms

A scheduling algorithm for periodic real-time tasks specifies an order in which all tasks will be executed while meeting all deadlines of all tasks. Most of the available hard-real-time scheduling algorithms are priority-driven and preemption-based algorithms. Preemption means that any task can be suspended at any time by a higher priority task and can resume later from where it was suspended. Different algorithms use different priority assignment policies. If the priority of a task is fixed and cannot be changed by time, it is called *static priority*. For example, *Rate-Monotonic* (RM) algorithm is of static scheduling where fixed priorities are given to the shortest periods. If the priority of a task is changing from time to time during the running of execution, then it is a *dynamic priority*. For example, *Earliest Deadline First* (EDF) is of dynamic scheduling where tasks with the nearest deadline are given highest priority, so the priority assignment of tasks changes from instant to another (Bertossi & Fusiello, 1997).

The scheduling algorithms decide if a set of arbitrary tasks are schedulable on a single processor or not by checking sufficient and/or necessary conditions. The scheduling problem is proven to be NP-complete and the only know test for general case requires simulating the schedule over an interval equal to the least common multiple of the tasks periods, which can run in exponential time (Bertossi & Fusiello, 1997). We will describe *Rate-Monotonic & Earliest Deadline First* (EDF) briefly in the following subsections.

### 2.2.1 Earliest deadline first (EDF)

As described earlier, *Earliest Deadline First* (EDF) is of dynamic scheduling algorithms in which a task having the nearest deadline is given the highest priority over all tasks. The algorithm can be described as follows,

- Whenever a new task arrives, resort the ready queue so the tasks closest to their deadlines are assigned the highest priority.
- After sorting, preempt the running task if it is not the first in the queue and run the task with the highest priority.

A given set of independent periodic tasks with deadlines equal periods ( $D_i = p_i$ ) for all  $i$  is schedulable by EDF algorithm *iff* the sum of utilization factor of the tasks in the set is;

$$u = \sum_{i=1}^n \frac{C_i}{p_i} \leq 1 \quad (2.1)$$

The previous inequality gave a necessary and a sufficient condition to schedule such a set of tasks using EDF algorithm. It has been proved by Liu and Layland that *Earliest Deadline First* (EDF) is an optimal priority-driven scheduling algorithm, in the sense of EDF can schedule the task set if any algorithm else can (Liu & Layland, 1973). Despite its optimality and simple schedulability test, EDF is not commonly adopted. Dynamic priority assignment is difficult to implement in practice due to the expense of sorting the queue online. Besides that, if any task fails to meet its deadline the next resulting schedule is not predictable. Therefore, it is often preferred to use *Rate-Monotonic* algorithm as described below instead of EDF.

### 2.2.2 Rate monotonic scheduling (RM)

Liu and Layland proposed a preemptive fixed-priority scheduling algorithm for a set of periodic tasks as follows (Liu & Layland, 1973);

- Assume a set  $S = \{s_1, \dots, s_n\}$  of periodic tasks each with deadline equals to its period ( $D_i = p_i$ ) for all  $i$  (implicit deadlines).

- Tasks are independent and preemptive.
- All tasks are always released simultaneously (critical instant).
- Priorities are assigned inversely to task periods, hence task  $s_i$  gets higher priority than task  $s_j$  if  $p_i < p_j$ .
- In order for a fixed priority assignment to be feasible, only the first deadlines of each task starting from a critical instant should be met.

*Rate-Monotonic* algorithm is optimal among *static* scheduling algorithms only, that is if a task set is schedulable with any fixed-priority scheduling algorithm, it is also schedulable by the *Rate-Monotonic* algorithm. *RM* algorithm has many schedulability tests as detailed in (Bertossi & Fusiello, 1997) and (Zapata & Alvarez, , 2005), two of them are described below;

#### 1. Utilization Bound (UB)

Based on the notion of critical instant, Liu & Layland (Liu & Layland, 1973) derived the following schedulability test for *Rate-Monotonic* algorithm. Given a set of  $S = \{s_1, \dots, s_n\}$  of periodic tasks each with deadline equals to its period ( $D_i = p_i$ ) for all  $i$ , the RM algorithm produce a feasible schedule based on priority assignment if,

$$u = \sum_{i=1}^n \frac{C_i}{p_i} \leq n(2^{1/n} - 1) \quad (2.2)$$

This bound depends only on the number of tasks and under this utilization bound *Rate-Monotonic* algorithm always yields a feasible priority assignment. The condition is *sufficient but not necessary*, hence, if a set of tasks meet the condition then all tasks will meet their deadlines. Nevertheless, there can be a case where the total utilization of its tasks is greater than utilization bound (2.2) and the set is still schedulable by the *Rate-Monotonic* algorithm. Thus, the test may be too conservative.

#### 2. The Completion Time Test (Exact Test)

An exact test was derived in (Joseph & Pandya, 1986) to find the worst-case response time for a given task assuming independent tasks, fixed priority and

deadlines less than periods ( $D_i \leq p_i$ ). The worst-case response time is given when a task is released simultaneously with all higher priority tasks. The following equation gives the worst-case response time  $R_i$  for a task  $s_i$ ;

$$R_i = C_i + \sum_{j \in hp(i)} \left\lceil \frac{R_i}{p_j} \right\rceil * C_j \quad (2.3)$$

$\left\lceil \frac{R_i}{p_j} \right\rceil$  is the number of task  $j$  instances during  $R_j$ ,  $\left\lceil \frac{R_i}{p_j} \right\rceil * C_j$  is the needed time to execute all instances of task  $j$  released within  $R_j$  and  $\sum_{j \in hp(i)} \left\lceil \frac{R_i}{p_j} \right\rceil * C_j$  is the time needed to execute instances higher priority tasks than task  $i$  released during  $R_j$ .  $R_j$  is the sum of the time required for executing task instances with higher priorities than task  $j$  and its own computing time. Solving equation (2.3) can be done by iteration as follows;

$$R_i^{k+1} = C_i + \sum_{j \in hp(i)} \left\lceil \frac{R_i^k}{p_j} \right\rceil * C_j \quad (2.4)$$

The iteration stops when either  $R_i^{k+1} > (D_i = p_i)$  (non-schedulable) or  $R_i^{k+1} = R_i^k$  and  $R_i^k < (D_i = p_i)$  (schedulable). This test should be repeated for all tasks of a given set, if all pass the set is schedulable, if some tasks pass they will meet their deadlines even the other don't.

### 2.3 Partitioned Multiprocessor Tasks Scheduling Algorithms

In order to schedule a set of real-time tasks over multiprocessor system it is necessary to find an allocation algorithm to allocate tasks between the available processors first, then schedule the allocated tasks for each processor using one of the scheduling algorithm described in section 2.2. Assuming the number of available processors is infinite, the allocation of a set of real-time tasks to these problems is analogous to the *bin-packing* problem. *The bin-packing* problem can be described as follows; we have a set of different objects each with different weight and an unlimited number of available bins all have the same capacity. We need to allocate these objects to these bins such that the minimum number of bins will be used. In our problem, real-time tasks represent the objects and the

utilization of each task represents object's weight whereas processors represent the bins with a capacity equals to utilization bound.

Since the *Bin-packing* problem is a well-known NP-hard problem, optimal algorithms for solving it cannot solve this in polynomial-time. Therefore, many heuristic algorithms have been proposed in the literature to solve the allocation problem. Most famous algorithms with their performance analysis are briefly described below.

- First-Fit (FF)

The First-Fit algorithm allocates a new object (task) to the lowest indexed non-empty bin (processor) such that the total weights (utilization) of the newly added object (task) along with the existing ones do not exceed the bin capacity (utilization bound). If there is no such a non-empty bin (processor), allocate the object (task) to a new empty bin (processor).

- Best-Fit (BF)

Best-Fit algorithm allocates a new object (task) to a bin (processor) among a set of non-empty bins (processors) which have the smallest capacity left (maximum total utilization). If two non-empty bins (processors) have the same capacity (total utilization) available allocate the object to the lower indexed bin (processor). If allocating the object (task) on all non-empty bins (processors) exceeds the bin capacity (utilization bound), allocate the object (task) to a new empty bin (processor).

- Next-Fit (NF)

Next-Fit algorithm allocates a new object (task) to the bin (processor) which the previous object (task) was allocated to. If it does not fit (a task not allocable), the new object (task) is allocated to a new empty bin (processor). This algorithm does not check if the previous bins (processors) can allocate the new object (task) or not.

The guaranteed performance of these heuristic algorithms is evaluated by the following equation;

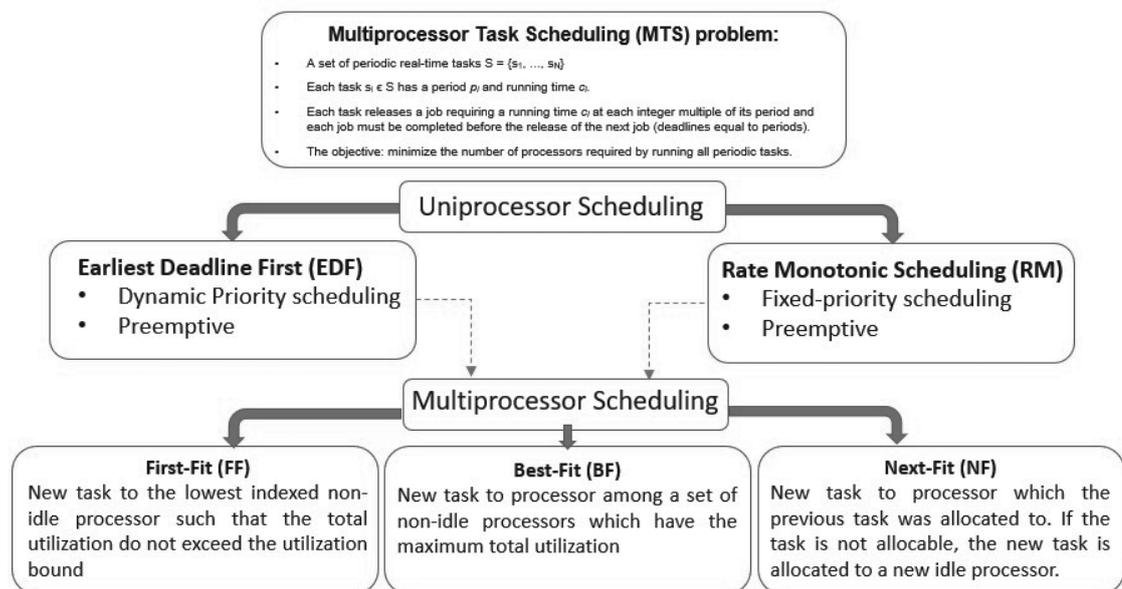
$$\mathfrak{R}_A = \lim_{N_{opt} \rightarrow \infty} \frac{N(A)}{N_{opt}} \quad (2.5)$$

Where  $N_{opt}$  denotes the number of used processors by an optimal algorithm, and  $N(A)$  is the number of used processors by algorithm A. Note that, the smaller (close to 1)  $\mathfrak{R}_A$  the value provided by algorithm A, the closer to the optimal solution. The following table (Table 2.1) presents the worst-case performance ratio for the previous allocation algorithms using RM or EDF as scheduling algorithms on single processor (Bertossi & Fusiello, 1997) (Zapata & Alvarez, , 2005).

**Table 2.1 - Worst-Case Performance Ratio For Task Assignment Heuristics**

Algorithm	RMFF	RMBF	RMNF	EDF-FF	EDF-BF
$\mathfrak{R}_A$	2.33	2.33	2.67	1.7	1.7

The following Figure (1.2) summarizes single and multiprocessor task scheduling algorithms.



**Figure 2.1 - Summary of Multiprocessor Task Scheduling Algorithms**

### **3. FLEXIBLE PHYSICAL LAYER ARCHITECTURE**

The impressive success of LTE-A mobile system in providing reliable, robust and spectral effective Mobile Broad Band (MBB) services was by means of using a customized numerology of OFDM as a basic waveform. There is a consensus that the main driving applications for 5G will need higher data rates, mobility, power efficiency, ultra-low latency, reliability and massive connectivity. Essentially, 3GPP has named three applications to be served, enhanced Mobile Broadband (eMBB), massive Machine Type Connectivity (mMTC) and Ultra-Reliable and Low Latency Communications (URLLC) (Ankarali et al., 2017) (Mansoor, et al., 2017). However, these highly diversified applications and heterogeneous services cannot be supported by one-for-all radio access technology (RAT) as in LTE and previous generations. Therefore, flexible radio design technologies and concepts are very important for future mobile generations. For that end, many types of research have been conducted to find either better waveform that overcomes OFDM weaknesses or readjust OFDM parameters in a service-based manner (Pedersen et al., 2016) (Pedersen et al., 2015), (Ankarali et al., 2017) (Mansoor, et al., 2017) (Zaidi, et al., 2016) (Sahin & Arslan, 2012) (Schaich et al., 2016) (Incorporated, Qualcomm, 2016).

The ultimate flexibility for any system can be obtained by playing with its very basic components and the bedrock for any wireless system is its physical layer structure. Physical layer building process starts with choosing a waveform which suits the targeted application of the wireless system. Then, waveform's parameters are mainly adjusted according to propagation channel characteristics and application's traffic and QoS requirements; a process called numerology design. Finally, the frame structure is drawn to contain data units generated by system users. Every stage of this process provides different flexibility level and we will briefly elaborate on some of them in the following sections.

### 3.1 Flexibility & Waveforms

Back to the main driving applications of 5G and the projects which have been initiated to draw its aspects and features (5GNow, METIS, FANTASTIC-5G...etc.), many waveforms have been investigated as potential candidates to serve wide variety and heterogeneity in services for 5G and beyond mobile systems (Roessler, A, 2016). Before talking about new waveform candidates, let us present main OFDM waveform drawbacks. CP-OFDM has become the dominant waveform for LTE system, Wi-Fi, and even many wireline communication systems such as digital subscriber line because of its optimal advantages in broadband applications. OFDM is a multicarrier transmission scheme which subdivides the available bandwidth into several subchannels called subcarriers. The spacing between these subcarriers is chosen such a way they are not frequency selective. OFDM-based access schemes benefit from the following advantages:

- Overlapped but orthogonal subcarriers provide high spectral efficiency.
- Introduced Cyclic Prefix CP increased robustness against Inter-Symbol-Interference (ISI) caused by multipath propagation.
- Since spatial interference from different antenna transmission is dealt with at a subcarrier level without extra complications of ISI, an excellent integration with MIMO system is offered using OFDM.
- OFDM make it possible not only to separate multiple users in the frequency domain using resource blocks (RB) but also scheduling these resource blocks in the time domain (every TTI) using Time Division Multiple Access (TDMA), altogether forming Orthogonal Frequency Division Multiple Access (OFDMA) scheme.

