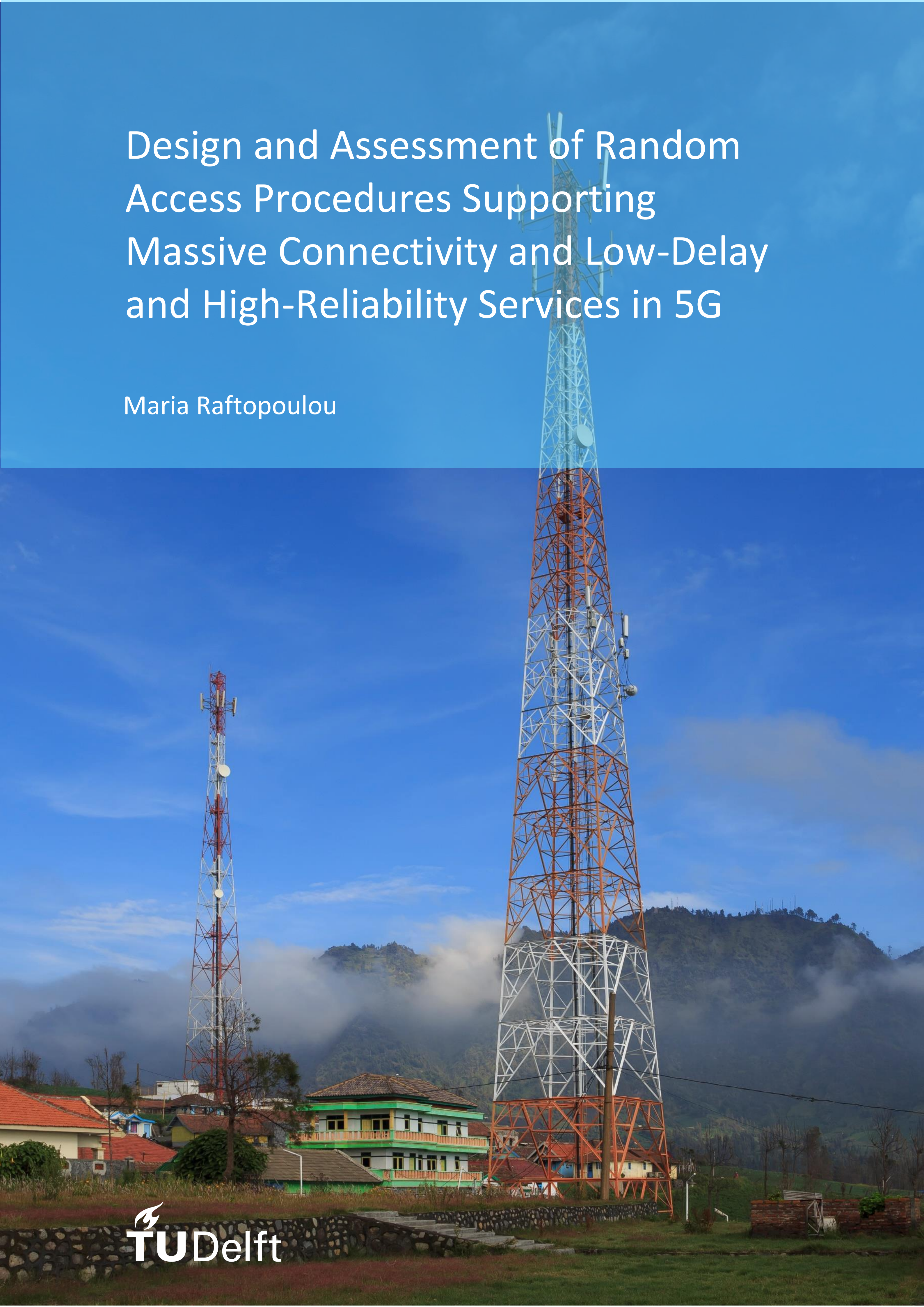
Design and Assessment of Random Access Procedures Supporting Massive Connectivity and Low-Delay and High-Reliability Services in 5G

Maria Raftopoulou



Design and Assessment of Random Access Procedures Supporting Massive Connectivity and Low-Delay and High-Reliability Services in 5G Networks

By

Maria Raftopoulou

in partial fulfilment of the requirements for the degree of

Master of Science
in Electrical Engineering
Track Telecommunications and Sensing Systems

at the Delft University of Technology,
to be defended publicly on Thursday September 20, 2018 at 10:00 AM.

Thesis committee:	Dr. Remco Litjens, MSc	TU Delft, TNO
	Dr. Ljupco Jorguseski,	TNO
	Dr. Ir. Jos Weber,	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Preface

This thesis marks the completion of my studies at the Delft University of Technology for the Masters of Science degree in Electrical Engineering in the track of Telecommunications and Sensing Systems. As this thesis was carried out at TNO in The Hague, I first want to thank the Networks departments at TNO for having me. My time at TNO was very enjoyable and provided me with many valuable experiences.

Additionally, I would like to thank Ljupco Jorguseski, my daily supervisor at TNO, who spent a lot of time guiding me and helping me with technical details and with the writings of this report. The quality of my work would have definitely not be in this level without his continuous constructive feedback. Also, I would like to thank Remco Litjens, my supervisor at TU Delft, as he was the one who brought me in contact with TNO and suggested me this thesis topic. Moreover, his feedback at critical parts of my work as well as his help with presentations and the writing of this report are of much importance as they helped me to improve my work significantly. Furthermore, I would also like to thank Haibin Zhang and Kallol Das who provided me with their inputs and feedback whenever needed, especially during the design of the new random access procedures for 5G. I also acknowledge the help that Johan Toh provided me, during the study of the propagation environment for this thesis.

As my thesis was part of the European/Taiwanese Clear 5G project (clear5g.eu) where TNO is a partner, I would like to acknowledge the feedback that was provided from the partners for this work. Also, I want to thank Ljupco Jorguseski, once again, as he gave me the opportunity to attend the telco meetings with the partners as well as encouraged me to regularly present my work during these telco meetings.

Finally, I would like to thank all my friends and family that supported me during this challenging for me work.

Maria Raftopoulou

The Hague, September 2018

Abstract

5G networks are expected to be used in many markets, one of which is the Factories of the Future (FoF). In the FoF, applications like regular monitoring and controlling of components (e.g. temperature) are introducing the need of massive deployment of sensors and actuators. Additionally, sensors which are monitoring time-critical components of the factory (e.g. pressure in power plants) should experience low delays with a high reliability. The current LTE-based technology for machine-type communications, namely Cat-M1, imposes limitations in supporting massive connectivity and application with low delay and high reliability requirements, primarily due to the so-called 'Random Access' (RA) procedure which is used by the devices to establish a connection with the base station, before the actual transmission of their data.

The key objectives of this study are to assess the RA procedure currently used in Cat-M1 networks as a baseline, and furthermore design and evaluate a new RA procedure for 5G networks, which targets typical FoF applications and their requirements. For the evaluation of the RA procedures studied, a system-level simulator was developed incorporating realistic characteristics of the factory and suitable propagation and traffic models. Additionally, for the best performing RA procedure, among those studied, a sensitivity analysis on the network load was carried out in order to determine the robustness of the procedure.

Based on the simulation assessment of the procedures studied, it was found that the so-called 'Two-Step RA procedure' is performing the best regarding end-to-end delay. In essence, this procedure allows devices to transmit their data right after they initiate the RA procedure and thus establishing a connection with the base station is not required. Specifically, it was found that e.g. in a FoF network with 3000 devices an end-to-end delay of 16 ms can be achieved with 99.99% reliability. Even though there is a gain of 64 ms compared to the end-to-end delay in Cat-M1 networks, the FoF requirement cannot be met as it indicates an end-to-end delay of 10 ms with 99.99% reliability. Moreover, we derive the most suitable configuration of the proposed RA procedure for different network loads yielding the best possible end-to-end delay performance.

Table of Contents

Table of Figures

Table of Tables

Chapter 1 Introduction

In this chapter, the required background is given as well as the objectives and contributions of this study. First, in Sections 1.1 and 1.2, an introduction to the Factories of the Future (FoF) and their requirements is provided respectively, as this study targets applications in FoF. Section 1.3, introduces 5G networks and correlates them with the applications in FoF. In Section 1.4, the Random Access (RA) procedure is presented as well as its importance and need in a network. Section 1.5 discusses the challenges in achieving the requirements of application in FoF due to the RA procedure that is currently in use. Section 1.6 presents the objectives of this study and then Section 1.7 gives an overview of studies found in literature regarding improvements of the RA procedure such that the challenging new requirements can be met. Finally, in Section 1.8 the contribution of this study is presented while Section 1.9 shows the approach followed to achieve the objectives of this study, as well as an outline of this study.

1.1 Factories of the Future (FoF)

In the FoF (or Industry 4.0 [1]), the factories will adapt a new ecosystem which integrates technologies such as autonomous robots, Internet of Things (IoT) and more, as presented in Figure 1-1. This new ecosystem aims for a clean, highly performing, environment friendly and socially sustainable factory [2]. For example, with the use of big data and IoT, the factory emissions can be monitored, through sensors, and a trend can be derived. Based on this trend, actions can be performed in order to reduce these emissions without degrading the quality and efficiency of the production. Another example which highlights the improvement of the factory’s performance are the use of autonomous robots as they could be used to perform complex tasks that human workers cannot. More examples can be found in [3].

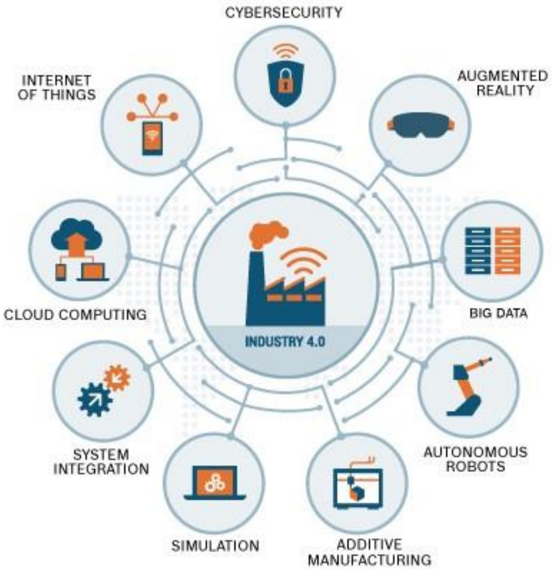


Figure 1-1 - Applications in FoF/Industry 4.0 [4].

These types of applications are introducing the need of deploying a massive number of sensors which will be responsible for measuring different components in the factory and actuators which will take actions whenever is needed. This kind of devices can be categorized in three groups:

- **Group 1:** Devices that transmit critical data very frequently. This type of devices is used in applications such as real-time machine control and therefore a continuous, reliable and fast communication is required.
- **Group 2:** Devices that transmit critical data incidentally, e.g. after an unexpected event. For example, sensors which are monitoring time-critical components of the factory (e.g. temperature or pressure in power plants) are detecting an abnormality and have to trigger an alarm or an action. Such unexpected events can trigger either a group of devices to report the event simultaneously or just a single one.
- **Group 3:** Devices that transmit non-critical data, typically not very frequently. These devices could be sensors which are monitoring regularly some non-critical components of the factory (e.g. room temperature in warehouses) and are deployed in large numbers.

For the transmission of critical data, it is important that the transmission will be fast and therefore for some applications Device-to-Device (D2D) communications can be used. D2D communication aims short range applications and can offer ultra-low delays, as the data are transmitted directly from the source to the target device, which makes it appropriate for localised real-time applications such as real-time machine control. As our study does not specifically aim at short range applications, we focus on communication between the device and the base station, hence not considering the potential of D2D communications in specific scenarios.

1.2 Requirements

The applications described in FoF are introducing new requirements and the need of distinction between devices, which is carried out based on the type of communication that is needed between the device and the base station. Devices that require transmission of critical data are distinguished as Ultra-Reliable Low Latency Communication (URLLC) (i.e. Group 1 and 2; see above) and devices that are massively deployed (and consequently do not handle critical data) are distinguished as non-URLLC (i.e. Group 3). The exact requirements of URLLC and non-URLLC devices in FoF are presented in Table 1-1, based on their definition by the 3rd Generation Partnership Project (3GPP) in [5]. These requirements concern the end-to-end delay, reliability and device density, where the end-to-end delay is defined as the time from the generation of data at the device until their correct reception at the base station.

Table 1-1 - Requirements for applications in FoF based on Table 5.3.8.1-1 in [5].

Device	End-to-End Delay	Reliability	Device Density
non-URLLC	50 ms – 1 s	> 99.9%	0.05 – 1 / m^2
URLLC	5 – 10 ms	> 99.9999%	0.05 – 1 / m^2

In this study these 3GPP requirements are adopted as follows. For the end-to-end delay it is obvious that the requirements will be completely different between URLLC and non-URLLC devices as the latter is transmitting only delay-tolerant data. It is defined in [5] that the non-URLLC data can handle from 50 ms to 1 s delay. In this study the delay requirement is set to 50 ms in order to take the most challenging target from the requirements range. On the other hand, the delay requirement for the URLLC traffic in this study is set to a maximum of 10 ms as the data are time-critical. For the URLLC traffic, it is noted that the end-to-end delay requirement is used as a reference to the maximum allowable end-to-end delay and the goal of this study is to reduce this end-to-end delay as much as possible.

Reliability is another important requirement for applications in FoF as it defines the percentage of data transmissions that should meet the delay requirements of e.g. 50 ms or 10 ms for non-URLLC or URLLC, respectively. For non-URLLC devices the reliability percentage requirement in this study is set to 99.9% while for URLLC devices the reliability percentage target is set to 99.99%. Note that it was decided to loosen up the very high reliability percentage of 99.9999% for URLLC devices due to practical evaluation limitations. For the evaluation, simulations are used and a reliability of 99.9999% would require simulations with an extremely high number of samples for the data transmissions and consequently excessively long simulation times.

The device density is also expected to be higher in 5G networks due to the massive deployment of sensors and actuators. From [5], it is defined that there will be 0.05 - 1 device per m^2 (assuming that all devices are deployed on the same floor) for both non-URLLC and URLLC devices. However, a small adjustment to this value was adopted in this study. The requirements for the URLLC devices from [5] were derived for all the URLLC devices that are deployed in the network (devices that handle frequent or incidental critical data as defined in Group 1 and Group 2 in Section 1.1). However, this study is focused only on URLLC devices that transmit data incidentally (see Section 1.4) and thus it was decided that the overall device density (URLLC and non-URLLC devices) should be at least one device per m^2 and that 5% of the total devices will be considered URLLC devices (as not many devices are expected to handle incidental critical data). It is noted that a short load analysis is included in Section 5.9, in order to derive conclusions for scenarios with lower and higher device densities while keeping the 95% and 5% of the total traffic to non-URLLC and URLLC traffic respectively.

The summary of the requirements considered in this study are presented in Table 1-2.

Table 1-2 - Final requirements used for this study.

Device	Number of Devices	End-to-End Delay	Reliability	Device Density
non-URLLC	95% of total devices	< 50 ms	> 99.9%	> 1/m^2
URLLC	5% of total devices	< 10 ms	> 99.99%	

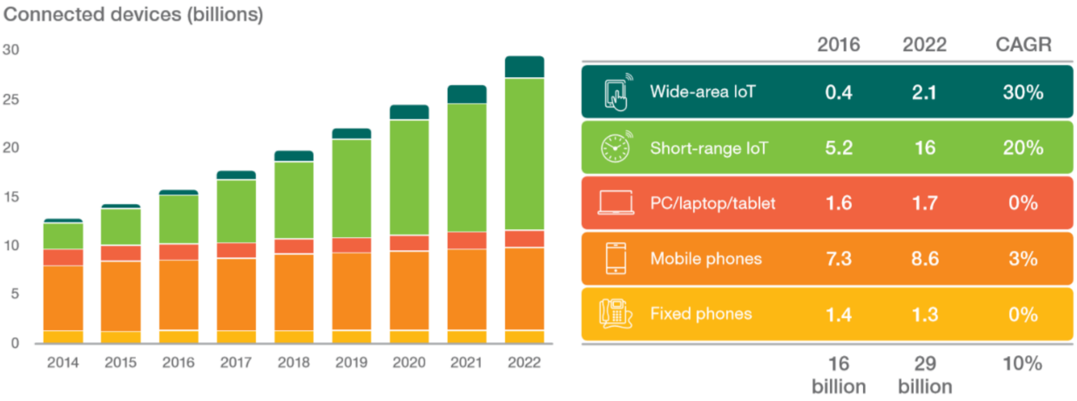
The wireless technology to be chosen in the network should address the above-mentioned requirements regarding end-to-end delay and reliability for the given device density. From all the available wireless technologies targeting these types of applications (e.g. LoRaWAN), the cellular (e.g. LTE) were the chosen ones to be studied, for the reasons given below (see also Section 2.1).

1.3 On the way towards 5G

The Long-Term Evolution (LTE) network is considered to be one of the most successful mobile communication network technologies because of its high data rates. Since its first deployment in 2009 by TeliaSonera, LTE underwent many enhancements such as the use of MIMO antennas and carrier aggregation which improved the overall performance of the network. Currently, bit rates of up to 300 Mbps can be observed with LTE, which highlights the success of increasing the throughput of data compared to previous cellular wireless communication networks [6].

In the coming years and by 2022, an annual increase of 45% in the total mobile data traffic is expected, 75% of the total mobile data traffic will concern video applications. Moreover, the further development of Internet of Things (IoT) networks is expected to have a great impact in mobile networks as by 2022, 18 out of 29 billion connected devices will concern IoT traffic [7]. A graph presenting the growth of IoT devices and other types of devices is shown in Figure 1-2. This rapidly growing IoT traffic is correlated with new types of applications which consequently introduce new requirements to the network such as energy efficiency, coverage, massive connectivity, low delay, high throughput and high reliability. These new requirements will therefore introduce the need of enhancing the current networks and create new technologies.

Currently, the LTE-based technologies that focus on IoT applications are Cat-M1 and NarrowBand IoT (NB-IoT). These technologies provide enhancements to the LTE standard in order to guarantee deeper coverage and a higher degree of connectivity compared to LTE, at the cost of lower throughputs and higher end-to-end delays. This study focuses on Cat-M1 as it is considered most suitable as a baseline technology to further enhance towards support of the requirements given in Table 1-2 for FoF applications (see also Section 2.1). It is very challenging to achieve low end-to-end delays with Cat-M1 (see Section 1.4), hence new technologies should indeed be designed in order to address the requirements of FoF, which will finally contribute to the standardization and implementation of 5G networks.



	2016	2022	CAGR
Wide-area IoT	0.4	2.1	30%
Short-range IoT	5.2	16	20%
PC/laptop/tablet	1.6	1.7	0%
Mobile phones	7.3	8.6	3%
Fixed phones	1.4	1.3	0%
	16 billion	29 billion	10%

Figure 1-2 - Detailed information in the growth of connected devices between the years 2014 to 2022 as presented in [7].

For the definition of 5G, three main service categories were introduced; enhanced Mobile Broadband (eMBB), Ultra Reliable and Low-Latency Communications (URLLC) and massive Machine Type Communications (mMTC). These three categories along with a few example applications are shown in Figure 1-3. The eMBB service category covers applications requiring high data rates (peak bit rates up to 20 Gbps) which are a requirement for applications such as ultra-high definition video and Virtual Reality (VR). For applications such as remote surgery and smart energy grids, that require the transmission of critical or emergency messages, the URLLC category applies as all these messages are time-critical and should reach their destination with an extreme degree of reliability. Finally, the mMTC category is defined mainly for IoT applications such as smart cities and smart sensors as they require the connectivity of billions of devices [8].

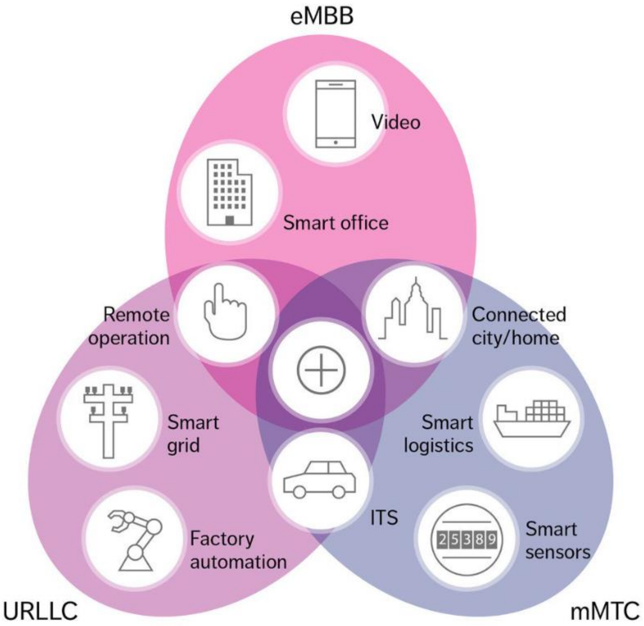


Figure 1-3 - 5G main categories with applications as shown in Figure 1 in [8].

In Section 1.1 it was defied that three groups of devices (Group 1, 2 and 3) are expected to be deployed in FoF. From these three groups, Group 1 and Group 2 can be categorized under the URLLC category presented in Figure 1-3 as they are handling time-critical data and thus require fast and reliable transmissions. Group 3 can be defined as a non-URLLC category (or an mMTC category based on Figure 1-3) as the data to be transmitted are more delay tolerant than in the URLLC category and the main challenge is to handle the large number of devices.

1.4 Connecting to the base station

As already mentioned, one of the requirements defined for FoF is related to the end-to-end delay, which is the time from the generation of data at the device until its correct reception at the base station. For the correct definition and evaluation of the end-to-end delay, it is important to understand the two so-called Radio Resource Control (RRC) modes, i.e. the 'RRC IDLE' mode and 'RRC CONNECTED' mode. More specifically, the RRC modes determine the state of the device which defines the amount of resources that is available to the device and its energy consumption.

In general, a device can switch between the two RRC modes based on its need to transmit/receive data. Specifically, when a device is inactive in the network (e.g. there is no transmission of data between the device and the base station), it is in RRC IDLE mode. Devices in RRC IDLE mode need to switch to RRC CONNECTED mode in order to transmit/receive data and consequently establish a connection with the base station as illustrated in

Figure 1-4. Additionally, devices that complete their transmission/reception of data, release their connection and switch to RRC IDLE mode after a pre-determined inactive time (e.g. 10 seconds), as also illustrated in

Figure 1-4. Therefore, for the end-to-end delay calculations, the time of connection establishment is needed (i.e. time needed for a device to switch from RRC IDLE mode to RRC CONNECTED mode) as well as the actual time of the data transmission, once the device is in RRC CONNECTED mode (see also Section 2.3).

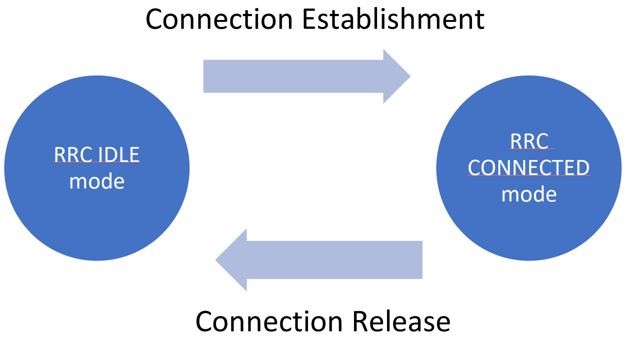


Figure 1-4 - Diagram of RRC modes.

In order to establish a connection with the base station, the device makes use of the so-called Random Access (RA) procedure, which is the procedure used to switch from RRC IDLE mode to RRC CONNECTED mode. To initiate the RA procedure, the device randomly selects one out of the 64 so-called preambles that are available, and transmits this preamble to the base station. All the preamble transmissions are carried out on a specific channel on the uplink, namely the Physical Random Access CHannel (PRACH), and multiple devices can make use of this channel simultaneously (see Section 2.2.1.2). In case the different simultaneous random access attempts of multiple devices use the same preamble, their transmissions may collide and hence not be properly heard by the base station. A reattempt would then be needed. After a successful reception of the preamble at the base station, an exchange of messages between the base station and the device is happening which finally leads to an established connection between the two.

Besides the potential collisions described, the RA procedure introduces multiple bottlenecks to the network, as described in Section 1.5, and a new RA procedure needs to be designed to support the stringent requirements of 5G applications. The key objective of this study is to assess the RA procedure currently used in Cat-M1[1] and design and evaluate a new RA procedure which addresses the requirements of applications in FoF.

[1] A detailed explanation of this particular RA procedure is given in Section 2.4.

As discussed in Section 1.1, there are three groups of devices that are expected to be deployed in the FoF. From these three groups, it is clear that devices in Group 1 are best served regarding end-to-end delay when they are configured in RRC CONNECTED mode all the time, as they continuously transmit data, and therefore they will not make use of the RA procedure. Devices in Group 2 and Group 3 are expected to spend most of the time in RRC IDLE mode, as their data transmissions are infrequent, and thus for every one of their transmissions they will have to go through the RA procedure. In line with our objective, for this study we will therefore focus on devices in Group 2 ('URLLC' devices) and Group 3 ('non-URLLC' devices). Details about the exact traffic generated by these devices in the conducted study, can be found in Section 4.2.

1.5 Challenges

The current LTE and LTE-A communication standards that are already available were designed in such a way that they efficiently serve Human-to-Human (H2H) or Machine-to-Human (M2H) communication such as transmission of voice or video. On the other hand, the most applications in the FoF require the support of Machine Type Communications (MTC). As mentioned, Cat-M1 is designed to support MTC but its performance cannot guarantee the performance requirements of the most demanding foreseen FoF applications. As the RA procedure has a key role in meeting the requirements for the FoF applications, enhancements to the currently used RA procedure need to be studied, as well as the development of new RA procedures.

One of the requirements for the FoF applications is that end-to-end delay for URLLC devices has to be at most 10 ms. However, the current LTE standard introduces a minimum of 24 ms delay only for the RA procedure [9] and therefore it is clear that the delay needs to be decreased by more than 50%, which is highly challenging. Another parameter that should be taken into account while re-designing the RA procedure is that the end-to-end delay for non-URLLC devices should still stay below 50 ms and thus procedures that favor URLLC devices but severely degrade the end-to-end delay for non-URLLC, should not be considered.

Guaranteeing that 99.9% and 99.99% of non-URLLC and URLLC devices, respectively, need to meet the end-to-end delay requirement is also challenging. The underperformance of just a few devices, e.g. due to poor propagation circumstances, may cause a violation of the reliability requirement of the network due to the highly stringent percentiles that apply. Additionally, the high reliability percentages are also adding a challenge to the evaluation of the RA procedure performance, as excessive computational resources will be needed. As the URLLC devices are just 5% of the overall devices, there is a need for few millions of samples for the data transmissions in order to have statistically reliable results for the high reliability values of 99.99%. Measuring with sufficient statistical accuracy such high reliability percentiles can be therefore limited by simulation time and/or computational restrictions.

In FoF and 5G networks, it is also expected that more devices will be deployed to the network than in the current LTE/LTE-A/Cat-M1 networks. This implies that a massive number of devices will have to transmit data and consequently will have to make use of the RA procedure in order to establish a connection with the base station. As mentioned before, with a high number of simultaneous RA attempts, preamble collisions may occur and all devices that transmitted that particular preamble fail to establish a connection. Those devices will re-initiate the RA procedure after a random back-off time (see also Section 2.4) with a new randomly chosen preamble and so on. Therefore, the deployment of a massive number of devices is expected to introduce a higher preamble collision probability to the network which can lead to the overload of the PRACH as devices will keep transmitting preambles without being able to establish a connection with the base station. Consequently, the new RA procedure should be designed is such a way that overloads of the PRACH will be avoided.

An additional problem that arises with the high preamble collision probability is the high access delays, where the access delay is defined as the time needed for a device to establish a connection with the base station (time needed to switch from RRC IDLE mode to RRC CONNECTED mode). The overload created on the PRACH forces the devices to transmit preambles multiple times (multiple attempts to establish a connection) until they manage to successfully transmit their preamble. This behaviour introduces extra delays in the network or even outages as devices are configured by the base station with a maximum number of connection attempts. Once a device reaches this maximum number of connection attempts without successfully establishing a connection, the device is considered to be in outage and it no longer tries to establish a connection. Therefore the challenging trade-off between massive deployment of devices and the access delay, and consequently the end-to-end delay, is highlighted.

1.6 Objectives

Based on the requirements derived for the applications in FoF and 5G networks and the challenges that arise, the objectives of the presented study are the following:

1. Assessment of the RA procedure in the current Cat-M1 networks as well as assessment of the latest RA procedure enhancements, targeting 5G applications, provided by 3GPP in Release 15.
2. Design of a new RA procedure for 5G networks which fulfils the requirements of the FoF applications regarding end-to-end delay and reliability and derive under which conditions (e.g. device density) these requirements can be met.

1.7 Related work

Improvements of the RA procedure have been applied in the recent LTE and LTE-A specifications by 3GPP. For example, the Access Class Barring (ACB) scheme can be applied to help with capacity bottlenecks during RA procedure. More specifically, multiple access classes can be defined in the network with a different access probability. Each time a device needs to initiate the RA procedure, it draws a random number and it actually initiates the RA procedure only if that number is lower than the access probability defined for the access class that it belongs to. Otherwise, the device has to back-off for a random time, based on a barrier timer defined for that particular access class. Therefore, ACB can prevent the PRACH overload but introduces higher access delays, especially to access classes with a low access probability and it is typically used to distinguish Machine-to-Machine (M2M) and Human-to-Human (H2H) traffic. In [10] and [11], more procedures that aim to differentiate M2M and H2H traffic are presented.

In [11], the problem of *massive access* is addressed by using the Distributed Queuing Random Access Protocol (DQRAP). This new protocol aims to guarantee access to the network to a large number of devices for M2M traffic while not degrading the access performance of H2H traffic. The main idea behind DQRAP is the use of mini-slots to reserve resources on the PRACH when the M2M traffic is high, while the H2H traffic uses the conventional LTE RA procedure. Additionally, DQRAP uses the conventional LTE RA procedure for both M2M and H2H traffic when the M2M traffic is low and only switches to a reservation protocol when the M2M traffic becomes high. From the results provided, it is observed that there is a gain in the number of devices that can access the network but comes at the cost of higher access delays. The authors in [12] introduce the clustering technique, in order to also address the problem of massive access, where devices are clustered based on their location and mobility information. With this procedure, devices of the same cluster communicate with a Cluster Head (CH) which is responsible to aggregate the data of the cluster and then forward the aggregated data to the base station. It is concluded that the RA procedure can be improved with this method even though access delay values were not presented. Delay analysis, in scenarios of clustering, have been studied in [13] where it is proven that for high traffic loads, the RA process contributes more than the aggregation process (at the CH) to the total access delay. Therefore, the clustering technique is promising for scenarios with high traffic loads but enhancements to the RA procedure are recommended as it contributes the most to the access delay and thus to the end-to-end delay.

Work which focuses on the *reduction of the RA delay* can also be found in literature. In [14], a method based on dynamic resource allocation is presented, where the base station uses a self-optimizing algorithm which adjusts the resources needed during the RA procedure based on the network's load. However, the authors did not include any kind of results to quantify the benefits of the method in a realistic scenario. Another method for delay reduction is the Distributed Queuing (DQ) scheme in combination with the so-called 'm-ary tree splitting' algorithm which is discussed in [15]. The main principle of this scheme is the organization of devices in virtual queues and each device continuously keeps the status of its position in the queue in order to know when it will get an access grant for transmission. Additionally, for the transmission of the preamble and thus for the initiation of the RA procedure, three mini-slots are being used on the PRACH which implies that only three devices can simultaneously transmit their preamble. Furthermore, the implementation of this scheme is limiting the number of available preambles from 64 (in LTE) to just six and thus the results should also be compared to networks where fewer preambles are available (e.g. networks with large cell radius). In general, the number of available preambles is correlated with the radius of the cell (a larger cell radius implies fewer available preambles) due to the way that the preambles are generated. The results of this method illustrate a significant decrease of the access delay compared to LTE with ACB but only for networks with cell radius larger than 5 km. For a small number of devices that need to transmit critical short messages in the context of industry applications, drastic delay reductions have been presented in [16]. The presented approach is based on the existing RA procedure with the difference that the data are transmitted during with the RA procedure which leads to delays of only 15 to 20 ms.

The improvement described in [17], namely Early Data Transmission (EDT), is one of the latest improvements adapted by 3GPP in Release 15 and it aims in reducing the end-to-end delay for devices that need to transmit small and critical messages. The main idea of this method is to transmit the data during the RA procedure as was also studied in [16]. In the same line of thought as EDT, more drastic techniques have also been studied and they aim in reducing the delay even further. The most popular is the two-step RA procedure which uses only half of the messages than the conventional RA procedure. This idea can be found in [18] and it was presented in one of the 3GPP discussion meetings. However, no exact implementation methodology is defined and no results are provided. This procedure has also been discussed in [19] where the authors present the challenges of implementation. Finally, in [20], the additional improvement of resource reservation is presented which is implied that can be implemented in combination with the two-step RA procedure in order to meet the 10 ms delay requirements that are introduced in 5G.

The conventional RA procedure and all the procedures that have been presented so far are assumed to use Orthogonal Frequency-Division Multiple Access (OFDMA). OFDMA suggests that all devices are allocated to orthogonal resources in the frequency domain such that their transmissions will not be causing interference to other simultaneous transmissions. In general, OFDMA is one out of the four *Orthogonal Multiple Access (OMA)* schemes that are generally used, as the orthogonality of resources can also be defined in the time (i.e. Time Division Multiple Access (TDMA)), code (i.e. Code Domain Multiple Access (CDMA)) and space (i.e. Multi-User Multiple Input Multiple Output (MU-MIMO)) domain. However, OMA limits the capabilities of the channel regarding spectral efficiency[2] as only one device can utilize the specific resources in the shared orthogonal domain, as illustrated in Figure 1-5. For example, devices in OFDMA share the time domain but use different resources in the frequency domain. Also, in CDMA and MU-MIMO, devices share the same resources in the time and frequency domain but different resources in the code and spatial domain respectively. Therefore, a new technique has been introduced in literature, namely the *Non-Orthogonal Multiple Access (NOMA)* [21] which allows multiple devices to transmit simultaneously on the same resources, without sharing an orthogonal domain. For example Figure 1-5 shows the power-domain NOMA which allows devices to share resources in the time and frequency domain (in contrast to OFDMA and TDMA) but they are multiplexed in the power domain (i.e. each device uses a different transmission power). Moreover, the power-domain NOMA is not comparable to CDMA as with NOMA it is implied that all devices use the same spreading code and thus the code domain is not orthogonal. Similarly, power-domain NOMA does not imply orthogonality in the spatial domain and thus it is also not comparable to MU-MIMO. Additionally, the authors in [21], state that NOMA is promising to achieve among other spectral efficiency, massive connectivity and reduce the transmission delay and signaling overhead at the cost of a more complex receiver, thus making NOMA appropriate for 5G applications.

[2] Spectral efficiency is defined as the amount of information than can be transmitted over a frequency band.

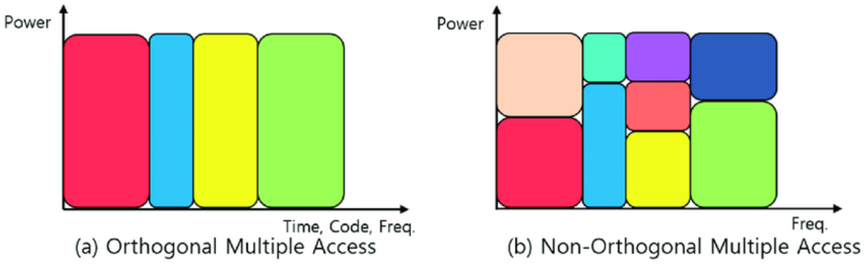

Power
Time, Code, Freq.
(a) Orthogonal Multiple Access
Power
Freq.
(b) Non-Orthogonal Multiple Access

Figure 1-5 - Assignment of resources in OMA (left) and NOMA (right), where each color represents a different device [22].

A comparison between Orthogonal Multiple Access (OMA) and NOMA is presented in [23] and it is clear that NOMA can handle more devices and provides better throughputs. It is also stated that there is no need for a RA stage as signaling and data can be transmitted together, and thus lower delays can be achieved. Additionally, the authors in [24] provide further results, while using the so-called NOMA scheme *contention-based Sparse Code Multiple Access (SCMA)*, which illustrate a higher reliability while supporting 2.8 times more devices in the system over the contention-based OFDMA. It is also worth mentioning that NOMA and uplink SCMA are applied in the European project METIS [25] for the definition of the 5G air interface with positive results. Last but not least, the authors in [26] designed a mechanism which combines the non-orthogonal transmissions with the conventional RA procedure such that it can co-exist with the current LTE system. Their technique makes use of power domain multiplexing and achieves a reduced access delay while supporting 30% more traffic than the current RA procedure.

Overall, procedures which allow the transmission of data during the RA procedure seem to be promising as the end-to-end delay can be improved significantly. EDT is one of these procedures as has been presented in [16] and [17], and has also been adapted by 3GPP Release 15. Furthermore, the two-step RA procedure is expected to reduce the end-to-end delay even further, as discussed in [18], but many open issues arise which are not addressed in literature so far. The most challenging part in implementing the two-step RA procedure is the allocation of resources to the devices as there is no prior communication about this with the base station and thus high interferences are expected (see also Section 3.2). Additionally, NOMA procedures are considered to be promising as they can support a high number of devices while achieving low end-to-end delays, as discussed in [23], [24] and [26]. However, NOMA procedures imply that there is no need of having a RA procedure, according to [23], and thus makes it incompatible with the current 3GPP standardized networks. Furthermore, there are many open issues in implementing the NOMA procedures in real networks, as discussed in detail in [26]. Two of these issues are the allocation of resources as there is no prior communication with the base station (as in the two-step RA procedure) and the appropriate power control that is critical for the effective use of such a non-orthogonal procedure. Finally, in [26], challenges in the performance evaluation through link and system level simulations of non-orthogonal transmissions are presented.

1.8 Contribution

This study investigates and quantifies the performance of enhanced RA procedures for 5G networks, via simulations, in order to meet the challenging 5G requirements regarding end-to-end delays up to 10 ms that are achieved with high-reliability (i.e. 99.99% of the URLLC data transmissions are with end-to-end delay lower or equal than 10 ms). Specifically, the EDT procedure presented in literature is evaluated along with a proposed enhanced version of EDT. Additionally, the two-step RA procedure has been used in this study, where a possible implementation is proposed and evaluated such that it can be applied for the FoF in the context of 5G. It is important to note that NOMA-based solutions were not studied for the reasons already discussed in Section 1.7, and as also considered and discussed in [26].

1.9 Approach and thesis outline

The high-level approach used in this study in pursuit of the objectives stated in Section 1.6, is presented below:

1. Derive the requirements for 5G in FoF.
2. Perform a literature review in order to obtain insight on the RA procedure and its shortcomings, challenges and possible solutions.
3. Study the latest technologies used for MTC and choose the most suitable one (Cat-M1) to be used as the reference technology.
4. Obtain deep understanding on the RA procedure used in LTE and Cat-M1 (reference technology), especially in the PHYsical (PHY) and Medium Access Control (MAC) layers.
5. Define details about the reference scenario concerning the network layout, traffic model and propagation environment.
6. Develop a system-level simulator that enables assessment of the RA procedure used in the chosen reference scenario and analyze the results.
7. Study the latest 3GPP enhancements regarding the RA procedure as well as new RA procedures proposed in literature for 5G networks.
8. Develop and integrate to the simulator new RA procedures that could possibly be used in 5G networks, either from literature or new ones.
9. Assessment of the new RA procedures based on the simulation results.
10. Repeat steps 8 and 9 in an iterative manner, every time with further enhancements on the RA procedure, until the set requirements are met.
11. Derive conclusions, recommendations, limitations and suggest further work based on the obtained results.

