ADAPTIVE HARQ (A-HARQ) FOR ULTRA-RELIABLE COMMUNICATION IN 5G

by

Emerson Cabrera

for the degree of Master of Research

Department of Engineering
Macquarie University

October 9, 2015

Supervisor: Dr Gengfa Fang

ACKNOWLEDGMENTS

I would like to acknowledge **Dr Gengfa Fang** for sharing his immense knowledge and giving me excellent feedback on my research work as my supervisor and mentor. He provided motivation and guidance during my time spent researching and writing this thesis. I would also like to acknowledge **Diep Nguyen** for providing assistance on how to set up and use a MATLAB LTE link level simulator, to be able to test HARQ schemes and resource allocation. Lastly, I would like to acknowledge **Xunqian Tong** for allowing me to be a co-author for his journal paper, which is related to my field of research on 5G cellular wireless networks, specifically on the use of clustering for the proposed A-HARQ.

STATEMENT OF CANDIDATE

I, Emerson Cabrera, declare that this report, submitted as part of the requirement for the award of Master of Research in the Department of Engineering, Macquarie University, is entirely my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualification or assessment at any academic institution.

Student's Name: Emerson Cabrera

Student's Signature:

Date: October 9, 2015

Publications and Awards

1. "Adaptive Hybrid ARQ (A-HARQ) for Ultra-Reliable Communication in 5G"
   - **Contribution**: Primary author
   - **Submitted to**: International Telecommunication Networks and Applications Conference (ITNAC) 2015
2. "An Energy-Balanced Routing Algorithm in Wireless Seismic Sensor Network"
   - **Contribution**: Co-author
   - **Primary author**: Xunqian Tong
   - **Submitted to**: Journal of Communications Technology and Electronics
3. Awarded the **Macquarie University Research Training Pathway Scholarship**, to supplement the research contained in this thesis.

ABSTRACT

This thesis addresses the need for motivating applications, such as mission critical industrial control and medical applications, to operate under Ultra-Reliable Communication (URC) mode in future 5th Generation (5G) cellular wireless networks, while also under strict Quality of Service (QoS) constraints such as ultra-low latency. Reliability has been shown to improve through the use of Hybrid Automatic Repeat reQuest (HARQ) for the retransmission (RTX) of erroneous packets during poor channel conditions. However, this can increase the delay to unacceptable levels if more than 1 RTX is required. Thus, an Adaptive HARQ (A-HARQ) scheme is proposed, where RTX are implemented on of better quality sub-bands, with resources dynamically allocated based on Channel Quality Indicator (CQI) reports. A-HARQ also increases the number of RTX within a 4 ms time period, by utilising Transmission Time Interval (TTI) bundling to decrease the delay incurred from many RTX. Performance analysis is conducted, by comparing A-HARQ and the legacy HARQ in terms of delay, where A-HARQ was shown to have about 35% lower delay than the legacy HARQ, with a slight decrease in throughput. Planned future work involves the field testing of A-HARQ and dynamic resource allocation for URC and A-HARQ, and optimising the balance between reliability, latency, and energy-efficiency.

Contents

List of Figures

List of Tables

List of Acronyms

3GPP	3rd Generation Partnership Project
5G	5th Generation
A-HARQ	Adaptive HARQ
ACK	Acknowledgement
AEI	Availability Estimation and Indication
AWGN	Additive White Gaussian Noise
BEP	Bit Error Probability
CEST	Channel Estimation
CDMA	Code Division Multiple Access
CQI	Channel Quality Indicator
CRC	Cyclic Redundancy Check
eNodeB	Evolved Node B
FDMA	Frequency Division Multiple Access
GBR	Guaranteed Bit Rate
GSM	Global System for Mobile Communications
H2H	Human to Human
HARQ	Hybrid Automatic Repeat reQuest
IMS	IP Multimedia Subsystem
IMT	International Mobile Telecommunications
IoT/IoE	Internet of Things/Internet of Everything
LLR	Log-likelihood Ratio
LTE	Long Term Evolution
M2M	Machine to Machine
MCS	Modulation and Coding Scheme
MIMO	Multiple Input Multiple Output

NAK	Negative Acknowledgement
NC-HARQ	Network-Coded HARQ
OFDM	Orthogonal Frequency Division Multiplexing
QAM	Quadrature Amplitude Modulation
QCI	QoS Class Identifier
QoS/QoE	Quality of Service/Quality of Experience
QPSK	Quadrature Phase Shift Keying
RACH	Random Access Channel
RAO	Random Access Opportunity
RAT	Radio Access Technology
RB	Resource Block
RB-HARQ	Reliability-based HARQ
RSC	Reliable Service Composition
RTL	Reliable Transmission Link
RTX	Retransmission
RV	Redundancy Version
Rx	Receiver
SINR	Signal-to-Interference and Noise Ratio
SISO	Soft Input Soft Output
SNR	Signal-to-Noise Ratio
TB	Transport Block
TDMA	Time Division Multiple Access
Tx	Transmitter
TTI	Transmission Time Interval
UE	User Equipment
UMTS	Universal Mobile Telecommunications Systems
URC	Ultra Reliable Communication
V2V	Vehicle to Vehicle
XOR	Exclusive OR

Chapter 1

Introduction

The latest generation of cellular wireless networks, referred to as the 5th Generation (5G), is currently under heavy research, and is expected to be rolled out between 2020 and 2030 [1–3]. By 2020, it is predicted that there will be approximately 50 billion devices interconnected for the Internet of Things (IoT)/Internet of Everything (IoE), which will naturally lead to an exponential increase in mobile data volume [1]. The previous 3rd Generation (3G) and 4th Generation (4G), also known as Long Term Evolution (LTE) and LTE-Advanced (LTE-A), respectively, were not designed to guarantee high reliability for the majority of the time [4]. This is why Ultra-Reliable Communication (URC) and Massive Machine to Machine (M2M) Communication are currently such hot research topics, and are identified as two of the major operating modes of 5G [1, 2].

There are many challenges to overcome on the road towards 5G. This thesis, however, focuses on one particular problem: How can the URC operating mode be enabled, while also satisfying strict Quality of Service (QoS) constraints, such as ultra-low latency? Mission critical industrial control and medical applications are used as motivations for the proposed solution.

One proven way of improving reliability of data transmission over cellular wireless networks is through the use of Hybrid Automatic Repeat reQuest (HARQ). HARQ consists of a combination of Forward Error Correction (FEC) coding at the physical (PHY) layer, and the Automatic Repeat reQuest (ARQ) at the medium access control (MAC) layer for error detection and recovery [5]. Type-I HARQ involves the complete retransmission (RTX) of erroneous packets, while Type-II HARQ adapts to the changing channel conditions by modifying the Modulation and Coding Scheme (MCS) for each RTX [6].

This chapter details the background, requirements, potential applications and challenges for 5G, and also the project specifics. Particularly, Section 1.1 gives an overview on the evolution of cellular wireless networks, Section 1.2 deals with some of the fundamentals of LTE/LTE-A, Section 1.3 introduces 5G, and Section 1.4 establishes the project scope, objectives and contributions. Chapter 2 provides the background on URC, Massive M2M Communication, HARQ schemes and the problems encountered in relation to the requirements. Chapter 3 presents the proposed Adaptive HARQ (A-HARQ) scheme and resource allocation for URC and A-HARQ. Chapter 4 analyses and compares the performance of the proposed A-HARQ with the legacy HARQ by simulations. Chapter 5 concludes this thesis, and establishes a pathway towards planned future research.

1.1 From 1st Generation to 3rd Generation

Cellular wireless networks have evolved from generation to generation at roughly one decade apart, where each successive generation has offered progressively higher data rates, and also additional features for cellular network subscribers.

The 1st Generation (1G) was rolled out in 1981. 1G involved cellular communication with analogue transmissions using circuit switching, which only allowed voice calls. Developed systems for 1G were largely independent from each other based on regions such as Japan, Europe, North America and Hong Kong [7].

The 2nd Generation (2G) or Global System for Mobile Communications (GSM), was rolled out in 1992. 2G was a significant advance from 1G, with the addition of Short Message Service (SMS) text messages, Electronic Mail (Email), and also allowed global roaming. Later additions to 2G were General Packet Radio Service (GPRS) and Enhanced Data rates for GSM Evolution (EDGE), commonly referred to as 2.5G and 2.75G respectively. 2G used the multiple access technologies of Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA) [7].

The 3rd Generation (3G) or Universal Mobile Telecommunications Systems (UMTS), was rolled out in 2001 with the addition of mobile internet. Later additions to 3G was High Speed Downlink/Uplink Packet Access (HSDPA/HSUPA), and High Speed Packet Access Evolution (HSPA+), commonly referred to as 3.5G, and 3.9G respectively, since they adhere to the IMT-2000 standard, but do not adhere to the IMT-Advanced standard. 3G uses the multiple access technology of Code Division Multiple Access (CDMA) [7].

1.2 LTE/LTE-Advanced

The 4th Generation (4G) was rolled out in 2011, and is supposed to be 'all-Internet Protocol (IP)' based. However, there is still some overlap with 2G and 3G networks, which means 4G is not 'all-IP' based yet. LTE from the 3rd Generation Partnership Project (3GPP) is commonly referred to as 4G, but is actually not true 4G, since it does not fully adhere to the IMT-Advanced standard. True 4G is referred to as LTE-A, which was standardised in Release 10 of the 3GPP specification series. LTE/LTE-A uses the multiple access technologies of Orthogonal Frequency Division Multiplexing (OFDM), TDMA and FDMA [7].

Some of the fundamentals of LTE/LTE-A [5, 7], such as resource blocks in Subsection 1.2.1, the resource scheduler and interference reduction in Subsection 1.2.2, and the major differences between LTE and LTE-A in Subsection 1.2.3, are detailed as follows.

1.2.1 Resource Blocks

LTE/LTE-A resources are split into the three dimensions of time, frequency and space, where the spatial dimension is measured in 'layers'. The spatial dimension can be accessed by the multiple antenna ports at the Evolved NodeB (eNodeB), where each antenna port has a reference signal (RS) that enables the user equipment (UE) to estimate the radio channel.

In the time domain, resources are split into frames of the length 10 ms, as shown in Figure 1.1 (on page 4). Each frame is split into 10 subframes of 1 ms each, with each subframe being further split into 2 slots of 0.5 ms each. Within a slot, there are 7 resource elements (RE) or OFDM symbols of the length 66.667 μs, which are all preceded by a cyclic prefix (CP) of 4.7 μs, except for the first OFDM symbol with a CP of 5.2 μs instead. The CPs are used to prevent inter-symbol interference, which are due to the varying lengths of the several transmission paths [5].

In the frequency domain, the total bandwidth is split into number of resource blocks (RBs) as shown in Table 1.1 (on page 4), which is the minimum resource unit that can be allocated to a UE. Each RB is further split into 12 subcarriers of 15 kHz spacing. The most common mode of operation, where the downlink and uplink have access to the same resources (but not simultaneously), is referred to as Frequency Division Duplexing (FDD) [7].

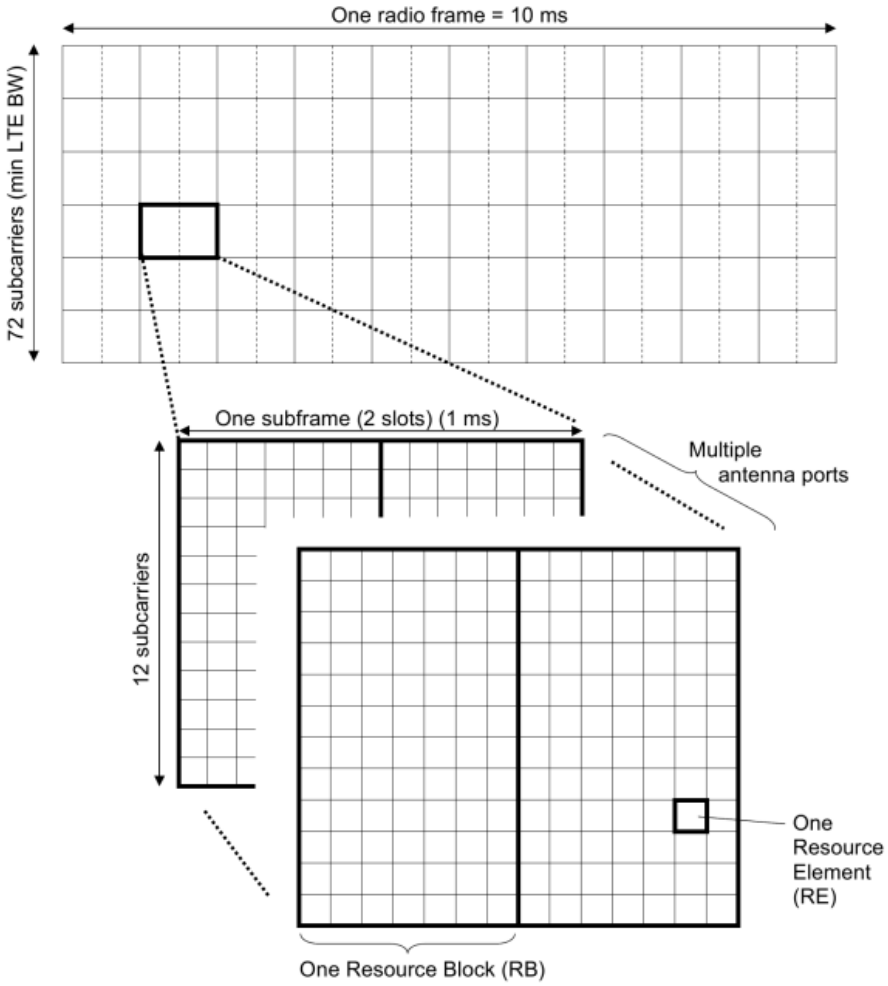

Figure 1.1: The division of resources within the spatial, frequency and time domains.