Showing all these advantages makes OFDM ultimate for LTE broadband services. However, some weaknesses are do exist of this waveform as listed below.

- High Peak-to-Average-power ratio (PARP): the summation of the individual uncorrelated subcarriers which have typically different phases but the same value at some instants leads to a peak value in output power. This peak value can

be very high compared to the average value. High PARP puts extra complications on power amplifiers lead to extra costs.

- **Poor spectral confinement:** side lobes of OFDM which uses rectangular pulse shape result in high out-of-band emissions and introduce the need of guard bands to ensure sufficient signal isolation. In addition, discontinuity of OFDM symbols creates spikes in the frequency domain at transition intervals. So, to overcome this problem, a windowing technique is used but at the cost of spectral efficiency. This drawback will cause unaccepted bandwidth utilization efficiency loss due to needed guard bands for co-existence of different 5G applications using OFDM.
- **The strict orthogonality and synchronism requirements of OFDM waveform** make it very sensitive to any frequency or time offsets. Therefore, Offsets caused by asynchronous access of massive M2M devices or by high Doppler shifts in vehicular applications need different waveform with relaxed orthogonality and synchronism constraints.
- **Cyclic Prefix Redundancy & Overhead:** CP is a copy of a symbol tail pasted at its beginning to reduce ISI caused by delay spread. Anyway, this copy is a redundant information and considered as overhead. Considering URLLC as one of the main applications of 5G, which its use cases have stringent latency and successful delivery requirements, CP will affect these applications negatively.

Referring to these limitations of OFDM waveform, 5G related projects looked for other candidates which can overcome these shortcomings. It can be noticed that the common idea between all suggested schemes is to 1) reduce out-of-band emissions by using different pulse shaping filters from OFDM, thus increase spectral efficiency and 2) introduce flexibility for the future heterogeneous mobile applications. Suggested waveform candidates may be categorized into two classes: subcarrier level filtering, and sub-band level filtering. The following part describes the main proposed waveforms briefly.

- **Filter-bank Multicarrier (FBMC) with OQAM:** this is a subcarrier-wise filtered waveform which allows choosing individual pulse shaping filter

(rectangular, raised cosine...) per subcarrier. This advantage alleviates strict synchronization requirement of OFDM and facilitate flexible adjustment of SC spacing and symbol duration within the same band which makes it appropriate for mMTC except for its inefficiency for burst transmission due to long filter tails. Unlike OFDM, there is no exist of CP in FBMC which make it more efficient in spectrum utilization.

- **Universal Filter Multicarrier (UFMC):** This is a sub-band-wise filtered waveform (a group of subcarriers) which decreases side lobes emissions like FBMC but with less overhead and suitability for burst and low latency transmission which makes it a better candidate for M2M communications. Instead of using CP, sub-band filters were introduced to reduce Out of Band Emissions (OBE) whereas Zero Prefix (ZP) provide protection against Inter-Symbol Interference (ISI).
- **Generalized Frequency Division Multiplexing (GFDM):** This waveform is particularly suitable for non-contiguous frequency bands offering empty frequency holes aggregation. It has lower PARP comparing to OFDM, but it needs more complicated receivers.
- **Flexibly Configured OFDM (FC-OFDM):** This waveform is considered as compromising solution between CP-OFDM and FBMC. Comparing to CP-OFDM, part of the cyclic prefix CP is sacrificed by additional windowing process for more spectrum confinement purpose. In the other hand, FBMC does not use the cyclic prefix at all.

There are some other waveform candidates which can be considered as an extension of the presented waveforms and share them the main advantages of reducing out-of-band OBE emission and provide flexibility for variant services, but each with the specific feature that may be more appropriate for some use cases. Among these candidates, FS-FBMC shows robustness against high delay spreads caused by asynchronous access of massive M2M devices but in a cost of increased Inter-Symbol-Interference ISI which hinder short burst transmission. Zero-Tail OFDM (ZT-OFDM) which has adjustable zero tail provide robustness against time and frequency dispersions but in the cost of high overhead scaling with tale length.

It is obvious from above discussion the research efforts paid to diversify options in front of system designers of 5G standard with all these waveform candidates to use them facilitating multi-service coexistence and pick the appropriate waveform for each application type. However, this degree of flexibility needs further study to ensure peaceful and smooth co-existence of different waveforms within the same frequency band.

### 3.2 Flexibility & Numerology Design

Every waveform has its own parameters which need to be determined to achieve the target application's QoS requirements taking channel conditions and service requirements into account. Setting values for these parameters all together is called numerology design. Talking about plain OFDM waveform which is used in the latest mobile generation (LTE/LTE-A), its parameters were chosen in a static and strict way to serve one main application, mobile broadband (MBB), in accordance with propagation environment. These parameters are mainly subcarrier spacing (15 kHz), a number of subcarriers (12 per RB) and cyclic prefix CP which is determined basically on maximum delay spread and Doppler spread in propagation channel (Normal CP= 5.2  $\mu$ s). This numerology design is optimum for LTE use cases but will not be so considering the future diversity of applications in the 5G system and beyond. Other waveforms; like FBMC, UF-OFDM/UFMC; have different parameters that can be used to draw several service-customized numerologies. We will briefly describe some waveform-specific flexibility parameters as follows:

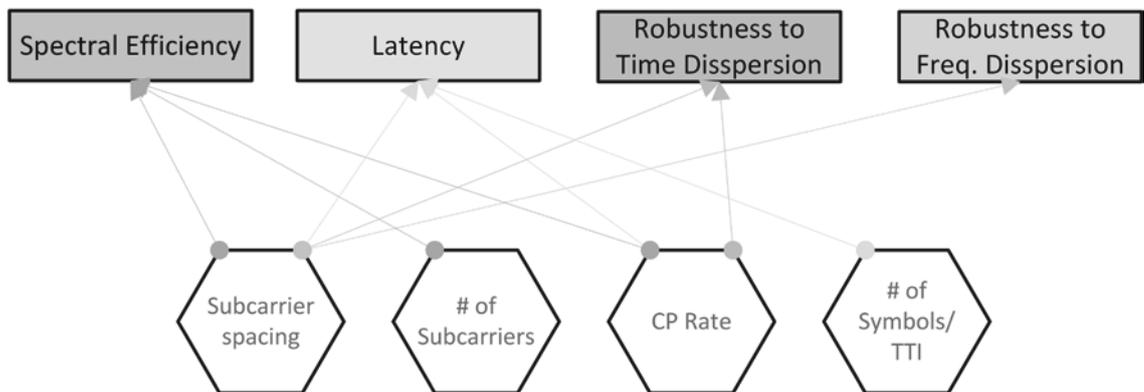
- **Subcarrier spacing:** since all proposed waveform candidates for future 5G and beyond mobile generations are of multi-carrier waveforms family, all of them share the parameter of the subcarrier spacing. The most important feature of multi-carrier waveforms is its high robustness against time and frequency dispersions of the channel, delay spread, and Doppler spread and frequency. This is done by dividing the available spectrum into smaller parallel subcarriers. The spacing of each subcarrier ( $\Delta f$ ) should be less than channel coherence bandwidth, which depends on delay spread, to ensure flat fading. In addition, increasing it significantly causes high cyclic prefix overhead. However, too small subcarrier spacing will increase symbol duration largely and make the system sensitive to

Doppler and phase noise. Therefore, subcarrier spacing must be large enough to keep symbol duration larger than the coherence time of the channel (Ankarali et al., 2017) (Zaidi, et al., 2016). These restrictions impose limits on subcarrier spacing to be chosen from between. In fact, the selection of subcarrier spacing thereafter depends on targeted service QoS requirements. Subcarrier spacing is in an inverse relationship with useful symbol duration; small subcarrier spacing means large symbol duration and vice versa. In LTE, this value of 15 kHz using CP-OFDM waveform results in symbol duration of 66.7  $\mu$ s, which was optimal for high data rates of mobile broadband services. Increasing subcarrier spacing to get shorter symbol duration is more suitable for low latency applications like tactile internet and URLLC. On the other hand, decreasing subcarrier spacing is preferable for massive connectivity, e.g. M2M applications. It also alleviates the effect of delay spread. Large symbol duration can also reduce overhead caused by cyclic prefix CP and thus increase spectral efficiency. In addition, cases like macrocells which have extended coverage over wide areas will profit from large symbol duration to overcome propagation delays and serve cell-edge users fairly. It is obvious from the previous discussion that sticking to one-for-all subcarrier spacing will not meet QoS requirement of this wide range of heterogeneous applications. Therefore, co-existence of different numerologies with different service-customized subcarrier spacings within the same assigned bandwidth is a strong candidate technique for future mobile networks. Anyway, applying this technique using CP-OFDM is spectral inefficient due to large guard bands needed for isolation between different numerologies. Thus, we need for more spectral localization using new candidates of multicarrier waveforms like UFMC. Using this degree of flexibility has been discussed in several recent types of research (Ankarali et al., 2017) (Zaidi, et al., 2016) (Schaich et al., 2016).

- **Number of Subcarriers:** as we know, the more data transmission rates the more bandwidth is needed. Therefore, to increase speed for a given subcarrier spacing, we need to increase the number of subcarrier per subchannel. This also may provide a relative degree of flexibility as a common parameter between all multicarrier waveforms.

- **Number of symbols & TTI:** This parameter plays a significant role to determine transmission time interval (TTI) duration for a given frame structure. Thus, another dimension of flexibility is introduced which can be used to serve different latency requirements of various users.

Figure 3.1 presents different adjustable waveform parameters. There are other waveform-specific parameters which may be used flexibly for designing co-existed multiservice numerologies. For example, since FBMC waveform applies filtering on subcarrier level, this allows for using different pulse shaping filters per subcarrier. Therefore, another flexibility aspect is introduced which can be used to meet different user requirements using different filter types. UFMC also has its specific parameters like sub-band filter length and Zero Prefix (ZP) length which can be used for flexible implementation for different scenarios like in (Ijaz, et al., 2016).

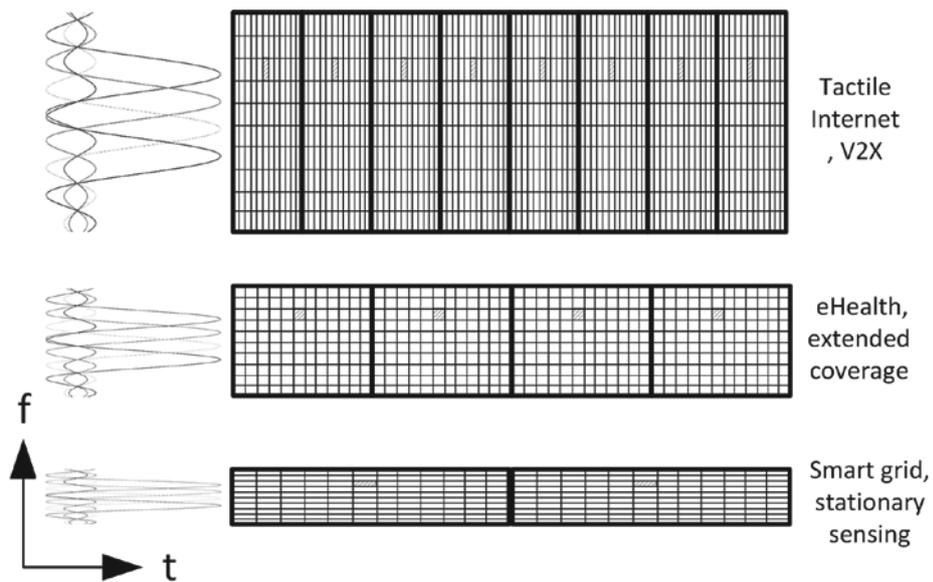


**Figure 3.1** - Different Waveform Parameters Providing Flexibility

### 3.3 Beneficial Use for M2M Communications

Machine Type Communication (MTC) as a main application targeted by 5G ongoing studies may effectively profit from this flexible approach. A significant part of MTC traffic comes from a large number of stationary sensors (e.g. smart homes, metering...) deployed over wide areas and produces the sporadic and small amount of data. For such kind of communications there is no Doppler effect, so using narrow subcarrier spacing is more convenient especially when the application is delay tolerant. In addition, instead of increasing power spectral density to extend coverage area, using smaller subcarrier

spacing stretch transmission over time allowing usage of cheaper and less energy-consuming battery-operated devices (Schaich et al., 2016). There are other MTC applications which have mobility characteristic (e.g. V2X) and using narrow subcarrier spacing will lead to high Doppler spread causing an increase in inter-carrier interference (ICI). Therefore, subcarrier spacing value should be wide enough to alleviate Doppler spread while keeping accepted CP overhead. There may be further MTC applications with stringent latency requirement which cannot be served by currently used TTI length. For such applications (e.g. e-Health), TTI length can be shortened by increasing subcarrier spacing which in turn shrinks symbol duration and TTI length while keeping the same number of symbols (Mansoor, et al., 2017) (Schaich et al., 2016) (Incorporated, Qualcomm, 2016). Figure 3.2 depicts the idea of flexible adjustment of subcarrier spacing to meet various M2M application requirements.