The requirements of 5G in FoF and the literature review have already been presented in Sections 1.2 and 1.7 respectively. 0 discusses the different MTC technologies, Cat-M1 and describes in detail the RA procedure in Cat-M1. Chapter 3, describes in detail the latest 3GPP enhancements regarding the RA procedure for 5G applications, as well as the new RA procedures that have been studied for application in 5G. The modelling aspects regarding the reference scenario including the propagation environment, traffic model and further simulation modelling are presented in Chapter 4. In Chapter 5, the results of all the RA procedures are presented along with comparisons to the reference scenario and to each other. Finally, Chapter 6, highlights the main conclusions, recommendations and limitations that are obtained throught this study, along with recommendations for future work.

Chapter 2 Overview of Cat-M1 (eMTC)

In this chapter, an extensive overview of Cat-M1 is presented as it is used as the reference technology upon which further technological enhancements are developed. After a brief high-level introduction in Section 2.1, the physical uplink and downlink channels are presented in Section 2.2. The RRC modes are discussed in Section 2.3. Finally, the Random Access (RA) procedure is described in detail in Section 2.4.

2.1 Introduction

The Internet of Things (IoT) communication systems have different requirements than the conventional human-oriented communication systems, as shown in Figure 2-1. The IoT devices should have reduced complexity, which will contribute to lower cost and longer battery life. Additionally, the requirements of IoT applications can vary significantly but some of those applications may require support for deeper coverage and also higher numbers of devices compared to legacy human-oriented communication systems.

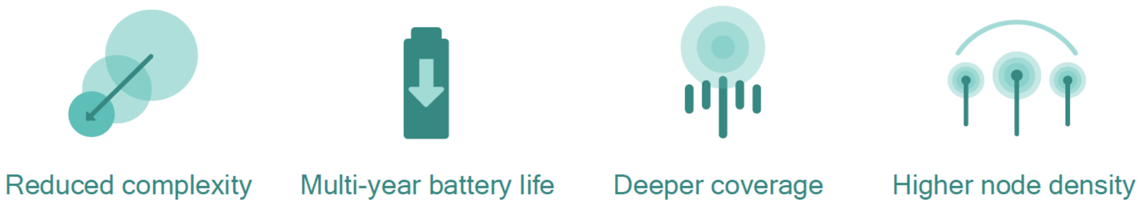


Figure 2-1 - Requirements of IoT networks [27].

The IoT communication systems that provide wide area coverage and support battery-operated IoT devices are also referred to as Low-Power Wide Area Networks (LPWAN). Several LPWAN communication technologies exist and operate in unlicensed spectrum, including LoRaWAN, SigFox, Symphony Link, Ingenu RPMA and Weightless. An extensive summary for each of these communication technologies is presented in [28], including an analysis of their respective advantages and disadvantages. Regardless of the availability of the above-mentioned LPWAN technologies, in this study we will consider only cellular 3GPP-based standards, for several reasons:

- Using globally standardized IoT communication technology in licensed spectrum implies a controlled interference environment and easier interworking between devices and systems that are essential requirements for mission- and business-critical applications.
- Cellular networks already support redundancy, end-to-end security and scalability that are important prerequisites for many IoT applications.
- It is expected that Mobile Network Operators (MNOs) worldwide will utilize the 3GPP-based IoT standards because this IoT overlay in their existing cellular networks could lead to possible cost savings as a new network will not have to be built from scratch.
- Cellular 3GPP-based standards will co-exist with other LPWAN standards that are currently in use and operate in unlicensed bands.

From the 3GPP-standardized User Equipment (UE) categories for LTE technology, UE Category 1 (Cat-1) was the first one to be included in the LTE specifications (3GPP Release 8) as the category used for machine-type communications (MTC) and cellular IoT traffic. However, in 3GPP Release 12, UE Category 0 (Cat-0) was standardized for IoT traffic. This new UE category supports lower bit rates than Cat-1 but still high enough to satisfy the requirements of the intended IoT applications. This bit rate reduction enabled the UE complexity to also be reduced by 50% and thus decrease the IoT terminal cost.

With 3GPP Release 13, a new UE category was standardized, namely Cat-M1 or eMTC. This new UE category has the capability of keeping the same peak bit rate (1 Mbps in both UL and DL) as Cat-0 while reducing the used bandwidth from 20 MHz to 1.4 MHz, and reduce the modem complexity by another 50% [29]. Moreover, Cat-M1 introduces a coverage gain of 15 dB (for 20 dB UE transmit power) compared to Cat-1, supporting a coupling loss up to 155.7 dB [27], and can be operated within any regular LTE band. 3GPP Release 13 further introduced the narrowband IoT (NB-IoT) technology and associated UEs, which reduces the modem complexity and bandwidth even further, and increases coverage by another 25 dB of acceptable coupling loss, compared to Cat-1 at a cost of lower bit rates.

Table 2-1, shows the different 3GPP-standardized UE categories that are meant for IoT traffic. For this study, our focus is on 3GPP Release 13 as it uses reduced bandwidth and UE complexity, and increased coverage compared to previous 3GPP releases. From the two new categories introduced in 3GPP Release 13, the one that will be used as a reference scenario is Cat-M1 as due to its larger bandwidth and higher bit rates it can provide lower end-to-end delays; a performance aspect which is very critical for Factories of the Future (FoF) applications. It is noted that for the rest of this study, a UE will be referred as a *device*.

Table 2-1 - Characteristics of each UE category [27] [30].

3GPP Release	UE Category	Maximum DL bitrate	Maximum UL bitrate	#Rx Antennas	Maximum Tx Power	Duplex Mode	Bandwidth	Complexity (vst. Cat-1)
13	Cat-M1 / eMTC	1 Mbps	1 Mbps	1	20 or 23 dBm	half/full duplex	1.4 MHz	25%
	NB-IoT	0.17 Mbps	0.25 Mbps	1	20 or 23 dBm	half duplex	0.2 MHz	<19%
12	Cat-0	1 Mbps	1 Mbps	1	23 dBm	half duplex	20 MHz	50%
8	LTE Cat-1	10 Mbps	5 Mbps	2	23 dBm	full duplex	20 MHz	100%

Cat-M1 (eMTC) provides enhancements to previous 3GPP releases and its design supports higher node density, deeper coverage, reduced complexity and energy efficiency in the cost of higher delay compared to LTE [27]. Cat-M1 is deployed in-band with LTE traffic as it operates in multiple narrowbands of 1.4 MHz within the frequency range of LTE spectrum as illustrated in Figure 2-2 (this feature is also applicable in NB-IoT).

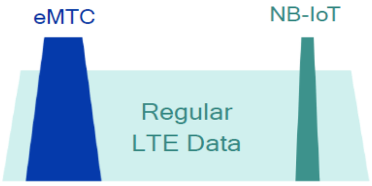


Figure 2-2 - In-band deployment for Cat-M1 (eMTC) and NB-IoT [60].

An additional feature introduced in Cat-M1 (and NB-IoT) is the different Coverage Enhancement (CE) modes, namely CE mode A and CE mode B. UEs are required to support CE mode A, while support for CE mode B is optional as the latter is used only in very poor coverage areas. The CE modes can be further divided to CE levels 0 and 1 for CE Mode A and CE levels 2 and 3 for CE Mode B, as shown in Figure 2-3. The main configuration parameter differs for the different CE levels is the number of repetitions that are applied every time a transmission is happening in either the UpLink (UL) or DownLink (DL) channel. For example, this mechanism in the UL channel, allows UEs in poor coverage areas to transmit their data multiple times, such that a higher aggregated Signal to Interference and Noise Ratio (SINR) is achieved at the base station (evolved NodeB (eNB) in LTE and next generation NodeB (gNB) in 5G) receiver and thereby enhances the likelihood of a successful transmission. A similar approach applies in the DL channel. Typically, the base station signals to the device the Reference Signal Received Power (RSRP) thresholds that are used by the device to decide in which CE mode/level it should operate. The device decides its CE level based on the measured RSRP value and on the signaled threshold values (defining the CE levels), which are broadcast by the base station in the System Information Block (SIB). Based on the measured RSRP value the device, during the RA procedure, announces to the base station in which CE level it will operate [31].

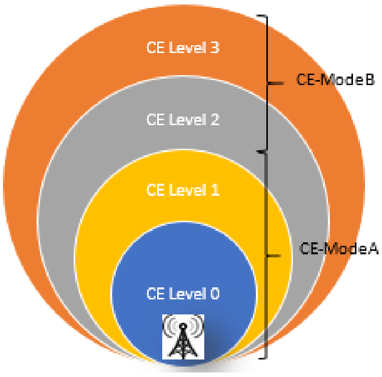


Figure 2-3 - Coverage Enhancement modes and levels in Cat-M1 [31].

2.2 Physical layer

In this section, an overview of the physical channels is presented. First, the channels used for the UL transmissions are discussed, followed by the channels used for the DL transmissions. As already mentioned, Cat-M1 is deployed in-band with LTE and therefore Cat-M1 uses only part of the total available physical (time-frequency) resources. In general, each channel in Cat-M1 uses one narrowband which is equal to six Physical Resource Blocks (PRBs) except for the Physical Uplink Control CHannel (PUCCH) which uses just two PRBs. Table 2-2 shows the number of PRBs and the maximum number of narrowbands that are allowed for different bandwidths of the LTE carrier. The values that are mentioned in the remainder of this study are assuming the LTE carrier bandwidth of 5 MHz. An extensive summary of the physical channels in LTE and Cat-M1 is given in [32] as well as a comparison between the two technologies.

Table 2-2 - Number of PRBs and Cat-M1 narrowbands for different LTE carrier bandwidths.

LTE carrier bandwidth (MHz)	Number of PRBs	Number of Cat-M1 narrowbands
1.4	6	1
3	15	2
5	25	4
10	50	8
15	75	12
20	100	16

2.2.1 Uplink channels

The three main physical uplink channels that are relevant for our study are the aforementioned PUCCH, the Physical Random Access CHannel (PRACH) and the Physical Uplink Shared CHannel (PUSCH). The amount of resources that are allocated for each channel is presented in Figure 2-4 and an overview of each channel is given below.

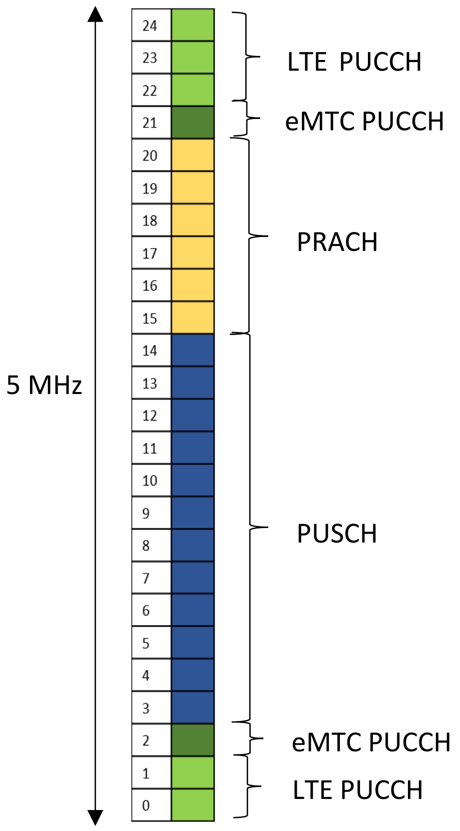


Figure 2-4 - Uplink channel resource allocation for a 5 MHz LTE carrier.

2.2.1.1 *PUCCH*

Based on [33], it is assumed that five PRBs are allocated for LTE PUCCH transmissions and two PRBs are allocated for Cat-M1 (eMTC) PUCCH transmissions (see also Figure 2-4). Focusing on the eMTC PUCCH channel, two types of control messages are transmitted: Scheduling Requests (SRs) and Hybrid Automatic Repeat reQuest (Negative) ACKnowledgements (HARQ-(N)ACK).

In general, a device transmits an SR in the PUCCH whenever it has UL data to transmit and therefore it must request dedicated resources from the serving base station for its UL data transmission. The SR occupies one PRB. Furthermore, a device transmits HARQ-(N)ACKs on the PUCCH after the reception of DL messages. The HARQ-(N)ACK is used to indicate whether the received DL message was corrupted and therefore whether a re-transmission of the message is needed.

Finally, frequency hopping between the two PUCCH PRBs is required such that diversity gain can be achieved.

2.2.1.2 *PRACH*

The PRACH is the channel used in order to initiate the Random Access (RA) procedure and it can be used for both eMTC and non-eMTC traffic. Basically, every time that a device needs to initiate the RA procedure, it randomly chooses a so-called Zadoff-Chu preamble and transmits it to the base station on the PRACH. In total there are maximum 64 orthogonal preambles which can be generated from prime-length Zadoff-Chu sequences and each one of those preambles needs six PRBs to be transmitted. More details about the preambles can be found in Chapter 17 in [6]. The *range of preambles* that a device can choose from is defined by the base station and different ranges can be used per CE level. Ideally, this preamble range should be set to maximum such that the probability of two or more UEs selecting the same preamble will be almost equal to zero. This probability is referred to the collision probability and it is a key performance metric when configuring the PRACH, since in case of collision, none of the UEs that are part of the collision will have a successful transmission. UEs with unsuccessful transmissions have to re-initiate the RA procedure, causing thus extra end-to-end delay. It is also noted that after a predefined (by the base station) number of unsuccessful RA attempts, the device considers itself in outage (see Section 2.4).

Apart from the preamble range, there are other parameters that can be configured and their configuration can also vary depending on the CE level. One of those parameters is the so-*called PRACH configuration index* which defines the time periodicity of the PRACH. This implies that only during specific subframes the PRACH is available. In those subframes no configured for potential PRACH transmission, the UL PRBs can be used for the PUSCH (see below). Generally, the periodicity is defined as the number of PRACH subframes within a 10 ms Radio Frame (RF), where a RF is defined as the set of ten consecutive subframes (see also Table 2-3). In total, there are 64 different configurations that can be used for the periodicity, which are defined in Table 5.7.1-2 in [34], but typically there will be one, two, three, five or ten PRACH subframes within one radio frame.

Because a preamble can be repeated multiple times in subsequent PRACH subframes due to the different CE levels, a *fresh* RA attempt can be initiated only in a subset of the PRACH subframes, leaving the remaining PRACH subframes only for repetitions. From [34], it is defined that the subframes available for fresh RA attempts are the subframes $jN_{start} + N_{rep}$ over the set of the PRACH subframes where N_{start} is an index signaled by the base station, N_{rep} is the number of repetitions of the preamble and $j = 0, 1, 2, \dots$. Moreover, N_{start} cannot be smaller than N_{rep} for each CE level. An example that illustrates this behavior is presented below:

> *Assuming a configuration index equal to three, based on Table 5.7.1-2 in* [34]*, the PRACH will be available in subframes 1, 11, 21, 31, 41, 51, 61 and so on. Setting the starting subframe (N_{start}) equal to 4 and the maximum number of repetitions per preamble (N_{rep}) equal to 2, we can calculate from the above-mentioned equation the subframe indices 2, 6, 10, 14. Therefore, the subframes in which a device can initiate a fresh RA attempt are the subframes 11, 51, 91, 131 and so on. In case a device initiates its RA attempt in subframe 51, it will repeat its preamble in subframe 61 since N_{rep} is equal to two. This behavior is also illustrated in* Table 2-3 *where as 'PRACH' are considered the subframes that a preamble can be transmitted and as 'Fresh' is considered the subset of PRACH subframes that a device can initiate its RA attempt. The subframes shown as 'Repetition', are the subframes to be used for a repetition of the preamble. The PRACH subframes that are not labeled as 'Fresh' or 'Repetition' (such as subframe 1), can be configured for use in a different CE level.*

Table 2-3 - Example presenting the available starting PRACH subframes with configuration index three, $N_{start} = 4$ and $N_{rep} = 2$.

	RF #0										RF #1									
Subframe	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
PRACH		√										√								
Fresh												√								
Repetition																				

	RF #2										RF #3									
Subframe	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**
PRACH		√										√								
Fresh																				
Repetition		√																		

	RF #4										RF #5									
Subframe	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**
PRACH		√										√								
Fresh												√								
Repetition																				

Finally, there is also the *frequency offset* parameter which indicates in which narrowband within the system bandwidth of the LTE carrier the PRACH is located. Defining a different frequency offset for each CE level can provide extra PRACH capacity to the network as there will be fewer preamble collisions when compared to the case of using the same narrowband for all CE levels. This PRACH capacity gain comes at a cost of fewer available UL resources for other UL data transmissions (PUSCH). As mentioned before, for the case that two or more CE levels share the same narrowband and subframes for the PRACH, disjoint ranges of preamble values (from the total of 64 preambles) can be configured for each of those CE levels such that there will be no preamble collisions among different CE levels.

It is important to also note that frequency hopping is mandatory for Cat-M1 UEs when transmitting on the PRACH.

2.2.1.3 *PUSCH*

The PUSCH is the channel used for the UL data transmissions. The device needs to request dedicated resources in the PUSCH (with a SR on the PUCCH) in order to transmit its UL data and those resources are signaled by the base station via a scheduling grant transmitted on the MPDCCH (see Section 2.2.2.1). Typically, the UL transmission starts three subframes after the device will receive the scheduling grant as also illustrated in Figure 2-5. Moreover, the base station signals to the device the number of repetitions for the UL data transmission, based on the CE mode of the device (in contrast with the PRACH, where the repetitions were defined per CE *level*). The available numbers of repetition are defined in Tables 8.2b and 8.2c in [35] for CE mode A and B, respectively (see also Section 5.7.1). There is a repetition value configured for the whole CE mode and based on that value, base station can chose the most appropriate number of repetitions per device.

The PUSCH can use any resources that are not used for any other UL channels. However, for a given Cat-M1 device, the assigned PUSCH resources must fall within a single narrowband and can thus be up to six PRBs. In the example of the 5 MHz LTE carrier bandwidth shown in Figure 2-4, there are twelve PRBs available for the PUSCH and thus two PUSCH narrowbands can be configured. Typically, only one narrowband is configured for Cat-M1 usage to ensure that there are also PUSCH resources available for non-eMTC LTE traffic. It is also worth mentioning that this resource separation between LTE and Cat-M1 is dynamic which implies that unused LTE PUSCH resources can be used by Cat-M1 UEs if needed and vice versa.

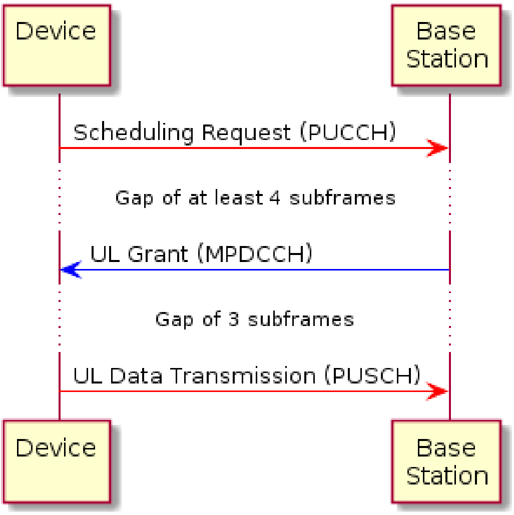

Device
Base Station
Scheduling Request (PUCCH)
Gap of at least 4 subframes
UL Grant (MPDCCH)
Gap of 3 subframes
UL Data Transmission (PUSCH)
Device
Base Station

Figure 2-5 - Signaling needed for an UL data transmission in Cat-M1.

The supported modulation schemes for Cat-M1 devices on the PUSCH are QPSK and 16-QAM, which implies a maximum Transport Block Size (TBS) of 1736 and 936 bits per narrowband (with 16-QAM) for CE mode A and B, respectively.

Finally, frequency hopping is mandatory, just like for the PUCCH and PRACH transmissions.

2.2.2 Downlink channels

Due to the shorter supported bandwidth in Cat-M1 (1.4 MHz), the DL physical channels are different than in the legacy LTE physical channels. Therefore, only three DL physical channels are available in Cat-M1 namely the MTC Physical Downlink Control CHannel (MPDCCH), the Physical Downlink Shared CHannel (PDSCH) and the Physical Broadcast CHannel (PBCH). The Primary Synchronization Signal (PSS), the Secondary Synchronization Signal (SSS) and the Reference Signal (RS) are fully reused by Cat-M1 devices [36]. The PBCH as well as the PSS/SSS and RS will not be discussed as they were not modelled in this study. However, it is noted that their role is considered as for example, the devices make use of the RSRP to define their CE mode/level. An overview of the MPDCCH and PDSCH is given below and their allocation on the downlink channel is illustrated in Figure 2-6.

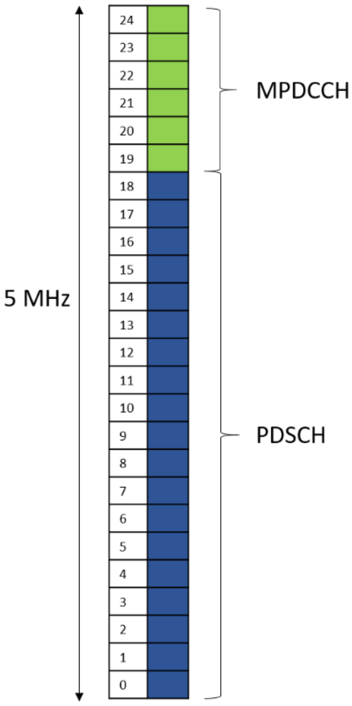


Figure 2-6 - Downlink channel resource allocation for a 5 MHz LTE

2.2.2.1 *MPDCCH*

MPDCCH is a new DL control channel introduced in Cat-M1 and it is meant to be used only by Cat-M1 devices while the non-eMTC devices continue to make use of the legacy PDCCH. This channel can carry different messages based on its Data Control Information (DCI) format. A high-level description of the different DCI formats is given in Table 2-4, while for a more extensive description of the fields of each format the reader is referred to Chapter 5 in [37]. It is important to note that there is no DCI format for power control regarding CE mode B as due to the poor coverage all devices with CE mode B transmit at full power. An MPDCCH is allocated in one narrowband (i.e. six PRBs are used) and it uses all OFDM symbols, as opposed to the legacy PDCCH which uses up to three OFDM symbols per subframe. Therefore, the MPDCCH and PDSCH cannot be multiplexed within the same PRBs and they are assigned in different narrowbands.

Table 2-4 - DCI format for MPDCCH.

DCI format	Description
3-3A	Power control for CE mode A
6-0A	Uplink grant for CE mode A
6-0B	Uplink grant for CE mode B
6-1A	Downlink scheduling for CE mode A
6-1B	Downlink scheduling for CE mode B
6-2	Paging

Each repetition of the MPDCCH uses an aggregation of one or several consecutive Enhanced Control Channel Elements (ECCEs), where each ECCE consists of multiple Enhanced Resource Element Groups (EREGs). The EREGs are used to map the channel to Resource Elements (REs) and there are 16 EREGs in total per PRB. The mapping to the REs can be either localized or distributed. The localized transmission is used when reliable subband Channel State Information (CSI) is available and the distributed transmission is used when the subband CSI is not reliable [38]. Moreover, it is configured that each ECCE will have four EREGs (based on Table 6.8A.1-1 in [34]) and therefore four ECCEs can be configured for each PRB and a total of 24 ECCEs in one narrowband. Depending on the type of transmission (localized or distributed), different MPDCCH formats can be defined as shown in Table 6.8B.1-2 in [34]. Each format indicates the number of ECCEs that consist the MPDCCH. For Cat-M1, the highest aggregation level of ECCEs is used which is 24.

A device should monitor a set of the MPDCCH candidates on one or more narrowbands such that it decodes any DL messages that are meant for it. These MPDCCH candidates are defining four search spaces namely type-0 Common Search Space (CSS), type-1 CSS, type-2 CSS and Device-specific Search Space. The candidates for each search space can be either on the same or different narrowband. All control messages, regarding the RA procedure, are transmitted in the type-2 CSS. The number of MPDCCHs to monitor in type-2 CSS is defined in Table 9.1.5.1b for CE mode A and Table 9.1.5.2b for CE mode B in [35]. This value depends on the aggregation level and the number of repetitions of the message in the MPDCCH. For Cat-M1, there is only one MPDCCH candidate regardless of the number of repetitions due to using an aggregation level of 24.

The repetitions used for the messages on the MPDCCH can be defined based on the CE level that the target device belongs to. This repetition value is pre-defined for each CE level but it is considered as a maximum value (r_{max}) since the base station can use a different value for each device within the same CE level. The base station uses r_{max} in combination with Table 9.1.5-3 in [35] and calculates four different repetition values $\{r_1, r_2, r_3, r_4\}$. From these four values, the base station selects the most suitable repetition value based on the signal strength of the device. Moreover, the starting subframe of the MPDCCH transmissions depends on the maximum number of the MPDCCH repetitions r_{max} and the factor G which is defined from higher layers. By using these two values, a period T can be defined as $T = r_{max} * G$ which indicates the starting subframe of the transmissions where the possible values for G are 1, 1.5, 2, 2.5, 4, 5, 8 and 10. Two examples showing this procedure are shown in Figure 2-7.

In the top part of Figure 2-7, the period T is equal to eight and thus a new cycle of MPDCCH messages can happen every eight subframes. In contrary, in the bottom part of the figure, there are four subframes that cannot carry any MPDCCH messages as the period T is equal to 12 and r_{max} is equal to eight. In both cases, messages that need to be transmitted one, two, four or eight times (due to repetitions) can be transmitted in subframes colored with light green, red, yellow or green respectively. It is important to note that only one MPDCCH transmission can happen in every subframe as there is only one MPDCCH candidate in Cat-M1. For example, in case that there is a transmission in subframe #0 targeting a device that needs one copy of the message, the next available subframe for transmission to devices that require two copies of the message, will be subframe #2 and not #1 (see also Figure 2-7).

Finally, frequency hopping is supported.

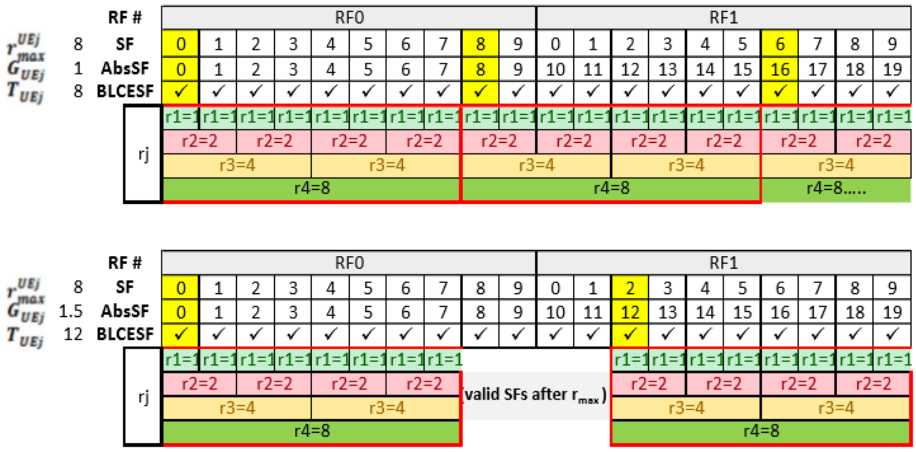


Figure 2-7 - Examples of valid starting subframes for the MPDCCH for r_{max}=8 in combination with G=1 (top) and G=1.5 (bottom) as presented in [39].

2.2.2.2 *PDSCH*

PDSCH is the channel carrying the actual DL data transmissions. Before a DL transmission on the PDSCH, there is always a message on the MPDCCH which indicates the narrowband of PDSCH, the number of used repetitions and the applied Modulation and Coding Scheme (MCS). The DL transmission always starts two subframes after the transmission of the message on the MPDCCH.

The number of repetitions for the DL data transmission is based on the CE mode of the device and the procedure is similar as in the PUSCH data transmission. The possible numbers of repetitions are defined in Tables 7.1.11-1 and 7.1.11-2 in [35] for CE mode A and B respectively. Like the PUSCH, there is a repetition value configured for the whole CE mode and based on that value, the base station selects the most appropriate number of repetitions per device.

The PDSCH for Cat-M1 devices is also defined in units of narrowbands and thus a DL transmission to one device can make use of a maximum of six PRBs. Furthermore, and similarly to the PUSCH, QPSK and 16-QAM are the only supported modulation schemes which lead to a maximum TBS equal to 1736 and 936 per six PRBs (with 16-QAM), for CE mode A and B respectively.

Finally, frequency hopping is supported.

2.3 Connection establishment for Cat-M1 device

In this study, all Cat-M1 devices are assumed to be in RRC IDLE mode and switch to RRC CONNECTED mode whenever they have to perform an UL data transmission. In such case, they will have to establish a connection with the base station and then transmit their data. After the transmission of the UL data, the device becomes inactive, which results in triggering the connection release procedure, typically after 10 seconds inactive time. The signaling involved it the connection establishment, transmission of UL data and connection release is shown in Figure 2-8.

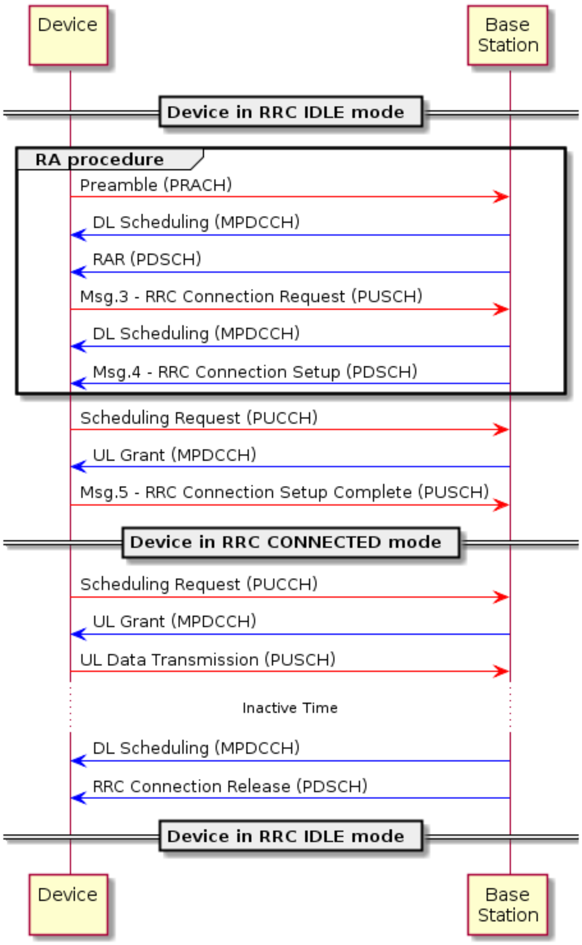

Device
Base Station
Device in RRC IDLE mode
RA procedure
Preamble (PRACH)
DL Scheduling (MPDCCH)
RAR (PDSCH)
Msg.3 - RRC Connection Request (PUSCH)
DL Scheduling (MPDCCH)
Msg.4 - RRC Connection Setup (PDSCH)
Scheduling Request (PUCCH)
UL Grant (MPDCCH)
Msg.5 - RRC Connection Setup Complete (PUSCH)
Device in RRC CONNECTED mode
Scheduling Request (PUCCH)
UL Grant (MPDCCH)
UL Data Transmission (PUSCH)
Inactive Time
DL Scheduling (MPDCCH)
RRC Connection Release (PDSCH)
Device in RRC IDLE mode
Device
Base Station

Figure 2-8 - Signaling for connection establishment, transmission of UL data and connection release.

First of all, the device initiates the RA procedure such that it will make the network aware that it wants to set up a connection. The signaling of the RA procedure is discussed in more detail in Section 2.4. After the completion of the RA procedure, the device can request dedicated UL resources with a 'Scheduling Request' (SR) sent on the PUCCH. In response, the base station sends a scheduling grant for the UL transmission through the MPDCCH, upon which the device transmits the 'RRC Connection Setup Complete' message on the PUSCH and finalizes its transition to the RRC CONNECTED mode.

While in RRC CONNECTED mode, the device again requests UL resources by following the exact same procedure as used for the transmission of the RRC Connection Setup Complete message (or Msg.5). However, this time it will transmit its UL data in the granted resources. In the case where only a part of the total data can be transmitted in the allocated resources, the same procedure should be repeated up until there is no more data in the buffer. This can happen in the case when the device is assigned a small TBS (due to poor coverage), in combination with the fact that the transmission be assigned no more than six PRBs. Typically 10 seconds (configurable value) after the completion of the data transmission, the base station will transmit the RRC Connection Release message which will finalize the transition of the device from RRC CONNECTED mode to RRC IDLE mode.

As a standardized addition to the above-mentioned procedure, Cat-M1 supports a feature of early data transmission for devices that need to transmit time-sensitive data. This feature basically allows devices to append their UL data (or part of them) to Msg.5, which finally leads to a lower end-to-end delay of UL data transmission. For this behavior, the device appends as much UL data as possible to Msg.5, such that the TBS is completely utilized, and then it sets a flag in the Radio Link Control (RLC) header of the message in order to indicate to the base station that there are also UL data within that specific message [40]. For scenarios that all UL data can fit in Msg.5, the end-to-end delay is reduced by at least 10 ms as the device will not have to undergo the procedure of requesting resources on PUSCH and wait for an UL grant after reaching the RRC CONNECTED mode (see also Figure 2-5 and Figure 2-8).

2.4 Random access procedure for Cat-M1 device

The signaling of the RA procedure (see also Figure 2-8) for Cat-M1 is similar to the RA procedure for 'regular' LTE with the difference that a different configuration of the RA resources and their usage can be applied for each CE level. The parameters that can be defined for each CE level and relate to the RA procedure are listed below [41].

- PRACH Configuration Index
- PRACH Frequency Offset
- PRACH Starting Subframe
- Preamble Range
- Number of Preamble Attempts
- Maximum Repetitions per Preamble
- MPDCCH Repetitions
- Random Access Response Window
- Contention Resolution Time

Figure 2-9 illustrates the signaling of the RA procedure in both LTE (right) and Cat-M1 (left). From this figure it is clear that the main difference between the two RA procedures is the introduction of two extra messages **(2)** and **(5)** on the new MPDCCH channel and the repetitions of messages (N1-N6). As a consequence, the RA delay in Cat-M1 is expected to be larger than in conventional LTE. Furthermore, in Figure 2-9, the values N1-N6 represent the number of repetitions per message and the values G1-G5 the time interval in terms of number of subframes. The N1 value refers to the repetitions of the preamble which are defined from the 'Maximum Repetitions per Preamble' parameter, the N2 and N5 values refer to the repetitions on the MPDCCH and are defined from the 'MPDCCH repetitions' parameter and the values N3, N4 and N6 are signaled in messages **(2)**, **(3)** and **(5)** respectively. The values of

G1-G5 are predefined or influenced from the processing delays that arise at the base station due to the network load. Both sets of parameters are defined in Table 2-5.

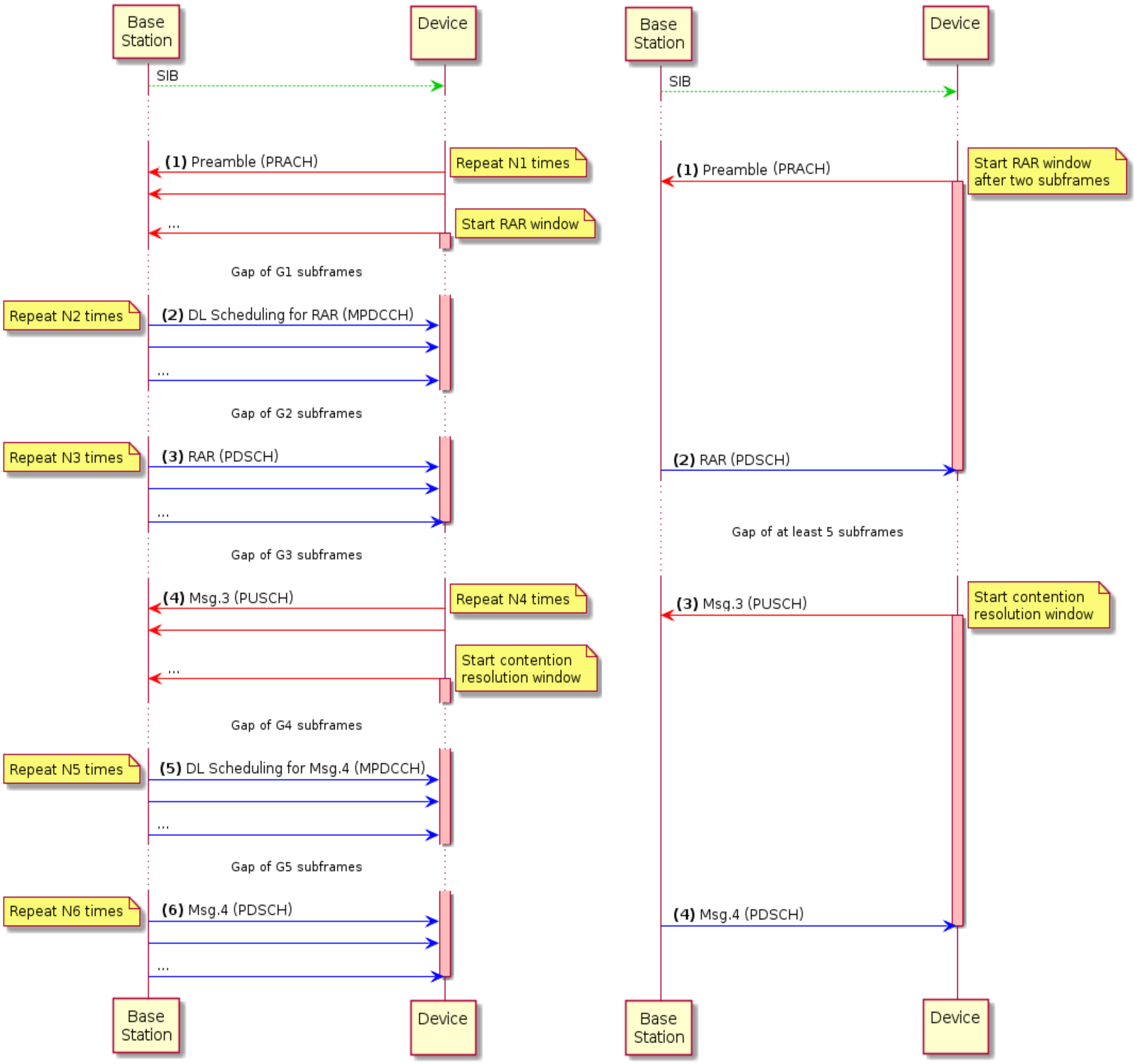


Figure 2-9 - Random Access Procedure in Cat-M1 (left) and LTE (right).

Table 2-5 - Definition of parameters shown in Figure 2-9.

Number of Repetitions		Gap in terms of number of subframes	
Value	**Description**	**Value**	**Description**
N1	Repetition of preamble is signaled in SIB per CE level	G1	Defined from the MPDCCH starting subframe which is influenced from the periodicity T
N2	See Section 2.2.1.1	G2	1
N3	See Section 2.2.2.2	G3	6 + (Δ), where Δ is defined in the RAR
N4	Defined in the RAR	G4	Defined from the MPDCCH starting subframe which is influenced from the periodicity T
N5	See Section 2.2.1.1	G5	1
N6	See Section 2.2.2.2	-	-

As already discussed in Section 2.2.1.2, for a device to initiate the RA procedure, it must randomly select a PRACH preamble value within the range of preambles allocated to its CE level and transmit it in repetition. The number of repetitions (N1) depends on the device's CE level and is signaled for each CE level in the SIB. Each preamble transmission **(1)** is taking place in consecutive PRACH subframes. After the preamble transmission, the device starts the Random Access Response (RAR) window and waits to receive the MPDCCH messages **(2)** which will indicate where and when exactly the device will receive an answer (RAR **(3)**) from the base station.