Table 1.1: Bandwidth resource allocation. 'Bandwidth Utilisation' refers to how much of the bandwidth is available for the transmission of data.

Bandwidth (MHz)	**# Resource Blocks**	**Bandwidth Utilisation (%)**
1.4	6	77.1
3	15	90
5	25	90
10	50	90
15	75	90
20	100	90

1.2.2 Resource Scheduling and Interference Reduction

The resource scheduler of LTE/LTE-A operates at the MAC layer of the IP stack. UEs queue for the available resources, which are in the form of RBs as detailed in Subsection 1.2.1. The MAC Scheduler allocates RBs to UEs waiting in the queue, which means that it applies equal treatment regardless of the content of the particular packet.

Such 'fair treatment' of incoming packets can be a problem for mission critical industrial control applications, as they require ultra-reliability and ultra-low latency. 4G cannot guarantee that these strict requirements will be satisfied, since it was not designed to provide such a level of service for the majority of the time. This is why 5G has been proposed as a way to provide ultra-reliability and ultra-low latency for UEs.

1.2.3 Major Differences between LTE and LTE-A

The major differences between LTE and LTE-A, where enhancements over LTE were introduced in Release 10 of the 3GPP specification series, are summarised below [7].

- LTE-A supports up to 100 MHz bandwidths, which is five times greater than the 20 MHz maximum in LTE, through carrier aggregation of five Component Carriers of 20 MHz bandwidth. Carrier aggregation enables backward compatibility with LTE.
- An enhanced downlink multiple antenna transmission is maintained by the number of antennas at the eNodeB and UE, which increases the number of antenna ports from 4 to 8 in Single-User Multiple Input Multiple Output (SU-MIMO).
- Uplink multiple antenna transmission is achieved by increasing the number of antenna ports from 1 to 4 in SU-MIMO, which increases the peak spectral efficiency to 15 bits per second per Hertz (bps/Hz) and also introduces transmit diversity for uplink control signalling.
- Relaying for parts of the network where wired backhaul is impractical. A relay node transfers the control information and data, to and from a donor cell.
- Support for heterogeneous network deployments, which consists of a layer of macrocells and a layer of small cells with at least one common carrier between them. Cross-carrier scheduling is used to avoid control channel interference between the macrocells and small cells, when control signalling is transferred between them.

1.3 5th Generation (5G)

The background and related work for the future 5G communications are covered as follows, with the design objectives in Subsection 1.3.1, the problems that need to be solved in Subsection 1.3.2, and the potential applications that could benefit from using 5G in Subsection 1.3.3.

1.3.1 Design Objectives

There are three major design objectives for 5G mobile communications, as proposed by Huawei in [8]:

1. Support for massive capacity and massive connectivity.
2. Support for the increasingly diverse set of services, applications and users.
3. Flexible and efficient use of all available non-contiguous spectrum.

The following specifications can be partially satisfied if those proposed design objectives by Huawei are successfully implemented [8]:

1. 10 Gigabits per second (Gb/s) speeds (fibre-like) for ultra high definition (UHD) video, multimedia interactions, and mobile cloud services.
2. Ultra-wide bandwidth, as in the 'millimetre wave' of the 30–300 GHz extremely high frequency (EHF) range [9], and sub-millisecond (< 1 ms) latency.
3. High reliability, in relation to QoS requirements for mission critical industrial control applications, Vehicle to Vehicle (V2V) communication and IoT applications, smart sensors, and text-based messaging for a smart city.
4. Zero-second switching between different Radio Access Technologies (RATs), to enable 'Ubiquitous' communications.

Before these specifications can be fully satisfied, there are many problems that need to be solved. The specific problems in relation to URC and Massive M2M Communication are summarised in the following subsection.

1.3.2 Current problems that need to be addressed

1. The scheduler in the MAC layer for LTE/LTE-A treats all packets equally regardless of their content, rather than giving higher priority to more important packets.

2. LTE/LTE-A cannot guarantee high reliability for the majority of the time, as the current network architecture is not designed to provide this level of service.

3. Energy-efficiency is not optimal enough for the energy limited UE. This can be a major problem for Massive M2M Communication, as these machines/devices need to be able to last for years on limited amounts of power.

4. Latency can increase to unacceptable levels during poor channel conditions, as the probability of one or more RTX is high, which can cause problems for applications with strict delay requirements.

Most of these problems are in relation to mission critical industrial control applications, but are also problems for other applications, as there is always a balance required between reliability, latency and energy-efficiency.

1.3.3 Potential Applications

There are many potential applications according to Osseiran *et al.*, in [1] that could benefit with the future 5G cellular wireless communications. These are:

1. e-banking, e-health, and e-learning

2. Augmented reality and virtual reality offices

3. Crowded areas such as events and shopping centres

4. Sensors and actuator networks

5. Smart-grid network

6. Traffic control systems

7. Security, logistics, automotive, and mission critical industrial control applications

Some of these applications require very high reliability, but others such as mission critical industrial control applications also require ultra-low latency. Energy-efficiency is important for limited power applications such as UE. This makes it very relevant for protocols being able to support the highly diverse sets of applications in the future.

1.4 Project Specifics

Scope

Mission critical industrial control and medical applications, have strict QoS requirements such as ultra-reliability and ultra-low latency, and will most likely communicate using cellular wireless networks. Consequently, this project focuses on proposing an improved HARQ scheme and dynamic resource allocation for this scheme.

Objectives

1. Build a solid understanding of LTE/LTE-A, HARQ schemes and resource allocation, by conducting a background literature review
2. Propose an Adaptive HARQ (A-HARQ) scheme and dynamic resource allocation for A-HARQ, to solve the reliability and delay problems for mission critical industrial control and medical applications
3. Verify the proposed A-HARQ scheme by using simulation software such as MATLAB

Contributions

Chapter 2 provides a thorough literature review on URC, the current HARQ schemes and resource allocation for Massive M2M Communication. Chapter 3 proposes an A-HARQ scheme and a dynamic resource allocation solution for A-HARQ, which ensures that the motivating applications can experience URC and ultra-low latency during poor channel conditions. Chapter 4, shows that the proposed A-HARQ has about 35% lower delay than the legacy HARQ within the expected poor Signal-to-Noise (SNR) range of 0 to 2 dB, with a slight decrease in throughput. A-HARQ also utilises some concepts from LTE/LTE-A, which would simplify the integration into the current LTE-A and also future 5G communications.

Chapter 2

Background and Related Work

This chapter presents a thorough literature review on the related work of URC and Massive M2M Communication, which are two of the major operating modes of the future 5G cellular wireless networks. It also discusses one of the proven ways of ensuring reliability for cellular wireless networks within the related literature. Section 2.1 introduces what URC is, Section 2.2 deals with Massive M2M Communication, Section 2.3 gives an overview on the LTE HARQ process, while Section 2.4 and Section 2.5 cover two categories of HARQ schemes called Reliability-based HARQ (RB-HARQ) and Network-Coded HARQ (NC-HARQ), respectively.

2.1 Ultra-Reliable Communication (URC)

Popovski [10] introduces one of the major operation modes for 5G communication systems, which is known within literature as URC [1, 2, 4, 10]. URC refers to a certain level of wireless mobile communication service that is available almost 100% of the time. To achieve URC, there has to be a way for the communication services to adapt the current requirements of a particular user, based on the current level of reliability that can be obtained.

5G wireless systems are expected to have five operating regions based on the data rate in bits/second (bps) and the number of users. However, based on the scope of this project, only two of these regions are relevant. These relevant regions are R3 and R4 (as shown in Figure 2.1 on page 10), and are summarised as follows:

- **R3** consists of Massive M2M Communication, with lowband data rates, and very

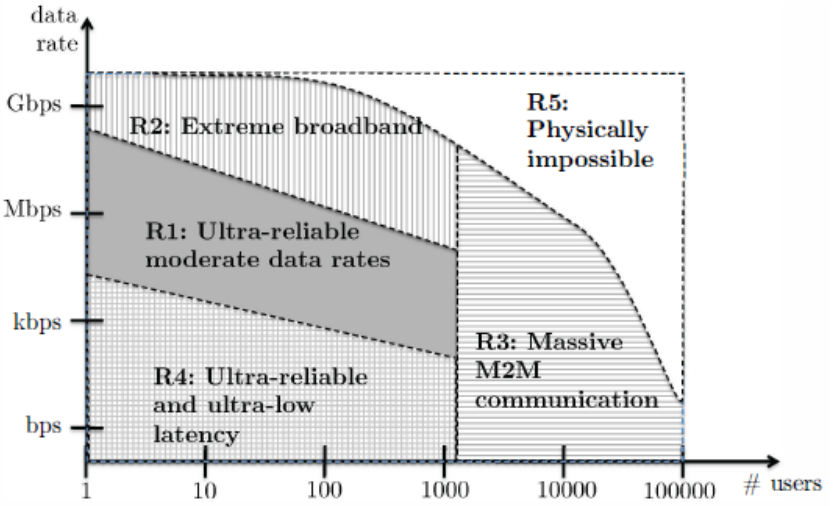

Figure 2.1: Operating regions of 5G wireless communication networks [10]

short messages exchanged. There is the challenge of supporting messages which are redundant or of low importance, and also messages of high importance (such as from mission critical industrial control applications) in R3.

- **R4** consists of lowband data rates with ultra-reliability and ultra-low latency, and very short messages exchanged. R4 requires new transmission techniques and access protocols to send ultra-reliable messages with ultra-low latency.

The following subsections introduces some of the proposed solutions from literature in relation to URC. The elements of URC are discussed in Subsection 2.1.1, while the Availability Estimation and Indication (AEI) mechanism is discussed in Subsection 2.1.2.

2.1.1 Elements of Ultra-Reliable Communication

The pre-condition to receive transmitted data is the metadata (or the header of a packet), as the recipient will not know what to do with the packet otherwise. Therefore, the header must be transmitted with high reliability.

The probability of successfully receiving the header and the data, $p_{s,(h,d)}$, is found using Equation (2.1), where the probability of error in receiving the header is $p_{e,h}$ and the probability of error in receiving the data is $p_{e,d}$. To achieve URC, $p_{s,(h,d)} \approx 1$, which means that $p_{e,h} \approx 0$ and $p_{e,d} \approx 0$.

$$p_{s,(h,d)} = (1 - p_{e,h})(1 - p_{e,d}) \tag{2.1}$$

Idea for URC

Popovski [10] suggests that the header and data should be enclosed into a single packet to increase $p_{s,(h,d)}$. The problem with this idea, is that it is not energy efficient for users for which the message is not intended for, as they must decode the packet before they can discard the message received. However, this idea is very relevant for services using short messages, such as those in regions R3 and R4 of Figure 2.1. This is due to the header size and data size being of comparable size, which means the header cannot be sent sub-optimally, by using repetition coding and a low header data rate, R_H.

Options to Consider

New transmission methods must consider the trade-off between high reliability and energy efficiency, by using full or partial joint encoding of the header and data. Reliability can be defined as "the probability of successful transmission of a certain amount of data from the transmitter (Tx) to the receiver (Rx) within a certain time frame". To be able to provide high system efficiency and high reliability, there are two options that can be considered as follows:

1. Pre-reserve resources that are idle most of the time.
2. Quality of Experience (QoE) degradation based on the communication condition, to which can be reliably supported. This is referred to as Reliable Service Composition (RSC), and is shown in Figure 2.2 [10].

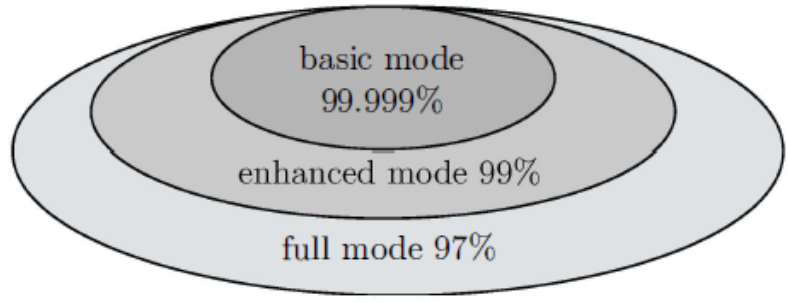

Figure 2.2: The different modes of RSC, based on the communication condition [10]

The first option can be tailored so that only mission critical industrial control applications have access to these pre-reserved resources, thereby being given higher priority than other packets. However, this option does not guarantee high reliability since transmission errors could still occur. The second option can be used specifically for mission critical industrial control applications in the event of worsening communication conditions. However, latency will increase on average since the reliable communication channel has lower availability in progressively 'higher' modes and therefore have to request for RTX more often.

Popovski's solution could work in practice, but the paper did not provide any simulation results or physical results to confirm the proposed idea. It also does not address the situation where the application is located at the cell edge, which may have little to no mobile internet reception. The solution is also for the physical (PHY) layer, which is out of the scope of this project.

2.1.2 Availability Estimation and Indication (AEI)

Reliability can be defined in a few different ways. Schotten *et al.* [4] defined link reliability as "when data can be transmitted successfully within a given deadline", while they defined system reliability as "where a system can accurately indicate the absence of link reliability and also provide link reliability as often as possible".