**Figure 3.2** - Flexible Subcarrier Spacing for Heterogeneous Service Requirements

For adjacent TDD networks which use different OFDM numerologies, it is desired that an integer number of subframes from one OFDM numerology fits into one subframe of other OFDM numerology with higher subcarrier spacing value to enable time aligned uplink and downlink periods. Otherwise, two adjacent TDD numerologies would require guard time to enable synchronous operation which is considered as the non-efficient use of resources (Zaidi, et al., 2016). Therefore, different subcarrier spacing values may be

generated using a base value  $\Delta f$  multiplied by a *scaling factor* of  $q_i = 2^{i-1}$ ,  $i \in \mathbb{N}$  (the set of natural numbers), such that a subcarrier spacing is an integer divisible by all smaller subcarrier spacing values;

$$f_i = 2^{i-1} \Delta f, \forall i \in \{1, 2, \dots, n\} \quad (3.1)$$

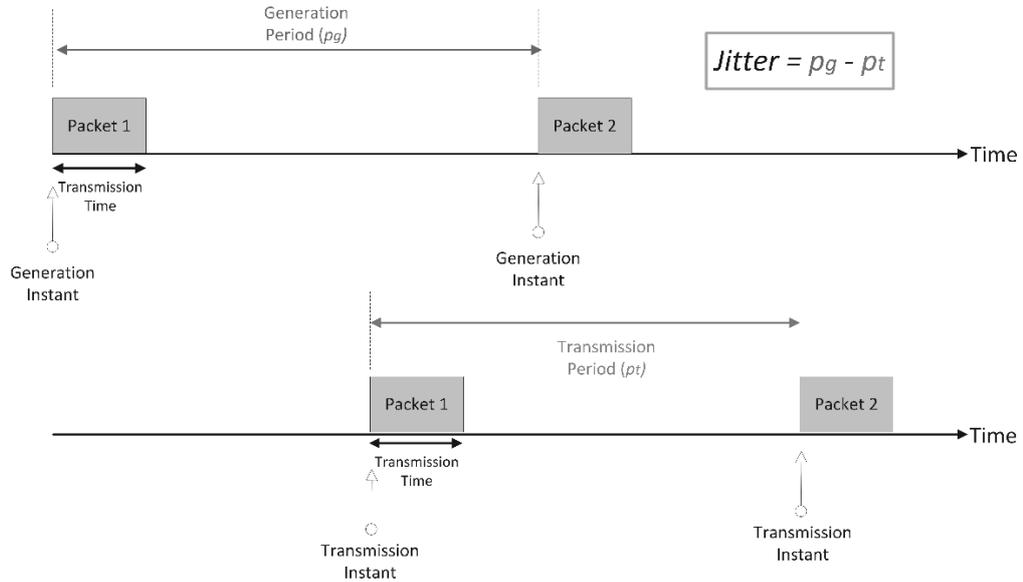
This idea of scaling subcarrier spacing values by  $2^i$  scaling factor is already applied in 3GPP standard for narrowband IoT NB-IoT where the commonly used subcarrier spacing value  $\Delta f = 15$  kHz of LTE-OFDM is downscaled by  $q = 2^2$  factor resulting in  $\Delta f = 3.75$  kHz.

### 3.4 System Model and Assumptions

The system model and assumptions are described as follows:

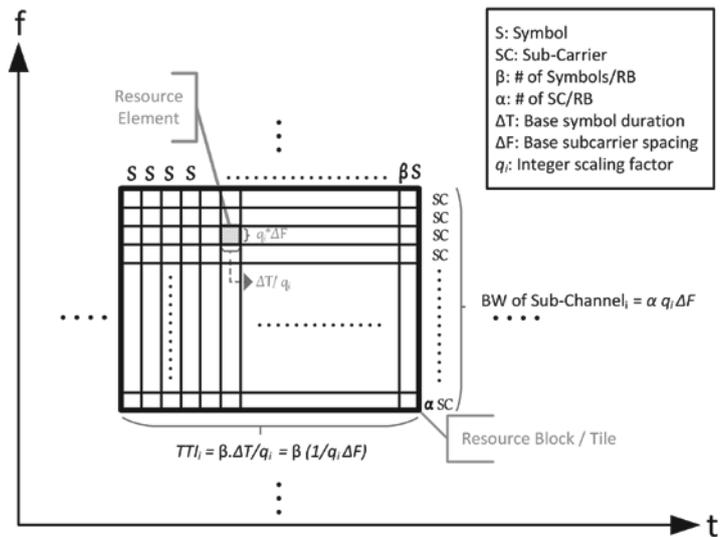
- 1) We consider a cellular system with a base station which serves a large number of M2M devices with diverse traffic characteristics in addition to H2H devices using different separated sub-bands for M2M devices and H2H UEs.
- 2) Most M2M applications involve time-triggered devices generating periodic data, in such applications as smart grid, e-health applications, intelligent transportation and industrial supply systems. This type of M2M devices generates a small amount of data (small packets) every pre-defined period  $p_i$ . The QoS requirements of time-triggered M2M devices can be captured by jitter. Jitter is defined as illustrated in Figure 3.3 by the time difference between the time of two consecutive packet departures and the time of two consecutive packet arrivals (Lien & Chen, 2011). Time-triggered M2M devices have a maximum allowable jitter that we call jitter tolerance  $\delta_i$ . The value of  $\delta_i$  for each time-triggered device can be at most equal to its period  $p_i$  such that the transmission of a packet has a deadline equals to its period, the case which we call implicit deadlines, otherwise the periodicity itself will be violated. This jitter tolerance can be determined by criticality of the application being served. Satisfying deterministic QoS requirements is critical in many applications, especially in safety-critical operations such as navigational data communications or health-care applications. There are different M2M applications involve event-

triggered non-periodic machines generating data at random intervals. For these devices, QoS requirements are captured mainly by latency.



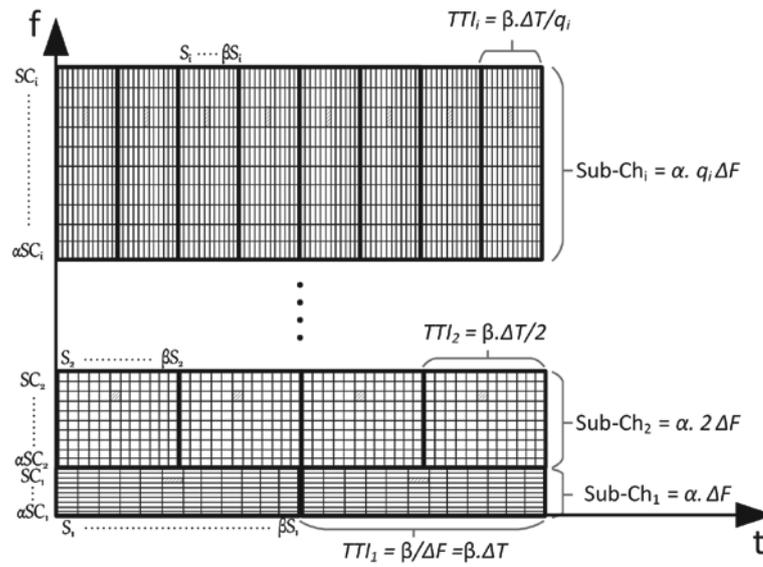
**Figure 3.3 - Jitter Requirement Definition**

- 3) Each M2M device is allocated a set of time-frequency radio resource elements forming together a tile called Resource Block RB. The structure of an RB is illustrated in Figure 3.4. As depicted, each RB has a certain number of subcarriers  $\alpha$ .SC, and time symbols  $\beta$ .S. These subcarriers have the same frequency width within RB, which is called subcarrier spacing ( $q_i \Delta f$ ); where  $q_i \in \{1,2,3,\dots\}$  is called *scaling factor* and  $\Delta f$  is a base subcarrier spacing value. This produces a useful symbol duration of  $\Delta T_i = 1 / (q_i \Delta f)$  identical for all subcarriers in one RB. The number of symbols  $\beta$  along with useful symbol duration  $\Delta T_i$  determines the length of one RB in time (Transmission Time Interval  $TTI_i$ ). In LTE, an RB is a time-frequency unit with  $\alpha=12$  subcarriers each with subcarrier spacing value of  $\Delta f=15$  kHz producing 180 kHz subchannel bandwidth in frequency, and  $\beta=7$  symbols in time producing  $TTI=0.5$  ms length in time (using CP-OFDM as a waveform which adds normal/extended cyclic prefix to TTI duration). RB-Based granularity is expected to be preserved in 5G cellular networks, even though the size may change.



**Figure 3.4 - Resource Block Structure**

- 4) Keeping the number of subcarriers and symbols,  $\alpha$  and  $\beta$ , in one RB constant, then changing the value of subcarrier spacing ( $q_i \Delta f$ ) by specifying different values of *scaling factor*  $q_i \in \{1,2,3,\dots\}$  yields different RBs which can carry the same amount of data but with different values of  $TTI_i = \beta \cdot \Delta T / q_i$  (wider in frequency but narrower in time) as illustrated in Figure 3.5. Maintaining an equal number of OFDM symbols per subframe for all numerologies simplifies scheduling and reference signal design.
- 5) For sake of simplification, we assume identical channel conditions and unified modulation scheme; e.g. BPSK, for all devices. In addition, we assume that each time-triggered M2M device sends only one packet per period and since the number of resource elements contained by each RB is constant for all scaled subcarrier spacing values ( $\alpha$  and  $\beta$  are constants), each M2M device can be scheduled on any subcarrier spacing value according to the scheduling algorithm. After scheduling a device on a subcarrier, all its data should be transmitted using this subcarrier only and its packets cannot migrate from one subcarrier to another. Transmission of any device's packet is independent of any other device transmission.



**Figure 3.5** - Multi-subcarrier spacing physical structure

- 6) If CP-OFDM waveform is used to generate different numerologies, there should be cyclic prefix (CP) duration added to each symbol time such that  $TTI_i = \beta (\Delta T + CP) / q_i$  (Zaidi, et al., 2016), but for the sake of simplicity and generality we will ignore adding this CP since it is divided by the same scaling factor  $q_i$  for all symbols within one RB.

## 4. MINIMUM BANDWIDTH RESOURCE ALLOCATION PROBLEM

Depending on the region, most probably 2 GHz of fragmented spectrum under 6 GHz is available for future mobile communications and it is worthy to note that less than half of this available spectrum is used today by mobile networks. The available spectrum is divided between FDD and TDD operations with domination for FDD in lower frequencies (under 3 GHz) due to the more favorable radio propagation conditions to provide wider area coverage and higher outdoor-indoor penetrations. The scarcity and non-contiguity of the available spectrum called for efficient utilization solutions. In LTE, spectrum aggregation is introduced in the form of carrier aggregation (cell aggregation). Motivated by the scarcity of the available spectrum for the wide heterogeneous provisioned applications in 5G, we describe the following problem.

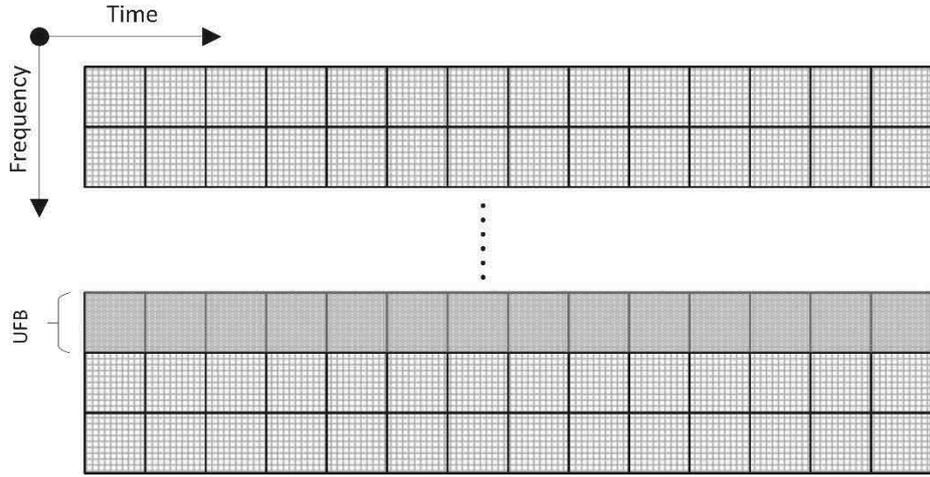
### 4.1 Problem Description

In this section, we describe the minimum bandwidth resource allocation (MB-RA) problem for machine-type communications in 5G and beyond cellular networks. The goal of the problem is to minimize the total bandwidth required by the allocation of a set  $S$  of time-triggered MTC devices. Each MTC device  $i \in S$  has a packet generation period  $p_i$  and maximum tolerable jitter  $\delta_i$  and must be allocated one RB each  $p_i$  to transmit its packet before the generation of the next packet without violating its jitter requirement.

**Definition 1.** A frequency band of 1 RB width is defined as a Unit Frequency Band (UFB). One UFB is considered as the minimum frequency allocation unit. (Figure 4.1)

Based on the above definition of UFB, the objective of the problem can be alternatively stated as to minimize the number of UFBs occupied by a set  $S$  of time-triggered MTC devices.

Recalling the physical layer time-frequency grid structure of RBs, we can say that if we were able to allocate  $N$  devices fully in a grid without leaving any single empty RB, then we reach the optimal point of minimum needed bandwidth and our bandwidth is fully utilized. Thus, the objective of minimizing the needed bandwidth for allocating a set of time-triggered M2M devices can be interpreted as maximizing the number of devices that can be scheduled on a single UFB. Here we can define the following metrics.



**Figure 4.1** - A Unit Frequency Band (UFB)

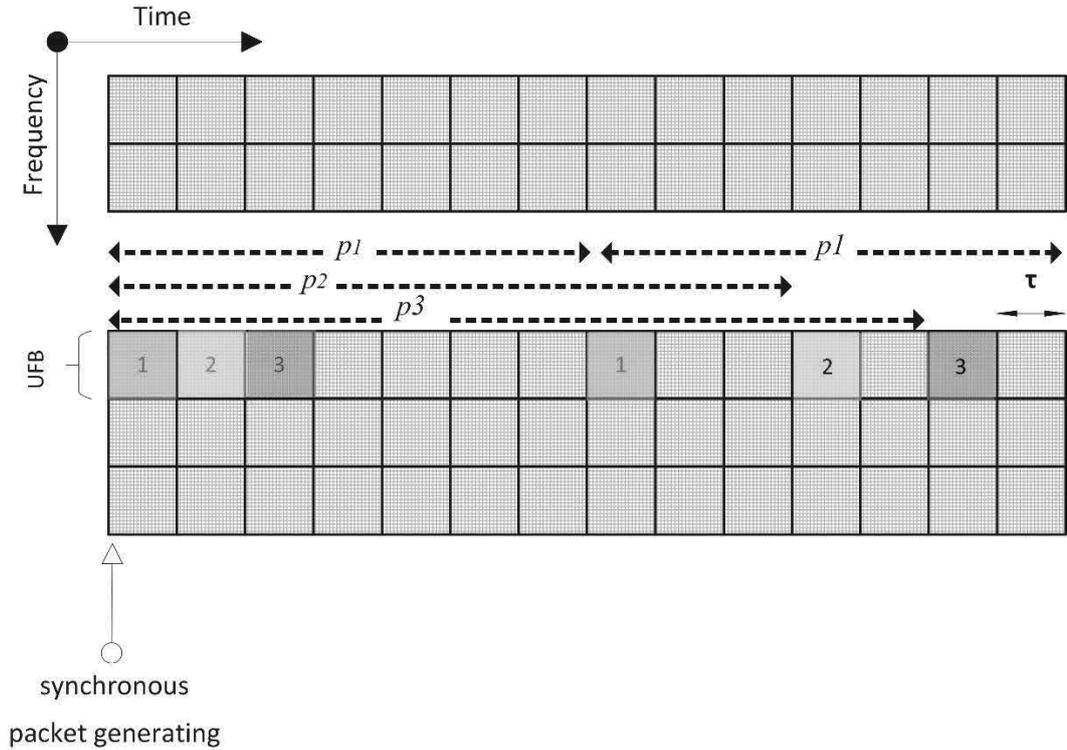
**Definition 2.** For any time-triggered M2M device with transmission time  $\tau_i$  and packet generation period  $p_i$ , its *band utilization* can be defined as,

$$u_i = \frac{\tau_i}{p_i} \quad (4.1)$$

Then, the *UFB utilization* of a set of time-triggered devices on a single UFB is defined as;

$$U_{UFB} = \frac{\tau_1}{p_1} + \frac{\tau_2}{p_2} + \dots + \frac{\tau_i}{p_i} = \sum_i u_i \quad (4.2)$$

In light of UFB utilization definition (Figure 4.2), we can alternatively define the problem of minimizing the needed bandwidth for a set of time-triggered M2M devices as maximizing the utilization of every single UFB while keep meeting each device QoS constraints.