However, there is a possibility that the base station does not receive the transmitted preamble, in which case it does not generate any response. This happens when the SINR of the preamble is less than the receiver's sensitivity and it can be caused in the following two scenarios:

1. The base station is not aware of the existence of the PRACH preamble transmission as the transmit power used by the device for the preamble is not high enough or the

channel between the device and the base station at that particular time experiences high propagation losses which significantly degrade the transmitted signal.

2. There are at least two devices which transmitted the same preamble and due to the interference that is caused, the base station does not receive the preamble of neither devices and therefore it does not generate a reply. This is called a collision of preambles. It is important to note that when the same preamble is used by multiple devices but only one is received by the base station, no collision of preamble will occur. The base station then transmits a response which however will be received by all the devices that used that preamble (since the reply is linked to the preamble) and not to a specific device. This, will then lead to a collision at the next UL transmission (Msg.3) of the devices. The consequence from such occurrences will be discussed later on.

If there is no RAR generated (and hence also no 'DL Scheduling for RAR' message in Figure 2-9), the RAR time window at the device will expire without the reception of the RAR. Therefore, the device will then have to re-initiate the RA procedure using a new preamble. This fresh RA attempt will be made only after waiting a random back-off time and with an increased transmit power compared to the previous attempt. The back-off time is sampled from a uniform distribution between zero and the back-off parameter value as signaled in the SIB. The transmit power is increased based on a 'ramping step' parameter which is also signaled in the SIB.

After the reception of the preamble, the base station needs to transmit the RAR. At first, it schedules a message on the MPDCCH (**(2)**; DCI format 6-1A/B based on Table 2-3) which gives information to the device about the RAR message which will follow in the PDSCH. The information in **(2)** includes the number of MPDCCH repetitions such that the device can calculate in which subframe the transmission on the PDSCH will start. Additionally, the MCS and number of repetitions to be used for the RAR on the PDSCH is included. Two subframes after the transmission of the message on the MPDCCH, the base station starts the transmission of the RAR **(3)** on the PDSCH. In essence, the RAR conveys the uplink grant to the device for its Msg.3 transmission, including an indication of the number of repetitions of Msg.3, the MCS and the resource allocation [35].

Upon reception of the RAR, the device transmits Msg.3 **(4)** on the PUSCH as instructed in the RAR. The first subframe that will be used to transmit these messages is defined by the parameter 'UL delay' in the RAR. If the 'UL delay' is set to 0 then Msg.3 will start to be transmitted at least after 6 subframes. In case that it is set to 1 then the transmission will start at least 6 + Δ subframes later, where Δ is set to the number of Msg.3 repetitions. After the device transmits all the repetitions on the PUSCH, it starts the Contention Resolution Timer and tries to decode the MPDCCH messages **(5)**, which will signal the control information for receiving Msg.4 **(6)** on the PDSCH. The procedure followed at the base station to transmit the MPDCCH messages and Msg.4 on the PDSCH is like the one used for the RAR.

However, and as already mentioned, a collision of Msg.3 messages from different devices may occur. Such case implies that two or more devices have transmitted their Msg.3 in the exact same UL resources as they had used the same PRACH preamble and consequently received the same RAR. When at least two of these Msg.3 transmissions are received at the base station, a collision happens as the base station does not know to whom it should reply. If only one of the Msg.3 transmissions is received at the base station, then the procedure continues normally just for that one device. All devices that do not receive the subsequently expected MPDCCH messages before the expiry of the Contention Resolution Timer, re-initiate the RA procedure with a new preamble, applying a randomly sampled back-off time and an increased transmit power compared to the previous attempt, as described before.

Finally, there is a maximum on the number of access attempts that a device can have and this number of attempts is also signalled in the SIB. In case a device fails to establish a connection, after using the maximum number of access attempts, it will configure itself to a higher CE level and initiate a new RA procedure with the new CE level configuration. Such a downgrade to a higher CE level (corresponding with weaker coverage) can subsequently happen if the RA procedure still fails until the device reaches CE level 3, which is the highest one. If there is no success in any of the access attempts in CE level 3 then the device is considered to be in outage and it does not initiate any further access attempts.

Chapter 3 Random Access (RA) procedure in 5G networks

In this chapter, the considered enhancements on the reference CAT-M1 RA procedure as well as the new two-step RA procedure are discussed. First, in Section 3.1, the Early Data Transmission (EDT) procedure is discussed as it is already standardized by 3GPP for Release 15, along with some further enhancements that can be applied. Then, the new two-step RA procedure will be discussed in Section 3.2. Finally, Section 3.3 describes the application of the two-step RA procedure in the scenario where mini-slots are supported. It is noted that the enhancements presented to EDT as well as the specific implementation of the two-step RA procedure are contributions of this study.

Figure 3-1 presents the signaling from the initiation of the RA procedure by the device (transmission of the preamble on the PRACH) until the UL data transmission for Cat-M1 (left), EDT (center) and two-step RA procedure (right). From Figure 3-1 it is clear that the signaling between the preamble transmission and the data transmission is decreased for the EDT (compared to the reference RA procedure used in Cat-M1) and eliminated for the two-step RA procedure. Consequently it can be expected that EDT will provide lower end-to-end delays compared to the RA procedure used in Cat-M1 and the two-step RA procedure will provide further reduction of the end-to-end delay, making it the best RA procedure among the three regarding the end-to-end delay.
It is noted that the procedures described and the estimated delay values, in this chapter, are assuming no repetitions of messages for the reasons explained later on in Section 4.5.

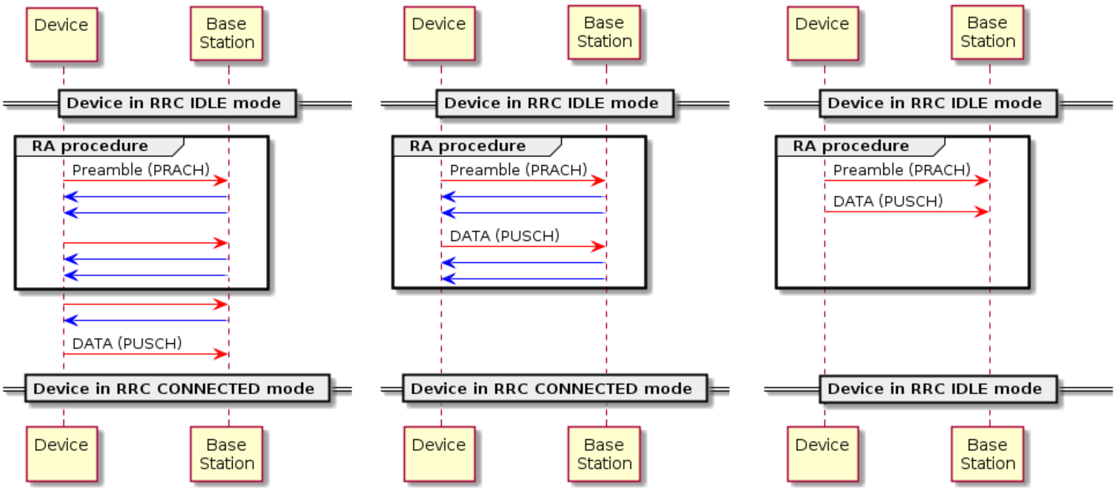


Figure 3-1 - Signaling from the initiation of the RA procedure until the UL data transmission for Cat-M1 (left), EDT (center) and two-step RA procedure (right).

3.1 Early Data Transmission (EDT)

The Early Data Transmission (EDT) is a new procedure that has been introduced by 3GPP in Release 15 [42] and aims the reduction of the end-to-end delay as the devices can transmit their data during the RA procedure (i.e. not necessarily going into RRC CONNECTED mode), under some restrictions. Section 3.1.1 discusses in detail this new procedure and then

Sections 3.1.2 and 3.1.3 discuss further enhancements that can be combined with EDT to further improve the end-to-end delay.

3.1.1 EDT based on 3GPP Release 15

The need of shorter end-to-end delays has driven the development of the EDT procedure which allows devices to transmit their data while undergoing the RA procedure. This is illustrated in Figure 3-2, in comparison with the reference RA procedure as standardized for CAT-M1. In the reference RA procedure the earliest possibility that a device could transmit its data via the PUSCH is in Msg.5 appended to the RRC Connection Setup Complete message (see also Section 2.4). However, with the EDT procedure the first occasion for UL data transmission is moved earlier i.e. with the transmission of the RRC Connection Setup Request message (or Msg.3).

In order to enable EDT according to the latest 3GPP standard, the base station broadcasts a new information to the devices, namely the *edt-TBS*. In case the UL data size of a device is lower than what the *edt-TBS* specifies, the device can initiate the RA procedure with EDT. Otherwise it initiates the conventional RA procedure (as in Cat-M1).

Furthermore, a device which initiates the RA procedure with EDT, uses a preamble from a group of reserved preambles (formed from a subset of the 64 available preambles in Cat-M1) such that the base station will be able to recognize an EDT attempt. Upon the reception of such an EDT-specific preamble, the base station responds with the Random Access Response (RAR) message which now includes an UL grant for a size of *edt-TBS*, even if the device may actually need less UL resources. The device then appends its data to Msg.3 and transmits the aggregated data in the assigned UL resources. Finally, the base station ends the RA procedure by replying to the device with the RRC Connection Setup message (or Msg.4), which now indicates to the device that it should stay in RRC IDLE mode.[3] It is noted that a failure to receive the Msg.4 forces the device to re-initiate the RA procedure as it implies that the base station did not receive successfully the data.

If the EDT procedure is compared with the conventional/reference RA procedure for Cat-M1, as presented in Figure 3-2, it is clear that the end-to-end delay can now be reduced by at least 23 ms (see Section 2.4 for the timings in RA procedure).

For the configuration of EDT, two parameters are crucial; the *edt-TBS* parameter and the number of reserved preambles for EDT. The trade-offs associated with the configuration of these parameters are explained below:

1. *edt-TBS:* As already mentioned in Section 2.2.1.3, the maximum number of PRBs that can be assigned to a device for an UL transmission is six. Therefore, the maximum value of *edt-TBS*, assuming six PRBs and the use of MCS 7 (which is the maximum allowed MCS for Msg.3), is 89 bytes (712 bits), see Table 4-3. However, the EDT procedure is expected to be used by devices which typically need to transmit even shorter messages, e.g. 40 bytes as assumed in this study (see Section 4.2). The transmission of data smaller than 89 bytes is effectively a waste of UL resources as the

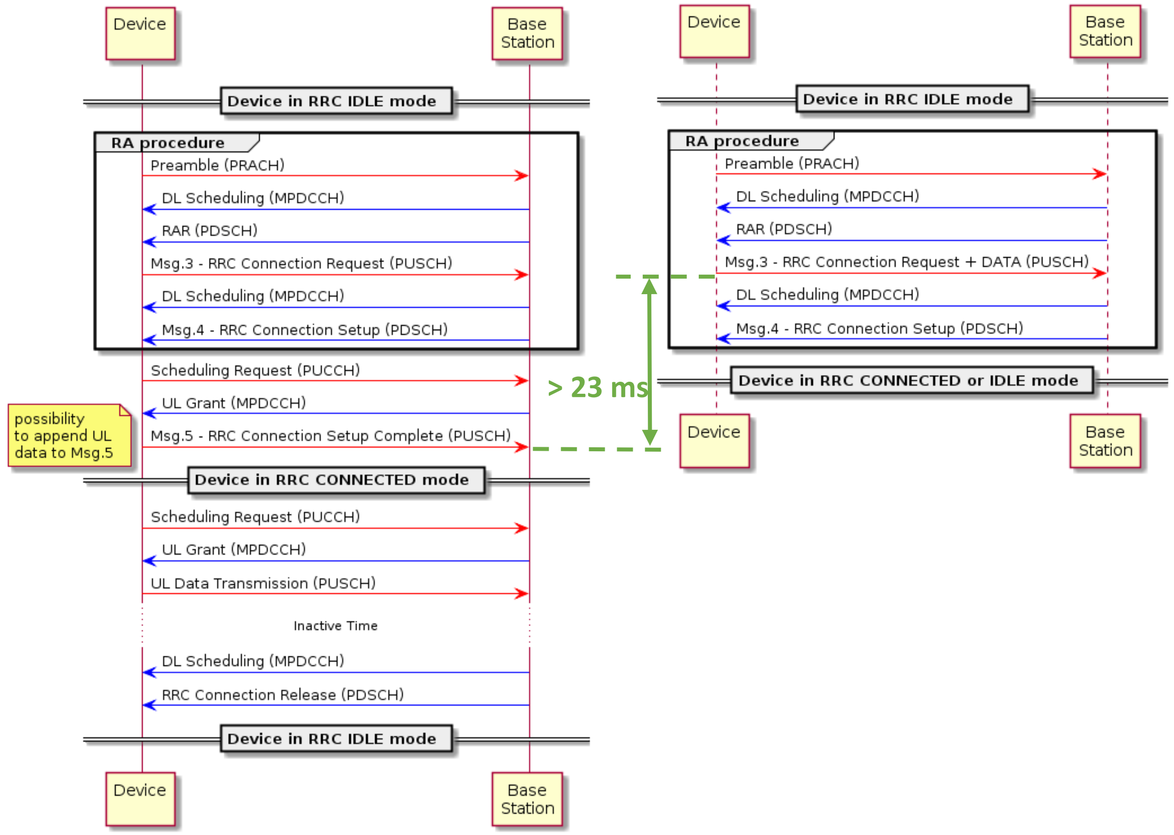
[3] In 3GPP Release 15, it is standardized that the base station can also instruct the device to switch to RRC CONNECTED mode.

devices will not have to make use of all six PRBs that are assigned to them. On the other hand, choosing a smaller *edt-TBS* can lead to under-performance of the procedure in terms of the delay enhancement, as some devices with larger messages may then in fact not qualify for the EDT procedure and hence not benefit from the potential delay improvement. For this study, it is chosen that the maximum TBS, hence 89 bytes, is used for the *edt-TBS* value, as for lower values, i.e. values corresponding to two or three PRBs, none of the devices of the studied scenario would be able to make use of the EDT procedure. This is because the size of Msg.3 of the RA procedure is six bytes and the smallest UL data used in the studied scenario is 40 bytes. With a total of 46 bytes to be transmitted in one TBS, six PRBs are needed (for MCS 7).

2. *Number of EDT-specific preambles:* The devices which intend to use the EDT procedure, use preambles from a subset of the total 64 preambles. The overall set of 64 preambles should thus be split into two groups; one for EDT transmissions and one for non-EDT transmissions. Reserving too few preambles for EDT can cause a high number of preamble collisions which will then lead to high delays or even outages[4] if the device reaches the maximum allowable number of access attempts. However, reserving too many preambles for EDT can have the described effect on the devices that do *not* use the EDT procedure. Therefore, in order to determine the optimal split of preambles, a sensitivity analysis is required (see Section 5.3).

Based on the above, for this study, it is expected that the URLLC devices will make use of the EDT procedure, as their UL data size indeed satisfies the limiting imposed by *edt-TBS*, while the non-URLLC devices will use the reference CAT-M1 RA procedure. It is also expected that the end-to-end delay for the EDT procedure, and consequently for the URLLC data, will be above the 10 ms requirement with 99.99% reliability for URLLC traffic. More specifically, an average delay of 2.5 ms is expected as waiting time for a preamble opportunity on the PRACH, considering that the PRACH is supported in two subframes in a radio frame of 10 ms (see Section 4.5.2). Additionally, on average and for devices experiencing no preamble collision, a delay of at least 14 ms is expected from the time of the preamble transmission until the successful reception of the data at the base station (see also Section 2.4). Thus, a total of an average 16.5 ms is expected for the end-to-end delay for URLLC devices that experience no preamble collisions. Therefore, the 10 ms requirement with 99.99% reliability can definitely not be met.

[4] Recall from Section 2.4 that a device is considered in outage when it does not succeed to establish a connection with the base station with the pre-defined number of tries.



Device
Base Station
Device in RRC IDLE mode
RA procedure
Preamble (PRACH)
DL Scheduling (MPDCCH)
RAR (PDSCH)
Msg.3 - RRC Connection Request (PUSCH)
DL Scheduling (MPDCCH)
Msg.4 - RRC Connection Setup (PDSCH)
Scheduling Request (PUCCH)
UL Grant (MPDCCH)
Msg.5 - RRC Connection Setup Complete (PUSCH)
possibility to append UL data to Msg.5
Device in RRC CONNECTED mode
Scheduling Request (PUCCH)
UL Grant (MPDCCH)
UL Data Transmission (PUSCH)
Inactive Time
DL Scheduling (MPDCCH)
RRC Connection Release (PDSCH)
Device in RRC IDLE mode
Device
Base Station
> 23 ms
Device
Base Station
Device in RRC IDLE mode
RA procedure
Preamble (PRACH)
DL Scheduling (MPDCCH)
RAR (PDSCH)
Msg.3 - RRC Connection Request + DATA (PUSCH)
DL Scheduling (MPDCCH)
Msg.4 - RRC Connection Setup (PDSCH)
Device in RRC CONNECTED or IDLE mode
Device
Base Station

Figure 3-2 - The conventional Cat-M1 RA procedure (left) in comparison with the EDT procedure (right).

3.1.2 EDT with priority and shorter windows

As already mentioned, on average, the end-to-end delay for the EDT procedure is expected to be 16.5 ms and thus the 10 ms requirement with 99.99% reliability cannot be met. Therefore, further enhancements are needed in order to improve this delay.

The first improvement that is studied, is the application of priority to devices following the EDT procedure. For those devices, the base station is scheduling its reply on the MPDCCH and PDSCH, for the RAR message and Msg.4, with priority, over devices which use the conventional RA procedure. Furthermore, the base station is granting access to the UL resources, on the PUSCH, again with priority to the devices using EDT over devices which use the conventional RA procedure. The priority given to devices which use the EDT procedure, introduces a trade-off for the end-to-end delay of non-URLLC and URLLC devices. This is because the EDT procedure is expected to be used by URLLC devices and thus the priority will benefit their end-to-end delay at the cost of the end-to-end delay of non-URLLC devices.

Because of the applied priority in scheduling of DL messages to the devices using EDT, it is expected that the RAR message and Msg.4 will be received earlier than for scenarios without priority. Therefore, the RAR window and the contention resolution window for the devices that use EDT can be configured with lower values. With this configuration, the devices which are experiencing preamble collisions, are able to recognize that they have to re-initiate the RA procedure earlier (due to the shorter windows) than previously. For this study, it was decided that both windows are set to half the values of the conventional/reference procedure, resulting in 10 ms and 40 ms for the RAR and contention resolution windows, respectively. However, the reduction of the RAR and contention resolution windows has a trade-off with the number of preamble retransmissions. For scenarios where very short windows and high network loads apply, a device might re-initiate the RA procedure while there was not an actual preamble collision and the base station was delaying its reply as it was busy handling other traffic in the network.

With the priority enhancement, it is expected than the end-to-end delay will be improved for devices using the EDT procedure, hence devices handling URLLC traffic. Moreover, due to the faster reactions upon preamble collisions, the end-to-end delay will be reduced significantly for devices that have to retransmit the preamble while using the EDT procedure. For example, reducing the RAR window from 20 ms (for the reference CAT-M1 RA procedure) to 10 ms gives a 10 ms delay gain per retransmission. However, it is still not expected that the end-to-end delay for URLLC traffic will drop below 10 ms with 99.99% reliability as devices that experience even one preamble collision will have an end-to-end delay higher than 10 ms, since just the value of the RAR window is already 10 ms.

3.1.3 EDT for all devices

The EDT procedures presented in Sections 3.1.1 and 3.1.2 are expected to be used by devices handling URLLC traffic. However, the delay gain that is experienced by URLLC devices will be at the cost of the delay of the non-URLLC devices, which are using the conventional RA procedure, as the URLLC devices have priority on both UL and DL resources. From the requirements derived in Section 1.2, it is clear that non-URLLC devices also have to meet an end-to-end delay requirement of 50 ms with 99.9% reliability. Therefore, further improvements on the EDT-based RA procedure should be considered such that the delay for non-URLLC devices should also be decreased.

As described in Section 3.1.1, the maximum *edt-TBS* that can be defined is 89 bytes (712 bits) due to the restriction of using a maximum MCS 7 for Msg.3. The restriction of MCS 7 was applied due to the importance of Msg.3, in the RA procedure, and thus more robust MCSs are preferred for this particular message. However, as a further EDT improvement, it is decided to relax the restriction of MCS and allow transmissions with up to MCS 15, which is the maximum allowable MCS for data transmissions in Cat-M1. Consequently, the maximum *edt-TBS* that can now be applied is 217 bytes (1736 bits) for six PRBs and MCS 15. The new and larger *edt-TBS* that is now chosen enables also some non-URLLC devices (the ones which can use MCS 15) to make use of the EDT procedure as non-URLLC devices need to transmit a total of 206 bytes (Msg.3 is 6 bytes and non-URLLC UL data are 200 bytes in this study).

Following the same line of thought, an additional enhancement can be applied such that all devices in the network will profit from the EDT procedure. The basic idea is to allow all devices transmit their UL data along with Msg.3, during the RA procedure, and transmit also a *buffer bit* which is indicating whether the device still has more UL data to transmit in its buffer. In case that the buffer bit indicates that more UL data needs to follow, the base station will signal to the device in Msg.4 to switch to RRC CONNECTED mode and thus the device can transmit (part of) the rest of its data during the transmission of Msg.5. If still not all data can be transmitted during the Msg.5 transmission, the device will transmit the remaining data from the RRC CONNECTED mode. In the alternative case that the buffer bit indicates that there is no more UL data to follow, the base station will signal through Msg.4 that the device will stay in RRC IDLE mode.

It is stressed that the priority enhancement and the shorter RAR and contention resolution windows are still applicable for URLLC devices such that their end-to-end delay will not be influenced by the new behavior of the non-URLLC devices.

For the implementation of the above-mentioned procedure, it is implied that there is no longer a need for advertising an *edt-TBS* as both non-URLLC and URLLC devices will make use of the EDT procedure. Therefore, the base station always includes in the RAR message an allocation of six PRBs on the PUSCH for the Msg.3 transmissions. Then, based on the MCS indicated in the RAR message, devices append to their Msg.3 as much UL data as possible such that they will make full use of the six PRBs that they are assigned.

A similar argument for the potential waste of PUSCH resources as noted above in Section 3.1.1, applies for this variation of EDT. This is because allowing for maximum MCS for Msg.3 from 7 to 15, the URLLC devices now might require fewer PRBs for their UL data transmission but they are assigned six PRBs anyway. Additionally, allowing non-URLLC devices transmitting (part of) their UL data during the Msg.3 transmission can lead to improvement of the UL resource utilization, as these devices will need fewer accesses on the PUCCH and PUSCH later on. This UL resource utilization improvement can be significant when there is a significant number of non-URLLC devices that can transmit all their UL data along Msg.3. From Figure 3-2, it is clear that a device which can transmit its full data along Msg.3, does not have to make use of the PUCCH for any scheduling requests and also Msg.5 no longer has to be transmitted (device will be in RRC IDLE mode). It is expected that the UL resource utilization of the network will be influenced mostly by the behavior of the non-URLLC devices as there are significantly more of them than of the URLLC devices.

Furthermore, improvement of the end-to-end delay is expected for the non-URLLC devices. Considering that there is a significant number of non-URLLC devices that can transmit all their UL data within Msg.3, the average end-to-end delay for non-URLLC devices will be comparable with the end-to-end delay for URLLC devices. It is noted that non-URLLC devices use longer RAR windows than URLLC devices and thus non-URLLC devices that experience preamble collisions, will have different end-to-end delays than the corresponding URLLC devices.

3.2 Two-step RA procedure

The two-step RA procedure, like EDT procedure, aims to decrease the end-to-end delay by enabling the transmission of UL data during the RA procedure. It is however noted that in the two-step RA procedure, the goal is to transmit the UL data quickly and not to establish a connection with the base station as this approach is meant to be used by devices handling infrequent URLLC data. Section 3.2.1 presents the basic two-step RA procedure and Section 3.2.2 presents an enhancement applying immediate repetitions of the UL data which may further improve the end-to-end delay performance. Finally, Section 3.2.3 presents another enhancement of the basic two-step RA procedure which makes use of feedback from the base station such that the end-to-end delay is improved further.

3.2.1 Basic two-step RA procedure

The basic two-step RA procedure aims to transmit the URLLC UL data as soon as possible after the transmission of the preamble, without waiting for the RAR message from the base station. Then, after the URLLC UL data transmission, the device remains in RRC IDLE mode. This behavior is illustrated in Figure 3-3. In general, the base station advertises an *edt-TBS* such that devices can decide whether they can make use of the two-step RA procedure (same concept as in Section 3.1.1). However, the two-step RA procedure is meant only for URLLC devices and thus the advertised *edt-TBS* should be set accordingly. More specifically, an *edt-TBS* of 69 bytes (552 bits) is a suitable choice as it is the smallest TBS that can transport 40 bytes (320 bits) within two PRBs when using MCS 15.

The RAR message indicates to the devices when they should perform their Msg.3 transmission on the PUSCH and which PRBs should be used for this transmission. As the RAR message is absent in the basic two-step RA procedure this creates two problems:

1. The URLLC devices do not know *when* they should perform their UL data transmission.
2. The URLLC devices do not know in which PRBs (*where*) on the PUSCH they should perform their UL data transmission.

In order to overcome the timing ('*when*') problem for the UL data transmissions, a pre-defined and fixed time value is assumed for the UL data transmissions such that the URLLC devices and the base station are time-synchronized. However, the issue at hand here is how long this pre-defined interval should be, expressed in number of subframes (SF; see Figure 3-3). In the regular RA procedure for Cat-M1, the minimum processing time that the base station needs in order to transmit the RAR message is 3 ms. However, during those 3 ms, the base station makes calculations about parameters (e.g. which narrowband on the PUSCH should be used by the device for its Msg.3 transmission) that need to be included in the RAR message. For the two-step RA procedure, the base station only needs to recognize a URLLC data transmission based on the value of the preamble and thus it is considered reasonable to assume that this processing time is a bit shorter, assumed at just 2 ms. Consequently, it is assumed that the UL data transmission from a URLLC device could be performed at least 2 ms after the transmission of the preamble.

Upon reception of a preamble from a URLLC device the base station transmits a flag on the MPDCCH during SF4 (see Figure 3-3), which indicates to all non-URLLC devices to back-off from utilizing PUSCH resources, which may have been allocated via previous UL grants, in the next subframe SF5. With this approach the base station ensures that all PUSCH resources will be available for use by URLLC devices[5], 3 ms after the transmission of the URLLC preamble. Consequently, the pre-defined time period for where URLLC devices are pre-configured to transmit their UL data on the PUSCH is decided to be 3 ms after their preamble transmission in SF5, as illustrated in Figure 3-3.

[5] In a 5 MHz bandwidth, a maximum of three narrowbands (18 PRBs) can be used for the PUSCH, for subframes without PRACH support.

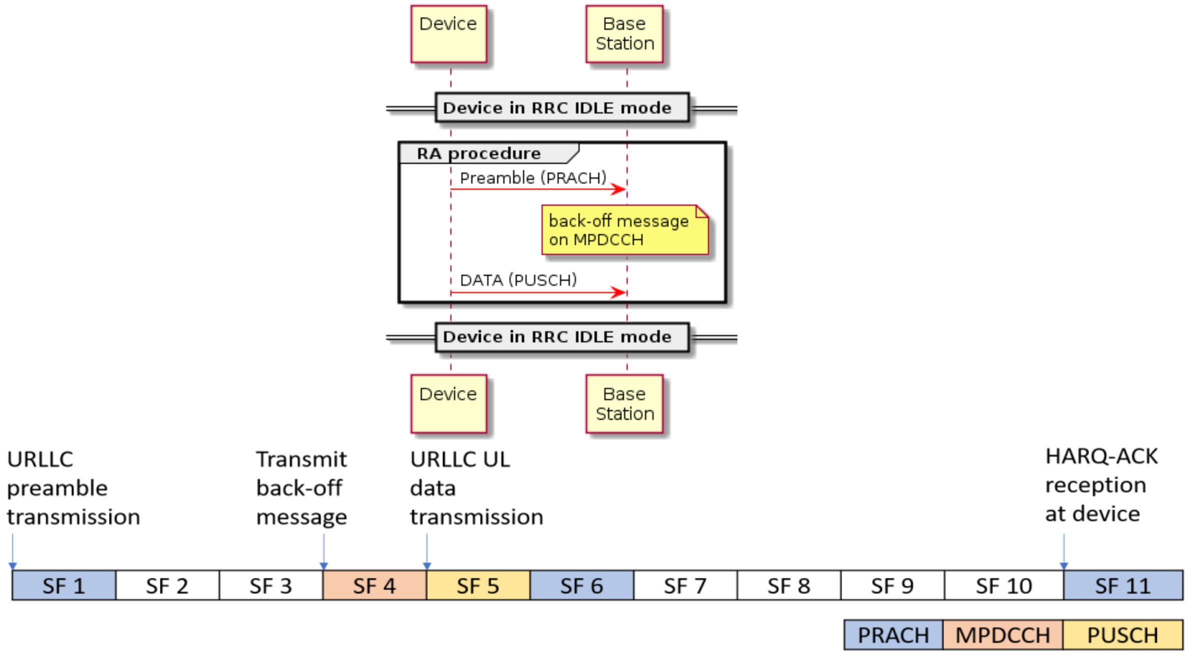


Figure 3-3 - Signaling and timeline for the two-step RA procedure, assuming two PRACH occasions (SF1 and SF6) within one 10 ms radio frame.

URLLC devices already know from *edt-TBS* that they can make use of two PRBs on the PUSCH and that with MCS 15 they will be able to transmit 69 bytes and thus all their UL data. However, *edt-TBS* only defines the *number* of PRBs (i.e. two PRBs) that are available for each URLLC device but not the mapping of those PRBs on the PUSCH (the '*where*' issue). Consequently, URLLC devices that choose to transmit their UL data on overlapping set of PRBs will experience collisions on data which results in unsuccessful data transmissions and an effective waste of UL resources. For this reason, it is very important that there is a matching of URLLC devices to different non-overlapping sets of PRBs. This matching can be signaled from the base station but a more realistic and less complex solution will be to introduce a matching of preambles to the non-overlapping sets of PRBs, which is derived by the base station and broadcasted in the SIB.

Figure 3-4 illustrates a potential matching of preambles to the UL resources for subframes that the PRACH is not supported and thus there are eighteen available PRBs for URLLC UL data transmissions. For the example case that nine preambles are reserved for URLLC transmissions, each preamble can be matched to a different set of PRBs (see left side of Figure 3-4) and thus devices which use different preambles will use different non-overlapping PRBs for their UL data transmission. For example, a device which used preamble #1 will transmit its UL data in PRBs 19 and 20, while a device which used preamble #8 will transmit its UL data in PRBs 5 and 6. However, for scenarios with more than nine preambles reserved for URLLC transmissions (e.g. fifteen reserved preambles; see right side of Figure 3-4), more than one preamble will match to a set of PRBs which can potentially lead to collisions on the UL data. For example, a device which transmitted preamble #1 and another device which transmitted preamble #10, will not experience a preamble collision as they are using different preamble values but will experience collision on the UL data as they will both transmit their UL data on PRBs 19 and 20.

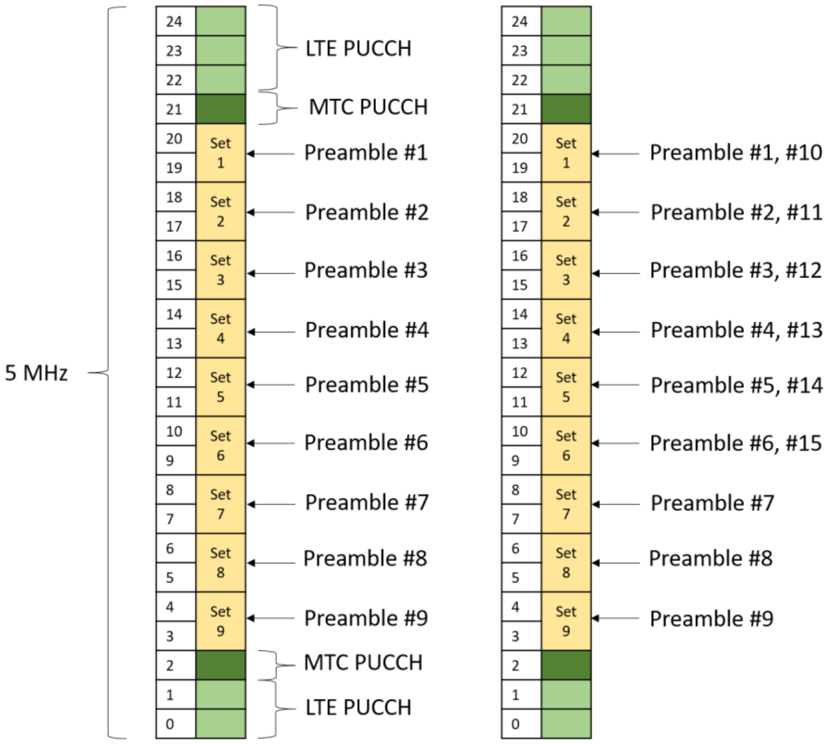


Figure 3-4 - Possible matchings of 9 (left) and 15 (right) preambles to non-overlapping PRB sets on the PUSCH in a 5 MHz carrier for URLLC transmissions in the two-step RA procedure.

After the transmission of their UL data on the PUSCH, URLLC devices start a timer of 5 ms[6] and start listening for an acknowledgement from the base station which will signal the correct reception of their data. In case the timer expires without receiving an acknowledgement, the devices have to restart the procedure. The two possible causes of forcing a device to restart the procedure are the collision of the preamble on the PRACH and the collision of the UL data on the PUSCH.

It is noted that for this specific method, the effects of the uncertain Time Alignment (TA) have not been taken into account as devices are considered to be fixed in a specific location. In general, the TA is a time offset which ensures that the UL data transmissions from different devices arrive approximately at the same time at the base station. The TA value is calculated at the base station and transmitted to the device in the RAR message, something that does not happen with the two-step RA procedure as there is no RAR message. The need of having a TA arises from the different propagation delays that devices have as they have different distances to the base station (see [43] for more details on TA). Therefore, for this study, it is considered that the TA is already known to the devices as due to their fixed location, their TA will be always the same.

With the two-step RA procedure, it is expected that the end-to-end delay for URLLC devices which experience no collisions, will be below 10 ms. This is because a URLLC device needs to wait for a PRACH opportunity for 2.5 ms on average, when two PRACH occasions occur per 10 ms. Then, 3 ms after the preamble transmission it can transmit its UL data and thus a total of

[6] This corresponds to the RAR window which is reduced even further to 5 ms for even faster reactions to collisions.

7.5 ms end-to-end delay is expected on average. Furthermore, the end-to-end delay for some non-URLLC devices will be increased by 1 ms compared to the EDT procedures due to the back-off time that they are forced to apply whenever there is a URLLC data transmission.

Note here that in the two-step RA procedure there are two kinds of collisions that can happen, namely a preamble collision or a data collision, and therefore the number of retransmissions in the network could be higher than in the EDT procedures. The higher number of retransmissions is also expected to increase the resource utilization of the UL channel compared to the EDT procedures as more access attempts may need to be carried out. Moreover, the reservation of all three narrowbands on the PUSCH, each time that a URLLC preamble is detected by the base station, also leads to an increased resource utilization of the UL channel. Essentially, we are trading a reduced resource efficiency for an enhanced delay performance for URLLC data, considering that the effective waste of UL resources may be bearable given the relatively low number of URLLC devices.

As already explained, the collisions during the UL data transmission can influence the performance of URLLC devices with respect to end-to-end delay significantly. Therefore, further enhancements of this basic two-step RA procedure are needed such that the probability of the UL data being received correctly at the base station is increased. Two enhancements, one with repetitions of the UL data and one with feedback from the base station, are explained in Sections 3.2.2 and 3.2.3 respectively and thus there is no need to numerically assess the basic two-step RA procedure presented in this section.

3.2.2 Two-step RA procedure with repetitions

As discussed above, in the two-step RA procedure there is a need to match the URLLC preambles to PRBs for the URLLC UL data transmissions. As a consequence, if URLLC devices use URLLC-specific preambles that match to the same PRB set of UL resources, then these URLLC devices will experience collisions on their UL data, and will have to re-initiate the procedure after the expiry of the ACK timer. In order to increase the probability of correct reception of the UL data at the base station, UL data repetitions could be applied as illustrated in Figure 3-5. It is noted that during SF6 in Figure 3-5 the PRACH is supported but there are also UL data repetitions happening on the PUSCH.

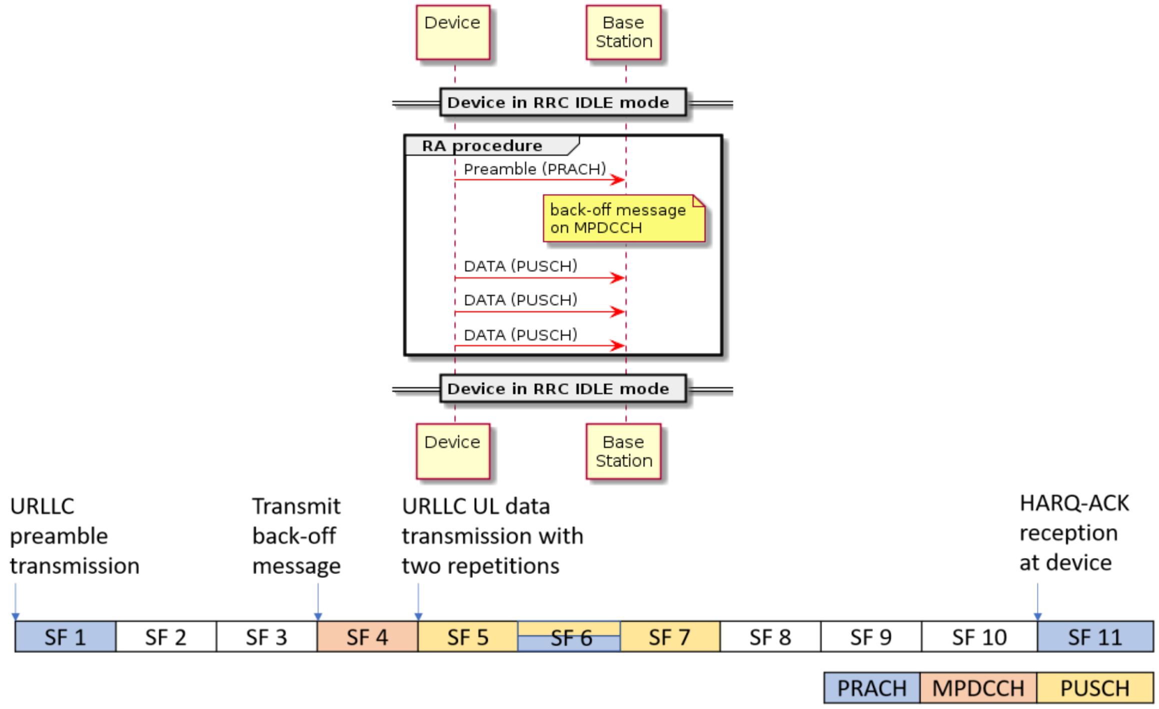


Figure 3-5 - Signaling and timeline for the two-step RA procedure with two repetitions of the UL data, assuming two PRACH occasions (SF1 and SF6) within one 10 ms radio frame.