AEI Proposal for URC

Link reliability requires a reliable transmission link (RTL), where Schotten *et al.* proposed an AEI mechanism which can supposedly predict the availability of RTL under certain conditions. The procedure is summarised as follows:

- The application sends an Availability Request (AR) to the AEI.
- The AEI evaluates the Signal-to-Interference and Noise Ratio (SINR) and/or the Acknowledgement (ACK)/Negative Acknowledgement (NAK) statistics of the RTX protocols used at the link level.
- The AEI then provides an Availability Indicator (AI) for the application based on the original requirements of the AR.

URC can be guaranteed if the link reliability is higher than a certain value, P_{UR}, based on the AR. URC is based on a conditional probability, as shown in Equation (2.2), where $P_{1|1}$ is the probability of a successful transmission ($RTL(\tau) = 1$) occurring, given the AI indicates that RTL is available ($AI(\tau) = 1$). In general, URC can be optimised by improving the modulation and coding techniques, or by improving the channel estimation and prediction techniques.

$$P_{1|1} = \Pr(RTL(\tau) = 1 | AI(\tau) = 1) \tag{2.2}$$

Proposed URC System

Schotten [4] *et al.* proposed a possible URC system with the use of the AEI, Channel Estimation (CEST), and the HARQ ACK/NAK (as shown in Figure 2.3). The procedure between two applications 'A' and 'B', is summarised as follows:

- 'A' sends an AR to 'B', where a CEST is performed for use by the AEI. The AEI receives the CEST after a few milliseconds from 'B', where the AEI predicts if RTL is available or not, based on the minimum tolerated delay from the AR.
- In this example, the AEI indicates that RTL is available to 'A' and to 'B'. 'B' then sends the required data over RTL, and then 'A' will try to decode the URC data.
- If there is a decoding error, then 'A' will send a NAK and 'B' will retransmit the data after 1 ms. If there is a decoding success, then the data is accepted, and the HARQ process sends an ACK back to 'B'.

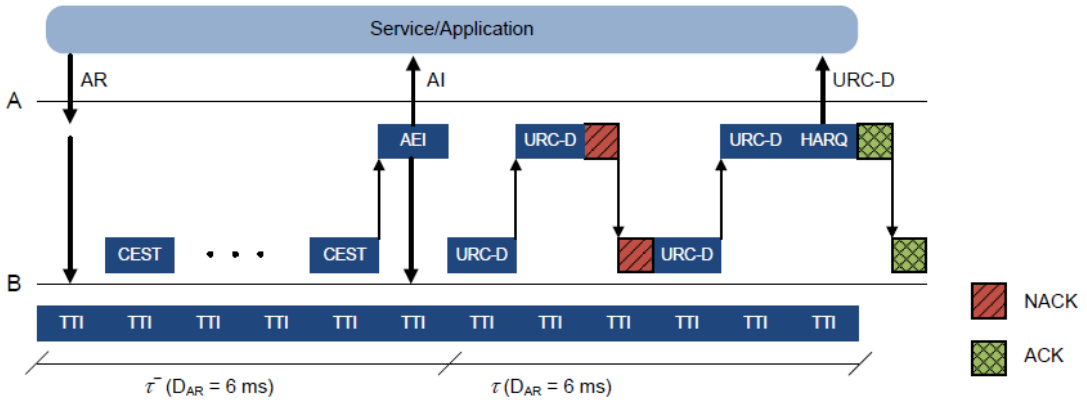

Figure 2.3: Schotten *et al.*'s proposed AEI mechanism in action [4]

Approaches in SINR Prediction

There are 3 different approaches that can be used by the AEI to predict the SINR [4]:

- **Benchmark (B)**: Benchmark considers perfect channel estimation, where it knows the expected SINRs of the 6 Transmission Time Intervals (TTIs) within time slot τ, in advance.
- **Simple (S)**: Simple predicts the SINR by considering the minimum SINR of the last 6 TTIs within time slot τ^{-}, but it does not consider channel variations between time slots due to fast fading.
- **Margin (M)**: Margin compensates for the likely SINR prediction errors while using the S approach, by adding a certain margin SINR to the predicted SINR, to ensure that $P_{1|1} > P_{UR}$.

Performance of the AEI Proposal

Schotten *et al.* [4] shows the performance of their proposed AEI mechanism using a RTL model for a V2V application, which has two cases of without HARQ RTX (QPSK coding rate 3/4, 2 bits per symbol) and with HARQ RTX (64QAM coding rate 5/6, 6 bits per symbol), at Doppler frequencies of 10 Hz, 33.3 Hz, 100 Hz and 200 Hz. The results are summarised in Table II of their paper, where some interesting things can be observed.

The B approach achieves the maximum reliability ($P_{1|1} = 100\%$) regardless of the Doppler frequency, and whether HARQ RTX are used or not. This should be expected due to having perfect channel estimation, however this is difficult to obtain in practice. The S approach has high reliability at lower Doppler frequencies, and can be increased if HARQ RTX are used. However, at higher Doppler frequencies (faster varying channels), the reliability is much lower than the requirement of URC, making this approach essentially useless for URC. The M approach is an improvement on the S approach, and always achieves URC ($P_{1|1} \geq 99.999\%$) by adding a certain SINR margin. However, this is at the cost of reducing the availability of RTL. Using RTX improves the RTL but at the cost of increasing the latency.

2.2 Massive M2M Communication

Massive M2M communications involves billions of machines/devices interconnected and exchanging short messages with each other [1, 2]. Non-network enabled devices can be indirectly connected to the network through Radio Frequency IDs (RFID), by a network enabled device using a RFID reader. This is the basic framework for the IoT/IoE, which is why this is one of the other major operating modes of 5G wireless communications. Mission critical industrial control applications are a subset of applications within the Massive M2M Communications category, which requires high reliability and low latency. These strict requirements pose a challenge, as there are also M2M devices that are not delay sensitive, and exchange redundant or lower important messages between each other.

The following subsections will cover some proposed solutions dealing with some of the problems encountered in Massive M2M Communication. These solutions are M2M-Aware Scheduling in Subsection 2.2.1 and Collision Resolution in Subsection 2.2.2.

2.2.1 M2M-Aware Scheduling

To increase the availability of resources for M2M communications, which include the mission critical industrial control applications, the most straight forward way would be to pre-reserve resources from the overall bandwidth for M2M communications only. However, to give more priority for these mission control industrial control applications, there has to be some scheduling mechanism which facilitates this.

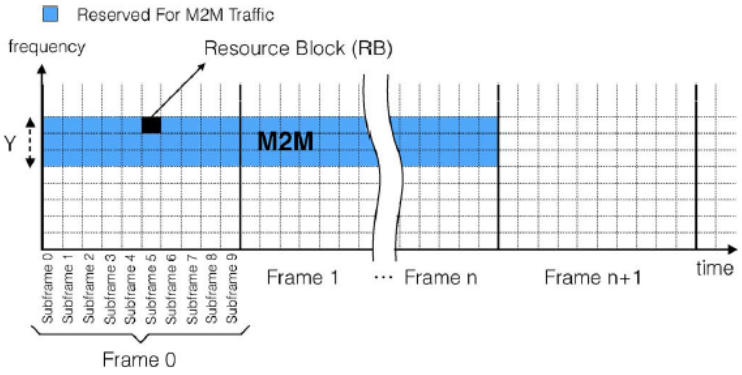

Figure 2.4: Y number of RBs are allocated for M2M communications during each subframe [11]

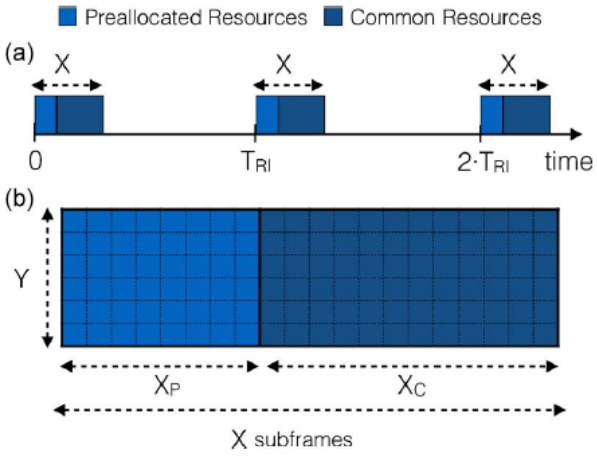

Figure 2.5: Resources reserved for M2M communications are further split into two, to separate general traffic and RTX traffic [11]

Pre-allocation of M2M resources

Madueno *et al.* [11] suggested an idea for increasing the reliability of M2M communications, by reserving resources for M2M applications (as shown in Figure 2.4 on page 15), which was an option mentioned in Subsection 2.1.1 by Popovski. They further suggested the splitting of these resources into preallocated and common resources (as shown in Figure 2.5). The preallocated resources are used for sending intended short messages, while the common resources are used for excess messages and RTX due to transmission errors.

This method relies on the feedback to allocate resources to each individual M2M device. The problem with this idea is that all M2M communications are treated equally and are assumed to require resources periodically. For mission critical industrial control applications, communication may not be periodic, and the equal treatment of the M2M devices may be detrimental.

Abdalla *et al.* [12] suggested a different method to address the reliability issues faced for M2M communications. They suggested an M2M-Aware Scheduler, which dynamically allocates a certain ratio of the available resources for M2M communications. Mission critical M2M communication is taken care of in this idea, through the use of M2M specific QoS Class Identifiers (QCIs). Mission critical M2M communications are within the guaranteed bit rate (GBR) QCI category for M2M communications.

Mission critical M2M devices are given priority over other M2M devices which may not be delay sensitive, which is highly desired. The problem with this idea is that non-M2M

Table 2.1: Example QCI mappings for delay sensitive and non-delay sensitive URC

CQI	Type	Priority	Packet Delay Threshold	Packet Error Loss Rate
10	GBR	1	10 ms	10^{-6}
11	nGBR	2	100 ms	10^{-6}

traffic is given a higher priority during peak periods. This can be a problem if there are many delay sensitive M2M devices requiring resources at the same time. This is due to the ratio of available resources for M2M communications not being allowed to be above a certain value, as shown in Equation (2.3). Human to Human (H2H) devices are allocated the rest of the resources as shown in Equation (2.4). pR is the certain percentage of the available resources that cannot be exceeded for M2M communication.

$$\sum_{j=1}^{m} M_j \leq pR \tag{2.3}$$

$$\sum_{i=1}^{n} U_i \geq (1-p)R \tag{2.4}$$

For the scope of this project, we can tailor the QCI mappings used in [12] for URC instead as shown in Table 2.1. Based on the QCI mappings for URC, the GBR URC will have the highest priority amongst URC traffic, with an acceptable packet delay threshold of 10 ms, and an acceptable packet error loss rate of 10^{-6}. A priority of 1 is equivalent to the priority given to non-GBR (nGBR) IP Multimedia Subsystem (IMS) signalling within the 3GPP specification series, while a priority of 2 is equivalent to a priority given to GBR conversational voice.

Clearly, Mission critical M2M devices should be given the highest priority even if there are non-M2M GBR applications requiring resources. This is due to the consequences being far greater for medical and mission critical M2M devices, if their strict requirements are not fulfilled compared to non-M2M devices.

2.2.2 Collision Resolution

One problem that could be encountered in the future for Massive M2M communications is the event where multiple M2M devices may want to access the network at the

same time. With the current LTE/LTE-A architecture, this would overload the Random Access Channel (RACH) of the uplink direction. This would cause many collisions, and therefore a significant waste of resources and energy. Collision resolution is one important aspect to consider for reliability in Massive M2M communications.

Madueno *et al.* [13] suggested a way to overcome this problem, by proposing a tree-splitting algorithm (as shown in Figure 2.6) which can resolve random access collisions rather than just trying to avoid these occurrences. This algorithm works on top of the RACH procedure and is only implemented when a RACH overload occurs. The example procedure is summarised as follows, where there are 6 devices and 4 preambles:

- Device 1 and 2 select preamble 'A', device 3 and 4 select preamble 'D', and device 5 and 6 select preamble 'C'.
- Collisions occur in subframe 7 due to the preambles 'A', 'C' and 'D', being selected by more than one device.
- The eNodeB now directs device 1 and 2 to compete for preamble 'A' and 'B' in subframe 11, and device 3 and 4 for preamble 'C' and 'D'.
- Device 1 and 2 select the same preamble again and another collision occurs. Device 3 and 4 select different preambles thereby having their contention resolved.
- Device 5 and 6 are directed to compete for preamble 'C' and 'D' in subframe 22, while device 1 and 2 compete again for the same preambles as before.