**Figure 4.2 - UFB Utilization Definition**

**Definition 3.** Bandwidth efficiency can be defined as the ratio between a given number of M2M devices  $N$  and the needed bandwidth to allocate them as follows;

$$\eta = \frac{N}{BW} \quad (4.3)$$

Considering this definition, the problem of minimizing the bandwidth can be stated as maximizing the bandwidth efficiency or in other words minimizing the average bandwidth allocated to each M2M device  $1/\eta$ .

## 4.2 NP-Hardness

**Theorem 1.** *The minimum bandwidth resource allocation problem is NP-hard.*

**Proof:** The bin-packing problem is a special case of MB-RA problem. The objective of the bin-packing problem is to find a feasible partition of a set of items with different sizes,  $s_i \in (0, 1]$  for  $i=1 \dots, N$ , into minimum number of bins such that the total size of the items in a bin cannot exceed 1, the size of a bin. Consider an instance of MB-RA problem such that a set  $S$  of devices generate packets periodically starting at the same time, i.e., synchronous packets, and maximum tolerable jitter values are equal to packet generation periods; i.e., implicit-deadlines case,  $p_i = \delta_i$  for all  $i \in S$ . Each device  $i \in S$  should be allocated one RB with duration  $\tau$  with period  $p_i$  where  $p_i$  is an integer multiple of  $\tau$ . Then, the equivalence of the problems arises after introducing the notion of utilization of a device on a band as  $u_i = \tau / p_i$ . A set of devices can be feasibly allocated in a band if and only if the total utilization of the devices on that band is less than 1, the capacity of a band. The objective of minimizing the number of bands in MB-RA problem is equivalent to the objective of minimizing the number of bins used in the bin-packing problem. Therefore, since the bin-packing problem is NP-hard, MB-RA problem is also NP-hard.

Proving the NP-hardness of the problem ensures that the MB-RA problem cannot be solved optimally using polynomial-time algorithms requiring a runtime polynomial in the size of the problem size. On the other hand, considering the massive machine connectivity envisioned in 5G and beyond cellular networks in which thousands of MTC devices are expected to be served by a single base station, exponential-time algorithms will be intractable. The radio resource allocation algorithms for MTC devices should be computationally simple besides being effective. In the following, we propose a fast and efficient polynomial-time algorithm with a guaranteed performance result with respect to the optimality for the implicit-deadlines case.

### 4.3 Optimization Problem

The optimization problem is formulated as follows:

$$\text{Minimize} \quad \sum_{k=1}^K z_k f_k \quad (4.4)$$

$$\text{S.t.} \quad \sum_{k=1}^K x_{ik} = 1, \forall i \in [1, M] \quad (4.5)$$

$$\sum_{i=1}^M x_{ik} \leq M z_k, \forall k \in [1, K] \quad (4.6)$$

$$\{x_{ik}, p_i, \delta_i, f_k\} \in S^{feasible}, \forall k \in [1, K] \quad (4.7)$$

$$\text{variables} \quad z_k \in \{0,1\}, \forall k \in [1, K]$$

$$x_{ik} \in \{0,1\}, \forall k \in [1, K], \forall i \in [1, M]$$

$$f_k \in \Delta f \times \{1, 2, \dots, 2^{N-1}\}, \forall k \in [1, K]$$

Where  $z_k$  is a binary variable taking the value 1 if any device is allocated to band  $k$ ,  $x_{ik}$  is a binary variable taking the value 1 if device  $i$  is allocated in band  $k$ , and  $f_k$  is a discrete variable representing the subcarrier spacing value for band  $k$ . Equation (4.4) represents the objective of minimizing the total bandwidth required by the allocation of  $M$  devices. Equation (4.5) states that each device  $i$  must be allocated in one band. Equation (4.6) represents the constraint that a band  $k$  is used in the schedule, i.e.,  $z_k = 1$ , if and only if at least one device is allocated in band  $k$ , i.e.,  $x_{ik}$  value must be equal to 1 for at least one  $i \in [1, M]$ . Finally, Equation (4.7) represents the schedulability constraint for a set of devices allocated in the same band.

The above optimization problem formulation is a Nonlinear Integer Optimization Problem which requires an exponential runtime effort to solve optimally. Moreover, it may be simply intractable to solve for large number of devices. In the following chapters, we propose fast and efficient polynomial-time algorithms.

## 5. FAST MINIMUM-BAND MAXIMUM-UTILIZATION ALGORITHM (SINGLE SUBCARRIER CASE)

In this section, we describe Fast Minimum-Band Maximum-Utilization (FMM) Algorithm and analyze its performance theoretically.

MTC devices with a common packet generation period are grouped into a cluster. Set  $S$  of time-triggered devices is grouped into  $M$  clusters  $\{C_1, C_2, \dots, C_M\}$ . We consider fixed priority among MTC clusters which dramatically reduces the complexity of an algorithm with respect to a dynamic-priority counterpart. Priorities are assigned to the clusters in decreasing order of the packet generation periods; i.e., a lower period implies a higher priority. The allocation of a packet of a higher priority cluster device is prioritized to the allocation of lower priority ones. Each cluster  $C_i$  includes  $N_i$  devices with a common packet generation period  $p_i$ .

### 5.1 Algorithm Description

FMM Algorithm, illustrated in Algorithm 1, is described as follows. MTC devices are grouped into  $M$  clusters, in decreasing order of priority, each having  $N_i$  devices. Each element  $N_i$  of vector  $\mathbf{N}$  specifies the number of unallocated devices of cluster  $i$  (Line 1). The algorithm keeps track of unallocated devices and terminates when all devices are allocated (Line 2).  $\mathbf{B}_k$  is an  $M$ -dimensional vector showing the allocation of devices from each cluster in UFB  $k$ ; i.e.,  $\mathbf{B}_k(i)$  is the number of devices from cluster  $i$  allocated in UFB  $k$ . For each band  $k$ ,  $\mathbf{B}_k$  is initialized to a zero vector (Line 4). The algorithm allocates the clusters in decreasing order of priority in each band. For each cluster  $i$  to be allocated in a band,  $u_i$  specifies the maximum waiting time due to the previously allocated higher-priority clusters in the same band (Lines 6-9). Each MTC device must be allocated one RB of duration  $\tau$  before the generation of its next packet and without violating the

maximum allowable jitter requirement. Therefore, remaining feasible allocation time  $u_{rem}$  is determined for each cluster accordingly (Line 10) and the number of devices that can be feasibly allocated from that cluster is determined considering that at most  $\lfloor \frac{u_{rem}}{\tau} \rfloor$  devices each needing one RB of duration  $\tau$  can be allocated in this remaining time (Lines 11-13). Note that the algorithm maximizes the utilization of each UFB by allocating the maximum number of devices from each cluster. After completing feasible allocations of all clusters in band  $k$ , the remaining devices from each cluster are updated (Line 15). The algorithm continues with the allocation of the next UFB if there are unallocated devices (Line 2).

---

**Algorithm 1** Fast Minimum-Band Maximum-Utilization (FMM) Algorithm

---

**Input:**  $\tau, p_i, \delta_i, N_i$  for  $i \in [1, M]$

**Output:**  $\mathbf{B}_k$  for  $k \in \{1, 2, \dots\}, k \in \mathbb{N}$  ;

```

1:  $k = 0, \mathbf{N} = [N_1, N_2, \dots, N_M]$ ;
2: while  $\mathbf{N} \neq \text{zeros}(1, M)$  do
3:    $k = k + 1$ ;
4:    $\mathbf{B}_k = \text{zeros}(1, M)$ ;
5:   for  $i = 1 : M$  do
6:      $u_i = 0$ ;
7:     for  $j = 1 : i$  do
8:        $u_i = u_i + \mathbf{B}_k(j) * \lfloor \frac{p_i}{p_j} \rfloor * \tau$ ;
9:     end for
10:     $u_{rem} = \delta_i - u_i$ ;
11:    if  $u_{rem} > 0$  then
12:       $\mathbf{B}_k(i) = \min(\mathbf{N}(i), \lfloor \frac{u_{rem}}{\tau} \rfloor)$ 
13:    end if
14:  end for
15:   $\mathbf{N} = \mathbf{N} - \mathbf{B}_k$ ;
16: end while

```

---

To explain the algorithm more in details, Figure 5.1 shows an example of how it works comparing to the Clustering-Based Algorithms (CBA) introduced by (Lien & Chen, 2011) - (Gotsis et al., 2013). There are three clusters need to be scheduled each with  $p_1 = 9$  ms,  $p_2 = 15$  ms and  $p_3 = 25$  ms period,  $N_1 = 15$ ,  $N_2 = 10$  and  $N_3 = 5$  number of devices and jitter tolerance equals to period ( $\delta_i = p_i$ ) for all clusters  $C_i$  (implicit dead-lines), respectively. Note that clusters are arranged in an increasing order according to their periods. FMM algorithm starts with the highest priority cluster  $C_1$  and allocates the maximum schedulable number of devices (9 devices) on the first UFB fully without violating devices' periodicity. The remaining number of devices from cluster  $C_1$  (6 devices) is scheduled on the second UFB along with the maximum schedulable number

of devices from cluster  $C_2$  (3 devices) and the maximum schedulable number of devices from cluster  $C_3$  (1 device) without violating periodicity of any device of any cluster. Schedulability is tested by checking remaining feasible allocation time  $u_{rem}$  (Lines 8) for each cluster after calculating the maximum waiting time due to the previously allocated higher-priority clusters in the same band (Lines 6-9). The algorithm is repeated until allocating all remaining devices from cluster  $C_2$  and  $C_3$  on the third UFB.

It can be noticed that from Figure 5.1 that the proposed scheduling algorithm utilize the available bandwidth more efficient than CBA algorithm while using the same QoS constraints (period, jitter) for both. Moreover, it is suggested to use an in-resource signaling scheme where control signals are transmitted once each certain number of consecutive transmission periods of a device which alleviates the signaling overhead significantly comparing to CBA algorithm over LTE. The performance evaluation and new devices admission ratio of FMM algorithm compared to CBA is investigated experimentally later on chapter 7.

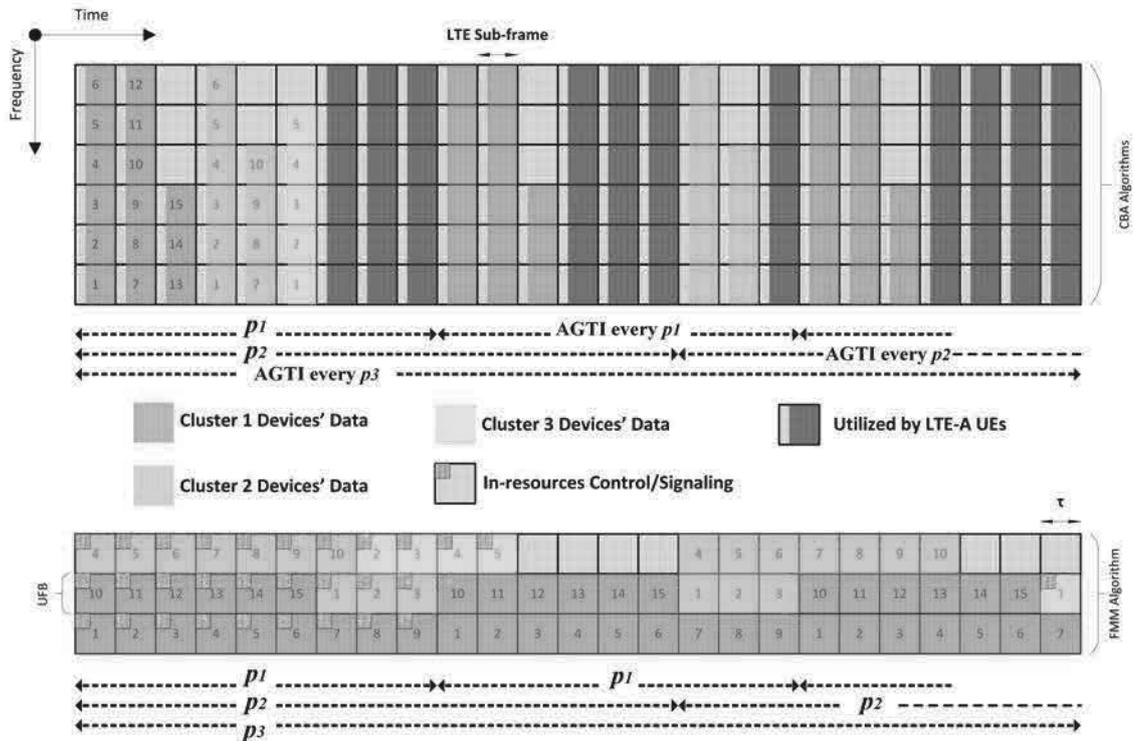


Figure 5.1 - FMM vs. CBA Algorithms

## 5.2 Approximation Ratio Performance

In this section, we present the approximation bound performance of the proposed FMM algorithm for the implicit-deadlines case in which packet generation period of each cluster is equal to the maximum allowable jitter. We first introduce the Multiprocessor Task Scheduling (MTS) problem. Consider a given set of tasks  $S = \{s_1, \dots, s_N\}$  where each task  $s_i \in S$  is characterized by its period  $p_i$  and running time  $c_i$ . Each task releases a job requiring a running time  $c_i$  at each integer multiple of its period and each job must be completed before the release of the next job. The objective is to minimize the number of processors required by running all periodic tasks. In the following, we use the approximation bound results provided for preemptive, fixed-priority, rate-monotonic scheduling policies proposed for MTS problem. We will build the equivalence between our FMM algorithm and the rate monotonic algorithms in two base points. The first point is the priorities used by both algorithms. FMM uses fixed priorities in decreasing order of the packet generation periods of the clusters. Equivalently, fixed priority rate monotonic scheduling policies designed for MTS problem allocate tasks in decreasing order of task periods without considering the running times. Second is the preemption used by these rate monotonic scheduling policies. Preemption indicates that the release of a higher priority task preempts the execution of lower priority tasks. In FMM algorithm, we do not use preemption meaning that we do not interrupt an ongoing packet transmission within an RB. However, we show in the following that preemptive allocation is not needed in RB-based resource allocation.