The main idea of this procedure is to match every URLLC preamble to a different set of UL resources at each transmission of UL data. With this way, the probability of two or more preambles, that are currently in use, match consistently to overlapping PRBs, will be reduced. Figure 3-6 illustrates an example of a matching of preambles to UL resources for two consecutive subframes. From this figure, it can be understood that the probability of a successful transmission of UL data for a device which uses preamble #1 - #3 or #10 - #15 is increased. Furthermore, devices which use preamble #4 - #9 are guaranteed that they will have a successful data transmission. In general, the probability of successful data transmission is correlated with the number of repetitions used, as an increase of the number of repetitions (and thus an increase of different matching combinations among preambles and PRBs) will reduce the probability of experiencing a data collision in every data transmission. It is noted that a data transmission is considered successful if the base station correctly receives the data at least once.

A detailed example (based on Figure 3-6) illustrating the above-mentioned behavior, for scenarios where two UL data transmissions (one fresh attempt, one immediate repetition) are configured, is presented below:

> *Three devices (D1, D2 and D3) are transmitting their UL data in the same subframe with preamble values #1, #10 and #13, respectively. Hence the preambles do not collide. During the first data transmission, however, D1 and D2 experience a collision as they are matched to the same set of PRBs (Set 1), while D3 has a successful data transmission. During the second transmission, preambles #1 and #13 are matched to the same UL resources (Set 1) while preamble #10 is matched to different resources (Set 7) and thus D2 will now experience a successful data transmission. Overall, the base station will receive correctly the UL data of D2 and D3 after these two consecutive transmissions while D1 will have to re-initiate the procedure, after the expiry of a timer.*

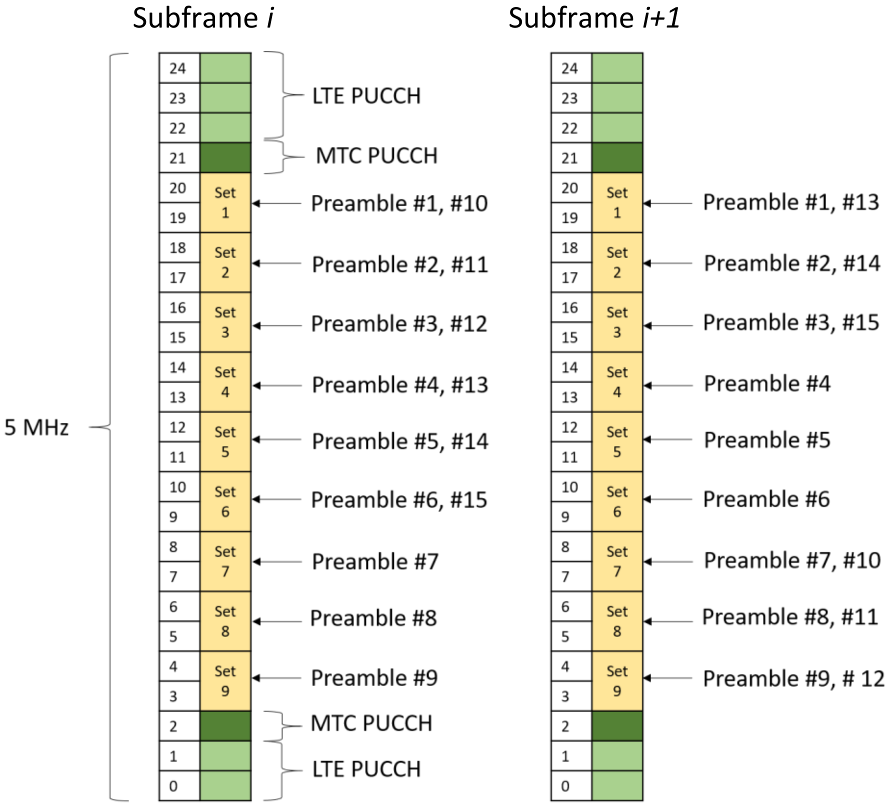


Figure 3-6 - Example of matching 15 preambles to UL resources while using two consecutive data transmissions.

From the above example, it is made clear that the number of transmissions of the UL data and the way that the preambles are matched to the UL resources is very important. Specifically, and in reference to the above example, in a setup where three transmissions would be applied, another matching of preambles to PRBs would have to be configured and the probability of collision would be reduced even further, providing D1 with a better chance to transmit its data successfully. Therefore, a sensitivity analysis should be carried out in order to define the number of UL data transmissions needed such that the probability of unsuccessful reception of the UL data to the base station due to data collisions will be reduced significantly, or ideally set to zero (see Section 5.5.2).

With this method, the probability that the UL data will be received correctly at the base station is increased. Of course, this comes with the cost of repetitions (each introducing 1 ms delay) and higher resource utilization of the UL channel as more traffic is generated in the network due to the multiple transmissions of the UL data. Moreover, the average end-to-end delay will be increased compared to the basic two-step RA procedure (without repetitions) but at the same time the number of attempts that a device will need in order to successfully transmit its data will be decreased. This decrease of number of attempts will strongly influence, positively, the higher percentiles of end-to-end delay, which is the focus of this study. Therefore, the improved delay performance for high percentiles of URLLC traffic comes at the cost of a reduced resource efficiency. It is also noted that the end-to-end delay for some non-URLLC devices (the ones transmitting their data during the period when URLLC transmissions also occur) will increase linearly with the number of URLLC data transmissions, as they will have to be back-off from the UL resources for more than 1 ms.

3.2.3 Two-step RA procedure with feedback

Increasing the probability of correctly receiving the URLLC UL data at the base station can also be achieved with the two-step RA procedure with feedback, as an alternative to the two-step RA procedure with repetitions discussed above. The goal of this procedure is to actually prevent collisions of the URLLC UL data transmissions such that the correct reception of the data at the base station will be achieved without the need of repetitions.

A way of implementing this procedure is for the base station to signal to URLLC devices, on the MPDCCH, a mapping of the used preambles to the corresponding sets of (two) PRBs that are to be used for the UL data transmission. As already mentioned in Section 3.2.1, nine such sets of two PRBs can be derived in the UL channel for URLLC data transmissions and thus nine devices can be transmitting at the same time their UL data in different PRB sets. In order for the base station to signal this mapping, at most nine pairs of preamble and PRB sets are needed and thus nine signaling messages of 10 bits[7] each have to be transmitted. This gives a total of 11.25 bytes (90 bits) to be transmitted on the MPDCCH. In order for this method to be applied, it is therefore required that 90 bits can indeed be transmitted on the MPDCCH along with the back-off message that is transmitted to non-URLLC devices, two subframes after the transmission of a URLLC preamble. In this study, the structure of the message to be transmitted on the MPDCCH was not studied and therefore for the rest of the study it will be assumed that it is indeed feasible to do this message transmission. Figure 3-7 presents the signaling and timeline of the two-step RA procedure.

Compared to the two-step RA procedure with repetitions, this procedure decreases both the UL resource utilization and the end-to-end delay for URLLC devices as their data are received successfully with one transmission. However, the downlink resource utilization is increased due to the extra bits transmitted on the MPDCCH and the maximum number of URLLC devices that is supported is nine as there are only nine sets of non-overlapping PRBs on the PUSCH per subframe for a 5MHz carrier. It is noted that for scenarios where more than nine URLLC devices transmit their preamble during the same PRACH subframe, some of them will not ‘read’ their preamble on the feedback message from the base station and therefore they will have to re-initiate the RA procedure.

[7] 6 bits are needed to signal a given choice from 64 preambles and 4 bits to signal a given PRB set from nine possible such PRB sets.

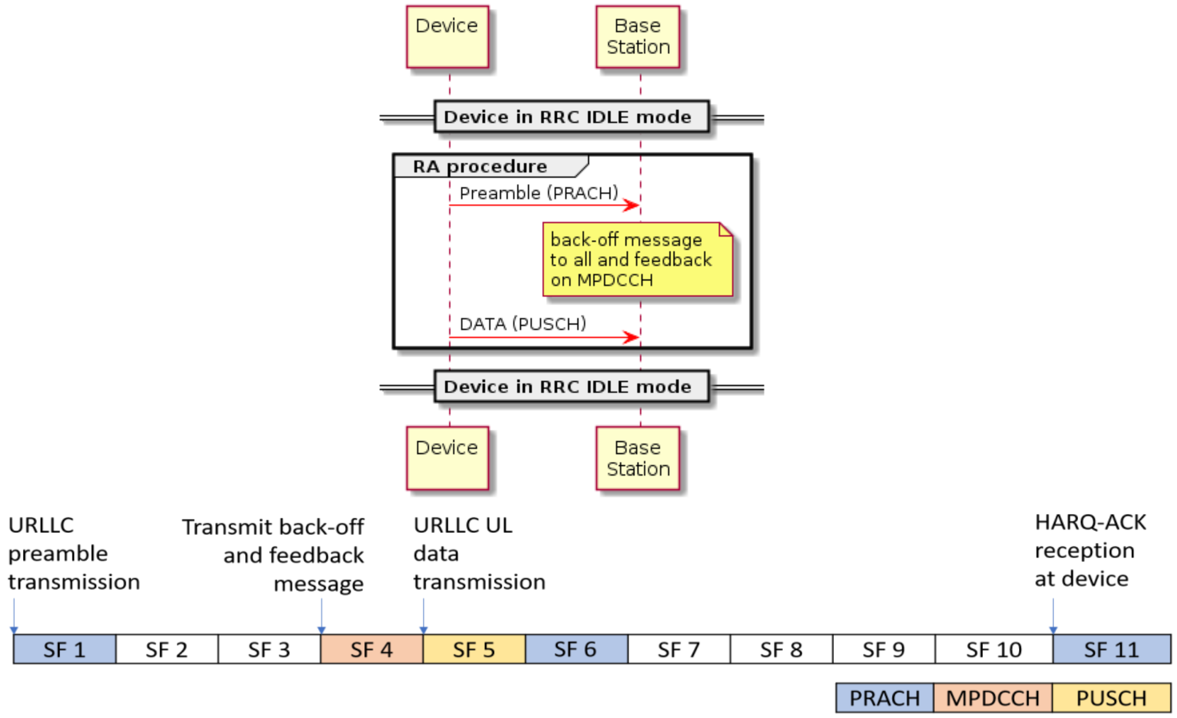


Figure 3-7 - Signaling and timeline for the two-step RA procedure with feedback, assuming two PRACH occasions (SF1 and SF6) within one 10 ms radio frame.

3.3 Mini-slots

The 5G New Radio (NR) physical layer will be supporting a new radio frame structure, which will have the option of having shorter subframes than those used in LTE [44]. For this study, it is considered that the subframes used so far can be reduced to half but at the same time the resources in the frequency domain for each PRB will have to be doubled in order to achieve the same resources (in term of Resource Elements (REs)) as with the normal PRBs. Figure 3-8 presents one LTE subframe which consists of 14 OFDM symbols and has a total duration of 1 ms and the respective PRB consists of 12 subcarriers with 15 kHz spacing. Additionally the new subframe used for 5G NR is also presented with the difference that the duration of the 14 OFDM symbols is half of that of the OFDM symbols in LTE, while the spacing of the 12 subcarriers of the corresponding PRB is now 30 kHz instead of 15 kHz. With this design, the amount of data that can be transmitted in 180 kHz in 1 ms in LTE can also be transmitted in 360 kHz in 0.5 ms in 5G NR. The support of these new mini-slots however offers the possibility to reduce the transmission delays involved in the RA procedures described and therefore a short analysis can indicate the influence of mini-slots on the end-to-end delay. This influence will be investigated only for the scenario of the two-step RA procedure as it is the procedure which provides the shorter end-to-end delays.

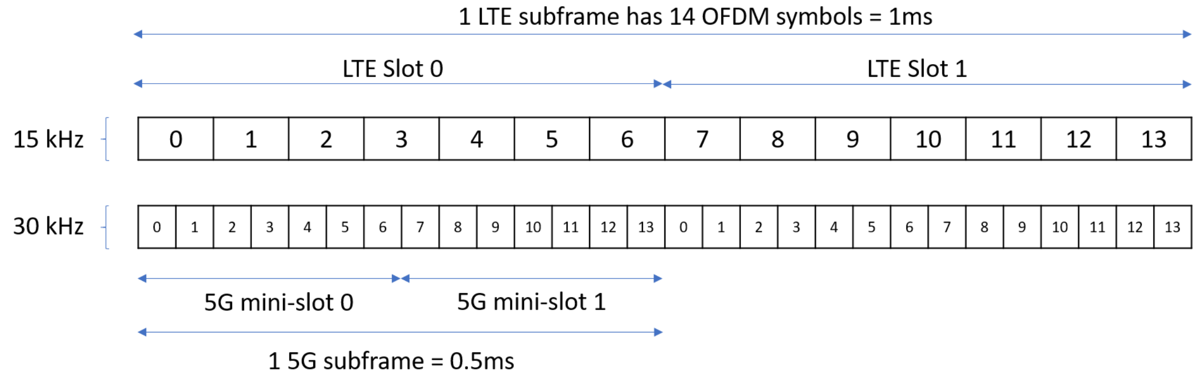
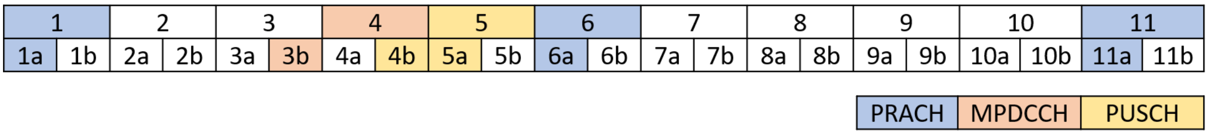


Figure 3-8 - Subframes and slots in LTE and 5G physical layer.

Figure 3-9 shows the messages exchanged in the two-step RA procedure with feedback in a time-line, for the scenarios of using the regular LTE subframes (top) and the new 5G subframes based on mini-slots (bottom). As already explained in Section 3.2.1, devices transmitting a URLLC preamble can transmit after 3 ms their UL data. Therefore, for the scenario of regular LTE subframes (and based on Figure 3-9), a device which transmits a URLLC preamble in SF #1, can transmit its UL data in SF #5. For the scenario where the 5G subframes are applied, a device can still transmit its preamble in SF #1a but its UL data transmission will only be possible six subframes later, in SF #4b. The increase from three (in LTE) to six (in 5G NR) subframes between the preamble and the UL data transmission is based on two factors:

1. The base station requires 2 ms processing time in order to identify that a URLLC preamble was transmitted. This 2 ms processing time translates to two regular LTE subframes and four 5G subframes.
2. After the identification of a URLLC preamble, the base station transmits a flag on the MPDCCH such that non-URLLC devices will back-off from the resources on the PUSCH. For this action it is considered that 0.5 ms is needed for the transmission of the flag and another 0.5 ms is needed as processing time at the non-URLLC devices. Therefore, this translates to one regular LTE subframe and to two 5G subframes.



1		2		3		4		5		6		7		8		9		10		11	
1a	1b	2a	2b	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b	8a	8b	9a	9b	10a	10b	11a	11b

PRACH	MPDCCH	PUSCH

Figure 3-9 - Time-line of two-step RA procedure with feedback using the regular LTE subframes (top) and the new 5G subframes based on mini-slots (bottom).

Furthermore, the transmission time of the UL data on the PUSCH cannot be reduced with the use of mini-slots. This behavior is relevant to the two-step RA procedure as the procedure requires the reservation of the full frequency resources for 1 ms (see Section 3.2.1). Thus, also in the scenario with 5G subframes, 1 ms is required for the UL data transmission.

Finally, for the scenarios where either a preamble or a data collision occurs, devices have to wait until the expiry of a 5 ms timer before they initiate a fresh try. This translates to five regular LTE subframes and to ten 5G subframes. Therefore, for both regular and 5G subframes, a device will start its fresh try on the PRACH at the same time (SF #11 for regular LTE and SF #11a for 5G in Figure 3-9).

From all the above, it is clear that there will only be 0.5 ms gain, on average, for the end-to-end delay for scenarios where 5G subframes are supported. Thus, applying the two-step RA procedure on 5G subframes provides limited gains, in terms of end-to-end delay, due to the high processing times that are needed throughout the procedure. It is also noted that the resource utilization of the UL channel is expected to be the same for both regular LTE and 5G subframes as the UL resources, in terms of REs, and the number of retransmissions in the network are not changing. Therefore, it can be concluded that the use of mini-slots will not be beneficial for the goals of this study and no numerical evaluation is carried out. It is also noted that these estimations were based on the assumption that the same physical layer modeling (e.g. transport block size) can be applied in both LTE and 5G NR.

Chapter 4 Simulation models

In this chapter, the simulation modeling is discussed including the assumptions and simplifications that are made. First of all, the details about the devices and the network topology in the considered Factory of the Future (FoF) are presented in Section 4.1. Subsequently, the traffic models used and the assumed propagation environment of the considered factory are presented in Section 4.2 and Section 4.3, respectively. The Radio Resource Management (RRM) mechanisms for the data transmissions are presented in Section 4.4. Then, Section 4.5 gives an overview on how Cat-M1 system was configured in the study. The chapter is concluded in Section 4.6 with an overview of the simulation flow.

4.1 Network topology

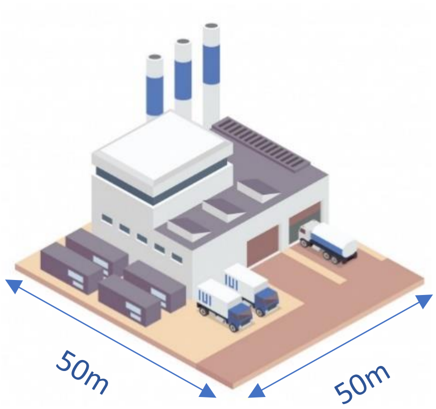


Figure 4-1 - Factory layout [58].

The presented study assumes a general layout that can be applied in many factories. The size of the factory, as illustrated in Figure 4-1, is chosen to be 50 m × 50 m and its height is 10 m which is a typical value for factories. It is important to note that the factory, regardless of its height, is assumed consist of only a ground floor.

The base station is located indoor in the center of the factory, mounted at the factory ceiling, as that can provide a better coverage than placing the base station outdoor. Furthermore, Table 4-1 shows the maximum transmit power, antenna gain, noise figure and height of the base station and the devices.

Table 4-1 - Details about the base station and devices in the network as also used in [45].

Parameter	Base Station	Device
Maximum Transmit Power	36.8 dBm	23 dBm
Maximum Antenna Gain[8]	2 dBi	0 dBi
Noise Figure	3 dB	5 dB
Height	10 m	1.5 m

Inside the factory, two types of devices are considered: devices handling URLLC traffic and devices handling non-URLLC traffic (see also Section 1.1 and 1.2). Both types of devices are uniformly distributed in the area of the factory at a height of 1.5 meter. All of these devices have the characteristics as given in Table 4-1.

An example layout of the factory, i.e. the deployed base station and the devices is shown in Figure 4-2. For this specific example, 6000 devices were used of which 95% are assumed to generate non-URLLC traffic and 5% generates URLLC traffic. The base station (eNB) is also shown in the center of the factory.

[8] For both base station and devices, omnidirectional antennas are considered

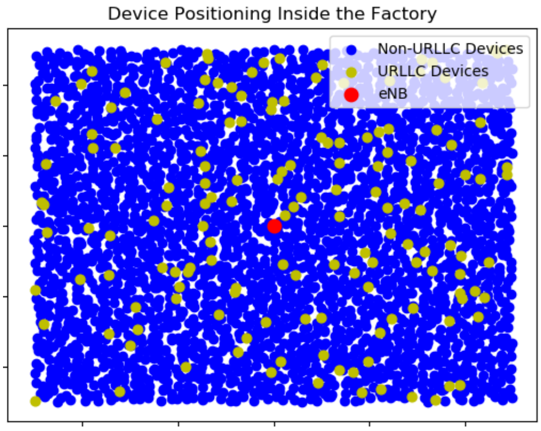


Figure 4-2 - One possible realization of placing 6000 devices (95% Non-URLLC and 5% URLLC) uniformly distributed in the factory area while keeping the base station (eNB) in the center of the factory.

4.2 Traffic models

As already mentioned, devices that generate URLLC and non-URLLC traffic are considered in the factory. Correspondingly, two distinct traffic models for the FoF are defined as shown in Table 4-2 (also see Section 1.2). This modelling is based on Table 5.3.8.1-1 and Table 5.3.8.1-2 in [46] and Table 6.1.1 in [47].

Table 4-2 - Traffic models.

Traffic Model	Number of Devices	Device Density	Arrival Distribution	Message Size
non-URLLC	95% of total devices	1.14 devices per m^2	Uniform over 60 seconds	200 bytes
URLLC	5% of total devices	0.06 devices per m^2	Beta over 10 seconds with α=3 and β=4	40 bytes

Most of the total traffic (95%) is expected to be of the non-URLLC type as it concerns regular reporting by sensors while only a few devices (5%) are assumed responsible for transmitting time-sensitive (URLLC) data as these devices will be triggered by an unexpected event. This study is carried out with the presence of a total number of 3000 devices in the factory such that the overall device density is 1.2 devices per m^2. It is noted that a sensitivity analysis on the network load is also carried out (see Section 5.9) in order to study the effects of lower and higher device densities.

Devices handling non-URLLC data are expected to be generating traffic once per minute with a randomized relative timing and therefore they were modelled to follow a uniform distribution over 60 seconds and the message size for each transmission is fixed to 200 bytes. For example, each device will first choose, in a uniform manner, a time-interval between 0 and 60 seconds to transmit its first 200 bytes and will then transmit new data with a periodicity equal to 60 seconds. On the other hand, the devices handling URLLC traffic will initiate a transmission based on an event and their message size is fixed to just 40 bytes (e.g. conveying critical information and/or alarms). These critical events can randomly occur and they are assumed to trigger all URLLC devices in the factory. The times of these transmissions are modelled to follow a beta distribution [47] over a period of 10 seconds with shape parameters α=3 and β=4 and the Probability Distribution Function (PDF) of this distribution is presented in Equation (4-1) and graphically in Figure 4-3:

$$\mathrm{PDF} = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \text{ where } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \text{ and } \Gamma(\cdot) \text{ is the gamma function} \qquad (4\text{-}1)$$

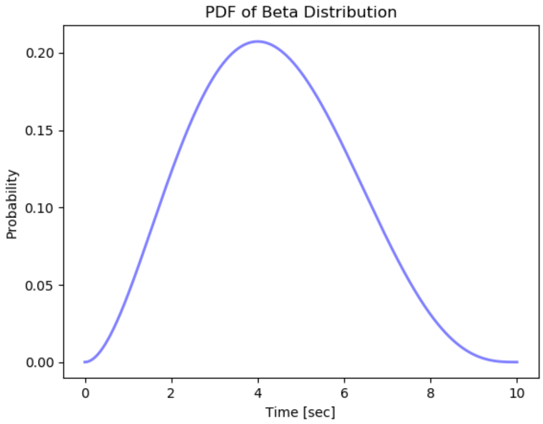


Figure 4-3 - PDF of Beta distribution with shape parameters α=3 and β=4 and scale parameter 10.

Naturally, the Human-to-Human (H2H) traffic can also be served on the frequency carrier (if resources are available). However, for the simplification of the simulator, it is assumed that all the traffic handled in the network is traffic from inside the factory and only from Cat-M1 devices, i.e. no H2H traffic is modelled. This traffic-related assumption makes it indeed plausible that the factory has deployed an indoor base station and uses a dedicated LTE carrier to handle the Machine-to-Machine (M2M) traffic as generated by/for the processes in the factory. This assumption was made because the main goal of this study is to quantify and minimize the end-to-end delay for machine type communications (MTC). Sharing the same physical time-frequency resources between the Cat-M1 and H2H traffic can potentially influence the results significantly, but an investigation of this effect is beyond the scope of our study.

4.3 Propagation environment

The propagation environment in a factory is based on the materials used in the factory as different kinds of materials can have different properties. For example, in a factory which uses many metallic machines and objects, multiple reflections of signals can be created which can degrade significantly or even entirely block the transmitted signal.

In this study a factory with reflective environment is studied and the modeling of such an environment is captured in the *path loss,* the *shadowing* and the *multipath fading* values. Based on [48], the Path Loss (PL) at a reference distance of 15 m ($PL(d_0)$) is 63.57 dB and the PL exponent (n) is 3.26. The parameters $PL(d_0)$ and n are used in Equation (4-2) in order to calculate the PL value of a device based on the distance d between the device and the base station, expressed in meters. Moreover, and based on [48], shadowing is modelled as normal fading (in dB) with the deviation value (σ) equal to 8.46 dB. It is noted that the multipath fading is not modelled in this study due to simplifications and the 2400 MHz band is considered.

$$PL(d)[dB] = PL(d_0)[dB] + 10nlog\left(\frac{d}{d_0}\right)[dB], where\ d_0 = 15m \qquad (4\text{-}2)$$

However, Equation (4-2) can only be used for distances larger than the reference distance d₀ as for shorter distances the model is not accurate. More specifically, 72% of the devices have distances larger than the reference distance d₀ and therefore for 28% of the devices Equation (4-2) does not apply. The blue curve in Figure 4-4 depicts Equation (4-2). In order to determine a suitable path loss model for the shorter distances for which Equation (4-2) is unsuitable, consider for instance an example distance of 1 m between the transmitter and receiver. Considering that the most favorable path loss for a reference distance in free space is given by Equation (4-3), where λ is the wavelength in meters, the actual path loss at a distance d = 1 m, should be at least 40.07 dB (on a 2400 MHz carrier; note that this value is also depicted in the figure, as a starting point for a correction to the path loss curve given by Equation (4-2)).

$$PL(d_0)[dB] = 20\log({}^{4\pi d_0}/_{\lambda}) \qquad (4\text{-}3)$$

Hence the path loss equation for distances smaller than 15 m should indeed be reconsidered. Figure 4-4, shows a linear and a logarithmic curve that can approximately describe the path loss for short distances. Between the two curves, the logarithmic one is chosen for modeling as it is more pessimistic than the linear one and hence leads to more conservative evaluations. Equation (4-4) is the final equation used for path loss (for reflective environment at 2400 MHz).

$$PL(d)[dB] = \begin{cases} PL(d_0)[dB] + 10nlog\left(\frac{d}{d_0}\right)[dB],\ d \geq 15m \\ 40.07[dB] +\ 19.98\,log(d)[dB],\ d < 15m \end{cases} \qquad (4\text{-}4)$$

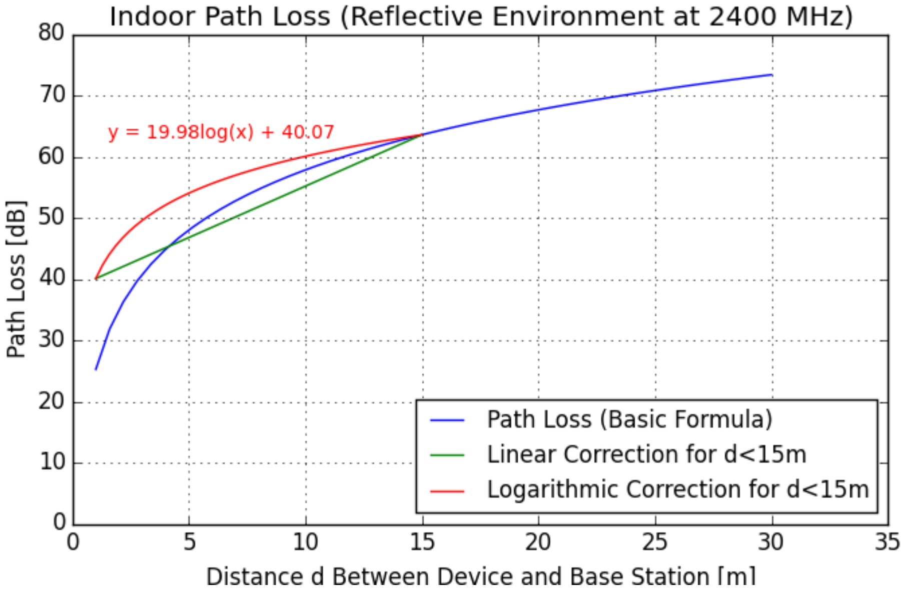


Figure 4-4 - Path loss values for reflective environment at 2400 MHz.

Another aspect characterizing the propagation environment is *multipath fading*. Multipath fading affects both static and mobile devices within the factory due to other objects in the factory such as machines, employees and robots. The multipath fading induces variability in the radio channel quality across the frequency range of the LTE carrier, although this variability is somewhat limited across the up to six PRBs utilized for Cat-M1 data transmission. Additionally, it causes for variability in the time domain for static devices due to other moving objects and for mobile devices due to both their own mobility and that of other moving objects in the factory.

As mentioned in Section 2.2, frequency hopping among different narrowbands is mandatory for PUCCH, PRACH and PUSCH transmissions or optional to the MPDCCH and PDSCH transmissions for Cat-M1 communication. It is expected that the use of frequency hopping, will eliminate the effect of multipath fading to some reasonable extend by introducing frequency diversity and 'averaging' the variability of the radio channel quality in both frequency- and time-domain for Cat-M1 communication. Consequently, to simplify the radio channel modelling in the simulation model it has been decided to model an average channel by taking into account only the path loss and shadowing components. This also makes the explicit modelling and implementation of frequency hopping for the downlink and uplink Cat-M1 communication channels unnecessary, as the effects of frequency hopping are *implicitly* modelled. This study only focuses on fixed devices and frequency hopping Cat-M1 communication system and thus it is considered justified to avoid modelling the multipath fading effects as these effects would have minimal influence on the different CE modes and levels used in Cat-M1 communications and thus the end-to-end delay which is the main Key Performance Indicator (KPI) of interest in this study.

4.4 Radio resource management mechanisms

The two main Radio Resource Management (RRM) mechanisms that are implemented are uplink (UL) *power control* and *scheduling*. Both mechanisms can influence the end-to-end delay and therefore it is important to describe these mechanisms in more detail.

4.4.1 Uplink power control

In the process of transmitting UL data while the device is in RRC IDLE Mode, the device undergoes two different settings for its UL power. The first setting is applied for the transmission of the preamble, in order to initiate the RA procedure, and is kept the same until the end of the RA procedure. The second setting is applied for the transmission of the UL data on the PUSCH after the completion of the RA procedure.

The setting of the UL transmit power for the preamble transmission in the RA procedure is derived from Equation (4-5) for devices that are transmitting in CE level 0, 1 and 2, while the maximum transmit power is used for devices transmitting in CE level 3 [49]. This UL transmit power is calculated as an aggregate value over six PRBs (one narrowband) as the preamble transmission utilizes all six PRBs. For devices that experience collisions or very poor coverage and thus need a new RA attempt, an increase of transmit power is applied per attempt. This increase of power is defined from upper layers and it is denoted as a 'power ramping' step. Furthermore, the preamble initial received target power that is shown in Equation (4-5), is also set by upper layers, while the preamble transmit counter is indicating how many RA attempts were made. In case repetitions of the preamble are required, the same transmit power will be used for all repetitions. Note from Equation (4-5), as specified in [49], that the preamble transmit power is trying to achieve the Preamble Received Target Power (*PRTP*) by compensating for the experienced channel losses while upper bounded by the maximum transmit power of the device.

$$PreambleReceivedTargetPower(PRTP)\ [dB] = \\ preambleInitialReceivedTargetPower[dB] + (preambleTransmitCounter - 1) * \\ powerRampingStep[dB] - 10\log(numRepetitionsPerPreambleAttempt)[\text{dB}]$$

$$P_{tx}[dB] = \min(P_{tx,max}[dB], PRTP[dB] + PathLoss[dB]) \tag{4-5}$$

Once the device completes the RA procedure (see Figure 2-8), it determines the transmit power of its data transmissions on the PUSCH. This transmit power is calculated per PRB as the transmissions on the PUSCH might not require the use of all six PRBs. Devices configured in CE Mode A calculate their transmit power per PRB based on their path loss and make use of this transmit power if it is lower than the maximum allowable transmit power per PRB. For devices that are configured in CE Mode B, the maximum transmit power per PRB is used. The definition of this UL transmit power per PRB is given by Equation (4-6) [50].

$$P_{tx}^{PRB}[W] = \begin{cases} \min(\frac{P_{tx,max}[W]}{6}, 10^{\frac{-80+0.8*PathLoss[dB]}{10}}), & CE\ mode\ A \\ \frac{P_{tx,max}[W]}{6}, & CE\ mode\ B \end{cases} \tag{4-6}$$

4.4.2 Scheduling

The channels that are modelled in detail and make use of scheduling schemes are the PUCCH, PUSCH and MPDCCH. In the downlink (DL), the PDSCH is not modelled in detail. This is driven by the assumption that the PDSCH resources are not limiting in carrying Cat-M1 traffic and thus the delays will not be affected by any congestion on the PDSCH, given that a dedicated 5 MHz LTE carrier is assumed to be deployed to handle the Cat-M1 traffic. This assumption in turn relies on the assumption that no conventional H2H LTE traffic is handled in the network and therefore the whole 5 MHz (25 PRBs) will indeed be available for the PDSCH and MPDCCH. It is reminded that the MPDCCH utilizes only six PRBs from the whole DL bandwidth and allows PDSCH to utilize the remaining three narrowbands (eighteen PRBs in total).

The scheduling schemes define the order in which devices will gain access to the resources as well as the number of PRBs that they should be assigned for their UL data transmissions on the PUSCH. For example the scheduler is the one to decide whether to first serve a request from a URLLC device over a request from a non-URLLC device. For each one of the above-mentioned channels, a description of the scheduling scheme used is provided below. It is important to note that in Cat-M1, which is used as the reference scenario, there is no distinction between the URLLC and non-URLLC devices, in any channel, and thus all devices are treated in the same way.

4.4.2.1 *PUCCH*

Refer back to Figure 2-8 to understand in detail all the messages that are exchanged between the device and the base station. From this figure it is clear that the PUCCH is only used for the Scheduling Requests (SRs). Moreover, as mentioned in Section 2.2.1.1, one PRB per TTI is available for the SRs in PUCCH and thus all devices that need to transmit an SR have to compete for that one PRB. However, no congestion is expected on the PUCCH as it is used after the RA procedure. Therefore, a simple First-Come First-Served (FCFS) approach is used for the scheduling.

4.4.2.2 *PUSCH*

The scheduling of PUSCH resources among the different devices is also done according to a FCFS scheduling discipline for simplicity reasons. However, as multiple PRBs are available in the PUSCH, multiple devices can transmit their data in the same Transmission Time Interval (TTI). The number of devices that can transmit simultaneously depends on the number of PRBs that each device needs for its transmission.[9] Moreover, the PRBs assigned to a device should be allocated within the same narrowband. It is noted that the scheduler always tries first to assign resources on narrowbands which are already meant to be used by other devices, before choosing to allocate resources in a different narrowband. An example illustrating this behavior follows.

> *Consider a PUSCH comprising twelve PRBs and thus two narrowbands; narrowband A and B. A device in the network already reserved four PRBs in narrowband A and another device reserved three PRBs in narrowband B. In the scenario where a third device needs to be assigned four PRBs, it will have to wait until the next TTI as there are available only two PRBs in narrowband A and three PRBs in narrowband B.*

The number of PRBs that each device is assigned for its UL transmission, is determined by the base station and is based on the measured SINR at the base station. The procedure which illustrates the method used is discussed below and depends on the transmit power and the size of data. Therefore, there is a slight difference in the PRB assignment for the RA procedure (*Msg.3*), for the transmission of *Msg.5* (RRC Connection Setup Complete message in Figure 2-8) and for the actual data transmission.

For *Msg.3*, the device transmits with the same power as it used for the preamble. However, the preamble power is defined for a transmission within six PRBs and the Msg.3 transmission might need a different number of PRBs as the size of this message is only six bytes in Cat-M1 [51]. Therefore, the transmit power per PRB is defined based on the transmit power of preamble from Equation (4-7). Moreover, based on the transmit power per PRB, an estimation of the SNR per PRB can be made based on Equation (4-8), which will then indicate the appropriate TBS index as shown in Figure 4-5.[10] For the Msg.3 transmission it is noted that devices in CE Mode A can use a maximum of TBS index seven while for devices in CE Mode B, the maximum TBS index is four. Furthermore, for the calculation of the SNR per PRB it is assumed that there is no interference as we consider a single-cell factory environment. This can possibly lead to a matching of a higher MCS and thus a higher TBS compared to environments where interference is assumed. Consequently, the probability of an error during a transmission on the channel will be higher than when using a shorter TBS and therefore the transmission might be unsuccessful.

$$P_{tx}^{PRB}[W] = \min\left(\frac{P_{tx}^{max}[W]}{6}, \frac{P_{tx}^{preamble}[W]}{6}\right) \qquad (4\text{-}7)$$

$$SNR^{PRB} = \frac{P_{tx}^{PRB}[W] * 10^{\frac{-CouplingLoss[dB]}{10}}}{N_0[W/Hz] * NF * BW[Hz]} \qquad (4\text{-}8)$$

[9] The maximum number of PUSCH PRBs that a device can be assigned is six.

[10] The method used for creating this figure is described in Appendix A .

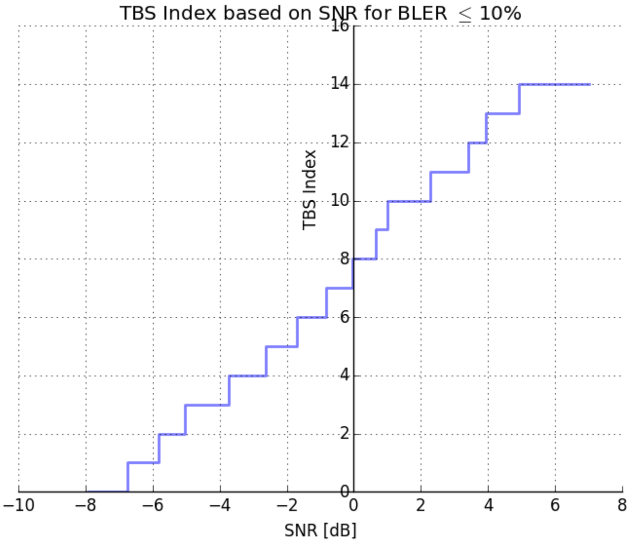


Figure 4-5 - TBS index indication based on SNR.

After the TBS index assignment, the TBS can be read from Table 4-3 for one PRB. In case the TBS is not big enough to transmit the data, the transmit power, SNR, TBS index and TBS will be re-calculated for a two-PRB transmission or transmission using even more PRBs, until the lowest number of PRBs is found that can convey the full message.

Table 4-3 - Matching of TBS index (I_{TBS}) with the TBS based on the number of PRBs (N_{PRB}) as presented in Table 7.1.7.2.1-1 in [50].

I_{TBS}	N_{PRB}					
	1	**2**	**3**	**4**	**5**	**6**
0	16	32	56	88	120	152
1	24	56	88	144	176	208
2	32	72	144	176	208	256
3	40	104	176	208	256	328
4	56	120	208	256	328	408
5	72	144	224	328	424	504
6	328	176	256	392	504	600
7	104	224	328	472	584	712
8	120	256	392	536	680	808
9	136	296	456	616	776	936
10	144	328	504	680	872	1032
11	176	376	584	776	1000	1192
12	208	440	680	904	1128	1352
13	224	488	744	1000	1256	1544
14	256	552	840	1128	1416	1736

For the transmission of *Msg.5* as well the actual *data* transmission, the same method is used to derive the appropriate number of PRBs with the only difference that the transmit power is defined per PRB as shown in Equation (4-6), and hence Equation (4-7) is not needed. Furthermore, the data is appended to Msg.5[11] in order to achieve a shorter delay in Cat-M1. However, the maximum TBS that can be achieved for a specific device using six PRBs may not be high enough to fully transmit both the Msg5 and the UL data at once. In this case, the maximum allowed TBS is used such that most of the data will be transmitted and then the device requests another UL grant in order to transmit the remaining data. This process carries

[11] The size of Msg.5 is considered in average equal to one byte based on [57].

on until all data is sent. Finally, for this kind of data transmissions, devices in CE Mode A can be assigned a maximum TBS index of 14 while for devices in CE Mode B, the maximum TBS index is equal to 9.