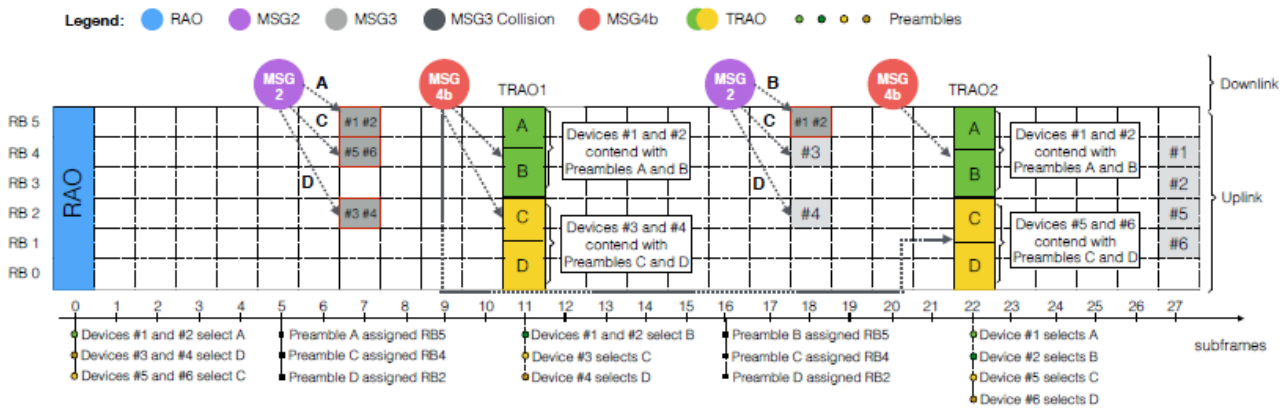

Figure 2.6: Madueno *et al.*'s proposed collision resolution algorithm in action [13]

- Device 1 and 2 select different preambles, and device 5 and 6 also does so, meaning that all contentions are now resolved by subframe 27.

2.3 Hybrid Automatic Repeat Request (HARQ)

HARQ is utilised as a Stop and Wait (SAW) process in LTE [7], where it has to wait for feedback before it can continue. A Transport Block (TB) is constructed as shown in Figure 2.7, by first passing a binary data sequence of length N_a through a turbo encoder, which results in a codeword of length N_b. The mother code rate is $N_a/N_b = 1/3$ for the turbo encoder. The encoded codeword is passed to a rate matcher, which outputs a codeword of length N_c, depending on the Redundancy Version (RV) number and the number of available physical RBs. The rate matched codeword is passed to a bit mapper for each OFDM symbol, to 2-bit Quadrature Phase Shift Keying (QPSK), 4-bit/6-bit Quadrature Amplitude Modulation (16QAM/64QAM). The transmitter (Tx) then sends the TB to the receiver (Rx) over a channel. A typical 8 ms LTE HARQ process follows as shown in Figure 2.8 (on page 20). It involves:

- The decoding of the received TB, followed by a Cyclic Redundancy Check (CRC), and then the encoding of an ACK for a decoding success or NAK for a decoding failure, within 3 ms.
- The ACK/NAK feedback is then transmitted to the Tx, over 1 ms.
- The Tx constructs the next TB if it receives an ACK or a RTX TB for a NAK, and then encodes the next TB or RTX TB within 3 ms.

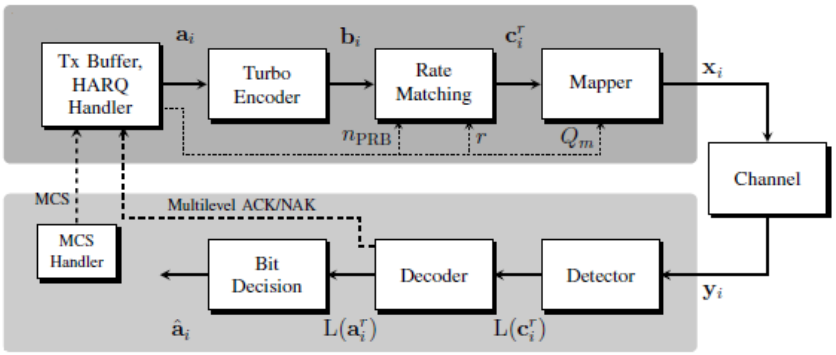

Figure 2.7: HARQ process utilising a turbo encoder and multi-level ACK/NAK. [14]

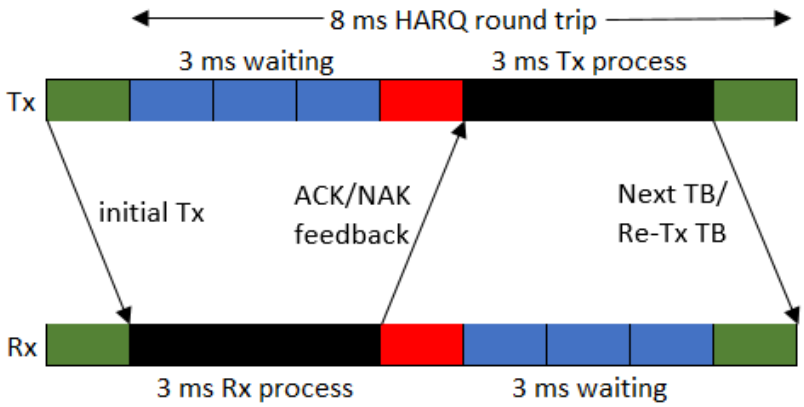

Figure 2.8: The typical 8 ms LTE HARQ process. Data Tx/RTX in green, Tx/Rx processing in black, ACK/NAK feedback in red, and the feedback waiting period in blue.

- The Tx then sends the next TB or RTX TB to the Rx, over 1 ms, to complete one 8 ms LTE HARQ round trip.

One problem with using a single HARQ SAW process is that the process has to wait for feedback during the majority of the 8 ms round trip, which can be observed in Figure 2.8. This is why LTE uses 8 parallel HARQ processes in the uplink direction, since it is of a synchronous nature, or up to 8 in the downlink direction, due to being of an asynchronous nature [7].

2.4 Reliability-based HARQ (RB-HARQ)

Within the related literature, many have been researching on ways to improve the current LTE HARQ scheme based on criteria such as reliability, throughput, etc. Subsection 2.4.1 will introduce what RB-HARQ schemes involve, Subsection 2.4.2 covers a proposed adaptation of RTX sizes based on multi-level ACK/NAK feedback, and Subsection 2.4.3 covers a proposed 'fast HARQ' scheme through the use of relaying.

2.4.1 What is RB-HARQ?

RB-HARQ schemes utilise the reliability estimates in the form of log-likelihood ratios (LLRs) generated by soft-input soft-output (SISO) decoders, which was first proposed by Shea in [15]. The RB-HARQ scheme involves the transmission of the bit positions of the most unreliable information bits as estimated by the SISO decoder. This additional

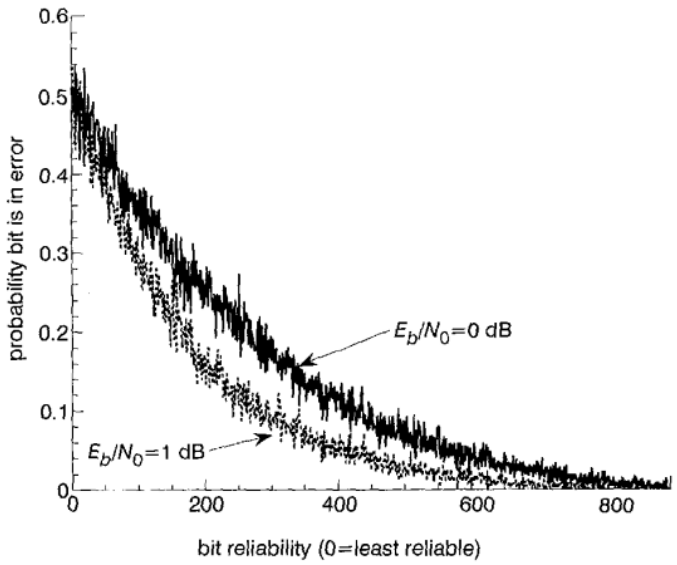

Figure 2.9: The ranking of information bits in terms of their BEP, which can be used to assist in the decision of which bits to RTX [15]

information can be used for additional decoding, and can enable performance close to channel capacity.

The bit error probability (BEP) and the average BEP can be determined by using the LLRs provided by the SISO decoder. The LLR for each information bit at position k can be found by using Equation (2.5), where $\hat{a}_{i,k}$ is the received information bit [14]. If $P(\hat{a}_{i,k} = 1) = P(\hat{a}_{i,k} = -1)$, then $L(a_i, k) = 0$, which represents the case where the information bit is the most unreliable.

$$L(a_i, k) = \log \frac{P(\hat{a}_{i,k} = 1)}{P(\hat{a}_{i,k} = -1)} = \log \frac{P(\hat{a}_{i,k} = 1)}{1 - P(\hat{a}_{i,k} = 1)} \tag{2.5}$$

The BEP can be found by using Equation (2.6) [16], where $L(a_i, k)$ is the LLR for each information bit at position k as in Equation (2.5). When $L(a_i, k) = 0$ for the most unreliable information bit, this equates to the bit contributing to errors 50% of the time, since $P_{b,k} = 1/(1+1) = 0.5$. This is illustrated in Figure 2.9, for the case where there are 900 information bits passed through a 1/3 code rate turbo encoder, for transmission over an Additive Gaussian White Noise (AGWN) channel.

$$P_{b,k} = \frac{1}{1 + e^{|L(a_i,k)|}} \tag{2.6}$$

The average BEP can be found by Equation (2.7), where $P_{b,k}$ is the BEP of an infor-

mation bit at position k as in Equation (2.6), and where N_a is the length of the original binary data sequence.

$$P_b = \frac{1}{N_a} \sum_{k=1}^{N_a} P_{b,k} \tag{2.7}$$

2.4.2 RTX size Adaptation

Woltering *et al.* [14] proposed the adaptation of RTX sizes based on either the Signal-to-Noise Ratio (SNR) or the reliability-based information from a SISO decoder, with the goal of optimising the throughput. Multi-level ACK/NAK is also utilised to adapt the RTX size, depending on which level of NAK feedback is received. The purpose of this scheme, is to reduce the overhead associated with the HARQ feedback. The scheme chooses the next RTX size $n_{\text{PRB}}^{(i+1)}$ by considering the average BEP from Equation (2.7), by using Equation (2.8), where $N_{\text{PRB}}^{\max}$ is the number of physical RBs available, and $\lceil \cdot \rceil$ is the ceiling function.

$$n_{\text{PRB}}^{(i+1)} = \left\lceil \frac{P_b}{0.5} N_{\text{PRB}}^{\max} \right\rceil \tag{2.8}$$

In this particular paper, $n_{\text{PRB}}^{(i+1)} = 5$ for NAK_1, $n_{\text{PRB}}^{(i+1)} = 14$ for NAK_2 and $n_{\text{PRB}}^{(i+1)} = 21$ for NAK_3, for a carrier bandwidth of 5 MHz ($N_{\text{PRB}}^{\max} = 25$ RBs). Obviously NAK_3 feedback signifies that a higher level of redundancy for the RTX is required since P_b ($0.4 < P_b \leq 0.42$) is almost at the maximum of $P_b = 0.5$. There are different sorts of RTX size mappings that can be chosen, but this is a trade-off between delay and throughput.

While the proposed HARQ scheme has increased the throughput relative to the current LTE HARQ, the results show that the RTX size adaptation scheme is not suitable for the motivating applications of this project in poor channel conditions, due to the unacceptable delay incurred by successive RTX.

2.4.3 Fast HARQ

Chung *et al.* [17] proposed a 'fast HARQ' scheme through the use of a partner in a cooperative diversity (CD) environment. This partner can be utilised for RTX purposes by storing a duplicate of the packet sent from the Tx through a broadcast. The Rx sends the same ACK/NAK feedback as in the current LTE HARQ schemes for a decoding

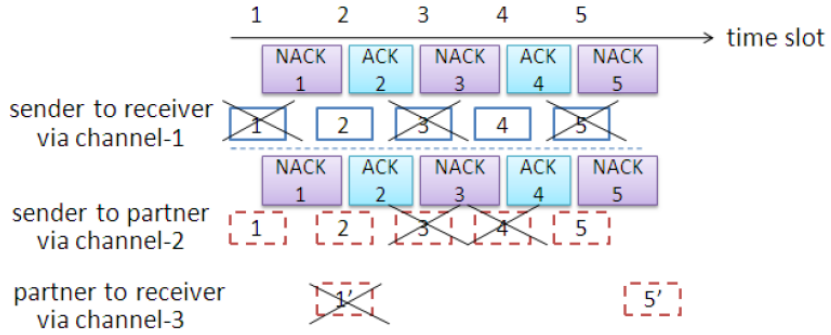

Figure 2.10: Chung *et al.*'s fast HARQ scheme, where the RTX of erroneous packets only occurs from the Tx.

success or decoding failure respectively to the Tx and the partner. Depending on factors such as the application layer protocol data unit (APDU) length and the channel SNRs amongst the Tx, partner and Rx, there are 2 RTX policies that could be chosen. Either RTX of the erroneous packet is done by the Tx only (as shown in Figure 2.10), or it is done by both the Tx and the partner (as shown in Figure 2.11).

This HARQ scheme uses redundancy in the form of storing a duplicate packet at the partner, to increase the reliability of transmission. However, this scheme can be a problem if there is no willing device to act as the partner in the event of the need for RTX. Other problems that can be observed, is that there are security concerns if data to be transmitted is of a sensitive nature, as it needs to be shared with the partner to enable this fast HARQ scheme. Lastly, limiting the scheme to only one RTX would fulfil the ultra-low latency requirement for the motivating applications, but it may not be suitable for the URC operating mode.