**Lemma 1.** *For a set of synchronous packets generating MTC devices  $a_1, \dots, a_N$  with uniform transmission time  $\tau$ , any resource allocation algorithm using preemption yields the same allocation as the equivalent non-preemptive algorithm.*

**Proof:** Assume that a packet from the device  $a_i$  starts transmission at time  $t=k\tau$  where  $k \in \mathbb{N}$ . Since all service requests from MTC devices are done on integer multiples of 1 RB, there can be no preemption in  $(k\tau, (k+1)\tau)$  interval. Since transmission times are equal to  $\tau$ , the packet from device  $a_i$  completes its transmission at time  $t=(k+1)\tau$ . Thus, any packet from an arbitrary device  $a_i$  cannot be preempted by other device implying that any resource allocation algorithm using preemption yields the same allocation as the

equivalent non-preemptive algorithm. Preemption assumption does not change allocation scheme for synchronous MTC devices having periods as integer multiples of RB duration.

Building the equivalence between the assumptions of FMM and the rate monotonic preemptive multiprocessor scheduling policies, next we state the approximation bound performance of the proposed FMM algorithm.

**Theorem 2.** *Let  $S$  be a set of MTC devices with implicit deadlines. Let  $N_{OPT}$  be the minimum number of UFB bands required to allocate the set  $S$  optimally and  $N_{FMM}$  be the number of UFB bands required to allocate the set  $S$  using FMM algorithm. Then, the following relation from (Zapata & Alvarez, , 2005) holds,*

$$\mathfrak{R} = \frac{N_{FMM}}{N_{OPT}} \leq \left[ 2 + \frac{(3 - 2^{3/2})}{2(2^{1/3} - 1)} \right] \approx 2.33 \quad (5.1)$$

**Proof:** Based on Lemma 1, we observe that the preemptive resource allocation of a set  $S$  is equivalent to its non-preemptive allocation. Thus, performance results for preemptive Rate Monotonic First Fit (RMFF) algorithm (Zapata & Alvarez, , 2005) designed for MTS problem can be used for the proposed FMM algorithm, since FMM is equivalent to RMFF except it does not allow preemption.

## 6. FAST MINIMUM-BAND MAXIMUM-UTILIZATION ALGORITHM (MULTIPLE SUBCARRIER CASE)

### 6.1 Multi-Subcarrier Effect Analysis

It is important for any future scheduling algorithm to be applicable and compatible with the flexibility concepts discussed in the literature. To that end, we present an effective analysis of using different subcarrier spacing values on the occupied bandwidth of our scheduling algorithm. Having different subcarrier spacing values  $f_i; i \in \{1, \dots, n\}$  directly affect single symbol duration and TTI size as well. This results in different capacities for each subcarrier spacing as follows;

$$N_i = \left\lfloor \frac{p}{TTI_i} \right\rfloor = \left\lfloor 2^{i-1} * \frac{p \Delta f}{\beta} \right\rfloor, \forall i \in \{1, \dots, n\} \quad (6.1)$$

It can be noticed from equation (6.1) that the number of schedulable devices on each band of subcarrier spacing  $f_i$  depends only on the used *scaling factor*, i.e. here  $2^{i-1}$ , while period  $p$ , base subcarrier spacing value  $\Delta f$  and number of symbols  $\beta$  are constants.

**Theorem 3.** For any two subcarrier spacing values,  $f_i$  and  $f_{i+j}, \forall i, j \in \{1, \dots, n\}$  calculated by equation (3.1), and assigned the same bandwidth such that,  $2^j * f_i = f_{i+j}$ , the following inequality is true:

$$2^j N_i \leq N_{i+j} \quad (6.2)$$

**Proof:** Using equation (6.1) and substituting  $\frac{p \Delta f}{\beta} = k$ , i.e.  $k$  is constant,  $N_i = \lfloor 2^{i-1} * k \rfloor$  and  $N_{i+j} = \lfloor 2^j * 2^{i-1} * k \rfloor$ . To get the occupied bandwidth by both subcarrier spacing values the same, we need to multiply the smaller one  $f_i$  by  $2^j$ , which means also multiplying the number of devices  $N_i$  by  $2^j$ . Then, we can easily notice that,  $2^j \lfloor 2^{i-1} * k \rfloor \leq \lfloor 2^j * 2^{i-1} * k \rfloor$  which is exactly inequality (6.2).

By using multiple subcarrier spacing values to allocate devices of the same periods, there will be a limit of  $N$  number of devices where, under this limit, if we use only one band of  $f_k$  subcarrier spacing value to allocate these devices, more bandwidth will be occupied than using any optimal combination of  $\{f_1, f_2, \dots, f_{k-1}\}$ . To calculate this limit, we need the following relation:

$$a_1 f_1 + a_2 f_2 + \dots + a_{k-1} f_{k-1} = f_k - f_{smallest} \quad (6.3)$$

using equation (3.1) to calculate  $f_i \forall i \in \{1, \dots, k\}$  and putting  $f_{smallest} = f_1$  we got the following,

$$a_1 2^0 + a_2 2^1 + \dots + a_{k-1} 2^{k-2} = 2^{k-1} - 2^0 \quad (6.4)$$

One of the possible integer solutions (we cannot use a fraction of a band) for equation (6.4) to get values of  $a_i \forall i \in \{1, \dots, k-1\}$  can be obtained by substituting  $a_1 = a_2 = \dots = a_{k-1}$ . An integer solution of this equation provides us with the maximum number of devices  $N$  that can be allocated optimally by  $f_1, \dots, f_{k-1}$  with total bandwidth less than one  $f_k$  as follows,

$$a_1 N_1 + a_2 N_2 + \dots + a_{k-1} N_{k-1} = \sum_{i=1}^{k-1} N_i \quad (6.5)$$

The resulted value from equation (6.5) represents a limit for the number of devices  $N \leq \sum_{i=1}^{k-1} N_i$  where for a less or equal value of it  $f_{k-1}$  is used, and for a greater value of it  $f_k$  is used. This limit value is calculated for each subcarrier spacing value  $f_k \forall i \in \{1, \dots, n\}$  creating ranges for number of devices  $[N_{Range_i}, N_{Range_{(i+1)}}]$  which from we can pick the optimal subcarrier spacing value.

## 6.2 Minimum Bandwidth Optimal Subcarrier Spacing Algorithm (OSC)

Based on the previous analysis, we propose a Minimum Bandwidth Optimal Subcarrier Spacing (OSC) Algorithm, given by algorithm 2 and it is described as follows. Let  $p$  be the period of  $N$  number of time-triggered M2M devices that are intended to be scheduled using  $n$  values of different subcarrier spacing  $f_i, i \in \{1, 2, \dots, n\}$  scaled from base  $\Delta f$  value by *scaling factor* of equation (3.1). For each subcarrier spacing value, the maximum

schedulable number of devices on each single band  $N_i$ ,  $i \in \{1, 2, \dots, n\}$  is calculated by  $N_i = \left\lfloor \frac{p}{TTI_i} \right\rfloor$  (equation (6.1)) (Lines 2-4). Based on equation (6.5), OSC algorithm creates ranges  $[N_{Range_i}, N_{Range_{(i+1)}}]$  for number of M2M devices  $N$  from which the optimal subcarrier spacing value is picked such that the minimum bandwidth is occupied (Lines 5-7). Based on the given number of devices  $N$  along with the created ranges  $N_{Range}$ , time-triggered M2M devices are allocated to the optimal subcarrier spacing value  $f_i$  and scheduled by its maximum schedulable number of devices on each single band  $N_i$ . If number of devices  $N$  is greater than  $N_{Range_n}$ , use the largest subcarrier spacing value to allocate devices then update the remaining number of devices  $N$  (Lines 9-15). If number of devices falls in a range less than or equal to  $N_{Range_n}$  and greater than  $N_{Range_2}$ , then find that range of devices  $[N_{Range_k}, N_{Range_{(k+1)}}]$  and pick the equivalent subcarrier spacing value for allocation (Lines 18-21). Update the number of devices  $N$  (Line 26). Finally, if number of devices  $N$  falls in  $[N_{Range_1}, N_{Range_2}]$ , use the smallest subcarrier spacing value for allocation (Lines 30-33). The algorithm is repeated until scheduling all  $N$  devices (Lines 8-34). The algorithm outputs the needed number of single bands  $a_i$ ,  $i \in \{1, 2, \dots, n\}$  of each subcarrier spacing value  $f_i$ .

---

**Algorithm 2** Minimum Bandwidth Optimal Subcarrier Spacing Algorithm (OSC)

---

**Input:**  $p, N, n, TTI_i, f_i, i \in \{1, \dots, n\}$ ;

**Output:**  $a_i, i \in \{1, \dots, n\}$ ;

```
1:  $a = N_{Range} = zeros(1, n)$ ;
2: for  $i = 1 : n$  do
3:    $N_i = \lfloor \frac{p}{TTI_i} \rfloor$ ;
4: end for
5: for  $i = 1 : n - 1$  do
6:    $N_{Range_{i+1}} = \sum_{j=1}^i N_j$ ;
7: end for
8: while  $N \neq 0$  do
9:   if  $N > N_{Range_n}$  then
10:     $a_n = a_n + \lfloor \frac{N}{N_n} \rfloor$ ;
11:    if  $\lfloor \frac{N}{N_n} \rfloor == 0$  then
12:       $a_n = a_n + 1$ ;
13:       $N = 0$ ;
14:    else
15:       $N = N - a_n N_n$ ;
16:    end if
17:  end if
18:  if  $N_{Range_2} < N \leq N_{Range_n}$  then
19:    for  $k = 2 : n$  do
20:      if  $N_{Range_k} < N \leq N_{Range_{k+1}}$  then
21:         $a_k = a_k + \lfloor \frac{N}{N_k} \rfloor$ ;
22:        if  $\lfloor \frac{N}{N_k} \rfloor == 0$  then
23:           $a_k = 1$ ;
24:           $N = 0$ ;
25:        else
26:           $N = N - a_k N_k$ ;
27:        end if
28:      end if
29:    end for
30:  else
31:     $a_1 = \lfloor \frac{N}{N_1} \rfloor$ ;
32:     $N = 0$ ;
33:  end if
34: end while
```

---

### 6.3 Multi-Subcarrier Fast Minimum-Band Maximum-Utilization Algorithm (FMM-OSC)

In this subsection we propose an algorithm that combines algorithms 1 and 2 heuristically to find a minimum scheduling bandwidth for  $M$  number of M2M clusters each with different QoS requirement and  $n$  number of scaled subcarrier spacing values  $f_i, i \in \{1, 2, \dots, n\}$ . The FMM-OSC algorithm is described as follows;

Period  $p_i$ , maximum jitter tolerance  $\delta_i$  and the corresponding number of devices of each cluster  $N_i$  are given arranged in an increasing order;  $i \in \{1, 2, \dots, M\}$ , as an input of the algorithm. A set of scaled subcarrier spacing values are also give as an algorithm input. This algorithm is designed to use a *scaling factor* of  $2^j, j \in \{1, 2, \dots, n\}$ . The output of this

algorithm is an overall schedule for allocating time-triggered M2M devices while each single band (UFB) schedule is given by  $\mathbf{B}_k$  and its corresponding used subcarrier spacing value is given by  $\mathbf{Bf}_k$  for  $k \in \mathbb{N}$ . Each element of vector  $\mathbf{N}$  represents the number of unallocated devices of cluster  $i$  (Line 1). The given scaled subcarrier spacing values are assigned to a row vector of size  $(1 \times n)$  called  $\mathbf{F}$  and  $\mathbf{Bf}$  is a column vector of size  $(k \times 1)$  specifies the used subcarrier spacing value for each  $\mathbf{B}_k$ ,  $k \in \mathbb{N}$  (Line 2). The first value of vector  $\mathbf{N}$  is assigned to  $NN$  variable (Line 3). The algorithm keeps repeating until allocating all devices in  $\mathbf{N}$  (Line 4). At each repeat, the number of remaining devices of the current cluster  $N_i$  is checked and if it equals zero the next cluster's number of devices  $N_{i+1}$  is assigned to  $NN$  (Lines 6-9). Based on the current cluster number of devices value ( $NN$ ) and the given set of subcarriers spacing, algorithm OSC is called and it returns the subcarrier spacing value and its corresponding TTI value that should be used for allocating the current  $\mathbf{B}_k$  band (Line 10). The resulted TTI value is assigned to  $\tau$  (Line 11) and  $f_j$  to  $\mathbf{Bf}_k$  (Line 12). For each band  $k$ ,  $\mathbf{B}_k$  is initialized to a zero-row vector of size  $(1 \times M)$  (Line 13). The rest of the algorithm is like FMM algorithm (Line 14-25).