4.4.2.3 *MPDCCH*

As already presented in Figure 2-8, the MPDCCH channel is used before the RAR message, Msg.4 and any UL transmission on PUSCH that occurs after the RA procedure. Due to the limited resources given to the MPDCCH, differentiated scheduling is carried out based on what message will follow (RAR message, Msg.4 or Msg.5/UL data).

First, priority is given to the signaling within the RA procedure and therefore to the RAR message and Msg.4, considering that the device will initiate a new RA attempt if it does not receive an answer from the base station within a specific time interval (RAR window). Moreover, between those two message types, further priority is given to the signaling of the RAR message as the RAR window is shorter than the contention resolution window and thus it affords less delay. For example, for the 'EDT for All Devices' procedure presented in Section 3.1.3, shorter RAR window applies to URLLC devices compared to Cat-M1 and therefore scheduling on the MPDCCH can affect the performance of URLLC devices. Finally, after the scheduling of all the signaling regarding the RA, the UL grants are scheduled. It is also important to note that for the procedures that do not apply priority to URLLC over non-URLLC traffic (i.e. the reference RA procedure in Cat-M1 and the 'EDT based on 3GPP Release 15' procedure), no differentiation on the MPDCCH scheduling is done between the two types of traffic. For example, if a URLLC and a non-URLLC transmission need a DL signaling for the RAR message on the MPDCCH, they will be served in a FCFS fashion.

4.5 Configuration of Cat-M1 system

As already mentioned, Cat-M1 supports deeper coverage by introducing four different coverage enhancement (CE) levels that use different power control and numbers of repetitions of messages. Therefore, the thresholds that distinguish the CE levels should be defined as well as the configuration of the number of repetitions of each CE level.

4.5.1 Definition of coverage levels

Based on [45], a coverage analysis (link budget analysis) is carried out to define the number of repetitions required to support an increase of the Maximum Allowable Coupling Loss (MACL), in order to achieve the more ambitious coverage targets of Cat-M1 using LTE as a baseline. Based on [45], the target MACL for the Cat-M1 system is 160.0 dB when less conservative noise figures and devices supporting 23 dBm transmit power are used. Table 4-4 presents the MACL that is allowed per channel (*baseline MACL*), as also defined for LTE and the higher target MACL for Cat-M1 (*target MACL*). This difference between the baseline MACL (for LTE) and the target MACL (for Cat-M1), will be achieved by repeating the different messages multiple times. The analysis in the table assumes a six PRB assignment for PUSCH and PDSCH, noting that it is also possible that fewer PRBs are assigned. Since the available transmission power budget needs to be split over the assigned PRBs, the use of fewer PRBs (provided that they suffice to convey the data) will lead to a higher *baseline MACL* and thus a lower number of repetitions will be required to achieve the *target MACL*.

Table 4-4 - Coverage Analysis for Cat-M1 CE Level 3.

CHANNEL	PUCCH	PRACH	PUSCH	PDSCH	MPDCCH
Transmitter					
(0) Max Tx Power [dBm]	23.0	23.0	23.0	36.8	36.8
(1) Power in channel bandwidth [dBm]	23.0	23.0	23.0	30.8	30.8
Receiver					
(2) Thermal noise density [dBm/Hz]	-174	-174	-174	-174	-174
(3) Receiver noise figure [dB]	3	3	3	5	5
(4) Occupied channel bandwidth [Hz]	180000	1080000	1080000	1080000	1080000
(5) Effective noise power = (2) + (3) + 10log((4))	-118.4	-110.7	-110.7	-108.7	-108.7
(6) Required SNR [dB]	-7.8	-10.0	-4.3	0.0	-0.7
(7) Receiver sensitivity = (5) + (6) [dBm]	-126.2	-120.7	-115.0	-108.7	-109.4
(8) Baseline MACL = (1) – (7) [dB]	149.2	143.7	138.0	139.5	140.2
(9) Target MACL	160.0	160.0	160.0	160.0	160.0

As mentioned, the *target MACL* for Cat-M1 is 160.0 dB which implies that devices with CL higher than 160.0 dB will be considered in outage and therefore 160.0 dB can be defined as the MACL for devices in CE level 3, which is the worst CE level in terms of coverage. It is noted that for this particular study, the CL for each device is measured based on the path loss, shadowing and antenna gains. Additionally, a MACL for the other CE levels should be defined. The smallest and hence most demanding *baseline MACL* among all channels is derived from the PUSCH while using six PRBs and it is equal to 138.0 dB, based on Table 4-4. This most demanding *baseline MACL* is defining the *target MACL* for the best CE level (CE level 0) as typically in CE level 0 no repetitions should be needed and hence no gain is achieved. Given that two more *target MACL* should be defined (for CE levels 1 and 2) the range between 138.0 dB and 160.0 dB is split to three intervals and therefore the two *target MACL* were chosen as 145.0 dB and 152.0 dB for CE level 1 and CE level 2 respectively such that the intervals will be almost equal. Table 4-5 shows the CL ranges that are allowed in each CE level and therefore, devices can choose their CE level based on their experienced CL and the *target MACL* of each CE level.

Table 4-5 - Target MACL for each CE level.

CE Level	Target MACL [dB]
0	138.0
1	145.0
2	152.0
3	160.0

Within each CE level, a number of repetitions per channel should be set such that the received SNR will be increased and thus the target SNR based on the target MACL of that channel and CE level is met. In order to quantify how many repetitions should be used per channel and CE level, the baseline MACL (MACL defined for the scenario with no repetitions and also used in LTE) and the CE level's target MACL should be compared, as it is known from [45] that doubling the repetitions of a message will result in 3.0 dB coverage gain. For example, for the PRACH a baseline MACL of 143.7 dB (as in LTE) applies but the target MACL for CE level 1 is 145.0 dB. Therefore, a coverage gain of 1.3 dB is required which in practice will be achieved by transmitting the PRACH preamble multiple times; for this case specifically, two transmissions (i.e. one fresh transmission and one repetition) of the preamble are sufficient. Therefore, devices that experience high CL, should use repetitions in order to boost their SNR and thus compensate for the higher coupling loss and still achieve the required SNR at the base station. The derivation of the required number of repetitions[12] per channel and per CE level is shown in Table 4-6.

Table 4-6 - Definition of repetitions per channel and CE level in Cat-M1.

		Channels				
		PUCCH	**PRACH**	**PUSCH**	**PDSCH**	**MPDCCH**
Baseline MACL [dB]		149.2	143.7	138.0	139.5	140.2
CE level 0	Target MACL [dB]	138.0	138.0	138.0	138.0	138.0
	Required Gain [dB]	0.0	0.0	0.0	0.0	0.0
	Required Repetitions	**0**	**0**	**0**	**0**	**0**
CE level 1	Target MACL [dB]	145.0	145.0	145.0	145.0	145.0
	Required Gain [dB]	-	1.3	7.0	5.5	4.8
	Required Repetitions	**0**	**1**	**7**	**3**	**3**
CE level 2	Target MACL [dB]	152.0	152.0	152.0	152.0	152.0
	Required Gain [dB]	2.8	8.3	14.0	12.5	11.8
	Required Repetitions	**1**	**7**	**31**	**31**	**16**
CE level 3	Target MACL [dB]	160.0	160.0	160.0	160.0	160.0
	Required Gain [dB]	10.8	16.3	22.0	20.5	19.8
	Required Repetitions	**15**	**63**	**255**	**127**	**127**

After this configuration of the CE levels, a short analysis is now presented in order to estimate how many devices will correspond to each CE level based on the network topology and

[12] The total number of message transmissions is one plus repetitions as the original message is not included in the number of repetitions.

propagation environment that is considered for the FoF scenario investigations. For this analysis, results from 100 snapshots are considered, where in each snapshot 3000 devices are uniformly distributed in the 50 m × 50 m area of the factory. The Cumulative Distribution Function (CDF) of the CL values is shown in Figure 4-6 and it is clear that all devices will be in CE level 0 as the highest experienced CL observed in the network is 90 dB. This behavior can be explained by the placement of the base station indoors, as indoor placement offers better coverage than outdoor placement, as well as by the small distances between the devices and the base station (the maximum distance, in this scenario, between a device and the base station in 71 m). Furthermore, recall that the CL in this study is measured based on the path loss, shadowing and antenna gains and that the multipath fading is ignored, something that would have contributed to the coupling loss.

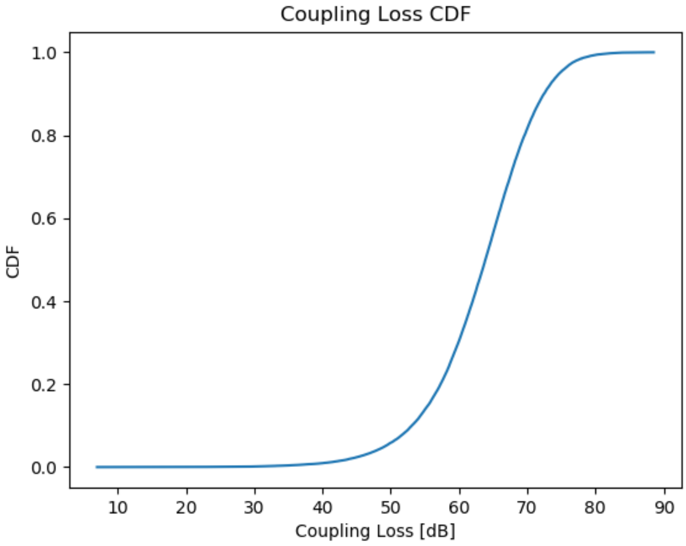


Figure 4-6 - CDF of coupling loss based on 100 simulations with 3000 devices.

4.5.2 PRACH

For each one of the CE levels, the PRACH parameters can be configured differently (see Section 2.2.1.2). However, as all devices will be configured in CE level 0, the PRACH configuration will only concern this CE level. The parameters are defined in the following way:

- **PRACH Frequency Offset:** One random narrowband in the UL carrier of 5 MHz.
- **Maximum Repetitions per Preamble:** Based on Table 4-6, the preambles need to be transmitted just once.
- **PRACH Starting Subframe:** As the maximum repetitions per preamble is equal to one, the starting subframe can also be configured to one. This means that in every PRACH subframe, a new PRACH attempt can be done.
- **Preamble Range:** As mentioned in Section 4.2, no Human-to-Human (H2H) LTE traffic will be modeled and thus no preambles need to be assigned for H2H traffic. Therefore, all available preambles (1-64) can be used for Cat-M1.

- **Maximum Number of Preamble Attempts:** It is configured to its minimum value, i.e. three attempts. Increasing the number of attempts would linearly increase the delay, which is undesirable as the goal of the study is to have low random access delay.
- **Random Access Response Window:** It is configured to its minimum value (20 ms) as the aim of this study is to achieve low delay and thus there is no need in having a larger value.
- **Contention Resolution Timer:** It is configured to its minimum value (80 ms) which is already long enough compared to the goals of this project.
- **PRACH Configuration Index:** This parameter defines the periodicity of the PRACH subframes. The tuning of this parameter involves a trade-off between the utilization of UL resources and the probability of preamble collision. A good configuration of the PRACH periodicity can be considered to be one which allows a maximum of 1% collision probability, on average, as otherwise the end-to-end delays and outage percentages will be high. It is noted that for the end-to-end delay performance targeted for URLLC devices, a collision probability of 1% can be considered high. However, at this point of the study, the network is configured based on the aggregated number of devices in the network and therefore the initial configuration is mostly based on non-URLLC traffic, which is the most common traffic of the network. In order to study the effect of the PRACH periodicity on the performance of the URLLC devices, an analysis in carried out in Section 5.8. Equation (4-9) shows an approximation of the average collision probability (as presented in [52] based on the definition of 3GPP in [47]) for uniform preamble arrival distributions, which is the arrival distribution followed by non-URLLC traffic in this study. In the numerator the number of preamble generations per second is given while the denominator contains the number of RA Opportunities (RAO) per second. The RAO is defined as the product of the total number of available PRACH preambles and the PRACH periodicity within a radio frame of 10 ms (i.e. there are in total 100 radio frames in one second).

  $$Pr_{collision} = \min\left(\frac{\#User\ arrivals\ /\ second}{\#RAO\ /\ second}, 1\right) = \min\left(\frac{\#Users\ in\ network\ /\ 60}{\#Preambles * PRACH_{periodicity} * 100}, 1\right) \quad (4\text{-}9)$$

  From Equation (4-9) it is observed that to keep the probability of collision below e.g. 1% the PRACH periodicity (number of PRACH subframes in a radio frame) should be configured based on the network load. In this study, 3000 devices will be present in the network (see also Section 5.1) and therefore for a 1% probability of collision, the PRACH periodicity should be set to one according to Equation (4-9). However, because URLLC traffic follows a beta distribution, Equation (4-9) does not apply to URLLC traffic and in fact a higher collision probability should be expected. It was therefore decided to set the PRACH periodicity to two.

4.6 Simulation flow

For the generation of statistically reliable simulation results, the short-term dynamic simulation method is used, where different snapshots are generated and simulated for a short time period. The length of this short time period is an important aspect as all devices will have to complete their UL data transmission within that period. This requires that each snapshot has a duration of at least 60 seconds, which is the period for the generation of the non-URLLC traffic. After the expiration of the 60 seconds, the traffic generation and transmission continues (i.e. the contention for Cat-M1 resources continues), but the end-to-end delay statistics are based only on the UL data that were generated within the first 60-second time interval. This implementation is selected in order to avoid the edge effect in the end-to-end delay performance after the expiry of the first 60 seconds. It is also noted that the event which triggers the generation of URLLC traffic, can randomly occur only within the first 50 seconds and the URLLC devices react to this event within a time interval of 10 seconds.

In each one of those snapshots, the location of the devices within the factory as well as their experienced shadowing loss is randomly sampled. In this way, the overall simulation results are based on many different device locations and not only on one specific and fixed location scenario. The choice of not using a fixed location scenario for all snapshots was based on the fact that conclusions should not be drawn upon a specific factory layout but for a more general scenario i.e. all factories of 50 m X 50 m size. Moreover, in each snapshot, every device chooses to transmit its UL data in a different time within the 60 seconds interval such that the final results will be applicable to an even more generalized scenario. With this approach, the conclusions of this study are applicable to factory layouts with similar sizes and placement of the base station as in this study and are uncorrelated from a specific data arrival pattern. Figure 4-7, illustrates the simulation process with the use of snapshots.

The simulation process shown in Figure 4-7 is applied for the evaluation of the RA procedure in Cat-M1 as well as the new RA procedures for 5G networks. Moreover, the number of snapshots that are needed in order to achieve the required reliability is directly correlated with the number of samples for the Key Performance Indicator (KPI) of interest (i.e. the number of devices in the factory). The method of calculating the minimum number of snapshots needed is illustrated through the following example.

> *From the definition of the URLLC traffic it is derived that the end-to-end delay should be guaranteed with 99.99% reliability as defined from the requirements of this thesis. In order to be able to achieve a meaningful estimation of the 99.99th end-to-end delay percentile for URLLC traffic, at least 10000 samples are needed. However, URLLC devices are only 5% of the total number of devices and thus in a network of 3000 devices, we can get only 150 samples. Therefore, it is needed to have at least 67 snapshots to achieve the required number of 10000 samples.*

In practice, a multiple of this minimum number of snapshots should be used in order to achieve a meaningful statistical confidence (e.g. a sufficiently narrow 90% confidence interval (see Section 5.1.2) in the derived estimate for the end-to-end delay percentile. An exact number of needed snapshots to achieve a high statistical confidence cannot be derived as it can vary between different system configurations (e.g. between Cat-M1 and 5G scenarios). Therefore, the confidence intervals are constructed continuously until the target confidence is achieved.

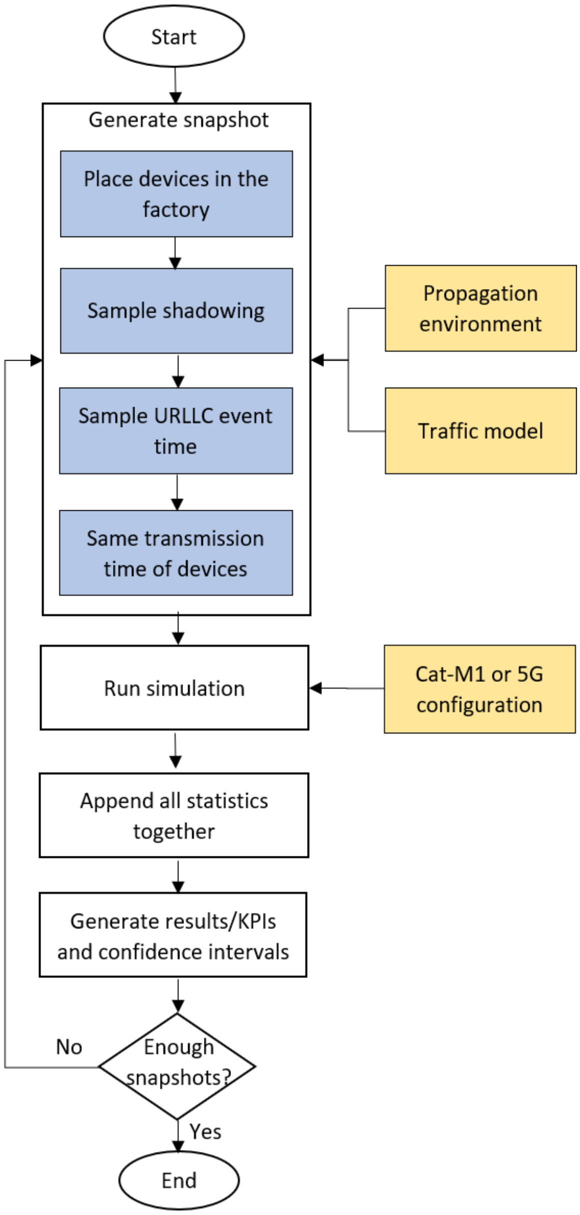

Start
Generate snapshot
Place devices in the factory
Sample shadowing
Sample URLLC event time
Same transmission time of devices
Propagation environment
Traffic model
Run simulation
Cat-M1 or 5G configuration
Append all statistics together
Generate results/KPIs and confidence intervals
No
Enough snapshots?
Yes
End

Figure 4-7 - Overview of the simulation process.

Chapter 5 Random access assessment

In this chapter, the assessment of the RA procedure for Cat-M1, the EDT procedure and the two-step RA procedure is presented. More specifically, the assessment is carried out from the generation time of the data at the device until their correct reception at the base station and thus the combination of the RA procedure and the UL data transmissions is assessed. Section 5.1 first presents the assessment methodology, followed by the quantitative assessment of the reference case of Cat-M1 in Section 5.2. Sections 5.3 and 5.4 then present the evaluation of the EDT procedures 'EDT based on 3GPP Release 15' and 'EDT for All Devices' [13] respectively. Subsequently the evaluation of the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback' is presented in Sections 5.5 and 5.6 respectively. An overall comparison of all RA procedures is presented in Section 5.7. Further analyses are then presented for the deemed best RA procedure, which is the two-step RA procedure (either with repetitions or feedback). For this solution, Section 5.8 presents a sensitivity analysis w.r.t. the configured PRACH periodicity, while Section 5.9 presents a sensitivity analysis w.r.t. the network load, assuming an optimized configuration of the PRACH periodicity of the two-step RA procedure. Finally, Section 5.10 discusses the effect of the data size data used in this study on the results and Section 5.11 concludes the chapter with a summary of the main results.

5.1 Introduction

The assessment of the RA procedures presented in Sections 5.2 to 5.8 is carried out for a network load given by the presence of 3000 devices in the network, 95% of which are non-URLLC devices and 5% are URLLC devices. With this configuration, the requirement of having at least one device per m^2 (see Section 1.2) is met. The different RA procedures are evaluated and compared in terms of the achieved end-to-end delay performance and the induced UL resource utilization. The Key Performance Indicators (KPIs) used for the assessment phase are defined in Section 5.1.1, while the methodology used to calculate the confidence intervals of the end-to-end delay percentiles is presented in Section 5.1.2.

5.1.1 KPIs

The main focus of this study is to design a RA procedure which guarantees end-to-end delays *with specific reliabilities* to non-URLLC and URLLC devices. Therefore, the main KPI of this study is the end-to-end delay matched with a reliability value. The end-to-end delay is defined as the time interval from the generation of the data at the device until their successful reception at the base station.

In order to derive this KPI, covering aspects of both end-to-end delay and reliability, the Cumulative Distribution Function (CDF) of the end-to-end delay is calculated. More specifically, the 99.9th and the 99.99th percentile of the end-to-end delay are selected as the KPIs for non-URLLC and URLLC devices, respectively. For completeness and easier comparison

[13] The 'EDT with Priority and Shorter Windows' procedure, discussed in Section 3.1.2, is not assessed as it is integrated in the 'EDT for all devices' procedure.

of the delay performance among the two types of devices, the 99.9th and 99.99th delay percentiles will be calculated for both device types. Moreover, the median (50th percentile) of the end-to-end delay is also presented for both types of devices in order to evaluate and indicate how challenging it is to guarantee end-to-end delay with such high degrees of reliability.

Another KPI used for the assessment of the RA procedures, is the utilization of the UL resources. The UL resource utilization is defined as a fraction of the available time-frequency resources that are used. It is noted that resources are considered as used when they are unavailable to conventional LTE traffic, meaning that resources which are reserved but are not used for an actual data transmission are still considered as used. Quantifying the UL resource utilization is important as some newly designed RA procedures may trade resource efficiency for enhanced end-to-end delays and this trade-off needs to be quantified, since the procedures should not have significantly higher UL resource utilization than that achieved by the reference Cat-M1 RA procedure. Furthermore, the UL resource utilization is a useful KPI to quantify the effects from different PRACH periodicity for a given RA procedure.

5.1.2 Confidence intervals for the end-to-end delay KPI

In order to indicate the degree of statistical confidence in the calculated KPIs, the calculation of Confidence Intervals (CI) is needed around the desired 99.9% or 99.99% end-to-end delay percentiles. For the mean value [53] [54] (rather than a percentile), the CI is calculated from the *N* KPI samples ($X_1, X_2, \ldots, X_N$) obtained over the various simulation snapshots assuming a standard normal distribution of the *N* samples around the 'calculated' average KPI value $\bar{X}$, as shown in Equation (5-1).

$$\bar{X} - Z_{\frac{\alpha}{2}}\frac{S}{\sqrt{N}} < M < \bar{X} + Z_{\frac{\alpha}{2}}\frac{S}{\sqrt{N}} \qquad (5\text{-}1)$$

In Equation (5-1), *M* is the unknown true value of the KPI, *S* is the standard deviation over the *N* KPI samples, and $Z_{\frac{\alpha}{2}}$ can be calculated from the standard normal distribution as α is defining the probability in each tail of the distribution. An example is shown in Figure 5-1 where α defines that the probability in each tail should be 0.025 as $\alpha = 0.05$, which will therefore provide a CI of 95% on the mean value and thus $Z_{\frac{\alpha}{2}} = Z_{0.025} \approx 1.96$ should be used in equation (5-1) in order to define the limits of the CI.

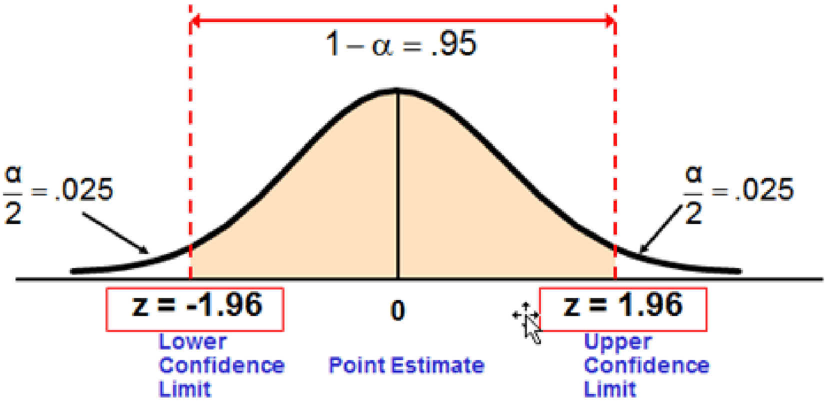


Figure 5-1 - Confidence interval of 95% for an average KPI value [54].

However, the end-to-end delay KPI is not a mean value but rather the end-to-end delay coupled with a reliability percentage (e.g. 99.9% or 99.99% reliability), and therefore calculating the confidence interval of the 99.9th or 99.99th percentile of the end-to-end delay requires a slightly different approach, as follows [55].

First, the *N* samples from the end-to-end delay are sorted in ascending order (Y_1, Y_2, …, Y_i, Y_{i+1}, …, Y_j, …, Y_N). Actually, these samples are used to calculate the empirical CDF of the end-to-end delay as illustrated in Figure 5-2. Figure 5-2 also presents how the 99.9th and 99.99th end-to-end delay percentiles can be extracted from the CDF and highlights how challenging it is to measure CI for such high percentiles as the values of the samples around the percentile can vary significantly (especially for the 99.99th percentile).

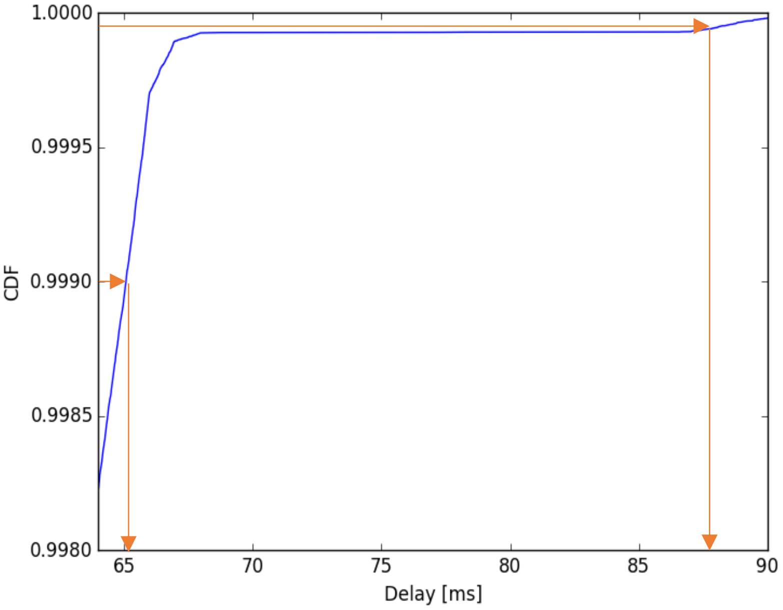


Figure 5-2 - Example CDF of end-to-end delay and the corresponding 99.9th and 99.99th percentiles.

Then, and as previously, α is defining the probability in each tail of the distribution (e.g. α = 0.10) and thus the targeted end-to-end delay percentile (T) is lying between sample-based percentiles Y_i and Y_j, as in Equation (5-2) with probability 1-α (e.g. 0.90). The samples (Y_i, …, Y_j) around the estimated delay percentile $\bar{T}$ (given by sample Y_T) are assumed normally distributed with mean (μ) and variance (σ^2) and can be calculated from Equation (5-3), where p is the targeted cumulative probability density value (e.g. 0.999 or 0.9999). From Equation (5-4), where $Z_{\frac{\alpha}{2}}$ can be calculated as previously, the indexes of the lower and upper samples Y_i and Y_j can be calculated that will define the CI.

$$1 - \alpha = P\left[Y_i < T < \ Y_j\right] \tag{5-2}$$

$$\mu = Np, \sigma^2 = Np(1-p) \tag{5-3}$$

$$i = \mu - Z_{\frac{\alpha}{2}}\sigma,\ j = \mu + Z_{\frac{\alpha}{2}}\sigma \tag{5-4}$$

A numerical example is presented below, which illustrates the calculation of the 90% CI of the 99.9th percentile, based on 5 million samples (N = 5 000 000). The CDF along with samples that define the estimated end-to-end delay value and the CI are presented in Figure 5-3.

1. All 5 million samples are sorted in ascending order (Y_1, Y_2, …, Y_i, …, $Y_{T,}$, …, Y_j , …Y_N).

2. The index of the 99.9th delay percentile can be calculated from $\mu = 5000000 * 0.999 = 4995000$ and its corresponding estimated delay value is $\overline{T}$ = $Y_{4995000}$ = 65.08 ms.
3. The samples around this percentile follow a normal distribution with given $\mu = 4995000$ and $\sigma = \sqrt{4995000 * 0.001} = 70.68$ ms.
4. Assuming a normal distribution for the samples around the desired end-to-end delay percentile, the 90% CI implies that the probability at the tails of the distribution is α = 0.1. Therefore, it is calculated that $Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.68$.
5. The lower index of the CI can then be derived from sample $i = 4995000 - 1.68 * 70.68 = 4994881$ with corresponding delay value Y_i = $Y_{4994881}$ = 65.02 ms.
6. The upper index of the CI can be derived from sample $j = 4995000 + 1.68 * 70.68 = 4995119$ with corresponding delay value Y_j = $Y_{4995119}$ = 65.16 ms.
7. Hence the CI of the 99.9th end-to-end delay percentile is [65.02, 65.16] ms.

A similar example can be derived for the calculation of the CI of the 99.99th percentile. In that case, the index of the 99.99th delay percentile is given by μ = 4999500 with corresponding end-to-end delay value $\overline{T}$ = $Y_{4999500}$ = 66.94 ms and the standard deviation $\sigma = \sqrt{4999500 * 0.0001}$ = 22.35. Therefore, the two samples, which determine the CI, are samples i = 4999462 and j = 4999538 with corresponding delay values Y_i = $Y_{4999462}$ = 66.90 ms and Y_j = $Y_{4999538}$ = 66.99 ms. Hence the CI is of the 99.99th end-to-end delay percentile is [65.90, 66.99] ms.

It is noted that with this approach of calculating the CI, the distance of the lower and upper bounds from the CI center value might differ significantly and lead to an asymmetric CI. Figure 5-2 presents an example of a CDF which might lead to such a CI for the 99.99th percentile as the end-to-end delay samples around this end-to-end delay percentile vary significantly.

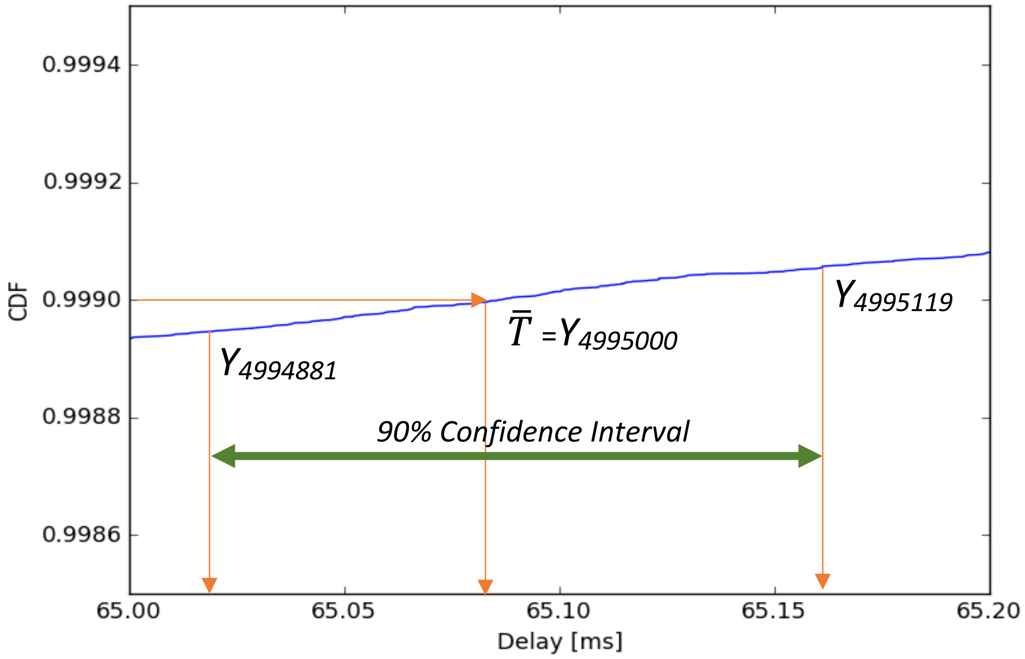


Figure 5-3 - Part of a CDF for the calculation of the 90% CI of the 99.9th end-to-end delay percentile.

In this study it was chosen to set the confidence level at 90% (and thus α = 0.10) and not to a higher (e.g. 95% or 99%) value as for a higher CI at the 99.9th or 99.99th end-to-end delay KPI percentile, more samples (and thus snapshots) are required, which implies longer simulation times.

5.2 Cat-M1

This section presents the performance evaluation of Cat-M1 as the reference scenario. Figure 5-4, illustrates the end-to-end delay for the 50th (median), the 99.9th and the 99.99th percentiles for non-URLLC devices on the left and for URLLC devices on the right as well as the corresponding CI for each percentile. Also, the grey dashed lines illustrates the requirements set for the 99.9th percentile for non-URLLC devices and the 99.99th percentile for URLLC devices.

For both non-URLLC and URLLC devices it appears that the median and the 99.9th percentile have the same end-to-end delay. The value of the median is calculated for both types of devices at around 40 ms (39.9 ms and 40.1 ms for non-URLLC and URLLC devices respectively) which is a value in accordance with the discussion in Section 2.4 assuming that devices established a connection during their first access attempt. Table 5-1 shows the probability of the number of access attempts required such that a device will establish connection to the network and illustrates that for more than 50% of the time both non-URLLC and URLLC devices indeed experienced no collisions and established a connection with base station during their first access attempt. Furthermore, it is observed that the end-to-end delay calculated for the 99.9th percentile is around 25 ms higher than the median. This increase of 25 ms is introduced due to devices that failed to connect during their first access attempt and thus they had to wait until the RAR expiry (20 ms) in order to restart the RA procedure and make a second access attempt.

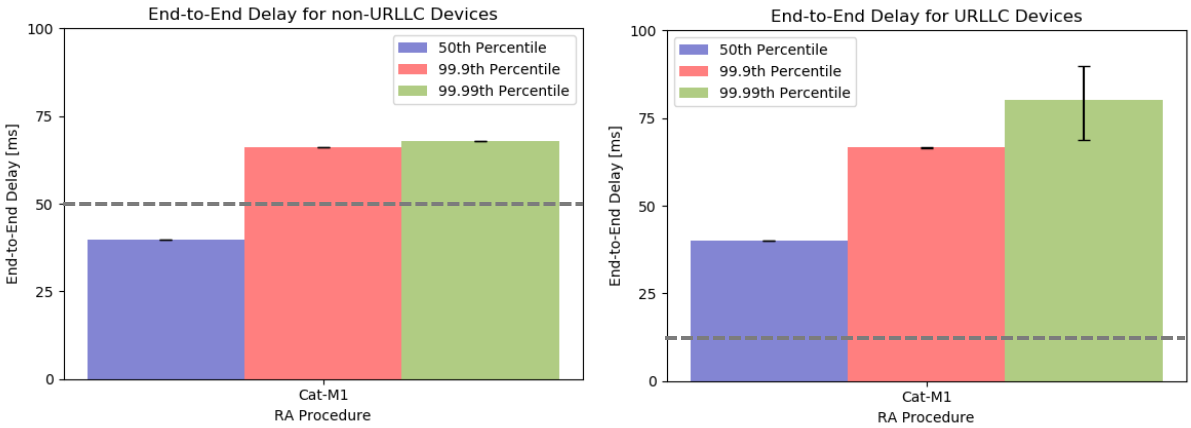


Figure 5-4 - End-to-end delay for Cat-M1 for non-URLLC devices (left) and URLLC devices (right).

In Cat-M1, there is no differentiation between non-URLLC and URLLC devices and this is the reason why the median and the 99.9th percentile are the same for the two types of devices. Additionally, the similarity of these delay percentiles between the two types of devices is due to fully transmitting the UL data along with Msg.5, as discussed in Section 2.3. However, the two types of devices make use of a different traffic model and the result of this difference is reflected in the 99.99th percentile.

In contrast to the 50th and the 99.9th percentile, the 99.99th end-to-end delay percentile is different for non-URLLC and URLLC devices. It can be derived that the 99.99th end-to-end delay percentile of non-URLLC devices is influenced only by devices that established connection during their *second* access attempt while for URLLC devices it is influenced by devices that had either two or three access attempts. This outcome is considered reasonable as the non-URLLC arrivals follow a uniform distribution while the URLLC arrivals follow a beta distribution. Therefore, the arrivals of URLLC transmissions are more clustered than the non-URLLC arrivals which implies a higher probability of preamble collision and consequently a higher number of access attempts. More specifically, Table 5-1 illustrates that the probability of having three access attempts is about two times higher for URLLC devices than for non-URLLC devices. Table 5-1 also shows that non-URLLC devices have one or two access attempts 99.9938% of the time while URLLC devices experience one or two access attempts just 99.9869% of the time, something that is reflected in the 99.99th end-to-end delay percentile. Recall that after three failed access attempts, the device is considered in outage.

Table 5-1 - Probability of access attempts required for non-URLLC and URLLC devices based on simulations.

Number of Access Attempts	non-URLLC devices	URLLC devices
One	99.6530%	99.4857%
Two	0.3408%	0.5012%
Three	0.0062%	0.0131%
Outage	0.0000%	0.0000%

From Figure 5-4 it is also observed that the CI for the end-to-end delay for the URLLC devices is 20 ms wide which highlights that the 99.99th percentile is the breaking point on the CDF, where devices can either have two or three access attempts. The sample, which determines the 99.99th end-to-end delay percentile, has the estimated delay value Y_{449955} = 80.0987 ms, and it is correlated with a device that had two access attempts.[14] 24 samples were needed to form the targeted 90% CI around the estimated delay value. Thus, the lower part of the CI was calculated from 12 samples with values in the range of 68.8607 ms to 79.7765 ms and all of these values are correlated to devices that had two access attempts.[15] The upper part of the CI was also calculated from 12 samples but with values in the range of 87.1127 ms to 89.5010 ms and all of these values are correlated to devices that had three access attempts. In contrast to the CI for URLLC devices, the CI for non-URLLC devices was constructed from 98 samples around the estimated delay with values within the range of 67.9143 ms to 67.9804 ms and all of these values are correlated to devices that had two access attempts. Both results are in accordance with the values presented in Table 5-1.

[14] Note that from the total N = 9 000 000 samples used only 5% (or 450 000) are from URLLC devices. In order to calculate the index of the 99.99th percentile we have μ = 450000 * 0.9999 = 449955.

[15] The standard deviation is $\sigma = \sqrt{449955 * 0.0001}$ = 6.71 so we have 1.68*6.71 = 11.27 and thus 12 samples below and above the 99.99th delay sample.

5.3 EDT based on 3GPP Release 15

The use of the EDT procedure requires that the preambles are split into two groups; one for non-URLLC devices and one for URLLC devices, as explained in Section 3.1.1. In order to derive the optimal split of preambles, a preamble analysis needs to be carried out and the corresponding results are presented in Figure 5-5, where the preamble split on the horizontal axes (e.g. 34/30) refers to the number of available preambles for the non-URLLC (e.g. 34) and URLLC (e.g. 30) transmissions. It is here noted that end-to-end delay percentiles that appear with 100 ms end-to-end delay (or more), are considered to have infinite end-to-end delay as the devices correlated to that percentile are in outage. Such example is the 99.99th end-to-end delay percentile for non-URLLC with preamble split 9/55, in Figure 5-5. This indication of outage is used for the rest of the study.