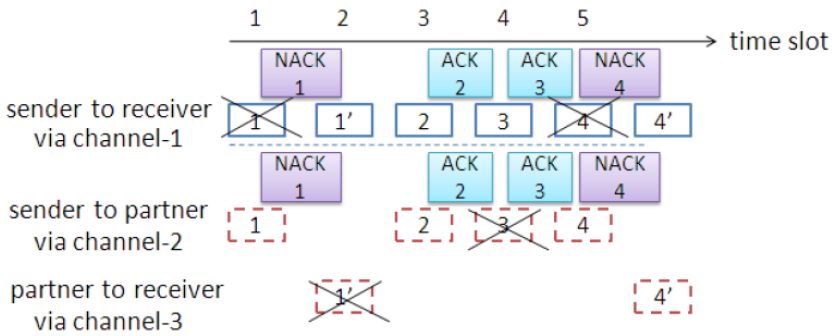

Figure 2.11: Chung *et al.*'s fast HARQ scheme, where the RTX of erroneous packets occurs from both the Tx and partner.

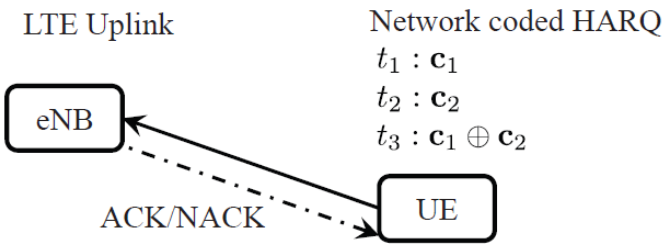

Figure 2.12: NC-HARQ where the network coded packet ($\mathbf{c}_1 \oplus \mathbf{c}_2$) is transmitted in timeslot t_3.

2.5 Network-Coded HARQ (NC-HARQ)

Lang *et al.* [18] proposed a HARQ scheme based on network coding, referred to as Network-Coded HARQ (NC-HARQ). The purpose of this scheme is to increase the throughput by applying network coding on 2 previously erroneous packets from 2 different parallel HARQ processes, $\mathbf{c}_1$ and $\mathbf{c}_2$, into one single packet of $\mathbf{c}_3$ by using the exclusive-OR (XOR) operation, as shown in Figure 2.12. This scheme is supposed to operate with a duration of 3 time slots (t_1, t_2 and t_3) as opposed to 4 time slots if the current LTE HARQ scheme is used.

The problem with using network coding is that a subset of systematic bits is retransmitted, which means that the reliability information provided by the SISO decoder must be reconstructed. This reliability information can be obtained from the network coded packet $\mathbf{c}_3$ by the use of the box plus operation or soft XOR operation, $L_1 \boxplus L_2$, where L_1 is the LLR of $\mathbf{c}_1$ and L_2 is the LLR of $\mathbf{c}_2$. The soft XOR of L_1 and L_2 can be found by Equation (2.9), and their approximation can be found by Equation (2.10), where sign$(\cdot)$ is the sign function.

$$L_1 \boxplus L_2 = 2\tanh^{-1}\left(\tanh\left(\frac{L_1}{2}\right)\right) \cdot \left(\tanh\left(\frac{L_2}{2}\right)\right) \tag{2.9}$$

$$L_1 \boxplus L_2 \approx \text{sign}(L_1) \cdot \text{sign}(L_2) \cdot \min(|L_1|, |L_2|) \tag{2.10}$$

Chapter 3

Proposed A-HARQ Scheme

In this chapter, an improved HARQ scheme called 'Adaptive HARQ' (A-HARQ) is proposed, to cater for mission critical industrial control and medical applications, which are used as motivations. A-HARQ involves pre-constructing and pre-encoding differing RVs with different combinations of systematic bits and parity bits, and storing them in the Tx buffer to ensure a quick RTX. TTI bundling is utilised to increase the number of RTX within a 4 ms time period, since the defined goal is to reduce the delay incurred by multiple RTX when using the legacy HARQ. Channel Quality Indicator (CQI) reports are also utilised to assist the eNodeB in selecting the best sub-bands to send the RTX, since the use of TTI bundling could still result in a decoding failure if the same poor quality channel is used. A resource allocation solution for URC when utilising A-HARQ, is also proposed in this chapter.

This chapter is organised as follows: Section 3.1 details all the design ideas that were considered prior to the current proposed A-HARQ and the proposed resource allocation for URC, and concludes with an outline on the proposed A-HARQ. Section 3.2 describes how A-HARQ operates and also shows an example of how A-HARQ is used for a specified carrier bandwidth. Section 3.3 outlines how resources are allocated for URC when utilising the proposed A-HARQ.

3.1 Design Ideas for A-HARQ

There were many modifications of HARQ presented before the current proposed A-HARQ and the proposed resource allocation for URC was decided upon. These modifications are

the pre-allocating of resources for URC in Subsection 3.1.1, the utilisation of licensed and unlicensed frequency bands in Subsection 3.1.2, the duplicating of transmitted packets to increase the reliability in Subsection 3.1.3, the splitting of resources for URC in Subsection 3.1.4, and the use of clustering in Subsection 3.1.5. The proposed A-HARQ is outlined in Subsection 3.1.6.

3.1.1 Pre-allocated Resources for URC

One of the first ideas explored for enabling the URC operating mode in 5G was to pre-allocate resources for use by URC applications only (as shown in Figure 3.1), similar to what was proposed by Madueno *et al.* [11] and Popovski [10] in Subsection 2.2.1 and 2.1.1, respectively. The pre-allocating of resources would mean that URC applications always have access to a portion of the available 'n' number of resource blocks for the transmission of data at almost 100% of the time.

However, there were some problems that could be observed with this idea, such as what if there were no URC application that requires resources at a particular time? Pre-allocating resources would lead to a waste of resources, which could be used by other non-URC applications. Other consideration was the energy efficiency of low power devices such as those in the periodic Massive M2M Communication category. These specific devices would need a dynamic way of resource allocation for URC, when they are currently not in their idle mode.

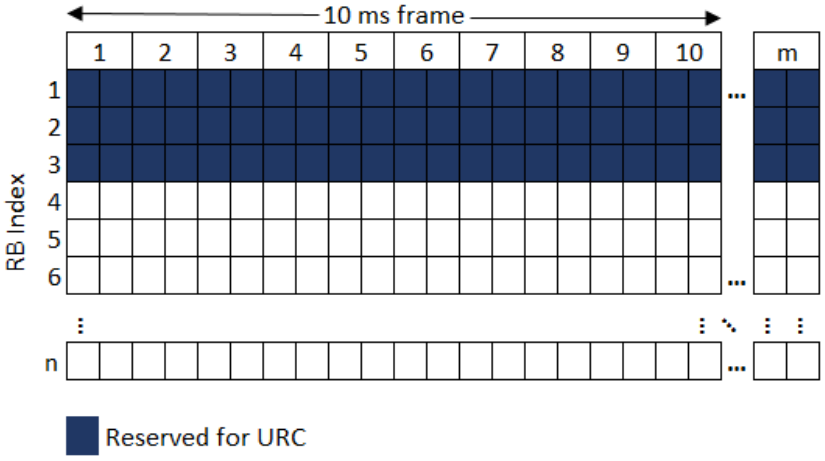

Figure 3.1: In this example, 3 RBs are reserved for URC applications for every slot in time. RBs are covered in Subsection 1.2.1 and Figure 1.1.

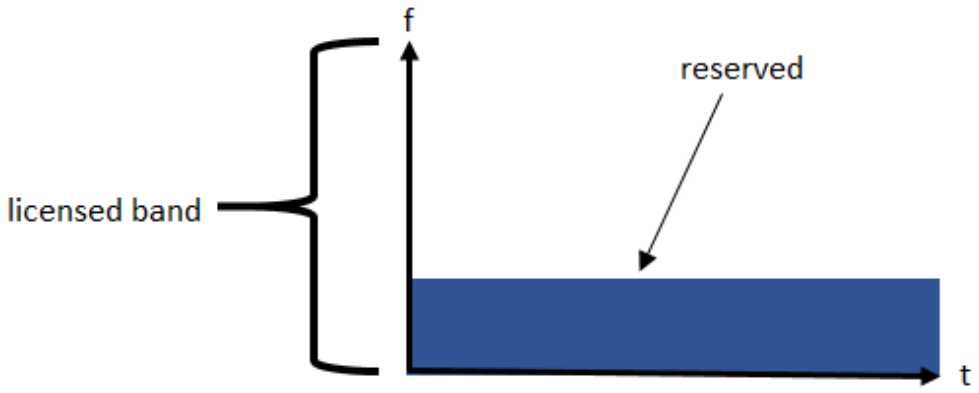

Figure 3.2: Resources are reserved for URC within the purchased licensed frequency bands.

3.1.2 Licensed and Unlicensed Frequency Bands

Another idea considered was to utilise licensed frequency bands for URC (as shown in Figure 3.2). Mission critical industrial control and medical application users could have more control over the interference experienced if data was transmitted over licensed frequency bands [10]. This can increase the availability of a reliable channel to transmit data, but the costs of purchasing licensed frequency spectrum is very high. However, as we envision what different components will be involved in future 5G communications, some degree of interference could still be experienced during Massive M2M Communication and the ultra-dense small cells, which currently have limited coordination.

Unlicensed frequency bands could also be used through the concept of offloading. Mobile cellular traffic of mission critical industrial control and medical applications can be offloaded to other network technologies such as Wi-Fi [19] (for outdoor traffic) or femtocells (for indoor traffic). Dual band devices (which can operate at either 2.4 GHz or 5 GHz) use the 5 GHz frequency band for the WiFi Institute of Electrical and Electronics Engineers (IEEE) 802.11a/ac/h/j/n standards. Offloading is especially useful in situations where the LTE/LTE-A reception is too poor to be able to transmit data with high reliability. It would also mean that mission critical industrial control and medical application users would not have to pay high prices for the use of a licensed frequency band. However, this is provided that the interference could be somehow minimised enough to ensure high reliability of data transmission.

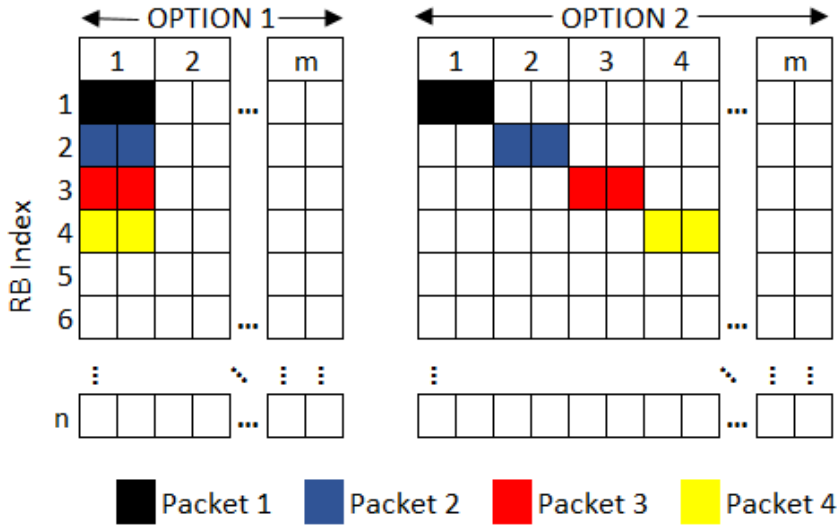

Figure 3.3: In this example, 4 packets with the same contents are sent by the Tx depending on which option is being considered.

3.1.3 Duplicating Packets

The duplicating of packets sent by the Tx was also considered to increase the reliability of data transmission. There were two options that were to be considered (as shown in Figure 3.3):

1. Have all the duplicate packets in the same subframe but occupy different portions of the available bandwidth

2. Have all the duplicate packets in different subframes and also occupy different portions of the available bandwidth

The first option was in relation to maintaining channel diversity, while the second option was in relation to maintaining time diversity. If the duplicate packets were in the same subframe and were occupying neighbouring channels within the available bandwidth, then the packets would experience interference and therefore would have a low probability of being transmitted successfully. The same thing would occur if the packets were in different subframes and also occupied the same channel. Another problem that could be observed is how to handle the situation when a HARQ ACK is received from the Rx, while the Tx is still sending a different duplicate packet. The Rx could discard all future duplicate packets as it still takes time for the Tx to stop sending duplicate packets, based on the ACK/NAK feedback.

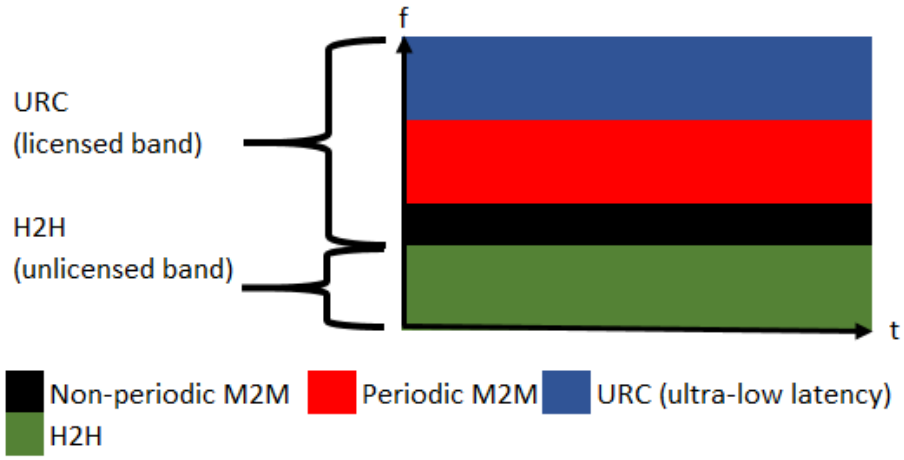

Figure 3.4: Resources split between URC and H2H, to handle different priorities from the use of standard QCIs, and to also simplify the scheduling of packets.