---

**Algorithm 3 Multi-Subcarrier Fast Minimum-Band Maximum-Utilization (FMM-OSC) Algorithm**

---

**Input:**  $p_i, \delta_i, N_i$  for  $i \in [1, M]$   $f_j$  for  $j \in [1, n]$

**Output:**  $\mathbf{B}_k, \mathbf{Bf}_k$  for  $k \in \{1, 2, \dots\}, k \in \mathbb{N}$

---

```

1:  $k = 0, \mathbf{N} = [N_1, N_2, \dots, N_M]$ ;
2:  $\mathbf{F} = [f_1, f_2, \dots, f_n], \mathbf{Bf} = \text{zeros}(k, 1)$ ;
3:  $N_{index} = 1, NN = \mathbf{N}(N_{index})$ ;
4: while  $\mathbf{N} \neq \text{zeros}(1, M)$  do
5:    $k = k + 1$ ;
6:   if  $NN == 0$  then
7:      $N_{index} = N_{index} + 1$ ;
8:      $NN = \mathbf{N}(N_{index})$ ;
9:   end if
10:   $[f_j, TTI_j] = \text{OSC}(NN, \mathbf{F})$ ;
11:   $\tau = TTI_j$ ;
12:   $\mathbf{Bf}_k = f_j$ ;
13:   $\mathbf{B}_k = \text{zeros}(1, M)$ ;
14:  for  $i = 1 : M$  do
15:     $u_i = 0$ ;
16:    for  $j = 1 : i$  do
17:       $u_i = u_i + \mathbf{B}_k(j) * \left\lceil \frac{p_i}{p_j} \right\rceil * \tau$ ;
18:    end for
19:     $u_{rem} = \delta_i - u_i$ ;
20:    if  $u_{rem} > 0$  then
21:       $\mathbf{B}_k(i) = \min(\mathbf{N}(i), \lfloor \frac{u_{rem}}{\tau} \rfloor)$ 
22:    end if
23:  end for
24:   $\mathbf{N} = \mathbf{N} - \mathbf{B}_k$ ;
25: end while

```

---

## 7. PERFORMANCE EVALUATION

In this section, we evaluate the performance of proposed algorithms in single and multi-subcarrier spacing scenarios as follows. Algorithms are run, and results are gotten using MATLAB.

### 7.1 Fast Minimum-Band Maximum-Utilization Algorithm (FMM)

In this section, we compare the performance of the proposed algorithm FMM to the previously proposed clustering-based algorithm (CBA) designed for LTE in (Lien & Chen, 2011) and (Lien et al., 2011) then to the optimality as well.

We consider the bandwidth reserved for M2M communications is 18 MHz which is divided into 100 UFBs each corresponding to a one-RB width frequency band. We consider a different number of clusters where each cluster has a random number of MTC devices uniformly distributed in the range [10, 100]. Performance results are averaged over 100 runs for each simulation scenario.

In Table 2.1, we illustrate the superiority of the proposed FMM over CBA algorithm. Simulations are performed for a different number of clusters with diverse QoS characteristics as given in Table 2.1. Bandwidth ratio is the ratio of bandwidth required by the proposed FMM algorithm to the bandwidth required by the previously proposed CBA algorithm. Note that, although the maximum allowable jitter requirement of each cluster is relatively tight with respect to the packet generation period, the bandwidth reduction is very significant. For maximum tolerable jitter values closer to packet arrival periods, bandwidth ratios are expected to be much smaller. The other parameter evaluated in the simulations is the admission gain defined as the percentage of the bandwidth required by the FMM algorithm which is suitable for the admission of the new MTC devices while are not possible in the schedule generated by CBA. For more than one

cluster, in the FMM algorithm, a frequency band corresponding to at least one UFB can support admission of new MTC devices which cannot be served by CBA. Hence, FMM outperforms CBA algorithm both in terms of the bandwidth efficiency and the flexibility of allocating new MTC devices.

**Table 7.1** - FMM Algorithm Performance over CBA

Scenario		Performance		
Cluster Index	$[p_i, \delta_i]$ (ms)	Scheduled Clusters	BW Ratio (%)	Admission Gain (%)
1	[10,2]	[1]	50.63	0
2	[20,4]	[1:2]	59.81	1.08
3	[20,6]	[1:3]	55.03	19.69
4	[40,12]	[1:4]	54.24	19.99
5	[100,50]	[1:5]	52.19	20.02
6	[100,60]	[1:6]	50.84	19.89
7	[200,80]	[1:7]	49.92	20.25
8	[250,100]	[1:8]	49.23	20.39
9	[500,150]	[1:9]	48.94	28.02
10	[500,200]	[1:10]	48.49	29.03
11	[1000,500]	[1:11]	48.13	51.62
12	$[10^5, 10^4]$	[1:12]	47.79	52.37

In Table 7.2, we illustrate the performance of FMM algorithm compared to the theoretical lower bounds for optimality. We use 2 different bounds. Bound 1 and bound 2 are specified as the minimum bandwidth required for allocation of the MTC devices considering that each UFB can feasibly support a set of MTC devices with a total utilization at most 1 and the utilization of a device is defined as the inverse of its period; i.e.,  $1/p$ , and inverse of its jitter  $1/\delta$ , respectively. Bound 1 and 2 are effective bounds for implicit deadlines and synchronous devices scenarios, respectively. Note that these theoretical bounds are lower bounds on the optimality for the corresponding scenarios. Simulations results are presented for a different number of clusters ranging from 1 to 12.

The packet generation periods of the clusters are same as given in Table 1. For both theoretical bounds, we consider the implicit deadlines case where periods are equal to maximum tolerable jitter values and the case where maximum tolerable jitter values are half of the periods. The approximation ratio values presented in the table give the ratio of the bandwidth required by the FMM algorithm to the bandwidth specified by these bounds. For implicit deadlines case, the FMM algorithm performs almost optimally. For a more stringent timing constraint in which maximum tolerable jitter values are half of the periods, FMM still achieves an approximation ratio bound less than 2. Considering the bounds are lower bounds on the optimality, FMM algorithm yields a much better approximation ratio in practice.

**Table 7.2 - Optimality Performance of FMM Algorithm**

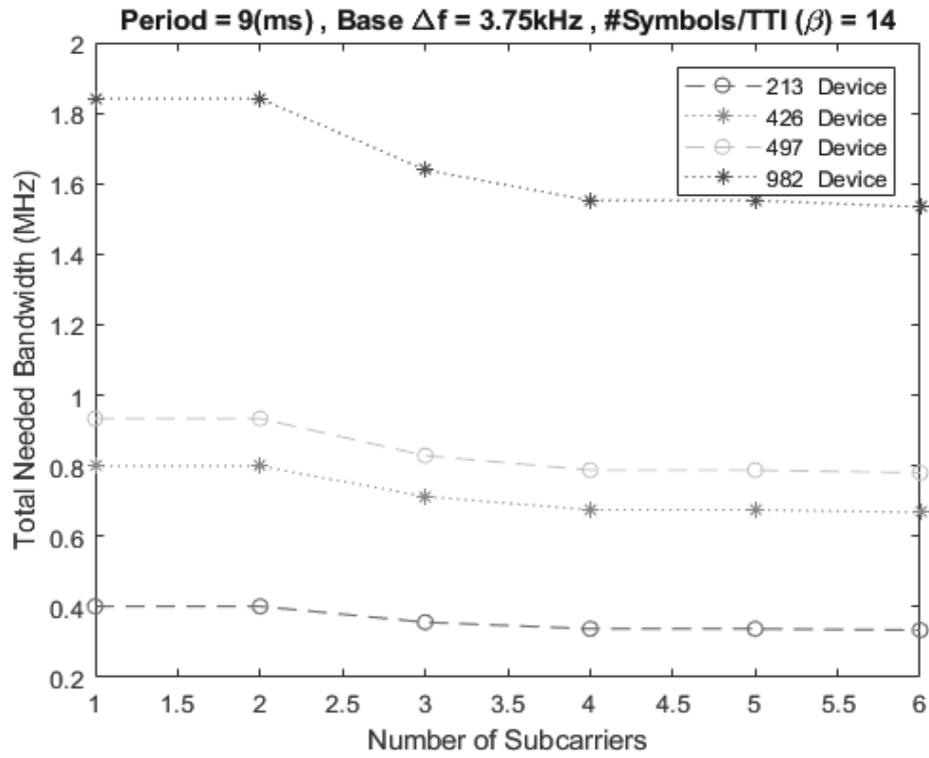
# of Clusters	Theoretical Bound 1		Theoretical Bound 2	
	$\delta = p$	$\delta = p/2$	$\delta = p$	$\delta = p/2$
1	1.00	1.87	1.00	1.01
2	1.01	1.93	1.01	1.02
3	1.01	1.95	1.01	1.02
4	1.02	1.95	1.02	1.03
5	1.01	1.96	1.01	1.03
6	1.02	1.97	1.02	1.04
7	1.01	1.96	1.01	1.03
8	1.01	1.98	1.01	1.03
9	1.01	1.96	1.01	1.04
10	1.01	1.96	1.01	1.04
11	1.02	1.96	1.02	1.05
12	1.03	1.97	1.03	1.05

## 7.2 Minimum Bandwidth Optimal Subcarrier Spacing Algorithm (OSC)

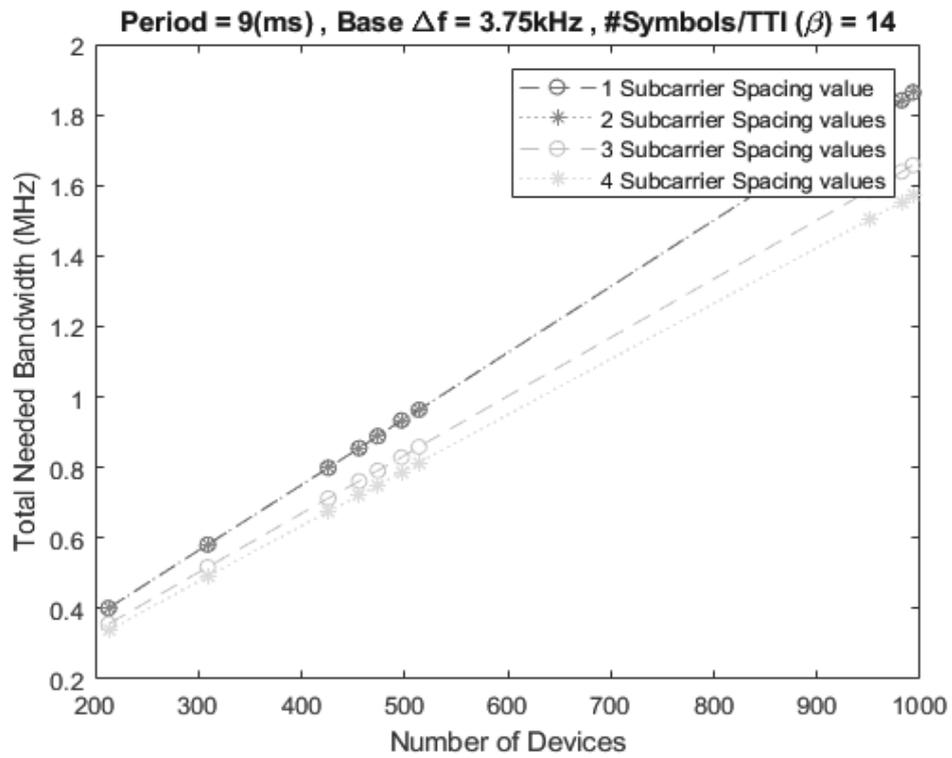
In this section, we evaluate the performance of the proposed algorithm in section 6.2 using different subcarrier spacing values. We used some real values of subcarrier spacing and

number of symbols on 5 different clusters each with different period  $p_i$ . Figure 7.1 through Figure 7.4 show the total needed bandwidth to allocate random values of number of devices  $N$  belongs to the same cluster using  $f_i$ ;  $i \in \{1,2,3,4,5,6\}$  subcarrier spacing value. Subcarrier spacing values are calculated by scaling up a base subcarrier spacing value ( $\Delta f = 3.75$  kHz) using scaling factor of  $2^{(i-1)}$ ; e.g.  $f_3 = 2^{3-1} * \Delta f = 15$  kHz. The number of symbols per TTI is constant ( $\beta=14$ ) and the number of subcarriers per RB is also constant for all sub-channels of different subcarriers spacing values ( $\alpha$  is constant). Results are obtained for 4 different clusters each with different period  $p$  picked randomly between  $[0,100]$  ms.

For a single cluster of period  $p = 9$  ms and random number of devices  $N$ , Figure 7.1 (a) shows that using more than one subcarrier spacing value (starting from 1 until 6 S.C) to allocate these devices helped to decrease the needed bandwidth for allocation. The bandwidth reduction is clarified more by noticing the trend of lines in Figure 7.1 (b) where each line represents the needed bandwidth to allocate random number of devices  $N \in [0,1000]$  using different subcarrier spacing values  $f_i$ ;  $i \in \{1,2,3,4\}$ ; 1 subcarrier spacing value means using  $f_1$  only for allocation (3.75 kHz), 2 subcarrier spacing values means using  $f_1$  and  $f_2$  (3.75 kHz and 7.5 kHz) for allocation and so on. Figure 7.2 (a) and (b) shows similar behavior but for period  $p = 26$  ms. Figure 7.3 and Figure 7.4 shows the worst-case performance of the algorithm when using different subcarrier spacing value for allocation does not introduce any reduction, or the reduction is limited, in the resulted bandwidth. This can be noticed clearly from Figure 7.3 (b) where all lines representing the resulted bandwidth from using different number of subcarrier spacing values are overlapping.