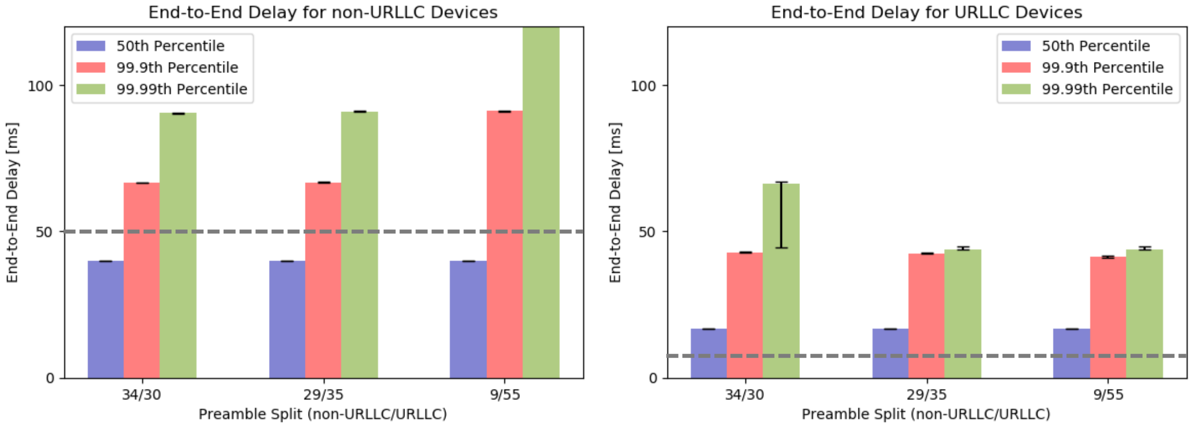


Figure 5-5 - Preamble analysis for 'EDT based on 3GPP Release 15' for non-URLLC devices (left) and URLLC devices (right).

The preamble analysis in Figure 5-5 illustrates the impact of different preamble splits to the end-to-end delay and it can be derived that the preamble split 29/35 can be considered a good performing split. This conclusion is based on the fact that the 99.99th end-to-end delay percentile of URLLC devices improves with this split compared to split 34/30 and does not improve further when reserving more preambles for URLLC transmissions. Furthermore, reserving 30 preambles for URLLC devices, introduces an uncertainty (i.e. a large CI represented by the vertical black line over the green bar in right graph in Figure 5-5) as the 99.99th percentile is influenced by devices that used two or three access attempts and therefore it is not chosen as the optimal split. This behavior is also illustrated in Table 5-2 which shows that 99.9883% of the time a device will have one or two access attempts for split 30 while for split 35, this percentage increases to 99.9926%. It is noted that preamble split 29/35 might not be the truly optimal split as the optimal number of preambles reserved for URLLC transmissions can in principle still lie in the interval of {31,…,35}. However, based on the results we expect that the end-to-end delay will not vary significantly between the truly optimal split and the selected (possibly even optimal) split 29/35.

Table 5-2 - Probability of access attempts required for URLLC devices when reserving 30 and 35 preambles for URLLC transmissions in the ‘EDT based on 3GPP Release 15’ procedure.

Number of Access Attempts	Split 34/30	Split 29/35
One	99.6255%	99.6773%
Two	0.3628%	0.3153%
Three	0.0117%	0.0074%
Outage	0.0000%	0.0000%

Figure 5-6, presents the results of the preamble split 29/35 for the ‘EDT based on 3GPP Release 15’ procedure (labelled as ‘EDT 3GPP’) along with the results obtained in the ‘Cat-M1’ procedure, and thus combines results shown in Figure 5-4 and Figure 5-5. Figure 5-6 also presents with dashed lines the end-to-end delay requirements defined for non-URLLC and URLLC devices. The negative impact of the ‘EDT based on 3GPP Release 15’ procedure and the applied preamble split on the performance of non-URLLC devices can be observed for the corresponding 99.99th end-to-end delay percentile, which is now increased to 91 ms (from 68 ms at ‘Cat-M1’) as a result of more non-URLLC devices having three access attempts instead of two compared to ‘Cat-M1’. This behavior is a result of the increased preamble collision probability as the lower number of preambles that are now available for non-URLLC devices (when compared to ‘Cat-M1’) are causing more preamble collisions and thus more retransmissions. More specifically, and according to Table 5-3 (where ‘EDT based on 3GPP Release 15’ is labelled as ‘EDT 3GPP’), the probability of having one or two access attempts decreases from 99.9938% (at ‘Cat-M1’) to 99.9690% (at ‘EDT based on 3GPP Release 15’) and the probability of having three access attempts is almost five times higher for ‘EDT based on 3GPP Release 15’ than for ‘Cat-M1’. The opposite behavior is observed for the URLLC devices as with ‘EDT based on 3GPP Release 15’ there are enough reserved preambles, just for URLLC devices, to decrease their preamble collision probability. This decreased preamble collision probability is reflected in Table 5-3 as the probability of having one or two access attempts is increased to 99.9926% compared to 99.9869% at ‘Cat-M1’. Therefore, the URLLC devices are experiencing lower 99.99th end-to-end delay percentile at ‘EDT based on 3GPP Release 15’ compared to ‘Cat-M1’.

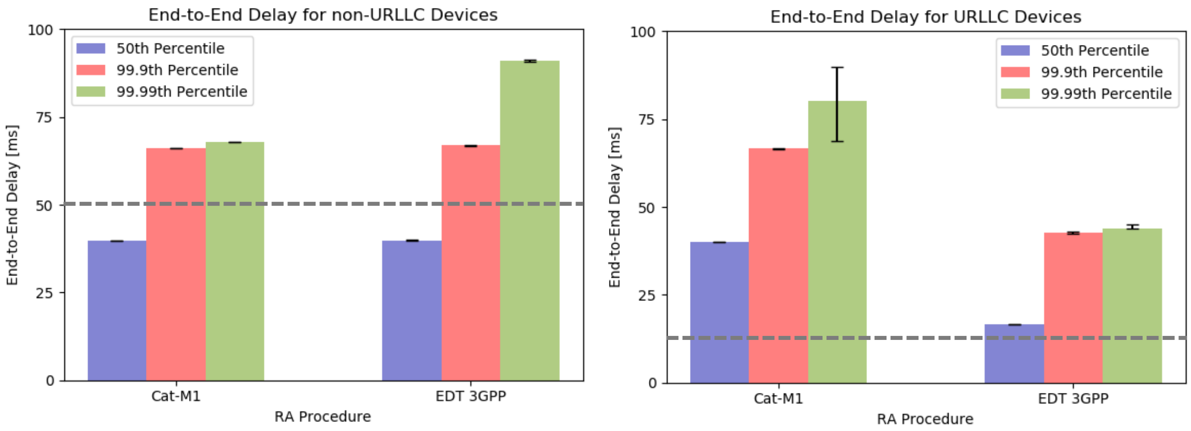


Figure 5-6 - End-to-end delays for ‘Cat-M1’ and ‘EDT based on 3GPP Release 15’ procedures for non-URLLC devices (left) and URLLC devices (right).

Table 5-3 - Probability of access attempts required for non-URLLC and URLLC devices in Cat-M1 and in the 'EDT based on 3GPP Release 15' procedure while reserving 35 preambles for URLLC transmissions.

Number of Access Attempts	non-URLLC devices		URLLC devices	
	Cat-M1	EDT 3GPP	Cat-M1	EDT 3GPP
One	99.6530%	99.2500%	99.4857%	99.6773%
Two	0.3408%	0.7190%	0.5012%	0.3153%
Three	0.0062%	0.0298%	0.0131%	0.0074%
Outage	0.0000%	0.0012%	0.0000%	0.0000%

Furthermore, the 50th and 99.9th end-to-end delay percentile for non-URLLC devices are *not* influenced by the reduction of the available preambles for non-URLLC transmissions (29 in 'EDT based on 3GPP Release 15' instead of 64 in 'Cat-M1') and thus the same delay values are observed as in 'Cat-M1'. However, the 50th delay percentile of URLLC devices drops to 16.8 ms which is roughly as expected and explained in Section 3.1.1. Moreover, the 99.9th end-to-end delay percentile for URLLC devices drops by around 23 ms compared to 'Cat-M1' due to the earlier transmission of the UL data, as explained in detail in Section 3.1.1 and illustrated in Figure 3-2.

5.4 EDT for all devices

This section presents the assessment of the RA procedure allowing all devices to use EDT along with priority and shorter RAR and contention resolution windows for URLLC devices. Therefore, a preamble analysis is again carried out in order to derive the optimal preamble split between non-URLLC and URLLC devices. The results of this analysis are presented in Figure 5-7.

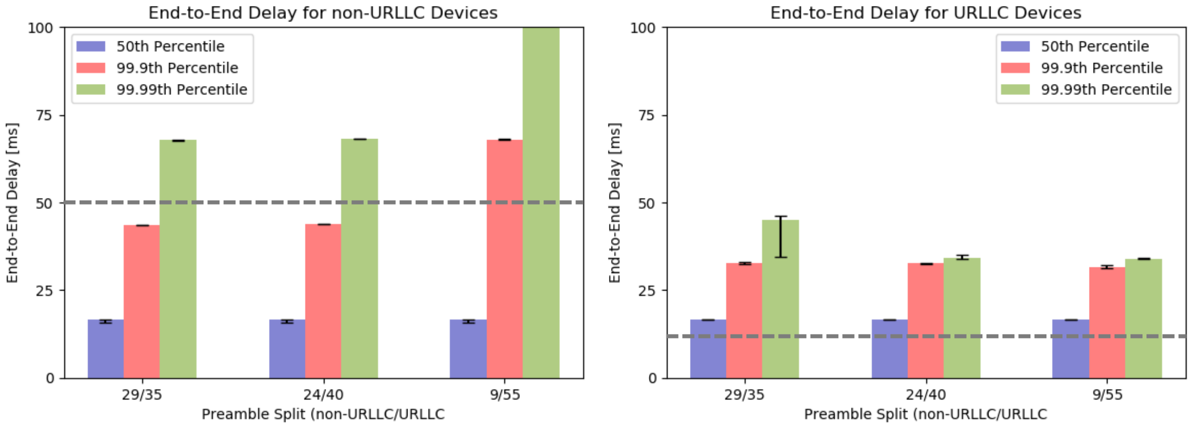


Figure 5-7 - Preamble analysis for the 'EDT for All Devices' procedure for non-URLLC devices (left) and URLLC devices (right).

From Figure 5-7, it is derived that the optimal number of preambles reserved for URLLC transmissions lays in the interval of {36,…,40} preambles. However, for the reasons explained in Section 5.3, the optimal split for this procedure is assumed to be 24 preambles for non-URLLC transmissions and 40 preambles for URLLC transmissions. Therefore, the optimal split is different than the one found in Section 5.3 for the 'EDT based on 3GPP Release 15' procedure, due to the increased probability of the URLLC devices having three access attempts. This probability is illustrated in Table 5-4 and it is calculated as 1.4 times higher in the 'EDT for All Devices' procedure compared to the 'EDT based on 3GPP Release 15' procedure. Due to the shorter RAR window that applies to URLLC devices at the 'EDT for All Devices' procedure, some of them have to retransmit their preamble due to a RAR window expiry, rather than a preamble collision. This can cause a higher number of retransmissions in the network which implies a higher preamble utilization per PRACH opportunity, and thus a higher probability of collision, which finally leads to the need of reserving more preambles for URLLC transmissions compared to the 'EDT based on 3GPP Release 15' procedure.

Table 5-4 - Probability of access attempts required for URLLC devices at the 'EDT based on 3GPP Release 15' and the 'EDT for All Devices' procedures while reserving 35 preambles for URLLC transmissions.

Number of Access Attempts	URLLC devices with split 29/35 at procedure:	
	EDT based on 3GPP Release 15	**EDT for all devices**
One	99.6773%	99.7007%
Two	0.3153%	0.2881%
Three	0.0074%	0.0107%
Outage	0.0000%	0.0005%

Figure 5-8 illustrates the results of the 'EDT for All Devices' procedure (with preamble split 24/40; labelled as 'EDT for all') along with those obtained before for the 'EDT based on 3GPP Release 15' procedure and for the 'Cat-M1' procedure and the requirements defined for non-URLLC and URLLC devices. From this figure it is observed that the end-to-end delays are decreased for both non-URLLC and URLLC devices. The delay for non-URLLC devices is reduced due to the fact that they can now make use of the EDT procedure and thus transmit their UL data earlier than before. For URLLC devices, the delay reduction is caused from the priority given to the devices on the UL and DL channels and the shorter RAR and contention resolution windows which enable them to detect preamble collisions faster.

Furthermore, from Figure 5-8 it is observed that the 50th and the 99.9th end-to-end delay percentiles for non-URLLC devices are comparable to those derived for URLLC devices for the 'EDT based on 3GPP Release 15' procedure. This outcome is expected, as now non-URLLC devices also follow the EDT procedure. However, their results are not comparable to the ones for URLLC devices of this particular procedure ('EDT for All Devices'), as URLLC devices have priority on the UL and DL channel and make use of shorter RAR and contention resolution windows. Regarding the delays for URLLC devices, the 50th end-to-end delay percentile reduces only by 0.2 ms. However, a significant reduction of 10 ms is observed for the 99.9th and the 99.99th percentiles which is an expected result from decreasing the RAR window from 20 ms to 10 ms. As the delay reduction for high percentiles is not larger than 10 ms and the reduction for the median is just 0.2 ms, it can be concluded that the priority given by the base station to the URLLC devices does not provide significant gains. This can be considered a reasonable outcome given that the non-URLLC traffic is uniformly distributed and thus the competition on the resources is mainly involving URLLC devices which have clustered traffic arrivals due to the beta distribution.

Overall, this enhanced EDT procedure provides better end-to-end delays than the 'EDT based on 3GPP Release 15' procedure and the end-to-end delay requirement for the 99.9th percentile for non-URLLC devices is now achieved.

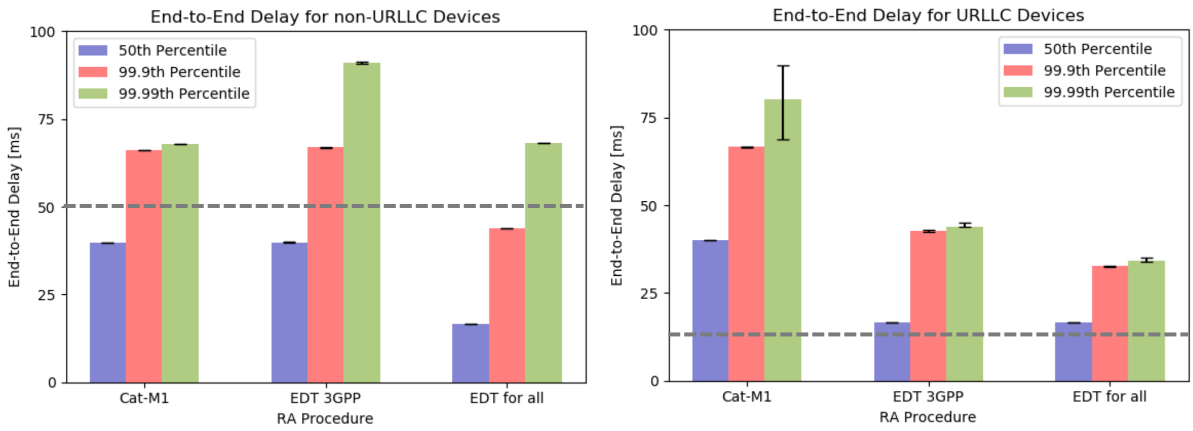


Figure 5-8 - End-to-end delays for 'Cat-M1', 'EDT based on 3GPP Release 15' and 'EDT for All Devices' procedures for non-URLLC devices (left) and URLLC devices (right).

5.5 Two-step RA procedure with repetitions

In this section, the results of the 'Two-Step RA procedure with Repetitions' are discussed. First, the preamble analysis is carried out, in order to derive the optimal number of preambles that should be reserved for the URLLC devices. However, and as discussed in Section 3.2.1, for this particular procedure collisions on the preamble as well as on the data can be experienced by the devices. In order to derive the optimal preamble split, the preamble analysis is carried out in an idealistic scenario without collisions of data transfer via the PUSCH such that only the effect of preamble splitting is captured in the results. This preamble analysis on the ideal scenario is discussed in Section 5.5.1. Furthermore, and as explained in Section 3.2.2, transmitting multiple copies of the same data can increase the probability that the data is received correctly at the base station. Section 5.5.2 presents an analysis on the number of repetitions required to achieve as lower as possible end-to-end delay for URLLC transmissions. It is noted that the optimization of the preamble split and the repetitions can be decoupled because of the introduction of the idealistic scenario without data collisions.

Recall that the number of available narrowbands for URLLC UL data transmission, 3 ms after the transmission of a URLLC preamble, should be calculated in order to derive the number of non-overlapping sets of two PRBs each, as also described in Section 3.2.1. As there are eighteen PRBs in total for the URLLC data transmissions on a 5 MHz LTE carrier, for subframes without PRACH transmission opportunity, nine two-PRB sets that are non-overlapping can be utilized for the URLLC UL data transmissions via PUSCH.[16]

5.5.1 Preamble analysis

As the idealistic scenario assumes no collisions in the data transmission phase via the PUSCH, the effect of the preamble split on the end-to-end delay only occurs through preamble collisions.

[16] The data size of the URLLC transmission is such that two PRBs are sufficient to transfer the URLLC data.

The results of the preamble analysis for the idealistic scenario are presented in Figure 5-9 and show that 35 out of the total 64 preambles should be reserved for the URLLC devices.[17] This is readily concluded as the 99.99th end-to-end delay percentile of URLLC devices improves when reserving 35 preambles for URLLC transmissions compared to when reserving 30 and does not improve further with a higher split value. The selected preamble split is in accordance with the previously studied procedures. In general, the optimal preamble split is influenced by the number of (re-)transmissions in the network and consequently by the number of devices, the distribution of the arrival of data and the number of PRACH opportunities. As none of these parameters are changed among the three procedures, the optimal preamble split should indeed always be the same (i.e. 35 preambles for URLLC transmissions and 30 preambles for non-URLLC transmissions). Recall that the reason for reserving more than 35 preambles for the URLLC devices in the 'EDT for All Devices' procedure is the shorter RAR window which forced more preamble transmissions in the network than necessary, something that is not applicable with the two-step RA procedure as the device transmits its UL data without awaiting the RAR reply from the base station.

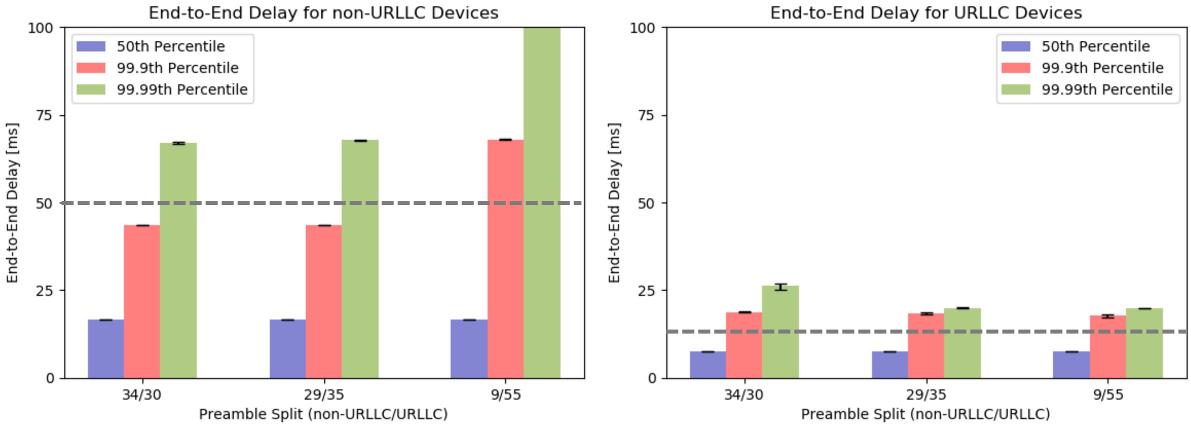


Figure 5-9 - Preamble analysis for the two-step RA procedure in the idealistic scenario for non-URLLC devices (left) and URLLC devices (right).

5.5.2 Repetition analysis

As mentioned, the effect of the URLLC data collisions on the end-to-end delay, can be alleviated by transmitting the URLLC data multiple times. Therefore, an analysis is carried out in order to derive the required number of URLLC transmissions via the PUSCH. The results of this analysis are presented in Figure 5-10, addressing the end-to-end delay while the impact on resource utilization is shown in Section 5.7.2.

[17] It is again noted that the actual optimal number of preambles that should be reserved for URLLC transmission can lay in the interval of {31,…,35} preambles.

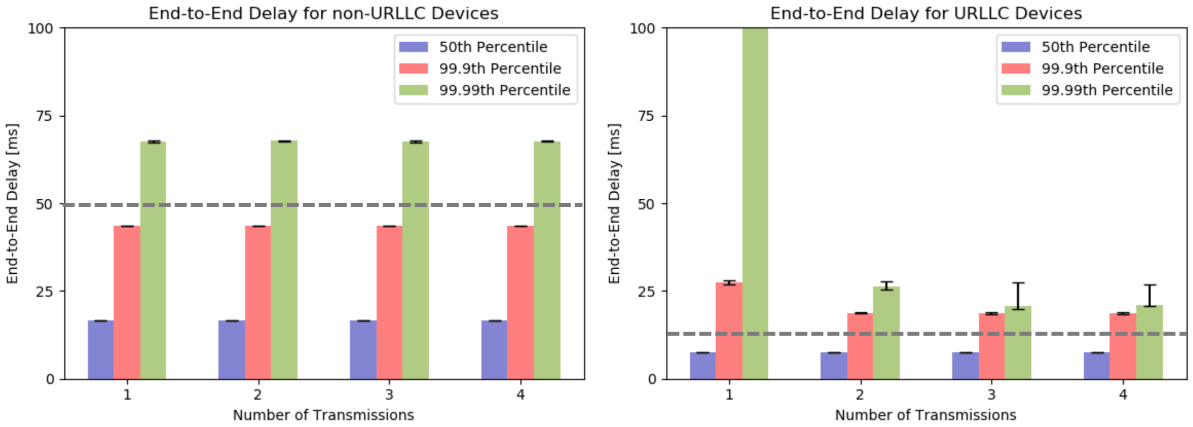


Figure 5-10 - Impact of transmitting the URLLC UL data multiple times on the end-to-end delay for non-URLLC devices (left) and URLLC devices (right).

From Figure 5-10 it is obvious that repetitions are needed as having no repetitions (i.e. one transmission in Figure 5-10) results in outage for the 99.99th end-to-end delay percentile for URLLC devices. Moreover, it can be concluded that the optimal number of transmissions, when reserving 35 preambles for URLLC transmissions, is three as there is no further improvement on the 99.99th end-to-end delay percentile for URLLC devices with a higher number of transmissions.

It is noted that the choice of reserving *fewer* than 35 preambles for the URLLC transmissions could have led to a lower number of needed repetitions in order to achieve the best result possible (for that particular preamble split). However, that best result would have still been worse than what is achieved with 35 preambles for URLLC transmissions and three repetitions as the end-to-end delay is already higher for fewer preambles even in the idealistic scenario and that can definitely not change. Alternatively, choosing to reserve *more* than 35 preambles for the URLLC transmissions could lower the preamble collision probability, but based on Figure 5-9 that preamble collision probability is not low enough to influence the end-to-end delay significantly. Further, more data collisions would occur as more preambles would have matched to the same UL resources on the PUSCH and thus more repetitions would have been required for compensation. It is also noted that an increase of the data collision probability would consequently increase the preamble collision probability as it would generate more preamble transmissions in the network. In conclusion, the introduction of the idealistic scenario enables the optimization of the two parameters (i.e. preamble split and repetitions) separately. Otherwise, the two parameters would have had to be jointly optimized.

Figure 5-11 presents the results of the 'Two-Step RA procedure with Repetitions' (labelled as 'Repetitions') along with the results of the previous procedures. Compared to the 'EDT for All Devices' procedure, the 50th end-to-end delay percentile for the URLLC devices is now reduced from 16.6 ms to 7.5 ms as calculated and explained in Section 3.2.1, due to less signaling involved. Furthermore, the 99.9th and 99.99th percentiles, for URLLC devices, are both decreased by around 14 ms and reached the delay values of 18.6 ms and 20.8 ms respectively. It is also worth mentioning that there is not a significant difference between the end-to-end delays for non-URLLC devices, when comparing the 'EDT for All Devices' procedure and the 'Two-Step RA procedure with Repetitions', as the introduced changes with the two-step RA procedure mainly influence the URLLC devices.

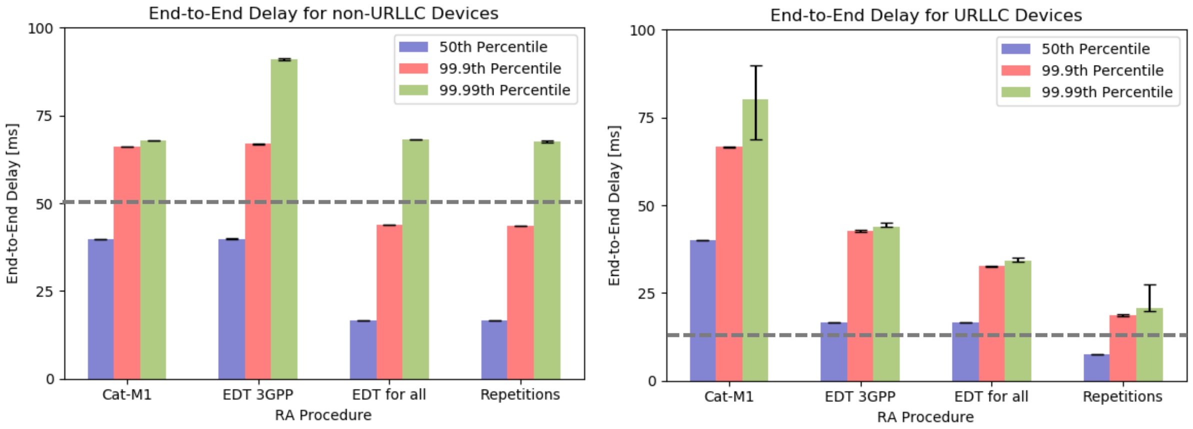


Figure 5-11 - End-to-end delays for 'Cat-M1', 'EDT based on 3GPP Release 15' and 'EDT for All Devices' procedures and 'Two-Step RA procedure with Repetitions' for non-URLLC devices (left) and URLLC devices (right).

5.6 Two-step RA procedure with feedback

In this section, the results regarding the end-to-end delay that can be obtained with the 'Two-Step RA procedure with Feedback' are presented. The impact on the resource utilization is presented in Section 5.7.2. Initially it is calculated, as for the 'Two-Step RA procedure with Repetitions', that nine non-overlapping sets of two PRBs can be utilized for the URLLC UL data transmissions and thus nine URLLC devices are able to transmit their data on these PUSCH resources simultaneously without colliding. Subsequently, this procedure considers that there can be successful URLLC UL data transmission with up to nine devices that transmit their URLLC UL data simultaneously in a particular subframe.

The results of the preamble analysis for the 'Two-Step RA procedure with Feedback' are presented in Figure 5-12 and are identical to those found for the 'Two-Step RA procedure with Repetitions'. This outcome is based on the fact that there were no more than nine devices transmitting their UL data simultaneously in a particular subframe. Therefore, the same observations and conclusions apply and consequently reserving 35 out of the total 64 preambles for the URLLC transmissions, is considered the best choice.

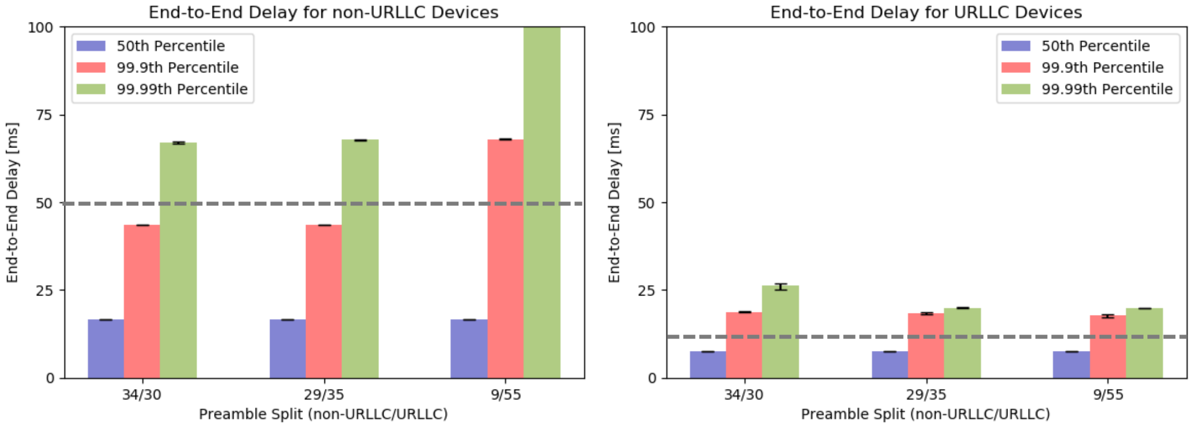


Figure 5-12 - Preamble analysis for the 'Two-Step RA procedure with Feedback' for non-URLLC devices (left) and URLLC devices (right).

Supposing that the required additional signaling on the MPDCCH can indeed be implemented, Figure 5-13 presents the results of the 'Two-Step RA procedure with Feedback' (marked as 'Feedback') along with the results calculated in the previously considered procedures.

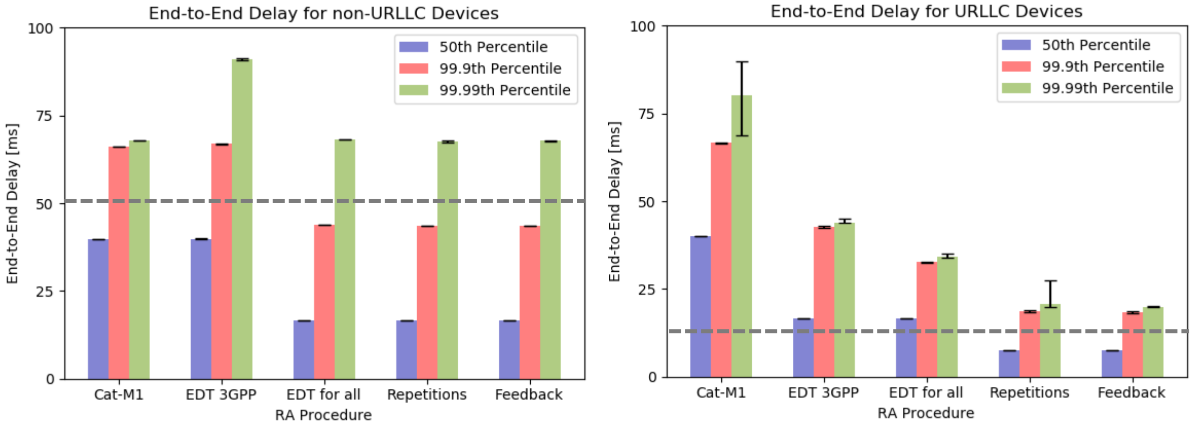


Figure 5-13 - End-to-end delays for 'Cat-M1', 'EDT based on 3GPP Release 15' and 'EDT for All Devices' procedures, 'Two-Step RA procedure with Repetitions' and 'Two-Step RA with Feedback' for non-URLLC devices (left) and URLLC devices (right).

Observe from Figure 5-13 that the CI for the 99.99th end-to-end delay percentile for URLLC devices for the 'Two-Step RA procedure with Repetitions' is 7 ms wide and considerably wider than that for the 'Two-Step RA procedure with Feedback'. Due to this uncertainty, it can be concluded that the 'Two-Step RA procedure with Repetitions' can only to some degree achieve the performance of the 'Two-Step RA procedure with Feedback'. It is worth mentioning that the uncertainty introduced for the 'Two-Step RA procedure with Repetitions', is correlated with the number of attempts to successfully transmit the URLLC UL data, which is jointly influenced by the preamble and data collision probability. Specifically, Table 5-5 illustrates that the probability of a URLLC device having one or two attempts is 99.9187% with the 'Two-Step RA procedure with Repetitions' while it increases to 99.9964% with the 'Two-Step RA procedure with Feedback'.

Table 5-5 - Probability of access attempts required for URLLC devices for the 'Two-Step RA procedure with Repetitions' and for the 'Two-Step RA procedure with Feedback', while reserving 35 preambles for URLLC transmissions.

Number of Access Attempts	URLLC devices with split 29/35	
	Repetitions	**Feedback**
One	99.6780%	99.6964%
Two	0.2407%	0.3000%
Three	0.0800%	0.0036%
Outage	0.0013%	0.0000%

Additionally, the 99.99th end-to-end delay percentile for URLLC devices for the 'Two-Step RA procedure with Feedback' is reduced by around 1 ms to 19.9 ms compared to the 'Two-Step RA procedure with Repetitions'. This outcome implies that for most URLLC devices, the UL data are received successfully at the base station with no repetitions being necessary while only for few devices a repetition was necessary.

It is also noted that the end-to-end delay for non-URLLC devices is the same between the two-step RA procedure with repetitions and with feedback, as also illustrated in Figure 5-13.

5.7 Comparison of RA procedures

In this section, an overall comparison of the reference 'Cat-M1' procedure, 'EDT based on 3GPP Release 15' procedure, 'EDT for All Devices' procedure, 'Two-Step RA procedure with Repetitions' and 'Two-step RA procedure with Feedback' is presented. The comparison is performed in respect to the end-to-end delay in Section 5.7.1 and the UL resource utilization in Section 5.7.2.

5.7.1 End-to-end delay

In Section 1.2 the requirements for this study were discussed and it was derived that the target values for the 99.9th end-to-end delay percentile for non-URLLC and the 99.99th end-to-end delay percentile for URLLC devices are 50 ms and 10 ms respectively. Table 5-6 presents the calculated values for these percentiles in respect to the studied procedures and it is derived that the best procedure is the 'Two-Step RA procedure with Feedback' as it provides the lowest end-to-end delays.

Table 5-6 - 99.9th and 99.99th end-to-end delay percentile for non-URLLC and URLLC devices respectively for PRACH periodicity two.

	Non-URLLC delay	URLLC delay
Cat-M1	66.0 ms	80.1 ms
EDT based on 3GPP Release 15	67.9 ms	44.0 ms
EDT for All Devices	43.7 ms	34.3 ms
Two-Step RA procedure with Repetitions	43.6 ms	20.8 ms
Two-Step RA procedure with Feedback	43.7 ms	19.9 ms

From Table 5-6 it can be deduced that the requirement set for non-URLLC devices can be met with the 'EDT for All Devices' procedure, the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback'. Initially, this end-to-end delay was 66.0 ms for 'Cat-M1' and it was slightly increased with the 'EDT based on 3GPP Release 15' procedure due to the preamble split that was used. Then, this delay was reduced to 43.7 ms with the 'EDT for All Devices' procedure as this particular procedure implies earlier data transmission for both non-URLLC and URLLC devices. The 'Two-Step RA procedure with Repetitions' and the 'Two-step RA procedure with Feedback' do not provide any improvement to the delay for non-URLLC devices as these procedures are focused on improving the delay of URLLC transmissions.

Regarding the URLLC devices, 'Cat-M1' offers an end-to-end delay of 80.1 ms which improves to 43.0 ms with the 'EDT based on 3GPP Release 15' procedure, as it allows devices to transmit their data sooner than before. Further enhancements were applied by the 'EDT for All Devices' procedure compared to the 'EDT based on 3GPP Release 15' which offer to URLLC devices a reduced end-to-end delay of 34.3 ms (mainly due to the applied shorter RAR window). The 'Two-Step RA procedure with Repetitions' enables URLLC devices to transmit their UL data even sooner than the EDT procedures and thus the end-to-end delay can be further decreased to 20.8 ms. Finally, the 'Two-Step RA procedure with Feedback' can reduce the end-to-end delay to 19.9 ms as the URLLC UL data are transmitted successfully during the first transmissions and there is no need for repetitions. Even though the 'Two-Step RA procedure with Feedback' offers a gain of 60.2 ms, the requirement of 10 ms cannot be met.

Table 5-7 presents for all studied procedures the estimated percentage of non-URLLC and URLLC devices that meet the end-to-end delay requirements of 50 ms for the 99.9th end-to-end delay percentile and 10 ms for the 99.99th end-to-end delay percentile, respectively. The table further presents for each procedure the absolute minimum end-to-end delay that can be achieved for non-URLLC and URLLC devices. This absolute minimum end-to-end delay is basically the delay that can be achieved with the corresponding technology, provided that the network is planned accordingly. From Table 5-7 observe that for non-URLLC devices, the 50 ms requirement can actually be achieved with 99.97% reliability in the considered scenarios, and that for both the reference RA procedure in Cat-M1 and all the other studied RA procedures, can indeed achieve a minimum end-to-end delay of less than 50 ms. On the other hand, for URLLC devices, only the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback' can offer a minimum delay below 10 ms. This is also reflected in the percentage of devices that experience less than 10 ms end-to-end delay as for 'Cat-M1', 'EDT based on 3GPP Release 15' and 'EDT for All Devices', that percentage is 0.00%. Furthermore, it is also observed that the 10 ms requirement is achieved by more devices for the 'Two-Step RA procedure with Feedback' than for the 'Two-Step RA procedure with Repetitions'. Finally, it can be concluded that only the 'Two-Step RA Procedure with Repetitions' and the 'Two-Step RA procedure with Feedback' can offer the possibility to achieve the requirements set in this study for the FoF, provided of course that the network is planned accordingly. An example of a different network plan such that the end-to-end delay requirements will be achieved could be a different network layout i.e. using two base stations.

Table 5-7 - Percentage of devices that meet the requirements set regarding end-to-end delay and the minimum end-to-end delay achieved in each procedure studied.

	Non-URLLC		URLLC	
	Percentage of devices	**Minimum delay**	**Percentage of devices**	**Minimum delay**
Cat-M1	99.66%	37.00 ms	0.00%	37.00 ms
EDT based on 3GPP Release 15	99.25%	37.00 ms	0.00%	14.00 ms
EDT for All Devices	99.96%	14.00 ms	0.00%	14.00 ms
Two-Step RA procedure with Repetitions	99.97%	14.00 ms	99.48%	5.00 ms
Two-Step RA procedure with Feedback	99.97%	14.00 ms	99.67%	5.00 ms

Considering that only the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback' can offer the possibility to achieve the end-to-end delay requirements set to 10 ms in this study for FoF for ca. 99.5% of the URLLC devices, a further analysis is carried out for both procedures to derive a better network planning which can possibly lead to achieving the end-to-end delay requirements. It is noted that the analysis is carried out on both procedures and not only on the best performing 'Two-Step RA procedure with Feedback' as for this particular procedure there is an assumption that the base station can transmit 90 bits on the MPDCCH, 2 ms after the detection of a preamble used by a URLLC device which might not be feasible in reality. For both the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback', it can be derived that the requirement of 10 ms for the 99.99th end-to-end delay percentile for URLLC devices is not achieved, which is primarily due to preamble collisions.[18] In order to reduce the probability of preamble collision even further, more PRACH opportunities need to be offered to the devices (see Section 4.5.2 for the PRACH initial configuration), which implies that a different PRACH periodicity should be applied. The PRACH periodicity investigation and its impact on the end-to-end delay and UL resource utilization is presented in Section 5.8.