3.1.4 Splitting of Resources for URC

The splitting of resources has already been considered before by Abdalla *et al.* [12] within Subsection 2.2.1, in terms of H2H and M2M traffic. However, for the scope of this project, the resources are proposed to be split between H2H and URC traffic instead (as shown in Figure 3.4). URC resources are further split as follows:

- URC with ultra-low latency
- periodic M2M (low priority/redundant messages)
- non-periodic M2M

Standard QCIs is also used, as mentioned in Subsection 2.2.1, to group application data in terms of priority and to simplify the scheduling of packets.

3.1.5 Use of Clustering

One problem with the idea in Subsection 3.1.4 is that for Massive M2M Communication in the uplink direction, there is the possibility of many devices trying to access the RACH at the same time. This would cause many Random Access Opportunity (RAO) collisions, leading to a waste of resources and device energy.

One possible solution for this problem is the use of clustering, as shown in Figure 3.5 (on page 30). Applications assigned with the standard QCIs can be grouped into clusters.

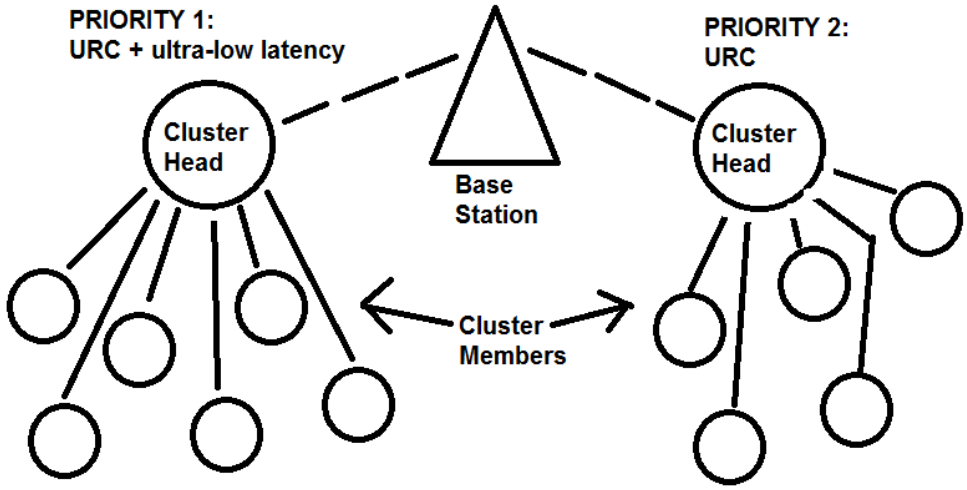

Figure 3.5: The use of clusters to reduce the RAO collusions for URC, with QoS Identifiers to determine the priority for a particular device.

One device in the cluster can act as the cluster head, thereby being responsible for communications between the eNodeB and all the other cluster members. This would lead to a reduction in RAO collisions, and savings of resources and energy. However, the cluster head would require high energy usage, which means it is necessary to dynamically assign the role of the cluster head amongst the other cluster members, and thereby enable the cluster as a whole to survive for a longer period of time.

3.1.6 Adaptive HARQ (A-HARQ)

Motivating applications would experience increased reliability when using the legacy HARQ. However, it can be observed that if 1 or more RTX are required to ensure a successful decoding, then 8 ms of delay is incurred for every RTX. This makes the current LTE HARQ process unsuitable for our motivating applications during poor channel conditions, since there is a high probability of RTX occurring. The proposed A-HARQ solves this problem of unacceptable delay during poor channel conditions for the motivating applications. CQI reports are used to assist the eNodeB in choosing the best sub-bands within the relevant carrier bandwidth to transmit data. In LTE, periodic UE-selected sub-bands CQI reports can be sent to the eNodeB over the Physical Uplink Control CHannel (PUCCH) every 2 ms. A particular CQI number corresponds to the MCS which ensures that the BLock Error Rate (BLER) is $\leq 10\%$ [7].

The eNodeB is set up to pre-construct and pre-encode multiple TBs of differing RVs,

and then store them in the Tx buffer. The eNodeB can freely schedule RTX of a different RV, within 1 ms after receiving a periodic CQI report, due to having already constructed the RTX TBs ahead of time, and being of asynchronous nature. TTI bundling is used in LTE to increase the amount of RTX within a 4 ms time period, where the eNodeB can send 4 TBs in consecutive subframes [7]. One advantage of using TTI bundling is that there is only one set of controlling signalling required for a TTI bundle. This reduces the signalling overhead required for resource allocation and HARQ ACK/NAK feedback.

3.2 Example of Proposed A-HARQ

The RTX process under the proposed A-HARQ is as follows:

- The eNodeB transmits RV #1 to RV #4, in consecutive subframes SF7 to SF10, using TTI bundling.
- The Rx receives RV #1 to RV #4 in consecutive subframes SF8 to SF11 and decodes each RV in consecutive subframes SF9 to SF12, and then stores the received TBs in the Rx buffer.
- After all the RTX TBs in a TTI bundle are stored in the Rx buffer, the Rx applies packet combining on the previously received erroneous TB and RV #1 to RV #4 in SF13 to attempt to decode the packet.

The probability of a successful decoding should be very high since there are many different combinations of systematic and parity bits being combined together. The proposed A-HARQ is shown in Figure 3.6.

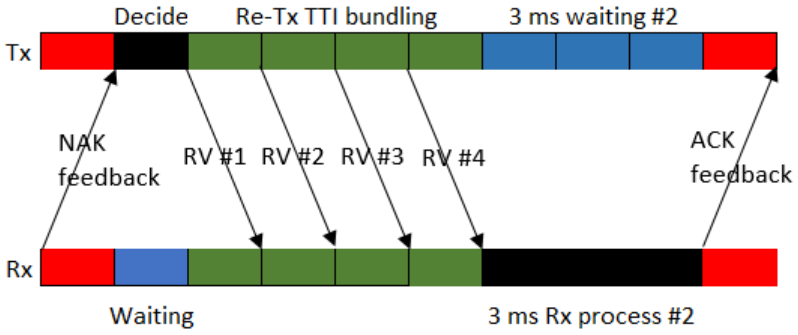

Figure 3.6: Proposed A-HARQ

One problem with using TTI bundling is that the probability of successful decoding under poor channel conditions will still be very low, since the initial TB and the 4 RTX TBs will have been sent on the same channel. Only if the channel conditions have changed to a more favourable condition, then the probability of a successful decoding will be high using the proposed A-HARQ. The proposed A-HARQ is extended, by utilising the CQI reports mentioned in Subsection 3.1.6, to assist the eNodeB in selecting the best sub-bands to send TTI bundles across to the Rx. Parallel HARQ processes are used within differing channels in the best sub-bands, where packet combining is applied to the erroneous TB and RTX TBs.

Using a 5 MHz carrier bandwidth (or $N_{\mathrm{RB}}^{\mathrm{DL}} = 25$ RBs), the sub-band size k for periodic UE-selected sub-bands CQI reports is 4 RBs (as shown in Table 3.1) [20]. This results in 7 sub-bands with index SB1 to SB7, where SB7 has only 1 RB according to Equation (3.2), since Equation (3.1) is true. $\lceil \cdot \rceil$ is the ceiling function, and $\lfloor \cdot \rfloor$ is the floor function, respectively.

$$\left\lceil \frac{N_{\mathrm{RB}}^{\mathrm{DL}}}{k} \right\rceil - \left\lfloor \frac{N_{\mathrm{RB}}^{\mathrm{DL}}}{k} \right\rfloor > 0 \tag{3.1}$$

$$N_{\mathrm{RB}}^{\mathrm{DL}} - k \times \left\lfloor \frac{N_{\mathrm{RB}}^{\mathrm{DL}}}{k} \right\rfloor \tag{3.2}$$

The 7 sub-bands are further divided into 2 bandwidth parts of B1 (consisting of SB1 to SB3) and B2 (consisting of SB4 to SB7), as shown in Figure 3.7. The UE reports a wide-band CQI value (which represents the average channel quality of the entire bandwidth),

Table 3.1: The division of the system bandwidth (BW) into 'N' sub-bands and further division into bandwidth parts.

System BW (MHz)	**System BW (RBs)**	**Sub-band size ('k' RBs)**	**BW parts (J)**
1.4	6	N/A	1
3	15	4	2
5	25	4	2
10	50	6	3
15	75	8	4
20	100	8	4

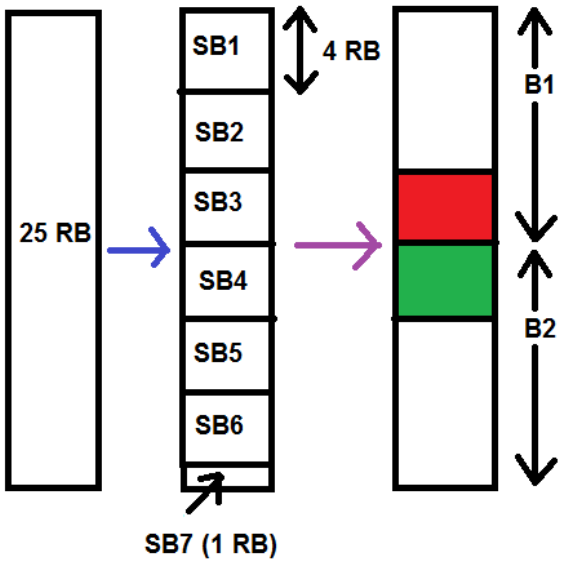

Figure 3.7: The carrier bandwidth is split into sub-bands and bandwidth parts for UE-selected sub-band CQI reports. The selected sub-bands for B1 and B2 are coloured in red and green respectively.

the CQI for the selected sub-band within each of B1 and B2, and the corresponding selected sub-band indexes to the eNodeB [7].

The eNodeB can now send multiple TTI bundles along the best selected sub-bands according to the UE-selected sub-band CQI reports, using parallel HARQ processes. The UE applies packet combining on RV #1 to RV #4, after the entire TTI bundle has been received, to attempt to decode the erroneous packet. The modified A-HARQ is shown in Figure 3.8, where it should lead to an increase in the probability of a successful decoding at the Rx and of meeting the low-latency deadline of the motivating applications, even under poor channel conditions.

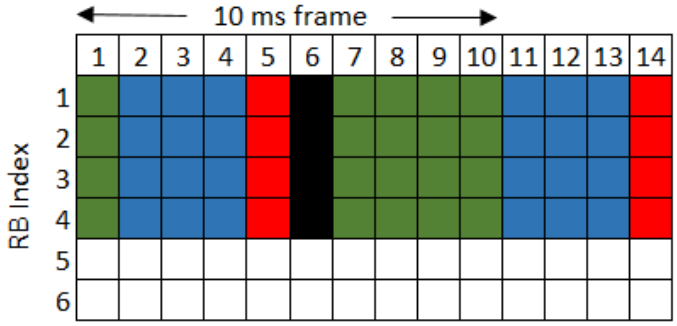

Figure 3.8: The proposed A-HARQ in the downlink direction.

3.3 Resource Allocation for URC and A-HARQ

The proposed resource allocation for URC and A-HARQ consists of a combination of some of the ideas developed in Section 3.1. The main idea is to dynamically allocate resources for the motivating applications of mission critical industrial control and medical applications. This done by utilising the UE-selected sub-band CQI reports sent by the UE to the eNodeB every 2 ms, as mentioned in Subsection 3.1.6. This is, however, unnecessary for cases where no RTX is required, since the motivating applications would already experience URC and ultra-low latency even under the legacy HARQ.

Based on the project scope in Section 1.4, the majority of motivating applications will be experiencing poor channel conditions. The best sub-band in each bandwidth part is selected for use to send TTI bundles under A-HARQ, as detailed in Section 3.2. This leads to a dynamic allocation of resources because these best sub-bands are reserved for use by the motivating application, during the period where the intended data sent by the eNodeB has not yet been correctly decoded. Motivating applications can also experience high resource availability by using licensed bands as mentioned in Subsection 3.1.2.

Since a sub-band consists of multiple RBs (this is 4 for a 5 MHz bandwidth for example), then the second option (as mentioned in Subsection 3.1.3) of duplicate packets in different subframes and occupying different portions of the sub-band can be used by multiple UEs as shown in Figure 3.9. This option would preserve channel diversity and time diversity, thereby reducing the interference experienced without diversity.

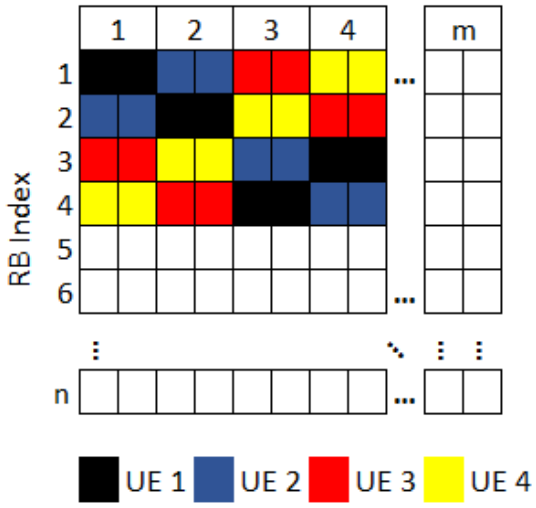

Figure 3.9: Proposed Resource Allocation for URC with 4 UEs.