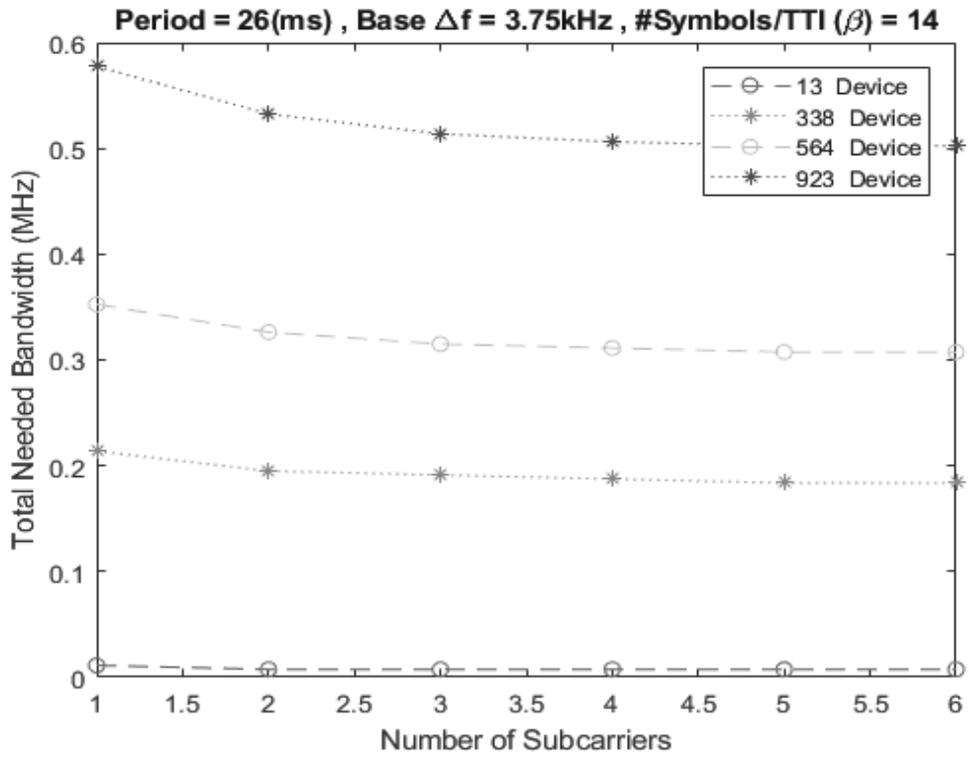


(a) Bandwidth vs. # of Subcarriers

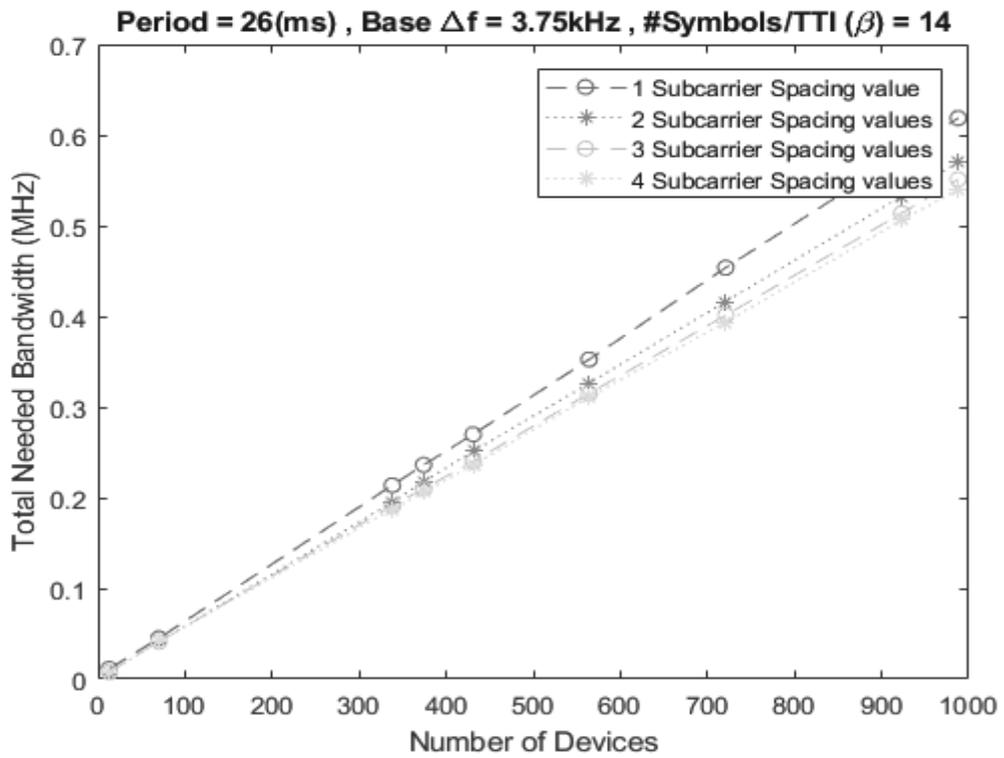


(b) Bandwidth vs. # of Devices

**Figure 7.1** – Bandwidth Reduction by Multi-Subcarrier Spacing Values for Devices of a Single Cluster with Period= 9 ms

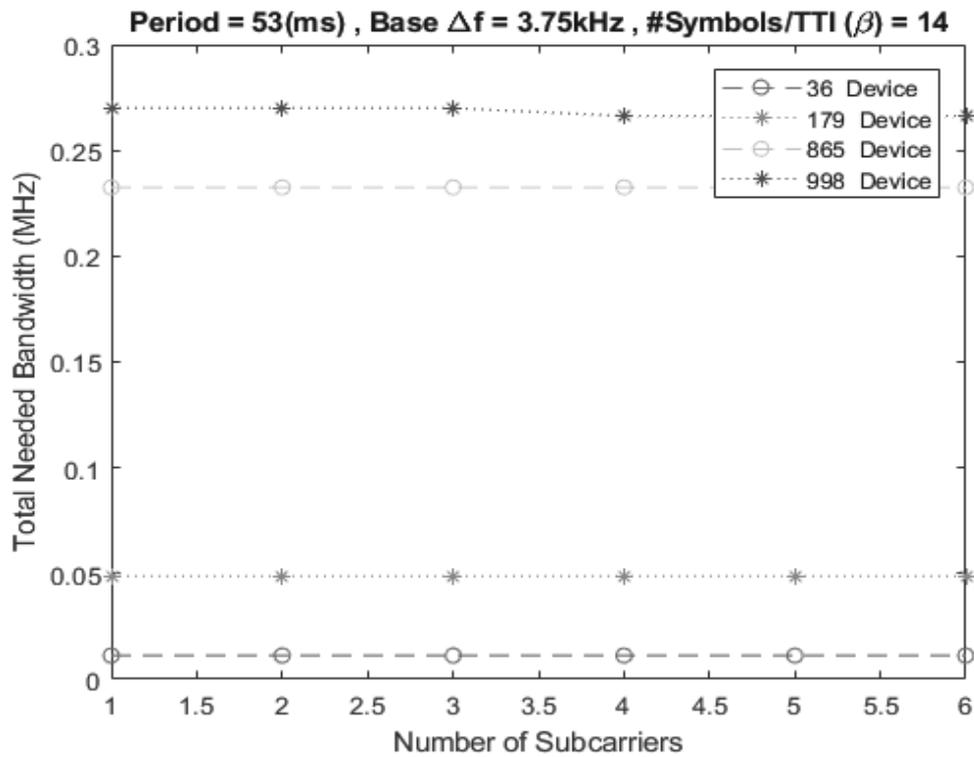


(a) Bandwidth vs. # of Subcarriers

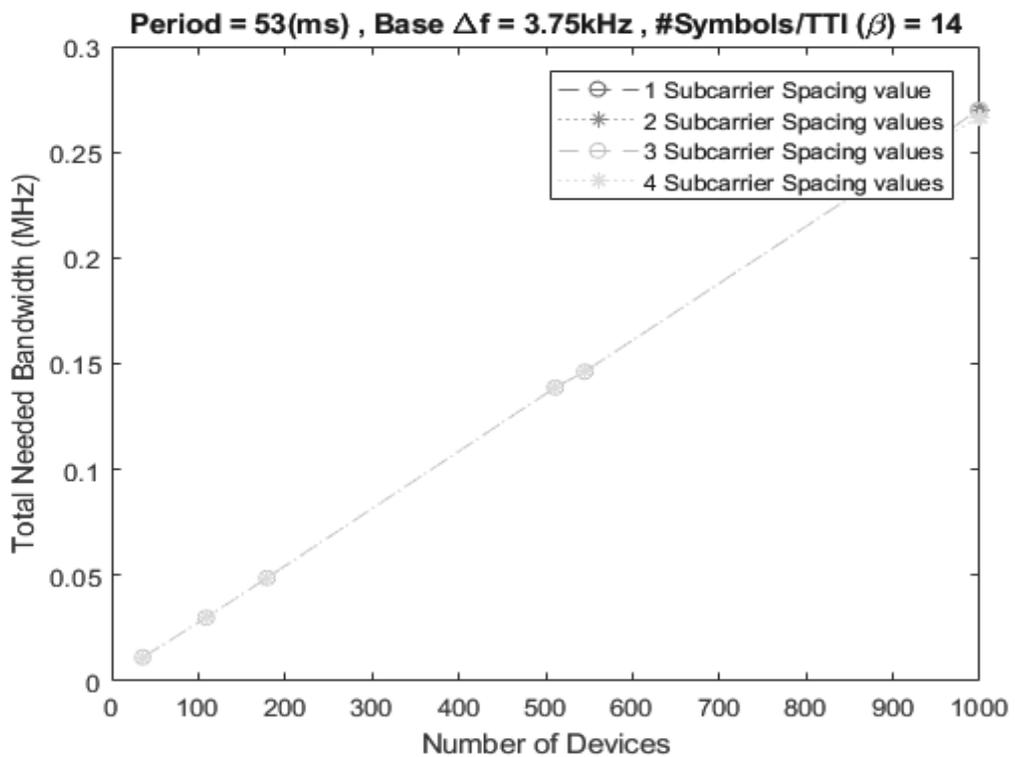


(b) Bandwidth vs. # of Devices

**Figure 7.2 - Bandwidth Reduction by Multi-Subcarrier Spacing Values for Devices of a Single Cluster with Period= 26 ms**

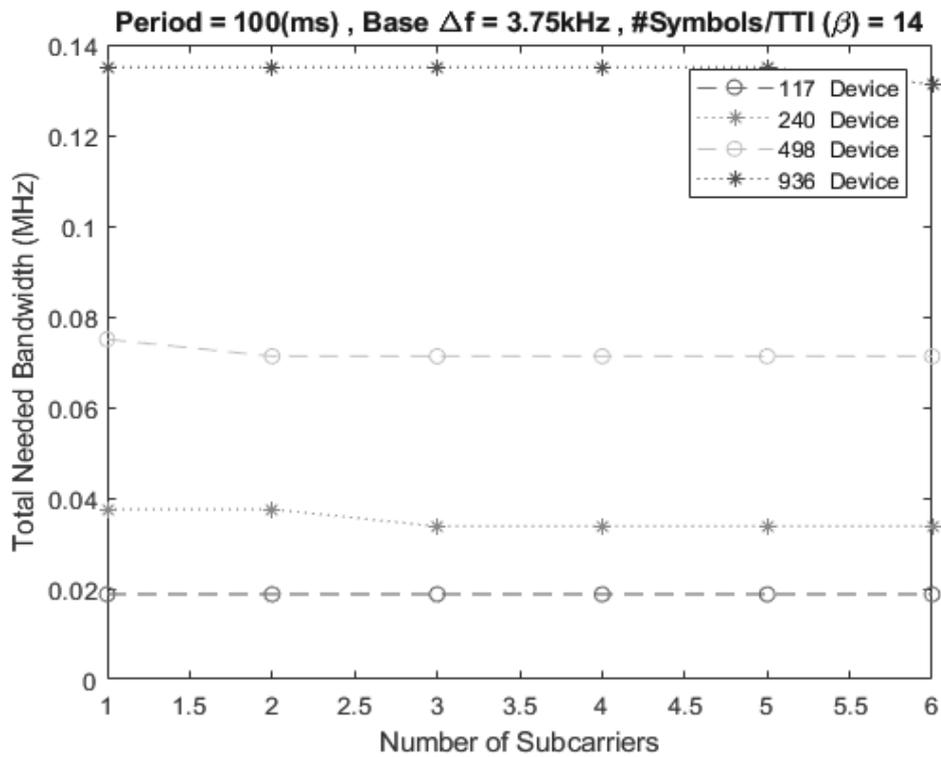


(a) Bandwidth vs. # of Subcarriers

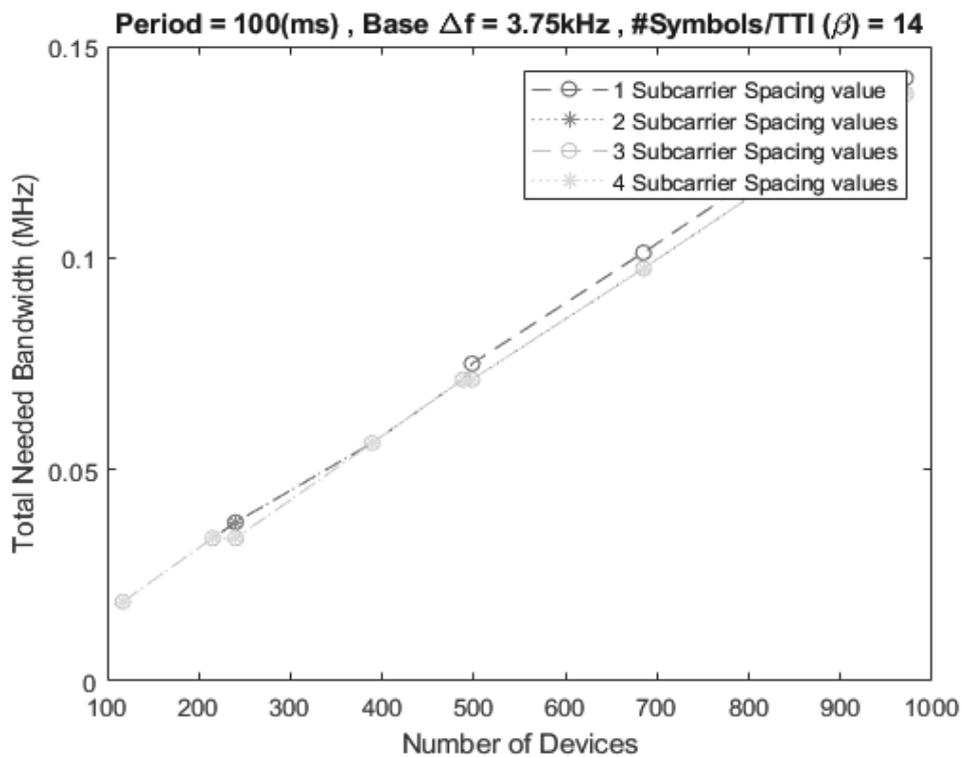


(b) Bandwidth vs. # of Devices

Figure 7.3 - Bandwidth Reduction by Multi-Subcarrier Spacing Values for Devices of a Single Cluster with Period= 53 ms



(a) Bandwidth vs. # of Subcarriers



(b) Bandwidth vs. # of Devices

**Figure 7.4** - Bandwidth Reduction by Multi-Subcarrier Spacing Values for Devices of a Single Cluster with Period= 100 ms

It can be noticed from Figure 7.1 to Figure 7.4 that for all periods  $p$  and for all used numbers of devices  $N$ , the total needed bandwidth to allocate these devices using more than one subcarrier spacing value; i.e.  $(f_1 + f_2)$  or  $(f_1 + f_2 + f_3)$ ....or  $(f_1 + f_2 + f_3 + f_4 + f_5 + f_6)$ , is minimized or at most equal to the needed bandwidth using only on subcarrier spacing value  $(f_1)$ .