5.7.2 UL resource utilization

In Section 5.7.1 it was discussed that the best RA procedure in terms of end-to-end delay is the 'Two-Step RA procedure with Feedback'. However, an evaluation regarding the UL resource utilization needs to be carried out in order to investigate whether the gain of end-to-end delay comes at a reasonable cost of UL resource utilization. Table 5-8 presents UL resource utilization for the studied RA procedures and it can be derived that there is no significant difference regarding the UL resource utilization between the RA procedures.

Table 5-8 - Uplink resource utilization for each studied RA procedure for PRACH periodicity two.

	UL resource utilization
Cat-M1	6.16%
EDT based on 3GPP Release 15	6.19%
EDT for All Devices	6.00%
Two-Step RA procedure with Repetitions	6.45%
Two-Step RA procedure with Feedback	6.11%

More specifically, and based on Table 5-8, the 'EDT based on 3GPP Release 15' procedure increases the UL resource utilization only slightly compared to the 'Cat-M1' procedure. This slight increase is caused by the *edt-TBS* value as there are UL resources reserved for URLLC devices that are not always fully used. The 'EDT for All Devices' procedure improves the UL resource utilization from 6.19% (in 'EDT based on 3GPP Release 15') to 6.00%. This improvement relies on the fact that non-URLLC devices are transmitting their data earlier than before and thus less messages need to be exchanged between the device and the base station, as also explained in Section 3.1.3. However, this reduction of message exchange and thus of UL resource utilization implies that the devices are never reaching the RRC CONNECTED Mode and that they are always in RRC IDLE Mode. For the 'Two-Step RA procedure with Repetitions' the UL resource utilization is 6.45% which is slightly higher than the other RA procedures. This is expected for this procedure as all the UL resources on the 5 MHz carrier are reserved for three consecutive subframes for URLLC transmissions, every time

[18] In Section 5.6 it was presented that for the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback', a URLLC device will use one or two access attempts 99.9187% and 99.9964% of the time, respectively. Devices which used two access attempts have experienced a preamble collision.

that a preamble reserved for URLLC transmissions is detected at the base station. The UL resource utilization for the ‘Two-Step RA procedure with Feedback’ is 6.11% which is less than the utilization in the ‘Two-Step RA procedure with Repetitions’ as the UL resources are now reserved only for one subframe instead of three. However, it has to be noted that the ‘Two-Step RA procedure with Feedback’ increases the resource utilization on the downlink, compared to the rest procedures, as the base station needs to transmit 90 bits on the MPDCCH every time that a URLLC transmission is expected. Finally, it can also be said that all values of the UL resource utilization for all procedures except the ‘Two-Step RA procedure with Repetitions’ are very similar and thus their differences might fall within the margins of simulation inaccuracy.

5.8 PRACH periodicity investigation

As discussed in Section 5.7.1, the procedure that is performing the best is the ‘Two-step RA procedure with Feedback’ due to the lowest achieved median, 99.9th and 99.99th end-to-end delay performance. However, it was also noted that the requirement set to 10 ms for URLLC devices cannot be met in the studied scenario.

Considering that the key reason for the delay is the occurrence of preamble collisions, for further reduction of the 99.99th delay percentile for URLLC transmissions in the ‘Two-Step RA procedure with Feedback’, more PRACH resources should be assigned. Consequently, the probability of preamble collision will be reduced further, which in turn will reduce the 99.99th end-to-end delay percentile. In order to increase the PRACH resources, a higher PRACH periodicity should be applied to the network, which will also increase the UL resource utilization. Therefore, it is again concluded that there is a trade-off between the end-to-end delay and the UL resource utilization.

Recall from Section 4.5.2 that the PRACH periodicity was chosen to be equal to two and therefore two PRACH subframes are available in a 10 ms radio frame. This PRACH periodicity can be increased to three, five or ten, and thus there can be three, five or ten PRACH subframes in a 10 ms radio frame, respectively. For the ‘Two-step RA procedure with Feedback’, an analysis is carried out to assess the influence on the end-to-end delay when increasing the PRACH periodicity to three, five and ten. Recall that the implementation of the ‘Two-Step RA procedure with Feedback’ is based on the assumption that the base station can indeed transmit 90 bits of feedback on the MPDCCH, 2 ms after the detection of a preamble used by a URLLC device. As the validity (or achievability) of this assumption has not been investigated in detail, the ‘Two-Step RA procedure with Repetitions’ is also studied under different PRACH periodicities, as it can provide similar end-to-end delays as the ‘Two-Step RA procedure with Feedback’ but with a simpler implementation. The bottom-line results are presented in Figure 5-14, nothing that the intermediate results regarding the optimization of the preamble split and the number of repetitions are discussed in Appendix B . It is noted that the results obtained for both procedures are very similar and thus in Figure 5-14 there are many results that are overlapping.

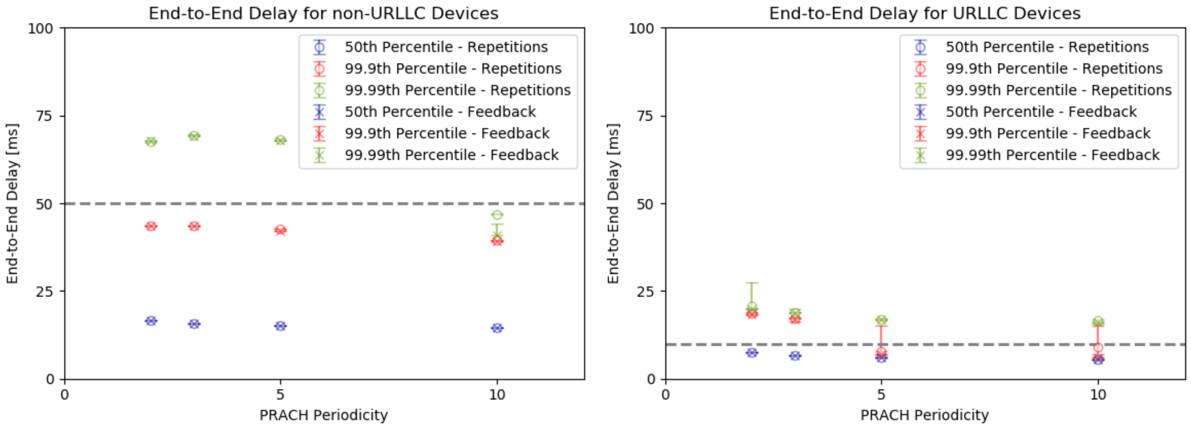


Figure 5-14 - End-to-end delay based on different PRACH periodicity for the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback' for non-URLLC devices (left) and URLLC devices (right).

From Figure 5-14, it is observed that the end-to-end delay for non-URLLC devices is almost the same for the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback' and thus the same reasoning applies for both procedures. From Figure 5-14 figure it can also be observed that the 99.9th end-to-end delay percentile reduces when the PRACH periodicity increases, as expected. Moreover, the 99.99th end-to-end delay does not vary significantly for PRACH periodicities two, three and five, but does improve significantly for a PRACH periodicity of ten, where it drops to 40.99 ms. Actually, it appears that the 99.99th end-to-end delay percentile for PRACH periodicity two, three and five have some relatively small differences which might be caused by the different number of preambles that are available for non-URLLC transmissions among the different PRACH periodicities (see Appendix B for the preamble analysis). For example, increasing the PRACH periodicity from two to five does not provide much end-to-end delay gain as the available preambles for non-URLLC transmissions are reduced from 29 to 14. The significant drop of the 99.99th end-to-end delay percentile for a PRACH periodicity of ten can be explained by the significant reduction of the collision probability, which allows more devices to transmit successfully their data with fewer tries and hence reduces the end-to-end delay. More specifically, and as illustrated in Table 5-9, the probability of a non-URLLC device having one or two access attempts is 99.9575% when PRACH periodicity five is used while it increases to 99.9955% with PRACH periodicity ten. Another explanation for the decrease of the 99.99th delay percentile for non-URLLC devices, is the reduction of the time that a device has to wait for a PRACH opportunity. Finally, it is noted that the requirement of 50 ms with 99.9% reliability can be met for all PRACH periodicities.

For the delay performance of URLLC devices the results for the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback' are almost the same with the only difference that the delays corresponding to the 'Two-Step RA procedure with Repetitions' have wide CI. The reasoning of the wide CI is similar to what was already explained in Section 5.6 and therefore it is omitted here. The reasoning of the delay performance in respect to the PRACH periodicity is given only for the 'Two-Step RA procedure with Feedback' as it is similar for both procedures. Figure 5-14 presents (for URLLC devices) that PRACH periodicities five and ten influence the 99.9th delay percentile and reduce it to 7.00 ms and 6.00 ms respectively, as more devices transmit their data with only a single attempt. More specifically, the corresponding probabilities of having just one attempt are

99.9195% and 99.9000% for PRACH periodicities five and ten respectively (according to Table 5-9). Regarding the 99.99th delay percentile for URLLC devices, its value is reduced from 19.92 ms (with PRACH periodicity two) to 15.94 ms with PRACH periodicity ten. However, the target of 10 ms is still not met.

Table 5-9 - Probability of access attempts required for non-URLLC and URLLC devices for the 'Two-Step RA Procedure with Feedback' for PRACH periodicities five and ten.

Number of Access Attempts	PRACH periodicity Five		PRACH periodicity Ten	
	Non-URLLC	URLLC	Non-URLLC	URLLC
One	99.3660%	99.9195%	99.3915%	99.9000%
Two	0.5915%	0.0805%	0.6040%	0.0956%
Three	0.0385%	0.0000%	0.0032%	0.0044%
Outage	0.0040%	0.0000%	0.0013%	0.0000%

Table 5-10 presents for the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA with Feedback' the percentage of devices that met the end-to-end delay requirements of 50 ms and 10 ms for non-URLLC and URLLC traffic, respectively. It is observed that for both procedures, the requirement for non-URLLC traffic is met with more than 99.9% reliability. It is however noted that this percentage non-monotonously varies over the different PRACH periodicities rather than continuously increasing, which is due to the different number of available preambles that are available for non-URLLC transmissions for the different PRACH periodicities. For URLLC devices, the percentage of devices that experience an end-to-end delay at most 10 ms nicely increases with an increase of the PRACH periodicity as expected, reaching 99.92% for PRACH periodicities five and ten for the 'Two-Step RA procedure with Feedback'. Recall that different numbers of preambles are reserved for the URLLC devices for the different PRACH periodicities. Simulation inaccuracies might also affect these percentages to some degree, but they are noted to be in accordance with the end-to-end delays presented in Figure 5-14.

The minimum achievable delays that can be achieved by the two procedures are also presented in Table 5-10, which inherently assume ideal circumstances in terms of zero load and a strong channel and hence are insensitive to the PRACH periodicities and the same for both procedures. This outcome is reasonable as the minimum delay is a feature of the technology and not of the network configuration and specifically, a good network configuration can possibly allow the achievement of the minimum delays provided by the technology. Based on this observation it can be concluded that both the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA with Feedback' are promising technologies as their minimum delay (5 ms) is significantly lower than the URLLC end-to-end delay requirement (10 ms). Considering the above, it can be further concluded that the requirement of 10 ms end-to-end delay with 99.99% reliability for URLLC devices can possibly be achieved under a different network layout, e.g. introducing another base station and thereby both enhancing link budgets and reducing cell loads. Recommendations on improving further the end-to-end delay are presented in Section 6.2.

Table 5-10 - Percentage of devices that meet the requirements set regarding end-to-end delay and the minimum end-to-end delay achieved in each PRACH periodicity studied for the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA with Feedback'.

PRACH periodicity	Two-Step RA procedure with Repetitions				Two-Step RA procedure with Feedback			
	Non-URLLC		URLLC		Non-URLLC		URLLC	
	Percentage	Minimum Delay	Percentage	Minimum Delay	Percentage	Minimum Delay	Percentage	Minimum Delay
2	99.97%	14.00 ms	99.48%	5.00 ms	99.97%	14.00 ms	99.67%	5.00 ms
3	99.98%	14.00 ms	99.75%	5.00 ms	99.98%	14.00 ms	99.80%	5.00 ms
5	99.95%	14.00 ms	99.90%	5.00 ms	99.96%	14.00 ms	99.92%	5.00 ms
10	99.99%	14.00 ms	99.91%	5.00 ms	99.99%	14.00 ms	99.92%	5.00 ms

Additionally, the trade-off between the end-to-end delay and the UL resource utilization for the two procedures is presented in Figure 5-15. As the delay results shown for both procedures are very similar, we limit our discussion to the results of the 'Two-Step RA procedure with Feedback', noting that an equivalent argumentation applies to the 'Two-Step RA procedure with Repetitions'. Also, this is the reason why the results in Figure 5-15 largely overlap for the two procedures.

The figure is derived based on the end-to-end delay and UL resource utilization for each PRACH periodicity and it is visible that the delay for URLLC devices reduces with an increase of the UL resource utilization (thus with a higher PRACH periodicity). It can also be derived that the end-to-end delay does not improve significantly from PRACH periodicity five to PRACH periodicity ten while the increase of the UL resource utilization between these two periodicities is significant (1.88 times higher). Table 5-11 presents a summary of the results. From the above, it cannot easily concluded which is the best configuration and RA procedure for the considered scenario with 3000 devices as there is a strong tradeoff between the end-to-end delay and the UL resource utilization for both the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback', as also presented in Table 5-11. Therefore, the choice of the PRACH periodicity value is left to the operator while the choice for the RA procedure is left to the feasibility of implementing the 'Two-Step RA procedure with Feedback'. Section 5.9 presents the performance analysis for different number of devices (i.e. network load) for both RA procedures and for PRACH periodicities five and ten.

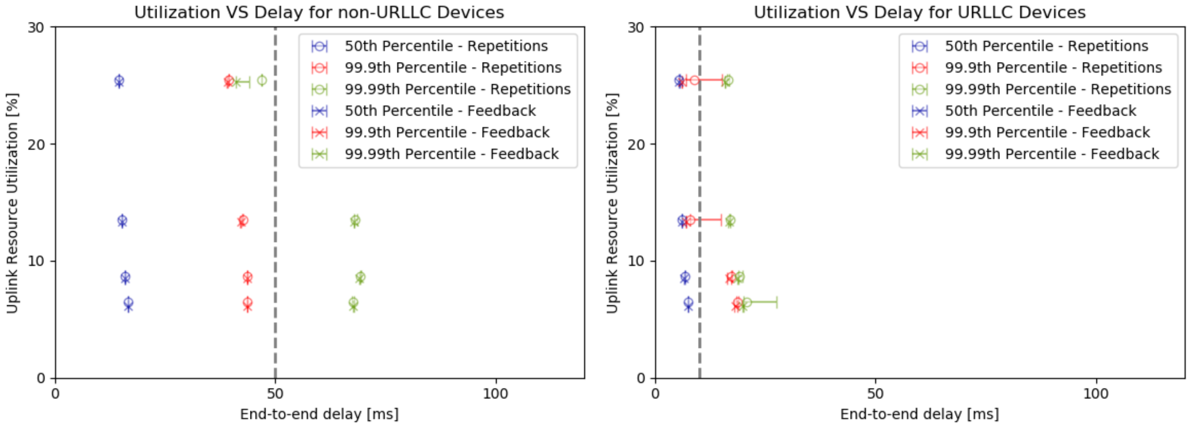


Figure 5-15 - Trade-off between the end-to-end delay and the UL resource utilization for different PRACH periodicities for the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback' for non-URLLC devices (left) and URLLC devices (right).

Table 5-11 - 99.9th and 99.99th end-to-end delay percentile for non-URLLC and URLLC devices, respectively, and UL resource utilization for the different PRACH periodicities for the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback'.

	Two-Step RA procedure with Repetitions			Two-Step RA procedure with Feedback		
PRACH periodicity	**Non-URLLC delay**	**URLLC delay**	**UL resource utilization**	**Non-URLLC delay**	**URLLC delay**	**UL resource utilization**
2	43.6 ms	20.8 ms	6.5%	43.7 ms	19.9 ms	6.1%
3	43.6 ms	19.1 ms	8.7%	43.6 ms	18.8 ms	8.5%
5	42.7 ms	17.0 ms	13.5%	42.1 ms	16.8 ms	13.3%
10	39.5 ms	16.6 ms	25.5%	39.1 ms	15.9 ms	25.3%

5.9 Sensitivity analysis w.r.t. the network load

In Section 5.8 the achieved end-to-end delay performance under the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA with Feedback' was presented for different PRACH periodicities. All results presented so far assumed a network load generated by 3000 devices, with 5% (150) of them URLLC devices. In this section a sensitivity analysis is presented of the delay performance w.r.t. different network loads. For this study, the PRACH periodicity five and ten are considered only as they are the configurations which perform the best regarding end-to-end delay for both the two above-mentioned procedures, assuming 3000 devices. Furthermore, it is noted that for this load analysis the preamble split found for 3000 devices will be used (see Appendix B) which indicates that 50 and 25 preambles will be reserved for URLLC transmissions for PRACH periodicity five and PRACH periodicity ten respectively. It was decided to keep the same preamble split among the different loads in order to study the influence of solely the load on the end-to-end delay. It is noted however that for an optimal configuration of the network, it is recommended that the preamble split should be derived based on the network load.

Figure 5-16 and Figure 5-17 present the results of the load analysis for PRACH periodicities five and ten for the 'Two-Step RA procedure with Repetitions' and for the 'Two-Step RA procedure with Feedback', respectively. The end-to-end delay from both figures is similar (with the exception that the 'Two-Step RA procedure with Repetitions' performs slightly worse than the 'Two-Step RA procedure with Feedback' for the reasons already explained) and therefore the argumentation will only be given for the 'Two-Step RA procedure with Feedback'.

From Figure 5-17 it can be concluded that PRACH periodicity five offers better end-to-end delays for URLLC devices in higher loads (i.e. more than 10000 devices) compared to PRACH periodicity ten which is reasonable as more preambles are available for URLLC devices for PRACH periodicity five. More specifically, the 99.99th end-to-end delay percentile for URLLC devices stays below 25 ms for up to 15000 devices for PRACH periodicity five, while for PRACH periodicity ten, the same percentile increases to around 25 ms with 10000 devices. However, it is noted that the requirement of 10 ms cannot be achieved even with just 1000 devices in the network (i.e. 950 non-URLLC and 50 URLLC devices), regardless of the PRACH periodicity. Additionally, the improved performance of URLLC devices under high loads and PRACH periodicity five, comes at a cost of an increased end-to-end delay for non-URLLC devices, when compared to PRACH periodicity ten. Specifically, for more than 10000 devices in the network, the 99.9th end-to-end delay percentile for non-URLLC devices exceeds the 50 ms target with PRACH periodicity five, while for PRACH periodicity ten the target is still met.[19]

For low to medium loads, both PRACH periodicity five and ten provide similar end-to-end delays for URLLC devices. However, for non-URLLC devices, the end-to-end delay is lower when PRACH periodicity ten is used instead of five. Specifically, with PRACH periodicity ten the 99.99th end-to-end delay percentile is lower than 50 ms for up to 6000 devices while with PRACH periodicity five it is around 20 ms higher. It is noted that the 50 ms target for the 99.9th end-to-end delay percentile for up to 6000 devices can be achieved with both PRACH periodicities.

It is therefore clear, that there is a strong trade-off between the performance of the non-URLLC and URLLC devices between the two PRACH periodicities. This tradeoff can be easily explained by the fact that more preambles are reserved for URLLC devices when PRACH periodicity five used than PRACH periodicity ten (see Appendix B for the preamble analysis). Thus, for low to medium loads the URLLC performance is similar for both PRACH periodicities as there is no bottleneck on the preamble collision probability but for higher loads PRACH periodicity five performs better as the preamble collision probability is lower than with PRACH periodicity ten. Similarly, for non-URLLC devices, there are fewer preambles available with PRACH periodicity five compared to PRACH periodicity ten and this is reflected on the 99.99th end-to-end delay percentile for up to 6000 devices and also on the 99.9th end-to-end delay percentile for more than 10000 devices.

It can be finally concluded that PRACH periodicity five is recommended for networks with high loads under the condition that the 50 ms end-to-end delay requirement for non-URLLC devices can be relaxed. For applications that this requirement cannot be relaxed, PRACH periodicity ten is recommended.

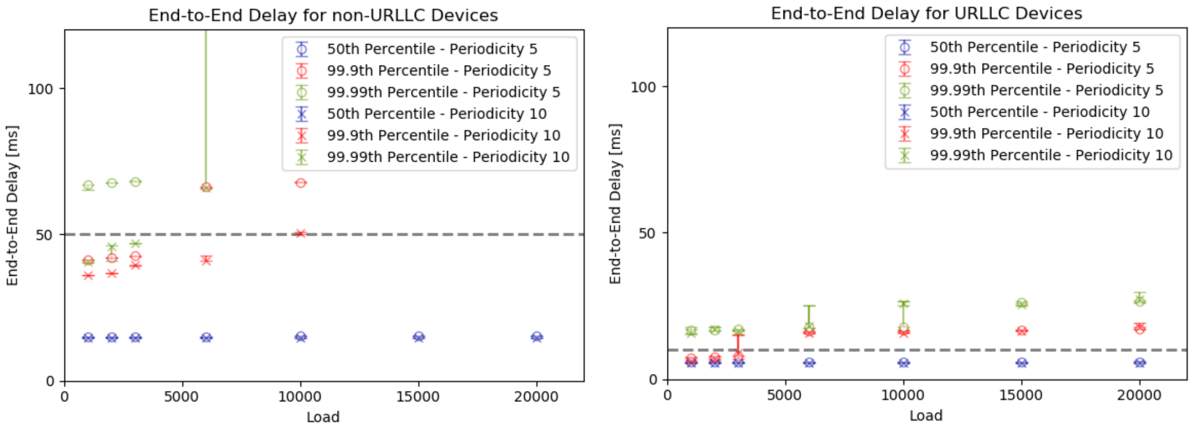


Figure 5-16 - End-to-end delay for non-URLLC (left) and URLLC (right) devices in 'Two-Step RA procedure with Repetitions' and PRACH periodicity five and ten.

[19] The end-to-end delays not shown in Figure 5-16 and Figure 5-17 imply that they are infinite and that the devices are in outage.

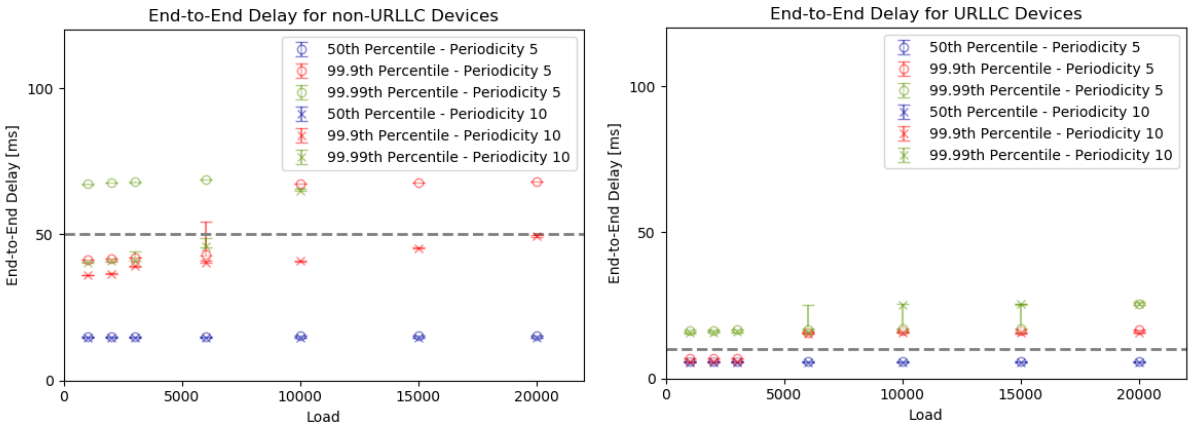


Figure 5-17 - End-to-end delay for non-URLLC (left) and URLLC (right) devices in 'Two-Step RA procedure with Feedback' and PRACH periodicity five and ten.

Furthermore, Figure 5-18 presents the UL resource utilization in respect to the load for the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback' and PRACH periodicities five and ten. First of all, it is observed that the UL resource utilization is increasing with an increase of the load, as expected, as more devices need to transmit their data.

Moreover, it can be derived that the UL resource utilization between the two procedures when using the same PRACH periodicity is similar. As already discussed in Section 5.7.2, the 'Two-Step RA procedure with Repetitions' has a slightly higher UL resource utilization than the 'Two-Step RA procedure with Feedback', as more consecutive subframes have to be reserved for the URLLC UL data transmissions. From Figure 5-18 it can be observed that this difference in resource utilization increases with an increase of the load. This is reasonable as under higher loads the number of URLLC devices (i.e. 5% of the total load) also increases, and hence so do the number of URLLC transmissions and, consequently, the UL resources reserved for URLLC transmissions. Furthermore, the PRACH periodicity influences the UL resource utilization as a higher PRACH periodicity implies the reservation of more UL resources for the PRACH.

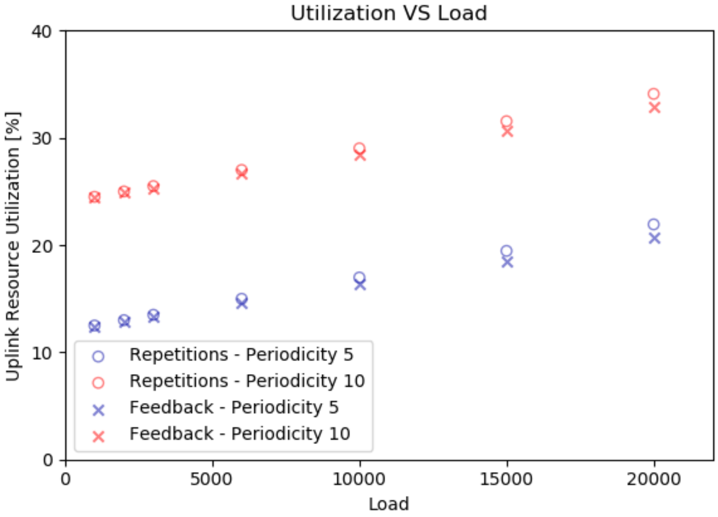


Figure 5-18 - UL resource utilization in respect to the load in the network for the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback' and PRACH periodicities five and ten.

5.10 Discussion on data sizes

This section presents a short discussion on the data sizes used in this study as they can influence significantly the performance of the procedures regarding both the end-to-end delay and the UL resource utilization. Recall that the assumed data sizes are 200 bytes for non-URLLC devices and 40 bytes for URLLC devices.

Recall further from Chapter 3 that the parameters (i.e. *edt-TBS*) used for the configuration of the EDT and two-step RA procedures were based on the data sizes assumed in this particular study. For example, for the two-step RA procedure, the UL resources were split to nine non-overlapping sets of two PRBs each as two PRBs can transport 40 bytes (with MCS 15). Therefore, shorter URLLC data imply that the UL resources can be split to more than nine sets. Specifically, for data sizes that can be transmitted in one PRB, eighteen sets of non-overlapping PRBs can be defined in the UL channel for URLLC transmissions which will improve the experienced end-to-end delay for URLLC devices as the collision probability of the data will be decreased significantly. Additionally, the UL resource utilization can possibly be decreased for the 'Two-Step RA procedure with Repetitions' as possibly fewer repetitions will be needed. On the other hand, a higher URLLC data size implies the opposite behavior and thus higher end-to-end delays and possibly higher UL resource utilization.

The 'EDT for All Devices' procedure was the procedure that introduced improvements for the non-URLLC devices, as it allows them to transmit their data earlier than normally. For this configuration, it was defined that the base station will be reserving six PRBs for both non-URLLC and URLLC devices such that the TBS will be big enough to transport the 200 bytes of non-URLLC devices (with MCS 15). Therefore, for a smaller non-URLLC data size, *edt-TBS* would have been smaller which implies that more devices would have been able to be granted access to the UL resources simultaneously. Consequently, the end-to-end delay for both non-URLLC and URLLC would be reduced as there would have been no extra waiting time for access to the UL resources. Additionally, the UL resource utilization would possibly decrease as the *edt-TBS* would be smaller. The opposite behavior is expected when the non-URLLC message size is larger than 200 bytes. Specifically, non-URLLC devices would not have been able to transmit the entire data in one TBS and they would have had to perform two data transmissions to fully convey the data.[20] This behavior would have led to significantly higher end-to-end delays for non-URLLC devices and possible higher UL resource utilization as more signaling would have been needed for the complete and successful transmission of the non-URLLC data.

5.11 Summary

This chapter presented the assessment of the different RA procedures as well as the effect of the PRACH periodicity and the network's load on the end-to-end delay for the best RA procedure among those considered in this thesis, which is the 'Two-Step RA procedure'. For the 'Two-Step RA procedure', both the 'Repetitions' and 'Feedback' flavors were studied. The analysis showed that the 'Repetitions' flavor does not perform as well as the 'Feedback'

[20] Recall that only six PRBs can be assigned to each device on the PUSCH and that no more than 200 bytes can be transmitted in six PRBs with the maximum MCS (MCS 15).

flavor, considering the higher degree of uncertainty visible in the derived CI. However, the analysis of the PRACH periodicity and load was carried out for both procedures as the 'Two-Step RA procedure with Feedback' can only be implemented under the assumption that a feedback message of 90 bits can be transmitted on the MPDCCH, 2 ms after a URLLC preamble transmission is detected at the base station; an assumption which we have been unable to verify at this stage. Table 5-12 presents an overview of the trade-offs that were observed in this study.

Table 5-12 - Trade-offs observed.

	Positive effects	Negative effects
High number of URLLC preambles	1. Decrease the probability of URLLC preamble collisions on PRACH. 2. Decrease the URLLC 99.99th end-to-end delay.	1. Increase the probability of URLLC data transmission collisions on PUSCH. 2. Increase the probability of collisions for non-URLLC devices
High PRACH periodicity	1. Decrease of the URLLC 99.99th end-to-end delay.	1. Increase the UL resource utilization

Section 5.7 provided an overall comparison of the RA procedures in a *reference* scenario, regarding end-to-end delay and UL resource utilization. It was derived that the 'Two-Step RA procedure with Feedback' can provide the lowest end-to-end delays, as in the considered scenario 43.7 ms and 19.9 ms can be achieved for the 99.9th end-to-end delay percentile for non-URLLC devices and the 99.99th end-to-end delay percentile for URLLC devices, respectively. However, the requirement of 10.0 ms for the 99.99th end-to-end delay percentile for URLLC devices is not met as only 99.67% of the devices have an end-to-end delay up to 10.0 ms.

While the above performance results were based on an assumed PRACH periodicity of two, in Section 5.8, the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback' were evaluated for different *PRACH periodicities* in order to examine whether the 10 ms requirement for URLLC devices can be met by adding more RA opportunities. Table 5-11 presented the obtained results. From these results, it can be concluded that none of the PRACH periodicities fulfill this particular requirement for the studied scenario, even though some delay enhancements have been observed. More specifically, the lowest 99.99th end-to-end delay percentile calculated for URLLC devices is 15.9 ms with PRACH periodicity ten which introduces a gain of 4.0 ms compared to PRACH periodicity two. However, this delay gain comes at the cost of increasing the UL resource utilization from 6.1% to 25.3% as the PRACH is then supported in every subframe. Due to this strong trade-off between the end-to-end delay and the UL resource utilization, it was concluded that is up to the operator to define which is the best configuration for the PRACH periodicity. Finally, it was also discussed that even though the target of 10 ms with 99.99% reliability for URLLC devices was not met, both the 'Two-Step RA procedure with Repetitions' and the 'Two-Step RA procedure with Feedback' can offer a *minimum* attainable end-to-end delay of 5 ms and thus with a different network planning this specific requirement can possibly be met e.g. by adding base station sites in order to enhance the link budget and reduce the effective cell loads.

An analysis on the *network load* was carried out in Section 5.9 in order to study the behavior of the two enhancements of the 'Two-Step RA procedure'. From this analysis it was derived that both PRACH periodicity five and ten can provide similar results for low loads in the network but for higher loads there is a trade-off between the end-to-end delay for non-URLLC and URLLC devices. In case of applications for which the end-to-end delay for non-URLLC devices can be sacrificed for an improved end-to-end delay for URLLC devices, PRACH periodicity five is recommended as it can support a higher number of URLLC devices. However, for applications that where non-URLLC and URLLC end-to-end delays are critical, each with their own noted requirement level, PRACH periodicity ten is recommended, coming at an obvious cost of higher UL resource utilization.

Finally, in Section 5.10 the impact of the non-URLLC and URLLC data sizes on the end-to-end delay and UL resource utilization was briefly discussed, noting that the size of data can influence significantly the performance of the studied RA procedures.

Chapter 6 Concluding remarks

In this study, we analyzed the end-to-end delay performance of small data transmissions in the context of massive IoT type applications in a Factory of the Future (FoF) deployment scenario. For the evaluation of the end-to-end delay, the focus was on the Random Access (RA) procedure as it contributes most significantly to this delay. Therefore, an evaluation of the RA procedure was carried out for the reference Cat-M1 access technology and it was quantified that the end-to-end delay experienced by both non-URLLC and URLLC traffic is significantly higher than the assumed requirements for FoF applications. Specifically, the 99.99th end-to-end delay percentile for URLLC traffic was found to be approximately 80 ms in the considered baseline scenario, while the assumed requirement is to have it below 10 ms. Specifically, the requirements for FoF applications is 50 ms for the 99.9th percentile for non-URLLC traffic and 10 ms for the 99.99th percentile for URLLC traffic, as shown in Section 1.2.

With an aim to improve the end-to-end delay performance in line with the requirements set for 5G applications, four different *enhanced* RA procedures were proposed and assessed, three of which are contributions of this study. The four enhancements studied in this thesis follow, along with their key enhancements.

- **Early Data Transmission (EDT) based on 3GPP Release 15:** This enhanced RA procedure is a contribution to 3GPP and allows URLLC data to be appended to the RA procedure signaling (appended to Msg.3) [17].
- **EDT for All Devices:** This RA procedure was introduced in this study as an enhancement to the 'EDT based on 3GPP Release 15' procedure in which non-URLLC data can be appended to the RA procedure signaling (appended to Msg.3) while furthermore a different configuration is applied to URLLC traffic (i.e. priority on resources and shorter RAR window).
- **Two-Step RA procedure with Repetitions:** This RA procedure is one possible implementation of the basic idea presented in [18]. The key principle is to allow consecutive URLLC data transmissions, 3 ms after the transmission of the preamble.
- **Two-Step RA procedure with Feedback:** This RA procedure is another possible implementation of the basic idea presented in [18]. For this implementation, feedback about the UL resources is given to devices handling URLLC traffic by the base station before the actual URLLC data transmission happens.

The main conclusions derived from the analysis are presented in Section 6.1 while Section 6.2 presents recommendations and future work that can be pursued as an extension to the presented work. It is noted that the derived conclusions are based on simulations in a factory of size 50 m × 50 m, where 95% of the devices generate non-URLLC traffic that follows a uniform traffic arrival distribution within a 60 seconds interval while the remaining 5% of the devices generate URLLC traffic that follows a beta traffic arrival distribution within a 10 seconds interval, triggered by an unexpected incident.

6.1 Conclusions

Based on the analysis described in Chapter 5, the main conclusions are as follows, where the end-to-end delay performance experienced by the non-URLLC devices is characterized by the 99.9th percentile, while for the URLLC devices it is given by the 99.99th percentile:

- **Cat-M1**

  The end-to-end delay performance experienced for the RA procedure in Cat-M1 was found to be too unsatisfactorily high for the considered non-URLLC and URLLC traffic. This relies on the fact that the measured end-to-end delays for non-URLLC and URLLC devices are 66.0 ms and 80.1 ms while the considered 5G application requirements are 50 ms and 10 ms, respectively.

- **EDT procedures**

  In 3GPP Release 15, the EDT procedure is introduced with the aim to reduce the end-to-end delays experienced in the network. However, even though the URLLC end-to-end delay is significantly improved (44.0 ms versus 80.1 ms for Cat-M1) the quantified end-to-end delays for both non-URLLC (e.g. 66.9 ms) and URLLC devices are still significantly higher than the delay requirements. Further enhancements have been studied for the EDT procedure ('EDT for All Devices') which yielded end-to-end delay values of 43.7 ms and 34.3 ms for non-URLLC and URLLC devices, respectively. It is therefore concluded that the EDT procedures cannot provide the delays required for the considered FoF applications, although it can provide significant delay improvements.

  Additionally, the EDT procedures introduce the need to distinguish non-URLLC and URLLC devices at the base station and therefore the preambles are required to be split into two groups; one for non-URLLC and one for URLLC. This split of preambles, however, introduces a trade-off between the end-to-end delay for non-URLLC and URLLC devices as reserving too many preambles for one group can introduce high delays to the other group. For this reason, a preamble analysis was carried out in all procedures that required a distinguish between non-URLLC and URLLC devices such that the optimal choice can be derived.

- **Two-step RA procedure**

  The two-step RA procedure enables URLLC devices to transmit their data 3 ms after their preamble transmission and therefore lower end-to-end delays can be achieved. Since this is arranged by applying a pre-defined matching of preambles to sets of non-overlapping PRBs used for the UL data transmission, multiple URLLC devices may transmit their data in the same UL PRBs, leading to *data collisions*. If not solved, these UL data collisions can significantly degrade the performance of the procedure resulting in a worse 99.99th end-to-end delay percentile for URLLC devices compared to the EDT procedures and Cat-M1, and even lead these URLLC to experience outages.

  Increase of the probability that the URLLC UL data will be transmitted successfully, can be achieved by transmitting the URLLC UL data multiple times which is referred as the 'Two-Step RA procedure with Repetitions'. However, such repetitions come to the cost of increasing the UL resource utilization as more traffic is generated in the network and therefore there is a clear trade-off between the number of repetitions and the UL resource utilization. A way to completely resolve the UL data collisions is to assign non-overlapping UL PRBs to URLCC devices by more explicitly (and costly) signaling the matching between the URLLC preambles and the PRBs by the base station. This signaling can be done, within the transmitted message on the MPDCCH, 2 ms after the transmission of a URLLC preamble and it is referred as the 'Two-Step RA procedure with Feedback'.

Assigning a high number of preambles to URLLC devices reduces the preamble collision probability but increases the data collision probability as more preambles match to the same PRBs. An increased probability of data collisions will require a higher number of repetitions of the URLLC UL data. Therefore, there is a trade-off between the number of preambles reserved for URLLC devices and the number of repetitions required.

- **Best performing RA procedure**

  The best results regarding end-to-end delay were achieved with the two-step RA procedure. Delays of 43.7 ms and 20.8 ms were measured for non-URLLC and URLLC devices, respectively, when transmitting the URLLC UL data three times in the 'Two-Step RA procedure with Repetitions'. Additionally delays of 43.7 ms and 19.9 ms for non-URLLC and URLL devices respectively were achieved when the base station used the feedback mechanism to indicate the matching of preambles to PRBs in the 'Two-Step RA procedure with Feedback'. It is noted that even though the two enhanced two-step RA procedures introduce significant end-to-end delay gains, the set requirements for the considered FoF applications can be achieved only for non-URLLC devices as the measured delay for URLLC devices is still almost twice as high as acceptable. However, it is worth mentioning that the minimum end-to-end delays that can be achieved with both procedures are 14 ms and 5 ms for non-URLLC and URLLC devices, respectively. It can be therefore be argued that the FoF requirements may potentially indeed be met with a more optimized 5G network deployment, e.g. by adding base station sites.