Chapter 4

Analysis and Simulation Results

In this chapter, a performance analysis is carried out based on the proposed A-HARQ scheme from Chapter 3. A MATLAB simulation was set up to compare the legacy HARQ with the proposed A-HARQ. This chapter is organised as follows: Section 4.1 provides the fundamentals required to analyse the performance based on the MATLAB simulation results, Section 4.2 shows how the MATLAB simulation was set up, while Section 4.3 presents the simulation results, comparing the legacy HARQ with the proposed A-HARQ based on specific criteria such as throughput vs SNR and delay vs SNR.

4.1 Analysis Fundamentals

4.1.1 Probability of Decoding Failure and CQI Value

The probability of decoding failure for a single HARQ process on the ith transmission is given by Equation (4.1), where $R_1 = Q_m R_c$ is the target rate per symbol of the first transmission.

$$P(R_1) = \text{Prob}(\text{ACMI}(\text{SNR}, i) \leq R_1) \tag{4.1}$$

Q_m is the modulation index of 2 for QPSK, 4 for 16QAM and 6 for 64QAM, and $R_c = N_a/N_c$ is the overall effective code rate, which is the ratio of the length of the original data sequence N_a and the length of output codeword from the rate mapper algorithm N_c. $\text{ACMI}(\text{SNR}, i)$ is the accumulated mutual information (ACMI) of the Signal-to-Noise Ratio (SNR) for the first transmission and subsequent RTX [14].

The CQI value reported by the UE is based on the measured SNR, which ensures that the BLER is $\leq 10\%$. This means that CQI is a function of SNR and the BLER, which is shown in Equation (4.2):

$$\text{CQI} = f(\text{SNR}, \text{BLER} \leq 10\%) \tag{4.2}$$

4.1.2 Trade-off between Throughput and Delay Reduction

The normalised throughput can be found from Equation (4.3), where b_c is the number of correctly decoded bits, and b_t is the number of bits transmitted, The normalised throughput is expressed as a % of the overall capacity. The normalised throughput can also be expressed in terms of bits per second (bps), by using Equation (4.4), where R is the maximum bit rate of the available bandwidth.

$$\eta = \frac{b_c}{b_t} \text{ [\%]} \tag{4.3}$$

$$\eta_A = \frac{b_c}{b_t} \times R \text{ [bps]} \tag{4.4}$$

R can be calculated as shown in Equation (4.5), where $S_R = 66.667\ \mu s$ is the OFDM symbol rate, b_M is 2 (QPSK), 4 (16QAM) or 6 (64QAM), and S_N is the number of subcarriers for the specified carrier bandwidth. In this case, $S_N = 25$ [RBs] $\times$ 12 [subcarriers/RB] = 300 for a 5 MHz carrier bandwidth, and $b_M = 4$ for 16QAM, which results in $R = 18$ Mbps.

$$R = \frac{1}{S_R} \times b_M \times S_N \text{ [bps]} \tag{4.5}$$

The delay at the Rx side, for the legacy HARQ can be found as shown in Equation (4.6), where r_c is the number of RTX. For the proposed A-HARQ, the delay can be found as shown in Equation (4.7), where r_p is the number of TTI 4 TBs RTX bundles.

$$d = 4 + 8 \times r_c \text{ [ms]} \tag{4.6}$$

$$d_A = 4 + 9 \times r_p \text{ [ms]} \tag{4.7}$$

Table 4.1: System parameters used for the LTE MATLAB simulation

Parameters	Setup
Carrier frequency	2.3 GHz
Bandwidth	5 MHz
Number of Physical Resource Blocks	25
Number of Transmit Antennas	1
Modulation	16QAM and QPSK
Turbo code rate	1/3
Max no. of RTX	4
Channel	AWGN

4.2 MATLAB Simulation

MATLAB was used to perform a simulation with the purpose of comparing the legacy HARQ with the proposed A-HARQ. The LTE System Toolbox functionality and some code examples on the Mathworks website were used to simplify the simulation process [21].

4.2.1 Legacy HARQ

The legacy HARQ was implemented using the MATLAB simulation parameters as shown in Table 4.1. The following codes within Code Snippets 1 to 4 show how the most relevant simulation parameters (bolded within the text) were set up for the simulations.

Code Snippet 1

```
%% Cell-wide Settings
enb.NDLRB = 25;              % No of Downlink RBs in total BW

%% PDSCH Transmission Mode Configuration
pdsch.NLayers = 1;             % No of layers to map the transport block
pdsch.TxScheme = 'Port0';      % Transmission scheme
pdsch.Modulation = {'16QAM'};      % Modulation
pdsch.RV = 0;                  % Initialize Redundancy Version
pdsch.PRBSet = (0:enb.NDLRB-1).'; % Define the PRBSet
```

The carrier bandwidth is set to 5 MHz (25 RBs), the number of transmission layers is set to 1, and the TxScheme is set for 'Port0' single antenna port transmission. The modulation is set to 16QAM, and the initial RV is 0 to indicate more systematic bits and less parity bits. The Physical RB Set is set by an array of index 0 to index 24.

Code Snippet 2

```
%% Downlink Coding Configuration
transportBlkSize = 4968;                    % Transport block size
[~,pdschIndicesInfo] = ltePDSCHIndices(enb,pdsch,pdsch.PRBSet);
codedTrBlkSize = pdschIndicesInfo.G;        % Available PDSCH bits
dlschTransportBlk = randi([0 1], transportBlkSize, 1); % DL-SCH data bits

redundancyVersions = 0:3; % Possible redundancy versions
```

The TB size in bits is set according to 16QAM modulation ($Q_m = 4$), specifically $I_{\text{MCS}} = 12$ and $I_{\text{TBS}} = 11$ within the 3GPP specification series [20], as shown in Figure 4.1. N_c from Subsection 4.1.1 can be obtained by calling **ltePDSCHIndices**$(\cdot)$ and then **pdschIndicesInfo.G**, which would result with $N_c = 12120$ bits for 16QAM modulation, 'Port0' Tx scheme, and 1 layer for the TB mapping. The number of RVs or maximum RTX allowed is set to 4 by an array of index 0 to 3.

MCS Index I_{MCS}	Modulation Order Q_m	TBS Index I_{TBS}
0	2	0
1	2	1
2	2	2
3	2	3
4	2	4
5	2	5
6	2	6
7	2	7
8	2	8
9	2	9
10	4	9
11	4	10
12	4	11
13	4	12
14	4	13
15	4	14
16	4	15

I_{TBS}	N_{PRB}	
	25	26
0	680	712
1	904	936
2	1096	1160
3	1416	1480
4	1800	1864
5	2216	2280
6	2600	2728
7	3112	3240
8	3496	3624
9	4008	4136
10	4392	4584
11	4968	5352
12	5736	5992
13	6456	6712
14	7224	7480
15	7736	7992

Figure 4.1: Left: I_{TBS} based on the I_{MCS} and Q_m. **Right**: Corresponding TB Size in bits based on the I_{TBS}.

Code Snippet 3

```
% Define soft buffer
decState = [];

blkCRCerr = 1; % TB CRC initial value
```

The soft buffer for Type-II HARQ Incremental Redundancy (IR) is initially defined as an empty array, which is later used to store the bits sent over the AWGN channel, in the event of a transmission failure. It allows the soft combining of the initial erroneous packet and additional RVs. The **blkCRCerr** is set to 1 for a *while loop*, which is used to check if a new RV is required for RTX, based on the CRC result after decoding.

Code Snippet 4

```
SNR = [0:0.1:7]; % Noise level in decibels (dB)

% Add noise to pdschSymbols to create noisy complex modulated symbols
pdschSymbolsNoisy = awgn(pdschSymbols,SNRIn);

% PDSCH receiver processing
rxCW = ltePDSCHDecode(enb, pdsch, pdschSymbolsNoisy);

% DL-SCH channel decoding
[rxBits, blkCRCerr, decState] = lteDLSCHDecode(enb, ...
pdsch, transportBlkSize, rxCW, decState);
```

After the PDSCH symbols (**pdschSymbols**) have been generated in previous steps within the code, then the symbols are sent through the AWGN channel at each of the SNR values, $m = \{0, 0.1, 0.2, \ldots, 7\}$. This is done by using a *for loop*, where **SNRIn** is set to each SNR value for each iteration. The received noisy complex modulated symbols (**pdschSymbolsNoisy**) are processed, and then decoding of the received bits is performed. Decoding involves comparing the **rxBits** contents with the **dlschTransportBlk** bits, and **blkCRCerr** is returned as 1 if there are any bits that do not match. This CRC failure results in the code continuing with the RTX *while loop* until the CRC succeeds or the number of RTX reaches the maximum amount of times allowed of 4.

Throughput and delay results ($R_{t(m,n)}$ and $R_{d(m,n)}$) for each of these SNR values were recorded in a column of the throughput and delay result matrices, respectively. The simulation was run 1000 times, $n = \{0, 1, 2, \ldots, 999\}$, with the results of each simulation run occupying a column in the result matrices. Each element of the result matrices is therefore a function of the SNR and simulation run index, as shown in Equation (4.8).

$$R_{t(m,n)/d(m,n)} = \begin{pmatrix} r_{0,0} & r_{0,1} & r_{0,2} & \cdots & r_{0,999} \\ r_{0.1,0} & r_{0.1,1} & r_{0.1,2} & \cdots & r_{0.1,999} \\ r_{0.2,0} & r_{0.2,1} & r_{0.2,2} & \cdots & r_{0.2,999} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{7,0} & r_{7,1} & r_{7,2} & \cdots & r_{7,999} \end{pmatrix} \tag{4.8}$$

The recorded results of the simulation in the result matrices were averaged out for each SNR value, to smooth out the variations in the simulation results. This is shown in Equation (4.9), where $\overline{R_{t(m,n)/d(m,n)}}$ is the average of a particular SNR value for all the 1,000 simulation runs. Results can vary at certain SNR levels where the number of required RTX differ by 1, since the noise added to the OFDM symbols within the simulated AWGN channel is random.

$$\overline{R_{t(m,n)/d(m,n)}} = \begin{pmatrix} \overline{r_{0,n}} \\ \overline{r_{0.1,n}} \\ \overline{r_{0.2,n}} \\ \vdots \\ \overline{r_{7,n}} \end{pmatrix} \tag{4.9}$$

4.2.2 A-HARQ

The legacy HARQ and A-HARQ are exactly the same in implementation if a RTX is not required. If a RTX is required, then the A-HARQ specific code is executed in MATLAB. One difference here is that 16QAM and QPSK modulation are both used for A-HARQ. Code Snippets 5 and 6 show the key differences between A-HARQ and legacy HARQ code in MATLAB.

Code Snippet 5

```
pdsch4.Modulation = {'QPSK'};      % Modulation
```

The modulation is set to QPSK if there is a RTX required under A-HARQ. This results in a smaller TB size due to the use of the more robust QPSK modulation. The resulting TB size is set according to the 3GPP specification series as before in Subsection 4.2.1 for a lower MCS index/TBS index [20].

Code Snippet 6

```
while (blkCRCerr >= 1) && (rvIndex <= 4)
        if(rvIndex == 0)
                %%%% The rest of the legacy HARQ code %%%%
        else
                %%%% The rest of the A-HARQ code %%%%
end
```

An *if statement* is used as above, where if the initial RV index is 0, then the usual legacy HARQ is used. If the RV index is > 0, then the A-HARQ part of the code is executed. The **rvIndex** is incremented by 1 for every iteration of while loop, which means that the exit condition is when 4 RTX has been requested.

Code Snippet 7

```
HARQMode = [1 2]; % 1 = legacy HARQ, 2 = proposed A-HARQ

for HARQ = HARQMode
        for SNRIn = SNR
                if(HARQ == 1)
                        for Test = 0:1:999
                        % legacy HARQ code
                else
                        for Test = 0:1:999
                        % A-HARQ code
                        end
                end
        end
end
```

An array **HARQMode**, as shown in Code Snippet 7, is used to be able to test the legacy HARQ and A-HARQ separately for each SNR value over 1000 simulation runs for each SNR value. This also enables the storage of results in the relevant arrays for comparison using Figure 4.2 and Figure 4.3, as shown in the following Section 4.3.

4.3 Analysis and Results

4.3.1 Probability of Decoding Failure and CQI Value

Referring to Equation (4.1), we can figure out what R_1 is by finding out what R_c is. N_c is found to be 12120 bits through the use of MATLAB code in Subsection 4.2.1, which is based on certain simulation parameters that have been set. This means that $R_c = 4968/12120 = 0.4099$, which leads to $R_1 = 4 \times (4968/12120) \approx 1.6396$.

The probability of decoding failure is now the probability that the ACMI at a particular SNR value of the first Tx and subsequent RTX is ≤ 1.6396. The higher the SNR value is for the AWGN channel, the lower the probability of decoding failure. At a lower SNR, the probability of decoding is of course higher, but it decreases as more RTX are requested, due to the increase in the ACMI from all the differing RVs. This means that for A-HARQ, the ACMI will increase in a shorter time frame, and therefore a higher likelihood of successful decoding as supported by Figure 4.2 and Figure 4.3 (on pages 43 and 44).

For the initial Tx, the eNodeB chooses an appropriate MCS based on the CQI value reported by the UE. If the actual SNR is lower than the measured SNR, then $P(R_1)$ will be highly likely, since the ACMI will be highly likely to be lower than R_1. In this case, when the proposed A-HARQ is used, the eNodeB will use TTI bundling over the preferred sub-band to ensure that the measured SNR is less than or equal to the actual SNR, since TTI bundling involves the use of the more robust QPSK encoding to send all the differing RVs in 4 consecutive subframes.