### 7.3 Multi-Subcarrier Fast Minimum-Band Maximum-Utilization Algorithm

In this section, we compare the performance of the proposed algorithm FMM to the algorithm proposed in section 6.3 (FMM-OSC) using different subcarrier spacing values. We consider 12 different clusters of M2M devices. The two algorithms were run 25 times each time with different set of a number of devices  $N$  generated uniformly in the range of  $[10,10000]$ . The algorithms were run using 4 different subcarrier spacings scaled up by the base subcarrier spacing value of 15 kHz. Number of symbols per TTI is constant  $\beta=14$ . Figure 7.5 illustrates both algorithms performance. It can be noticed that both algorithms have almost the same occupied bandwidth.

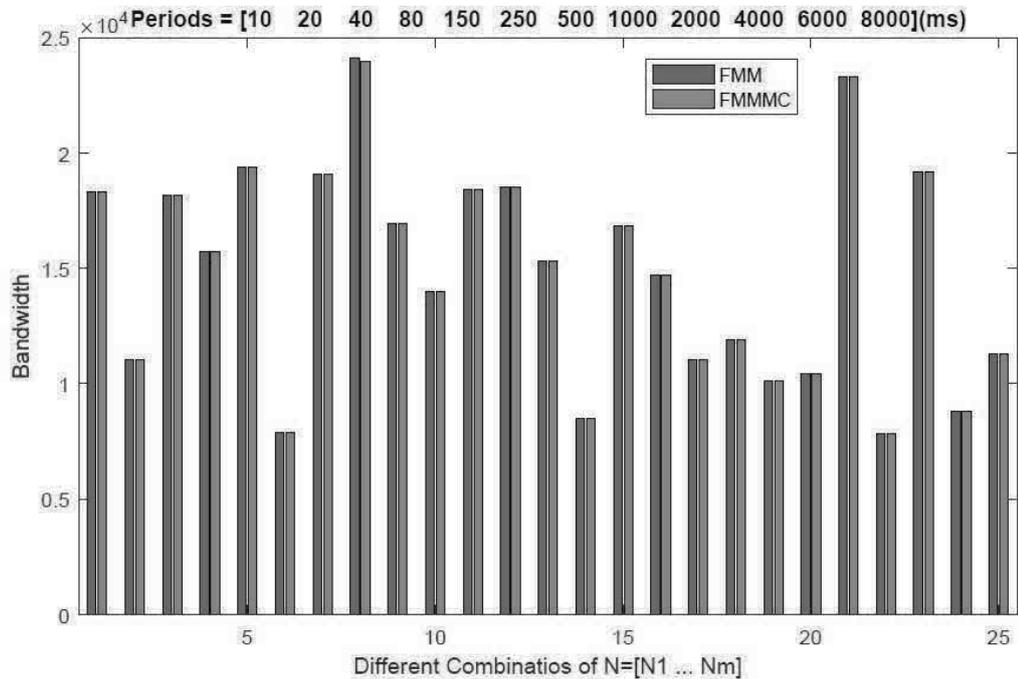
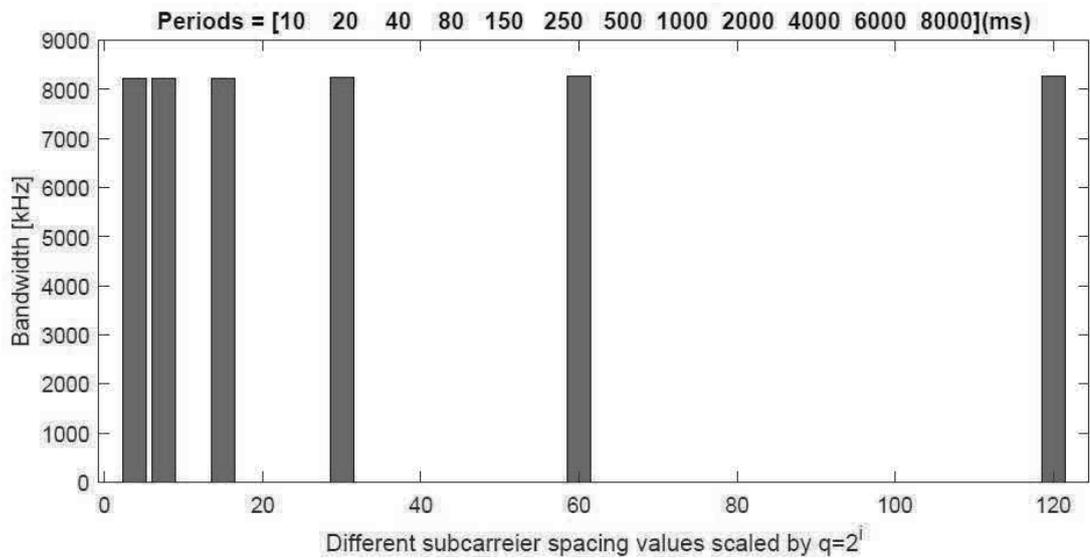


Figure 7.5 - FMM Algorithm/FMM-MC Algorithm Performance Comparison

We further check the performance of FMM algorithm using different values of the subcarrier spacing. We consider again 12 clusters of M2M devices each with  $N_i$  number of devices generated uniformly in the range of [10,10000]. The algorithm was run 6 times each with different subcarrier spacing value taken from a set of scaled subcarrier values from a base value of 3.75 kHz. Figure 7.6 illustrates the resulted bandwidth for each run.



**Figure 7.6** - FMM Algorithm using 6 different scaled subcarrier spacing values

It can be noticed from Figure 7.5 and Figure 7.6 that our proposed algorithm has almost similar performance when different subcarrier spacing values being used. Therefore, the advantages of using different subcarrier spacing values presented previously can be realized without affecting the bandwidth efficiency.

## 8. CONCLUSION

In this thesis, we propose a novel resource allocation algorithm for M2M communications in future cellular networks. The proposed FMM algorithm minimizes the bandwidth required for the periodic allocation of MTC devices while meeting their QoS requirements. Then, we investigated the same problem for the flexible physical layer architecture envisioned for 5G and beyond cellular networks in which different subcarrier size and transmission time interval values are used in different subchannels to meet the QoS requirements of M2M and H2H traffic more efficiently. The proposed algorithm utilizes the available bandwidth much better than the clustering-based algorithms proposed in previous works in literature. We show a similar behavior of the suggested algorithm when multiple subcarrier spacing values are used to meet heterogeneous service requirements for M2M communication projected in 5G.

## REFERENCES

- Ankarali, Z. E., Pekoz, B., & Arslan, H. (2017). Flexible Radio Access Beyond 5G: A Future Projection on Waveform, Numerology, and Frame Design Principles. *IEEE Access*, 5, 18295--18309.
- Bertossi, A. A., & Fusiello, A. (1997). Rate-monotonic scheduling for hard-real-time systems. *European Journal of Operational Research*, 96(3), 429--443.
- Chen, Kwang-Cheng. (2012). Machine-to-machine communications for healthcare. *Journal of Computing Science and Engineering*, 6(2), 119--126.
- Dhillon, H. S., Huang, H. C., Viswanathan, H., & Valenzuela, R. A. (2013). Power-efficient system design for cellular-based machine-to-machine communications. *IEEE Transactions on Wireless Communications*, 12(11), 5740--5753.
- Dhillon, H. S., Huang, H., Viswanathan, H., & Valenzuela, R. (2014). Fundamentals of throughput maximization with random arrivals for M2M communications. *IEEE Transactions on Communications*, 62(11), 4094--4109.
- Elhamy, A., & Gadallah, Y. (2015). BAT: A balanced alternating technique for M2M uplink scheduling over LTE. *Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st*, 1--6.
- Fan, Z., Haines, R., & Kulkarni, P. (2014). M2M communications for E-health and smart grid: an industry and standard perspective. *IEEE Wireless Communications*, 21(1), 62--69.
- Ghavimi, F., & Chen, H.-H. (2015). M2M communications in 3GPP LTE/LTE-A networks: Architectures, service requirements, challenges, and applications. *IEEE Communications Surveys & Tutorials*, 17(2), 525--549.
- Gotsis, A. G., Lioumpas, A. S., & Alexiou, A. (2012). M2M scheduling over LTE: Challenges and new perspectives. *IEEE Vehicular Technology Magazine*, 7(3), 34--39.
- Gotsis, A. G., Lioumpas, A. S., & Alexiou, A. (2012). Evolution of packet scheduling for machine-type communications over LTE: Algorithmic design and performance analysis. *Globecom Workshops (GC Wkshps), 2012 IEEE*, 1620--1625.
- Gotsis, A. G., Lioumpas, A. S., & Alexiou, A. (2013). Analytical modelling and performance evaluation of realistic time-controlled M2M scheduling over LTE

cellular networks. *Transactions on Emerging Telecommunications Technologies*, 24(4), 378--388.

- Hasan, M., Hossain, E., & Niyato, D. (2013). Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches. *IEEE communications Magazine*, 51(6), 86--93.
- Ijaz, A., Zhang, L., Grau, M., Mohamed, A., Vural, S., Quddus, A. U., . . . Tafazolli, R. (2016). Enabling massive IoT in 5G and beyond systems: PHY radio frame design considerations. *IEEE Access*, 4, 3322--3339.
- Incorporated, Qualcomm. (2016). Numerology requirements. *3GPP TSG RAN WG1 Meeting*, 8.1.5(11).
- Jang, H. S., Kim, S. M., Park, H.-S., & Sung, D. (2016). Message-embedded random access for cellular M2M communications. *IEEE Communications Letters*, 20(5), 902--905.
- Jiang, D., Wang, H., Malkamaki, E., & Tuomaala, E. (2007). Principle and performance of semi-persistent scheduling for VoIP in LTE system. *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, 2861--2864.
- Joseph, M., & Pandya, P. (1986). Finding response times in a real-time system. *The Computer Journal*, 29(5), 390--395.
- Karrenbauer, A., & Rothvo, T. (2010). A 3/2-approximation algorithm for rate-monotonic multiprocessor scheduling of implicit-deadline tasks. *International Workshop on Approximation and Online Algorithms*(Springer), 166--177.
- Laya, A., Alonso, L., & Alonso-Zarate, J. (2014). Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives. *IEEE Communications Surveys and Tutorials*, 16(1), 4--16.
- Lien, S.-Y., & Chen, K.-C. (2011). Massive access management for QoS guarantees in 3GPP machine-to-machine communications. *IEEE communications Letters*, 15(3), 311--313.
- Lien, S.-Y., Chen, K.-C., & Lin, Y. (2011). Toward ubiquitous massive accesses in 3GPP machine-to-machine communications. *IEEE Communications Magazine*, 49(4).
- Lioumpas, A. S., & Alexiou, A. (2011). Uplink scheduling for machine-to-machine communications in LTE-based cellular systems. *GLOBECOM Workshops (GC Wkshps), 2011 IEEE*, 353--357.

- Liu, C. L., & Layland, J. W. (1973). Scheduling algorithms for multiprogramming in a hard-real-time environment. *Journal of the ACM (JACM)*, 20(1), 46--61.
- Lonn, H., & Axelsson, J. (1999). A comparison of fixed-priority and static cyclic scheduling for distributed automotive control applications. *Real-Time Systems, 1999. Proceedings of the 11th Euromicro Conference on*, 142--149.
- Mansoor, S., Molisch, A. F., Smith, P. J., Haustein, T., Zhu, P., De Silva, P., . . . Wunder, G. (2017). 5G: A tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE Journal on Selected Areas in Communications*, 35(6), 1201--1221.
- Mehmood, Y., Haider, N., Imran, M., Timm-Giel, A., & Guizani, M. (2017). M2M communications in 5G: state-of-the-art architecture, recent advances, and research challenges. *IEEE Communications Magazine*, 55(9), 194--201.
- Mostafa, A. E., & Gadallah, Y. (2017). A statistical priority-based scheduling metric for M2M communications in LTE networks. *IEEE Access*, 5, 8106--8117.
- Pedersen, K., Frederiksen, F., Berardinelli, G., & Mogensen, P. (2015). A flexible frame structure for 5G wide area. *Vehicular Technology Conference (VTC Fall), 2015 IEEE 82nd*, 1--5.
- Pedersen, Klaus I and Berardinelli, Gilberto and Frederiksen, Frank and Mogensen, Preben and Szufarska, Agnieszka. (2016). A flexible 5G frame structure design for frequency-division duplex cases. *IEEE Communications Magazine*, 54(3), 53--59.
- Pepper, R. (2015). The Rise of M2M Devices. *3rd BEREK Stakeholder Forum*.
- Roessler, A. (2016). 5G waveform candidates application note. *Rohde & Schwarz, Munich, Germany, Tech. Rep. IMA271*.
- Sahin, A., & Arslan, H. (2012). Multi-user aware frame structure for OFDMA based system. *Vehicular Technology Conference (VTC Fall), 2012 IEEE*, 1--5.
- Schaich, F., Wild, T., & Ahmed, R. (2016). Subcarrier spacing-how to make use of this degree of freedom. *Vehicular Technology Conference (VTC Spring), 2016 IEEE 83rd*, 1--6.
- Schlienz, J., & Raddino, D. (2016). Narrowband Internet of Things Whitepaper. *IEEE Microwave Magazine*, 8(1), 76--82.

- Seo, J.-B., & Leung, V. (2011). Design and analysis of backoff algorithms for random access channels in UMTS-LTE and IEEE 802.16 systems. *IEEE Transactions on Vehicular Technology*, 60(8), 3975--3989.
- Si, P., Yang, J., Chen, S., & Xi, H. (2015). Adaptive massive access management for QoS guarantees in M2M communications. *IEEE Transactions on Vehicular Technology*, 64(7), 3152--3166.
- Wang, Z., & Wong, V. W. (2015). Optimal access class barring for stationary machine type communication devices with timing advance information. *IEEE Transactions on Wireless communications*, 14(10), 5374--5387.
- Wu, G., Talwar, S., Johnsson, K., Himayat, N., & Johnson, K. (2011). M2M: From mobile to embedded internet. *IEEE Communications Magazine*, 49(4).
- Zaidi, A. A., Baldemair, R., Tullberg, H., Bjorkegren, H., Sundstrom, L., Medbo, J., . . . Da Silva, I. (2016). Waveform and numerology to support 5G services and requirements. *IEEE Communications Magazine*, 54(11), 90--98.
- Zapata, O. U., & Alvarez, P. (2005). Edf and rm multiprocessor scheduling algorithms: Survey and performance evaluation. *Seccion de Computacion Av. IPN*, 2508.
- Zhang, Y. (2014). Tree-based resource allocation for periodic cellular M2M communications. *IEEE Wireless communications Letters*, 3(6), 621--624.