- **PRACH periodicity investigation**

  The bottleneck of the two-step RA procedure is the occurrence of preamble collisions which cannot be reduced enough to ensure that devices have a successful transmission at their first access attempt. A way of reducing the probability of preamble collision is to increase the PRACH opportunities by applying a higher PRACH periodicity, even though this will clearly increase the UL resource utilization and take resources away from other UL data transmissions. An increased PRACH periodicity implies that each device will experience shorter delays until the first available PRACH subframe, in order to transmit its preamble, and further the load per PRACH subframe will be lower and thus the probability of a preamble collision will be reduced. Specifically, the 99.99th end-to-end delay percentile for URLLC devices can be reduced to 15.9 ms with the 'Two-Step RA procedure with Feedback' in the cost of increasing the UL resource utilization from 6.1% (with PRACH periodicity two) to 25.3% (with PRACH periodicity ten). Furthermore, it was observed that a gain of just 1 ms can be achieved for the more demanding URLLC devices when the PRACH periodicity is increased from five to ten and the UL resource utilization is consequently almost doubled.

- **Load analysis**

  A load analysis was carried out for both enhanced two-step RA procedures, for PRACH periodicities five and ten, in order to assess the performance of the procedure experienced under different load conditions. The analysis revealed that the requirement of 10 ms for URLLC devices cannot be met even with as few as 1000 devices (950 non-URLLC and 50 URLLC devices) in the network for either of the considered PRACH periodicities and procedures. It was further concluded that for higher network loads, a PRACH periodicity five can be a preferred option (when compared to PRACH periodicity of ten) if the cost of increasing the delay of non-URLLC devices well above 50 ms is acceptable. Otherwise, using PRACH periodicity ten for low to medium loads, offers better results for non-URLLC and URLLC devices, compared to PRACH periodicity five, at the cost of increasing the UL resource utilization. This trade-off between the delay of non-URLLC and URLLC devices is correlated with the used split of preambles. It is noted that, the same general conclusions can be drawn for both the ‘Two-Step RA procedure with Repetitions’ and the ‘Two-Step RA procedure with Feedback’, noting that the former performs slightly worse than the latter.

- **Mini-slots**

  A theoretical analysis of the two-step RA procedure was carried out in order to assess whether networks that support mini-slots can provide even lower end-to-end delays. It was concluded that the use of mini-slots can only provide slightly improved end-to-end delays, considering the fact that the delay is mainly influenced from processing delays and the used RAR window. Due to this conclusion, no numerical analysis was carried out for mini-slots.

Overall, in the considered FoF scenario, the ‘Two-Step RA procedure with Feedback’ can support an end-to-end delay of 50 ms with 99.9% reliability for non-URLLC devices (as defined by the FoF requirements) while for URLLC devices, an end-to-end delay of 15.9 ms can be achieved with 99.99%. Therefore, the FoF requirements for non-URLLC devices can be met while for URLLC devices they cannot (even though there is a significant gain of 64.1 ms), as the target is 10 ms. It is noted that an end-to-end delay of 10.0 ms with 99.9% reliability can be achieved for URLLC devices.

6.2 Recommendations and future work

Based on the results and conclusions derived we recommend that the ‘Two-Sep RA procedure with Feedback’ can replace the conventional RA procedure of Cat-M1 in order to (partially) support 5G applications for FoF. However, in order to fully achieve the requirements of such applications, new improvements or different network layout (i.e. addition of a base station or a carrier) should be considered. It is also noted that even though the requirements for this particular services in FoF were not met, it can be possible that the ‘Two-Step RA procedure with Feedback’ can support other types of applications in 5G which do not have such stringent end-to-end delay requirements.

For the application of the 'Two-Step RA procedure with Feedback' the *edt-TBS* should be defined as well as the number of preambles to be reserved for URLLC transmissions. It is reminded that *edt-TBS* defines the number of devices that can transmit their UL data on non-overlapping PRBs on the PUSCH, in one subframe. The size of *edt-TBS* is defined based on the URLLC UL data size and the propagation environment, which eventually will influence the MCS of each device. For example, for a propagation environment with good channel conditions (maximum MCS can be achieved) and URLLC UL data of 40 bytes, the *edt-TBS* can be set to 69 bytes, which will allow nine URLLC devices transmitting simultaneously on the PUSCH on non-overlapping PRBs. The number of preambles that need to be reserved for URLLC devices for an optimum end-to-end delay performance is related to the choice of the PRACH periodicity. Therefore, we recommend that analysis regarding the *edt-TBS* and the preamble reservation should be carried out before an actual deployment of the two-step RA procedure as a different network layout and traffic model might significantly change the configuration.

For applications where low to medium load is expected, the 'Two-Step RA procedure with Feedback' is recommended to be used with PRACH periodicity ten as this periodicity offers the lowest end-to-end delays regarding the 99.9th and 99.99th end-to-end delay percentile for non-URLLC and URLLC devices respectively, with the cost of high UL resource utilization. Additionally, for applications where high load is expected and the end-to-end delay requirement for non-URLLC devices can be relaxed, the 'Two-Step RA procedure with Feedback' is recommended to be used with PRACH periodicity five as this periodicity offers lower end-to-end delays for URLLC devices compared to PRACH periodicity ten in the cost of increasing the end-to-end delay for non-URLLC devices.

Overall, the 'Two-Step RA procedure with Feedback' improves significantly the end-to-end delay for both non-URLLC and URLLC devices even though the target of 10 ms delay with 99.99% reliability for URLLC devices cannot be achieved for this particular scenario. The bottleneck of the two-step RA procedure is the collision probability during a PRACH opportunity and therefore the target can be achieved only if this collision probability is further reduced, e.g. with a wider carrier that supports multiple PRACHs. Additionally, the preamble collision probability does not apply to networks that support NOMA as the RA stage can be completely eliminated (as also presented in the literature Section 1.7) and thus for such demanding applications (where the 99.99th end-to-end delay percentile cannot be relaxed to 16 ms), NOMA schemes are recommended instead of the two-step RA procedure (see also future work below).

Based on the above, some suggestions for *future work* are given:

- **Propagation environment**
  For this study the simplification of ignoring multipath fading was made and therefore the variability of the propagation loss is somewhat limited which consequently lead to a better coverage in the network than in reality. It is thus recommended to integrate a more realistic propagation environment to the simulator and study the effect of a harsh industrial environment to the studied RA procedures. Furthermore, different factory environments can be studied e.g. factories where highly absorbent materials are used. It is reminded that the propagation environment influences the choice of *edt-TBS* in the two-step RA procedures which can influence the performance of the procedures significantly.

Due to the good coverage that was derived from the propagation environment used in this study, all devices were configured with the same coverage level (i.e. CE level 0) as they were experiencing low propagation losses. However, in scenarios with worse coverage, there could be devices in different CE levels and thus a different configuration, with repetitions of each message, for the RA procedure will be applicable. It is therefore expected that the end-to-end delay will be significantly higher and thus a study about the performance of the studied RA procedures is recommended in order to quantify how robust to different configurations these procedures are.

- **Optimizing the two-step RA procedure**
  The end-to-end delay in the two-step RA procedure is mainly influenced by the processing delays and the ACK timer/RAR window. Therefore, it is recommended to develop a new faster receiver implementation at the base station in order to reduce the processing delay. This decreased delay (especially of the ACK timer/RAR window), is expected to reduce significantly the end-to-end delay experienced by URLLC devices. It is noted that a faster receiver at the base station can of course benefit all devices in the network, even the non-eMTC ones, although the principal benefit is of course experienced by devices supporting delay-sensitive applications.

  Furthermore, an extra study can indicate a consistent and efficient way of matching the preambles to the PRBs on the UL channel such that the probability of collision on the data transmission phase will be minimized.

  Additionally, wider carriers (e.g. 100 MHz at 3.5GHz band) can support the simultaneous transmission of more than nine URLLC devices on non-overlapping PRBs on the PUSCH and thus the probability of collision during the data transmission will be reduced. This reduction of collision probability will consequently reduce the end-to-end delay for URLLC devices. Wider carriers can also be combined with multiple Radio Access Technology (multi-RAT) networks, where devices can make use of the PRACH on different carriers. In general, with wider carriers and multi-RAT, an offload of non-URLLC traffic can be achieved during peaks of URLLC transmissions which will finally result in better performance.

- **Non-Orthogonal Multiple Access (NOMA)**
  In Section 1.7, NOMA was introduced and based on literature it was presented as a promising approach for applications that require massive connectivity and ultra-low delays. Therefore investigations on NOMA procedures are recommended for future work in order to quantify the performance of such procedures and compare it to the one achieved with the OMA procedures studied.

References

[1] Ordinal Software, "Industry 4.0, factory of the future, factory 4.0, smart factory, ...," Ordinal Software, [Online]. Available: https://www.ordinal.fr/en/industry-4-0-smart-factory.htm. [Accessed 22 August 2018].

[2] European Commission, "Factories of the Future," [Online]. Available: http://ec.europa.eu/research/industrial_technologies/factories-of-the-future_en.html. [Accessed 20 June 2018].

[3] D. Küpper, K. Kuhlmann, S. Köcher, T. Dauner and P. Burggräf , "The Factory of the Future," 6 December 2016. [Online]. Available: https://www.bcg.com/publications/2016/leaning-manufacturing-operations-factory-of-future.aspx. [Accessed 20 June 2018].

[4] G. Sheader, "SME Manufacturers Adopting Industry 4.0 Technologies," 8 March 2018. [Online]. Available: https://www.manufacturersalliance.co.uk/2018/03/08/sme-manufacturers-adopting-industry-4-0-technologies/. [Accessed 20 June 2018].

[5] 3GPP, "Technical Specification Group Services and System Aspects; Sudy on Communication for Automation in Vertical Domains (TR 22.804 v2.0.0)," 2018.

[6] S. Sesia, I. Toufik and M. Baker, LTE-The UMTS Long Term Evolution From Theory to Practice, Second ed., Sussex: John Wiley & Sons Ltd, 2011.

[7] Ericsson, *Ericsson Mobility Report,* 2016.

[8] O. Teyeb, G. Wikstrom, Stattin Magnus, T. Cheng , S. Faxer and H. Do, "Evolving LTE to fit the 5G future," 31 January 2017. [Online]. Available: https://www.ericsson.com/en/ericsson-technology-review/archive/2017/evolving-lte-to-fit-the-5g-future. [Accessed 16 May 2018].

[9] A. I. A. Jabbar and Y. A. Fawaz, "Effect of RACH Procedure on the Performance of LTE-based M2M Communication," *International Journal of Computer Applications,* vol. 147, no. 5, pp. 12-17, August 2016.

[10] A. Biral, M. Centenaro, A. Zanella, L. Vangelista and M. Zorzi, "The Challenges of M2M Massive Access in Wireless Cellular Networks," *Digital Communications and Networks,* vol. 1, no. 1, pp. 1-19, 2015.

[11] A. Samir, M. M. Elmesalawy, A. S. Ali and I. Ali, "An Improved LTE RACH protocol for M2M applications," *Mobile Information Systems,* pp. 1-11, 2016.

[12] K. Chatzikokolakis, A. Kaloxylos, P. Spapis, N. Alonistioti, C. Zhou, J. Eichinger and O. Bulakci, "On the Way to Massice Access in 5G: Challenges and Solutions for Massive Machine Communications," *Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2015 M. Weichold et al. (Eds.): CROWNCOM 2015,* vol. 156, pp. 708-707, 2015.

[13] M. Vilgelm and W. Kellerer, "Impact of Request Aggregation on Machine Type Connection Establishment in LTE-Advanced," in *Wireless Communications and Networking Conference (WCNC)*, San Francisco, 2017.

[14] A. Lo, Y. Law, M. Jacobsson and M. Kucharzak, "Enhanced LTE-Advanced Random Access Mechanism for Massive Machine-to-Machine (M2M) Communications," *27th Wireless World Research Forum (WWRF27),* no. 2011, 2011.

[15] A. Laya, L. Alonso and J. Alonso-Zarate, "Efficient Contention Resolution in Highly Dense LTE Networks for Machine Type Communications," in *IEEE Global Communications Conference (GLOBECOM)* , San Diego, 2015.

[16] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro and J. Sachs, "Enhanced Radio Access and Data Transmission Procedures Facilitating Industry-Compliant Machine-Type Communications over LTE-based 5G Networks," *IEEE Wireless Communications,* vol. 23, pp. 56-63, 2016.

[17] Intel Corporation and Ericsson, *Introduction of EDT for eMTC and NB-IoT in Rel-15 TS 36.321,* Sanya: 3GPP, 2018.

[18] SONY, *Discussions on 2 Steps RACH Procedure,* Spokane: 3GPP, 2017.

[19] L. Tian, C. Yan, W. Li, Z. Yuan, W. Cao and Y. Yuan, "On Uplink Non-Orthogonal Multiple Access for 5G: Opportunities and Challenges," *China Communications,* vol. 14, no. 12, pp. 142-152, 2017.

[20] Y.-J. Chen, L.-Y. Cheng and L.-C. Wang, "Prioritized resource reservation for reducing random access delay in 5G URLLC," in *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Montreal, 2017.

[21] L. Dai, B. Wang, Y. Yuan, S. Han, C. L. I and Z. Wang, "Non-Orthogonal Multiple Access for 5G: Solutions, Challenges, Opportunities and Future Research Trends," *IEEE Communications Magazine,* vol. 53, pp. 74-81, 2015.

[22] J. Cheon and H.-S. Cho, "Power Allocation Scheme for Non-Orthogonal Multiple Access in Underwater Acoustic Communications," *Sensors,* vol. 17, no. 11, 23 October 2017.

[23] M. Shirvanimoghaddam, M. Dohler and S. Johnson, "Massive Non-Orthogonal Multiple Access for Cellular IoT: Potentials and Limitations," *IEEE Communications Magazine,* vol. 55, no. 9, pp. 55-61, 2016.

[24] K. Au, L. Zhang, H. Nikopour, E. Yi, A. Bayesteh, U. Vilaipornsawai, J. Ma and P. Zhu, "Uplink Contention Based SCMA for 5G Radio Access," in *IEEE GLOBECOM Workshops*, Austin, 2014.

[25] METIS, "Components of a New Air Interface - Building Blocks and Performance," 2014.

[26] Y. Liang, X. Li, J. Zhang and Z. Ding, "Non-Orthogonal Random Access for 5G networks," *IEEE Transactions on Wireless Communications,* vol. 16, pp. 4817-4831, 2017.

[27] I. Qualcomm Technologies, *Paving the path to Narrowband 5G with LTE Internet of Things (IoT),* 2016.

[28] C. Mclelland, "Comparison of LPWAN Technologies - Which is Best for Me?," 29 December 2016. [Online]. Available: https://www.leverege.com/blogpost/comparison-of-lpwan-technologies. [Accessed 24 April 2018].

[29] Y. Hwang, "Cellular IoT Explained: NB-IoT vs. LTE-M vs. 5G and More," [Online]. Available: https://medium.com/iotforall/cellular-iot-explained-nb-iot-vs-lte-m-vs-5g-and-more-8f26496df5d4. [Accessed 29 May 2018].

[30] 3GPP, "User Equipment (UE) radio access capabilities (3GPP TS 36.306 version 13.8.0)," 2018.

[31] L. Grover and A. Gupta, "eMTC- LTE for Connected Devices," 24 August 2017. [Online]. Available: https://connect.aricent.com/2017/08/emtc-lte-for-connected-devices/. [Accessed 16 January 2018].

[32] M. Elsaadany, A. Ali and W. Hamouda, "Cellular LTE-A Technologies for the Future Inter-of-Things: Physical Layer Features and Challenges," *IEEE Communications Surveys and Tutorials,* vol. 19, no. 4, 2017.

[33] B. Pansner, "IoT radio access technologies," in *ITG 5.2.4 Workshop Cellular Internet of Things*, Munich, 2017.

[34] 3GPP, "Physical Channels and Modulation (TS 36.211 v13.8.0)," 3GPP, 2018.

[35] 3GPP, "Physical layer procedures (TS 36.213 v13.8.0)," 3GPP, 2018.

[36] B. Panzner, "IoT Radio Access Technologies," in *ITG 5.2.4 Workshop "Cellular Internet of Things"*, Munich, 2017.

[37] 3GPP, "Multiplexing and Channel Coding (TS 36.212 v13.7.0)," 3GPP, 2018.

[38] Yipig, "What are Localized and Distributed Transmissions for EPDCCH?," 21 July 2014. [Online]. Available: http://lteuniversity.com/get_trained/expert_opinion1/b/ywang/archive/2014/07/21/what-are-localized-and-distributed-transmissions-for-epdcch.aspx. [Accessed 24 January 2018].

[39] A. Chakrapani, "Efficient resource scheduling for eMTC/NB-IoT communications in LTE Rel.13," in *IEEE Conference on Standars for Communications and Networking (CSCN)*, Helsinki, 2017.

[40] R. He, "IoT: LTE-BL/CE(CAT-M1) Attach Procedure (RRCConnectionSetupComplete and following up signaling)," 11 October 2016. [Online]. Available: http://riverheltenotes.blogspot.com/2016/10/iot-lte-blcecat-m1-attach-procedure.html. [Accessed 8 July 2018].

[41] 3GPP, "Radio Resource Control (RRC); Protocol Specification (TS 36.331 v13.8.1)," 3GPP, 2018.

[42] Intel Corporation and Ericsson, "Introduction of EDT for eMTC and NB-IoT in Rel-15 TS36.321," 3GPP, Sanya, 2018.

[43] K. Swamy, "Timing Advance and Time Alignment Timer," 2 July 2014. [Online]. Available: http://howltestuffworks.blogspot.com/2014/07/timing-advance-and-time-alignment-timer.html. [Accessed 25 7 2018].

[44] A. A. Zaidi, R. Baldemair, M. Andersson, S. Faxér, V. Molés-Cases and Z. Wang, "Designing for the future: the 5G NR physical layer," 24 July 2017. [Online]. Available: https://www.ericsson.com/en/ericsson-technology-review/archive/2017/designing-for-the-future-the-5g-nr-physical-layer. [Accessed 2018 July 26].

[45] Altair Semiconductors, *Coverage Analysis of LTE-M Category-M1,* 2017.

[46] 3GPP, "Study on Communication for Automation in Vertical domains (CAV) (TR 22.804 v1.0.0)," 2017.

[47] 3GPP, "RAN Improvements for Machine-type Communications (TR 37.868 v11.0.0)," 2011.

[48] E. Tanghe, W. Joseph, L. Verloock, L. Martens, H. Capoen, K. v. Herwegen and W. Vantomme, "The Industrial Indoor Channel: Large-Scale and Temporal Fading at 900, 2400 and 5200 MHz," *IEEE Transactions on Wireless Communications,* vol. 7, no. 7, pp. 2740-2751, 2008.

[49] 3GPP, "Medium Access Control (MAC) Protocol Specification (TS 36.321 v13.7.0)," 2017.

[50] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures (TS 36.213 v13.8.0)," 2018.

[51] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC) (TS 36.331 v13.8.1)," 2018.

[52] R.-G. Cheng, C.-H. Wei, S.-L. Tsao and F.-C. Ren, "RACH Collision Probability for Machine-Type Communications," in *Vehicular Technology Conference (VTC Spring), 2012 IEEE 75th*, Yokohama, 2012.

[53] P. Rod, "Confidence Intervals," Math Is Fun, 15 August 2018. [Online]. Available: http://www.mathsisfun.com/data/confidence-interval.html. [Accessed 27 August 2018].

[54] F. Train, "Confidence Interval for a Popuation mean, with a known Population Variance," [Online]. Available: https://financetrain.com/confidence-interval-population-mean-known-population-variance/. [Accessed 19 August 2018].

[55] The Pennsylvania Statie University, "Lesson 47: Distribution-Free Confidence Intervals for Percentiles," [Online]. Available: https://newonlinecourses.science.psu.edu/stat414/node/317/. [Accessed 1 August 2018].

[56] C. Mehlfuhrer, M. Wrulich, J. C. Ikuno, D. Bosanska and M. Rupp, "Simulating the Long Term Evolution Physical Layer," in *17th European Signal Processing Conference (EUSIPCO 2009)*, Glasgow, 2009.

[57] R. Khastur, "LTE RRC Timer & RRC Connection Setup Complete message evolution," 8 April 2016. [Online]. Available: https://www.slideshare.net/RayKhastur/lte-rrc-timer-rrc-connection-setup-complete-message-evolution. [Accessed 19 March 2018].

[58] "https://br.freepik.com/vetores-gratis/fabrica-isometrica-branco_1086486.htm," [Online]. [Accessed 29 May 2018].

[59] J. Sachs, O. Liberg, M. Sundberg, E. Wang and J. Bergman, Cellular Internet of Things: Technologies, Standards and Performance, Elsevier, 2017.

[60] Qualcomm Technologies, Inc., *Leading the LTE IoT evolution to connect the massive Internet of Things,* 2017.

Appendix A SNR-TBS curves for Cat-M1

The curves that define the TBS based on the SNR are based on BLER curves such that BLER ≤ 10%. The BLER curves, as shown in Figure A-1, are based on [56] and indicate the maximum SNR value per Channel Quality Indicator (CQI) such that BLER ≤ 10%. However, these curves are drawn for LTE and not for Cat-M1 and thus further processing was needed.

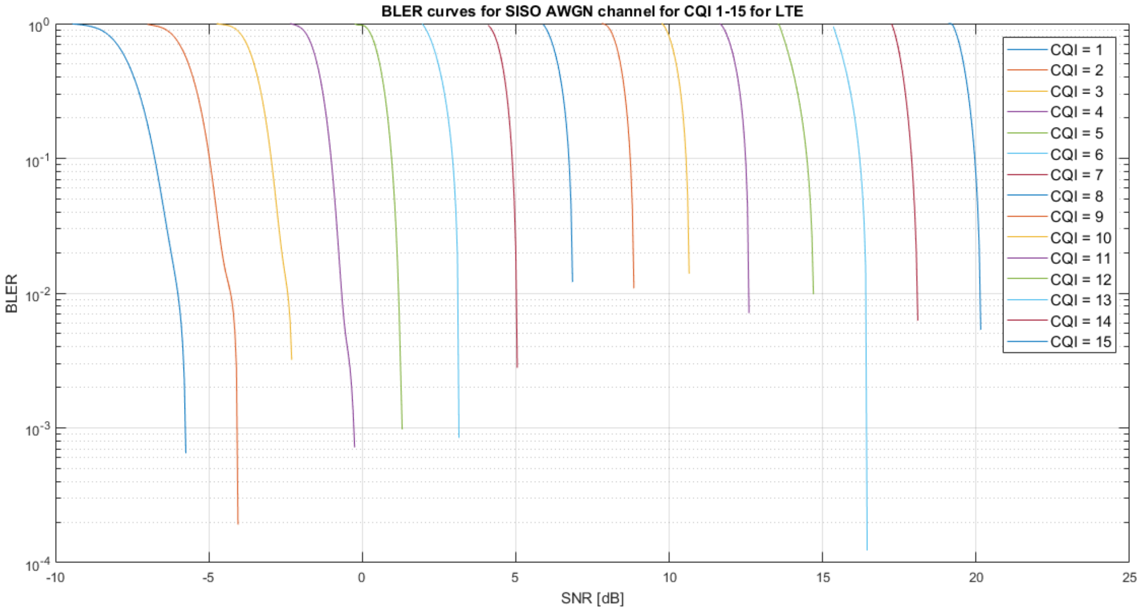


Figure A-1 - BLER curves for SISO AWGN channel in LTE for CQI 1-15.

For those values and with the use of Table 7.2.3-1 in [50], the bit rate is derived based on the code rate and it is shown in Figure A-2 (LTE curve). Furthermore, it can also be derived that the LTE CQIs are comparable to Cat-M1 MCS 0-15 and thus to TBS Index 0-14, by using Table 8.6.1-2 and 7.2.31 in [50]. Therefore, a polynomial curve is fitted based on the LTE CQI bit rates which matches the TBS with the SNR. Then, with the use of the TBS for Cat-M1 and the fitted curve, the SNR values were derived and the final SNR-TBS curve for Cat-M1 is defined as shown in Figure A-2.

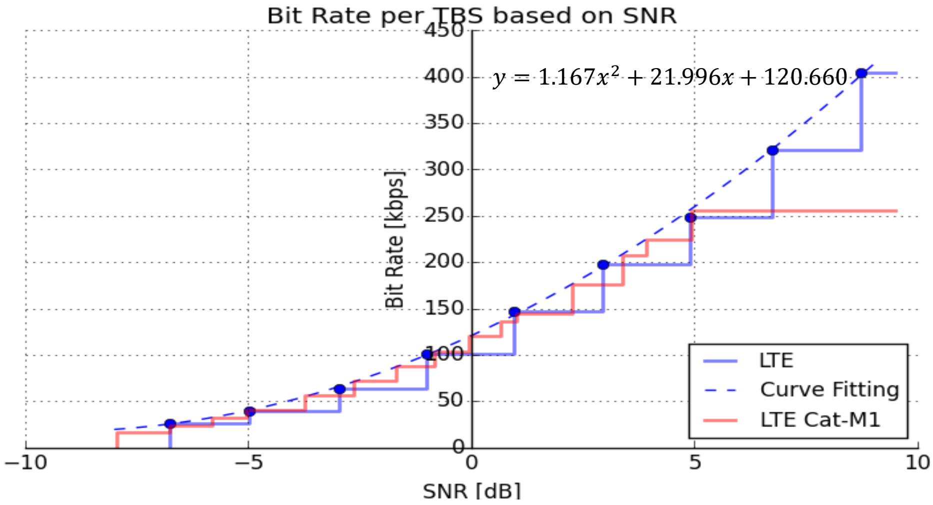


Figure A-2 - Bit Rate/TBS based on SNR in LTE and Cat-M1.

Appendix B Configuration of the two-step RA procedure

All the intermediate results used for the definition of the end-to-end delays with different PRACH periodicities in the ‘Two-Step RA procedure with Repetitions’ are presented in this Appendix. For each of the PRACH periodicities, the preamble split between non-URLLC and URLLC devices is derived based on a preamble analysis on an idealistic scenario. Then, based on that particular preamble split, the number of required number of repetitions is derived as in Section 5.5.2. It is noted that the preamble analysis needed for the ‘Two-Step RA procedure with Feedback’ is similar to the one for the ‘Two-Step RA procedure with Repetitions’.

- **PRACH Periodicity 3**

The preamble split analysis for PRACH periodicity three is presented in Figure B-1 and it can be derived that the best choice is to reserve 35 preambles for URLLC devices and 29 preambles for non-URLLC devices. Furthermore, Figure B-2 presents the impact of repetitions of URLLC UL data on the end-to-end delay and it can be deduced that two consecutive transmissions of the URLLC UL data can provide the same results as in the idealistic scenario where no collisions on the data were happening.

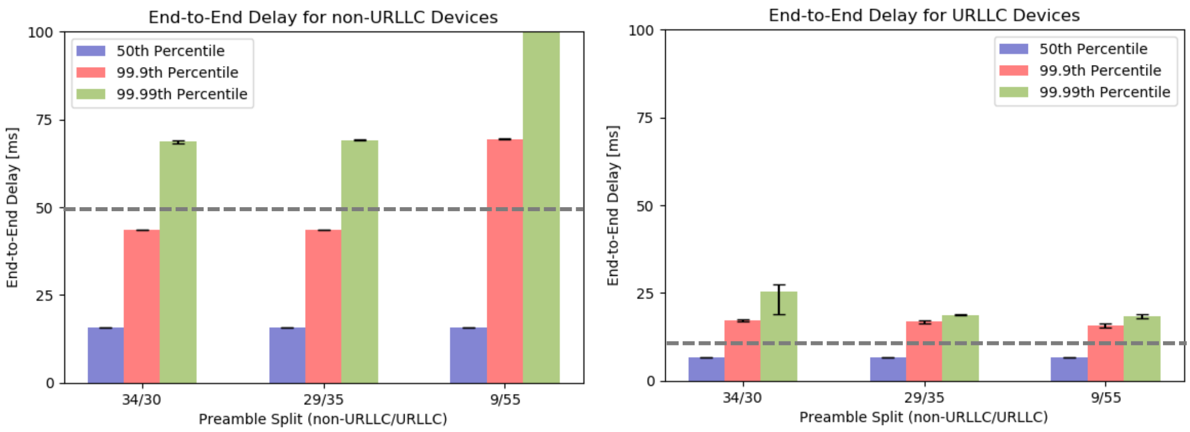


Figure B-1 - Preamble analysis for the ‘Two-Step RA procedure with Repetitions’ and PRACH periodicity three for non-URLLC devices (left) and URLLC devices (right).

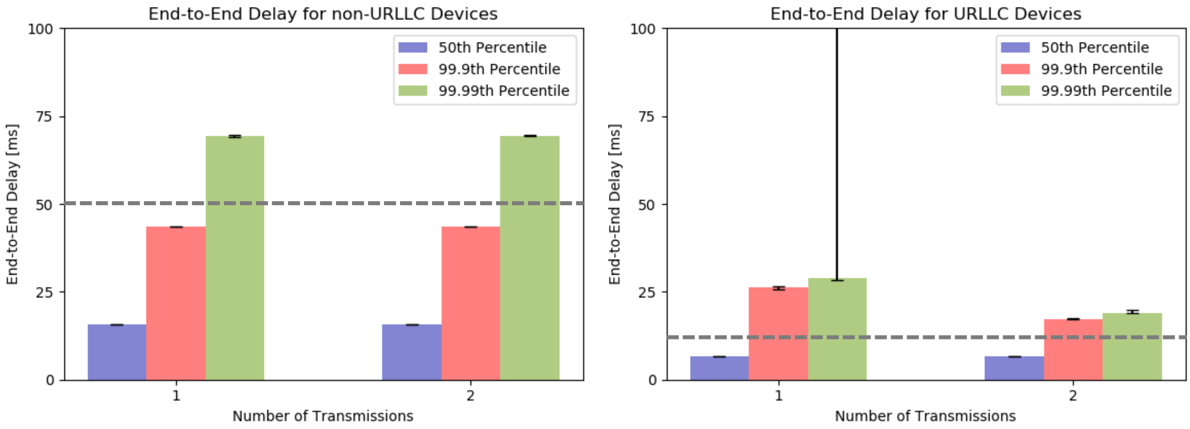


Figure B-2 - Impact of transmitting the URLLC UL data multiple times on the end-to-end delay for non-URLLC devices (left) and URLLC devices (right) while using PRACH periodicity three.

- **PRACH Periodicity 5**

For PRACH periodicity five, the results of the preamble analysis are illustrated in Figure B-3 and it is deduced that the best choice is to reserve 50 preambles out of the available 64 for the URLLC transmissions. It is noted that a lower 99.9th percentile can be achieved compared to PRACH periodicity three, although this requires a higher preamble split than for PRACH periodicity three. However, due to the high number of preambles reserved for URLLC devices, more data collisions are expected to occur on the PUSCH as more preambles match to the same PRBs compared to PRACH periodicity three. This behavior is reflected on the repetitions analysis in Figure B-4, as with two consecutive transmissions of the URLLC UL data the 99.9th end-to-end delay percentile is uncertain (as a wide CI applies). It can also be that using three repetitions increases the delay instead of reducing it. This increase of delay is caused from the fact that there are data collisions occurring on the PUSCH from devices that transmitted their preambles in different PRACH subframes as the PRACH periodicity is increased. An example illustrating this follows:

> *The PRACH subframes when devices are able to transmit their preambles are subframes #1, #3, #5, #7, #9, #11, … . A URLLC device which transmits its preamble with value 10 in subframe #1 will transmit its UL data three times, in subframes #5, #6 and #7. Another URLLC device, transmits its preamble with value 10 in subframe #3 and thus it transmits its UL data in subframes #7, #8 and #9. Therefore, the UL data of both devices will collide during subframe #7 even though they initiated the RA procedure in different subframes.*

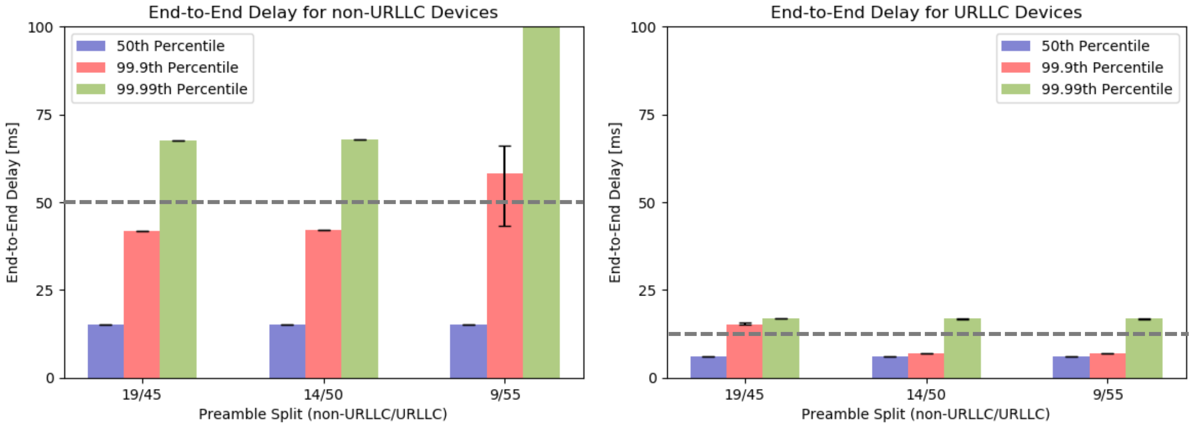


Figure B-3 - Preamble analysis for the 'Two-Step RA procedure with Repetitions' and PRACH periodicity five for non-URLLC devices (left) and URLLC devices (right).

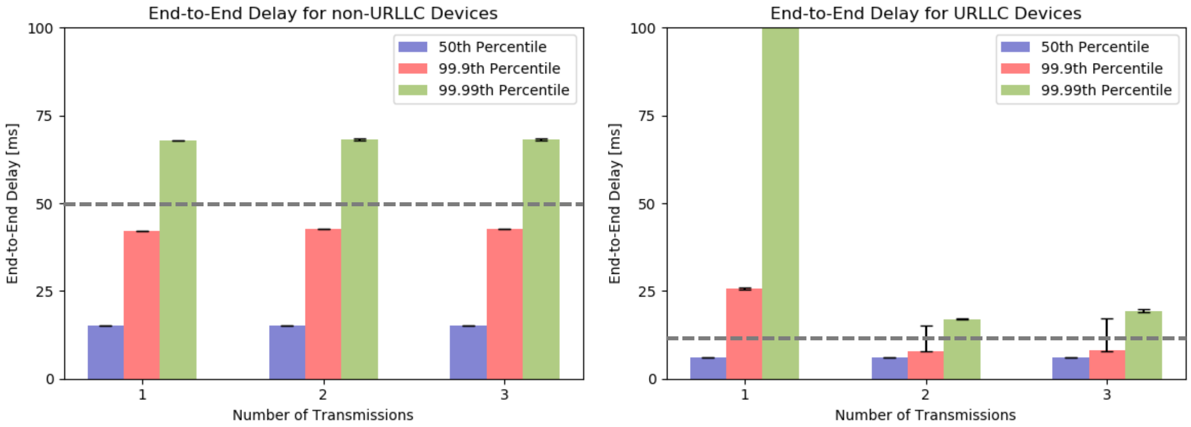


Figure B-4 - Impact of transmitting the URLLC UL data multiple times on the end-to-end delay for non-URLLC devices (left) and URLLC devices (right) while using PRACH periodicity five.

- **PRACH Periodicity 10**

The best preamble split for PRACH periodicity ten, is 25 preambles for URLLC transmissions and 39 preambles for non-URLLC as shown in the preamble analysis in Figure B-5. For this preamble split, the repetition analysis is carried out with the results presented in Figure B-6. The results obtained from transmitting the URLLC UL data twice achieves a 99.9th end-to-end delay percentile with uncertainty (wide CI). This behavior is due to the high value of the PRACH periodicity as there are data collisions on the PUSCH between devices that transmitted their preamble in different PRACH subframes (same behavior as with three transmissions for PRACH periodicity five). Therefore, using three transmissions for the URLLC UL data increases further the delay instead of decreasing it, as more collisions on the data happen, as also illustrated in Figure B-6. Thus, the best number of times for transmitting the URLLC UL data is two.

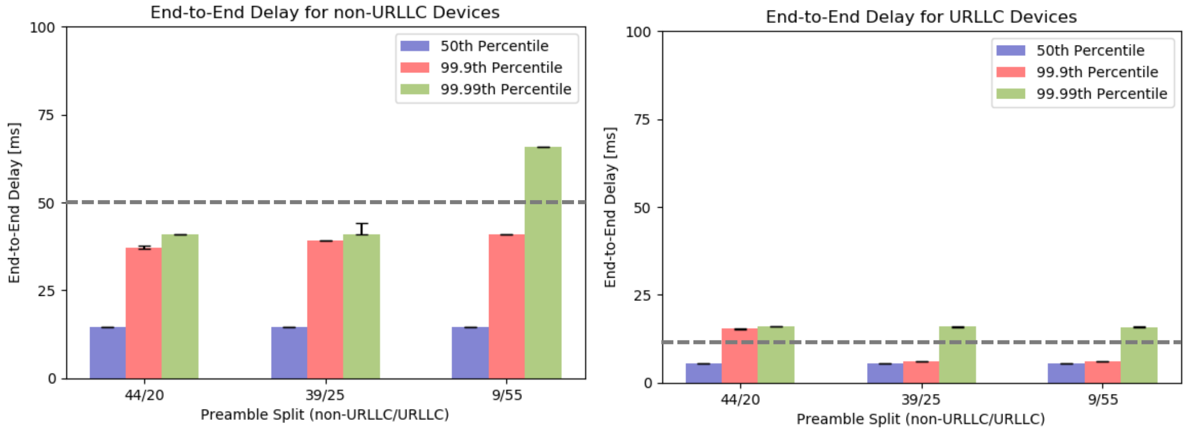


Figure B-5 - Preamble analysis for the 'Two-Step RA procedure with Repetitions' and PRACH periodicity ten for non-URLLC devices (left) and URLLC devices (right).

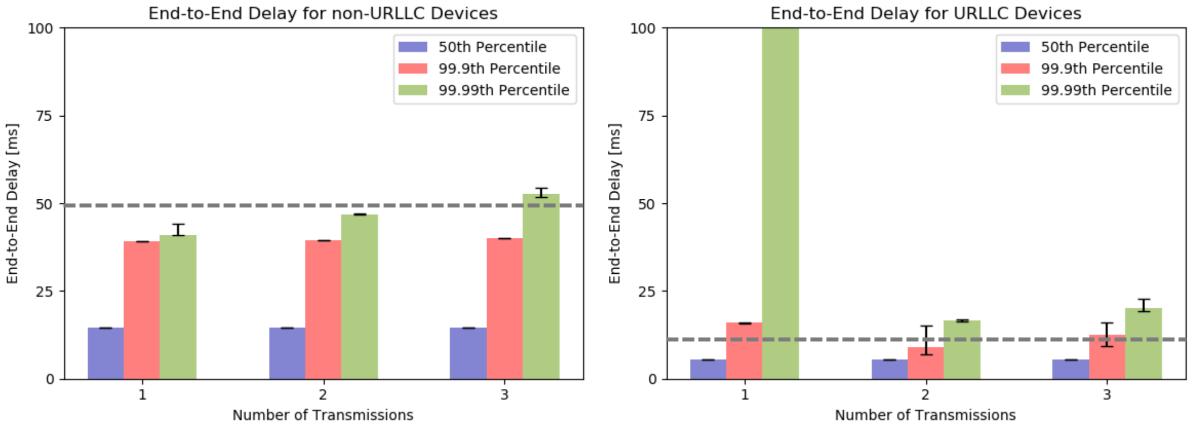


Figure B-6 - Impact of transmitting the URLLC UL data multiple times on the end-to-end delay for non-URLLC devices (left) and URLLC devices (right) while using PRACH periodicity ten.