This will result in the ACMI increasing in a shorter time frame due to the TTI bundled RTX TBs within a time period of just 4 ms, and the ACMI will be most likely larger than R_1. This increase in the ACMI equates to the decrease of the probability of decoding failure after the use of TTI bundling RTX to ensure ultra-reliability.

4.3.2 Delay vs SNR

The legacy HARQ is compared with the proposed A-HARQ in terms of delay vs SNR, by referring to Figure 4.2. The delay is higher for our proposed A-HARQ by 1 ms ($d = 12$ ms and $d_A = 13$ ms) if there is 1 RTX required before the successful decoding and packet combining of the erroneous TB and the RTX TB. However, if the number of RTX for the legacy HARQ is 2 for example, then $d = 20$ ms, and $d_A = 13$ ms, still. The proposed A-HARQ sends 4 differing RVs in only 4 ms, and by using packet combining on the 4 different RVs and the erroneous TB, the probability of successful decoding is higher than the legacy HARQ.

At low SNR (0 to 2 dB), which is what the motivating mission critical industrial control and medical applications experience in very harsh environments, the delay is $\approx$ 35% lower than the legacy HARQ. At mid SNR, the delay is $\approx$ 8.3% higher, while at high SNR, the delay of both schemes converge to the same value because there is only 1 Tx required before successful decoding.

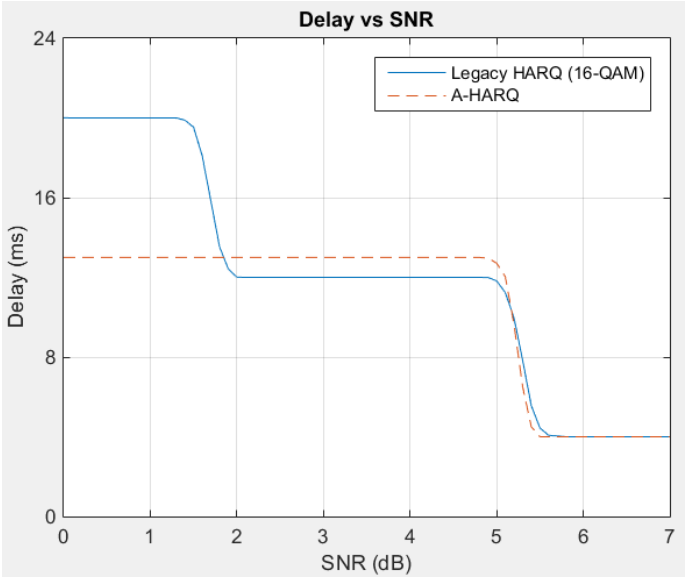

Figure 4.2: Comparison between the legacy HARQ and the proposed A-HARQ in terms of delay vs SNR

4.3.3 Normalised Throughput vs SNR

The legacy HARQ is also compared with the proposed A-HARQ in terms of the throughput vs SNR, by referring to Figure 4.3. The throughput is calculated by using Equation (4.3) for the legacy HARQ and A-HARQ if a RTX is not required. For A-HARQ, the throughput is calculated by using Equation (4.10), where b_{AHARQ} is the number of accepted packets using A-HARQ. $b_{\text{AHARQ}} = 5 - b_A$, where b_A is the number of packets in a TTI bundle needed for successful decoding. For $b_A = \{1, 2, 3, 4\}$, $b_{\text{AHARQ}} = \{4, 3, 2, 1\}$. Therefore $\eta = \{80\%, 60\%, 40\%, 20\%\}$ for A-HARQ.

$$\eta = \frac{b_{\text{AHARQ}}}{b_t} \ [\%] \tag{4.10}$$

By decreasing the delay incurred through the use of the proposed A-HARQ when a RTX is required, the normalised throughput decreases since at least 5 TBs (the initial erroneous TB and the 4 TTI bundle TBs) are sent using A-HARQ. This illustrates the trade-off associated with throughput and delay reduction for general HARQ schemes.

At low SNR (0 to 2 dB), the A-HARQ throughput is $\approx 24.2\%$ lower than the legacy HARQ, but the delay is $\approx 35\%$ lower as a result. At mid SNR (2 to 5 dB), the throughput is $\approx 34\%$ lower than the legacy HARQ, while at high SNR, the throughput of both schemes converge to the same value because there is only 1 Tx required before successful decoding.

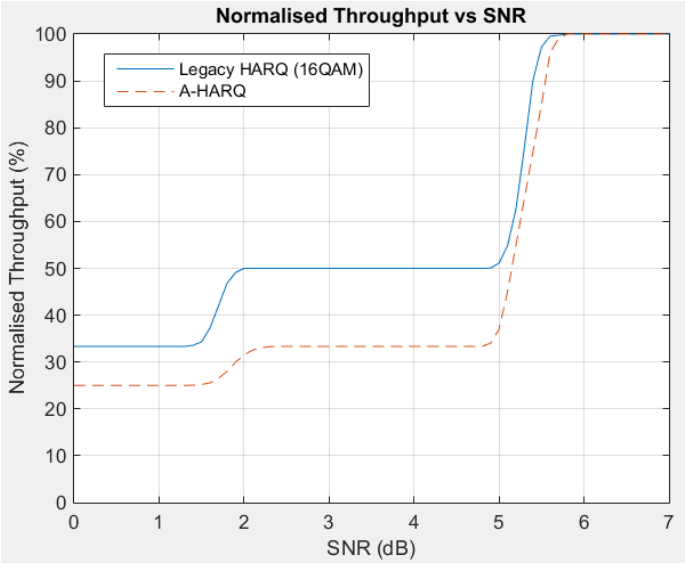

Figure 4.3: Comparison between the legacy HARQ and the proposed A-HARQ in terms of throughput vs SNR

Chapter 5

Conclusion and Future Work

5.1 Conclusions

1. In Chapter 3, an A-HARQ scheme was proposed to cater for applications that will require URC over the future 5G cellular wireless networks. These applications may also have the added QoS constraint of ultra-low latency. A-HARQ involves pre-constructing and pre-encoding different RVs and storing them into the Tx buffer for quick RTX.

2. TTI bundling is used to increase the number of RTX within a time period of 4 ms, and UE-selected sub-band CQI reports are used to ensure that the TTI bundle occurs over the best sub-bands. An example of how A-HARQ operates for a 5 MHz carrier bandwidth is shown in Section 3.2.

3. In Subsection 3.3, dynamic resource allocation is also proposed for URC and A-HARQ, by utilising the mentioned CQI reports, to ensure channel and time diversity for the multiple applications within the best sub-bands. This was illustrated in Figure 3.9 for 4 UEs.

4. In Section 4.3, A-HARQ has been shown to incur $\approx 35\%$ less delay than the legacy HARQ, with a slight decrease in throughput. This was within the low SNR range of 0 to 2 dB, which makes the proposed A-HARQ scheme very useful for the motivating applications within very harsh signal environments.

5.2 Future Work

5.2.1 Testing of A-HARQ and Dynamic Resource Allocation

Currently, the proposed A-HARQ has been shown to incur less delay than the legacy HARQ scheme through simulation over an AWGN channel. The AWGN channel model involves the distribution of 'white Gaussian noise' equally amongst the associated frequency band in the shape of a normal distribution [22]. This is similar to the randomness that occurs in nature, and it simplifies the simulation complexity for this project.

However, the AWGN channel model does not take into account other factors that could affect the communication over LTE/LTE-A, such as 'pathloss' when an UE is mobile, 'shadowing' when there are obstacles within the signal propagation path, and 'fading' due to the multi-path propagation environment [23]. Fast fading especially can cause SINR prediction errors while using the legacy HARQ as mentioned in Subsection 2.1.2.

Thus, it is planned to use or develop a more advanced channel model to account for the other factors that could affect the communication over LTE/LTE-A. However, simulation results is the ideal case for the proposed A-HARQ, which means that practical testing within the laboratory and also in many different signal environments is required. The proposed A-HARQ would be compared against the legacy HARQ over the LTE/LTE-A network to be able to realistically evaluate how it would operate in the real-world.

5.2.2 A-HARQ Optimisation

As mentioned in the previous subsection, practical testing could be used to aid in the optimisation of the proposed A-HARQ scheme from Chapter 3. The goal of this project was to ensure that the URC and the ultra-low latency requirements are met for the motivating applications. However, it does not consider the energy-efficiency of applications utilising the proposed A-HARQ.

Applications in one related category of Massive M2M Communication will require high energy-efficiency due to the need to operate for many years on limited power. This is why it is planned to optimise the proposed A-HARQ to be as energy-efficient as possible, while not violating the URC and ultra-low latency requirements of the motivating applications. This illustrates the challenges that lie ahead in striking a balance between reliability, latency and energy-efficiency for future applications.

Bibliography

[1] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch *et al.*, "Scenarios for 5G Mobile and Wireless Communications: the Vision of the METIS Project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, May 2014.

[2] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, Feb 2014.

[3] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang *et al.*, "Cellular Architecture and Key Technologies for 5G Wireless Communication Networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122–130, Feb 2014.

[4] H. Schotten, R. Sattiraju, D. Gozalvez Serrano, Z. Ren, and P. Fertl, "Availability indication as key enabler for ultra-reliable communication in 5G," in *2014 European Conference on Networks and Communications (EuCNC)*, June 2014, pp. 1–5.

[5] M. Sauter, *From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband*. John Wiley & Sons, Ltd, 2011.

[6] H. Ding, S. Ma, C. Xing, and Z. Fei, "Performance analysis of incremental redundancy hybrid ARQ in mobile ad hoc networks," in *2014 IEEE International Conference on Communications (ICC)*, June 2014, pp. 5759–5764.

[7] S. Sesia, I. Toufik, and M. Baker, *LTE-The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. John Wiley & Sons, Ltd, 2011.

[8] "5G: A Technology Vision," Huawei Technologies Co. Ltd., Jan 2014. [Online]. Available: http://www.huawei.eu/files/publications/pdf/huawei_5g_white_paper_en_20140129.pdf

[9] "Nomenclature of the frequency and wavelength bands used in telecommunications," International Telecommunications Union (ITU), Recommendation ITU-R V.431-8, Aug 2015. [Online]. Available: http://www.itu.int/rec/R-REC-V.431-8-201508-I/en

[10] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *2014 1st International Conference on 5G for Ubiquitous Connectivity (5GU)*, Nov 2014, pp. 146–151.

[11] G. Madueno, C. Stefanovic, and P. Popovski, "Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling," *IEEE Wireless Communications Letters*, vol. 3, no. 4, pp. 429–432, Aug 2014.

[12] I. Abdalla and S. Venkatesan, "A QoE preserving M2M-aware hybrid scheduler for LTE uplink," in *2013 International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT)*, Aug 2013, pp. 127–132.

[13] G. C. Madueno, S. Stefanovic, and P. Popovski, "Efficient LTE access with collision resolution for massive M2M communications," in *2014 Globecom Workshops (GC Wkshps)*, Dec 2014, pp. 1433–1438.

[14] M. Woltering, D. Wubben, A. Dekorsy, V. Braun, and U. Doetsch, "Link Level Performance Assessment of Reliability-Based HARQ Schemes in LTE," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*, May 2014, pp. 1–5.

[15] J. Shea, "Reliability-based hybrid ARQ," *Electronics Letters*, vol. 38, no. 13, pp. 644–645, Jun 2002.

[16] M. M. Butt, J. C. Fricke, and P. A. Hoeher, "Reliability-Based Packet Combining with Application to Interleave-Division Multiple Access," in *2006 4th International Symposium on Turbo Codes Related Topics; 6th International ITG-Conference on Source and Channel Coding (TURBOCODING)*, Apr 2006, pp. 1–6.

[17] Y.-L. Chung and Z. Tsai, "Cooperative Diversity with Fast HARQ for Delay-Sensitive Flows," in *2010 IEEE 71st Vehicular Technology Conference (VTC 2010-Spring)*, May 2010, pp. 1–5.

[18] Y. Lang, D. Wubben, A. Dekorsy, V. Braun, and U. Doetsch, "Improved HARQ based on network coding and its application in LTE," in *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr 2012, pp. 1958–1963.

[19] S. Dimatteo, P. Hui, B. Han, and V. Li, "Cellular Traffic Offloading through WiFi Networks," in *2011 IEEE 8th International Conference on Mobile Adhoc and Sensor Systems (MASS)*, Oct 2011, pp. 192–201.

[20] "Periodic CSI Reporting using PUCCH; UE Selected subband feedback," 3rd Generation Partnership Project (3GPP), Technical Specification 36.213 V12.6.0, Sep 2015. [Online]. Available: http://www.etsi.org/deliver/etsi_ts/136200_136299/136213/12.06.00_60/ts_136213v120600p.pdf

[21] "DL-SCH HARQ Modeling - MATLAB & Simulink Example," Mathworks, 2015. [Online]. Available: http://au.mathworks.com/help/lte/examples/dl-sch-harq-modeling.html

[22] "Add white Gaussian noise to signal - MATLAB awgn," Mathworks, 2015. [Online]. Available: http://au.mathworks.com/help/comm/ref/awgn.html

[23] R. Sattiraju and H. Schotten, "Reliability Modeling, Analysis and Prediction of Wireless Mobile Communications," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*, May 2014, pp. 1–6.