Radio Resource Management for Ultra-Reliable Low-Latency Communications in 5G

Gerardino, Guillermo Andrés Pocovi

DOI (link to publication from Publisher):
10.5278/vbn.phd.tech.00029

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

RADIO RESOURCE MANAGEMENT FOR ULTRA-RELIABLE LOW-LATENCY COMMUNICATIONS IN 5G

BY
GUILLERMO ANDRÉS POCOVI GERARDINO

DISSERTATION SUBMITTED 2017

AALBORG UNIVERSITY
DENMARK

Radio Resource Management for Ultra-Reliable Low-Latency Communications in 5G

Ph.D. Dissertation

Guillermo Andrés Pocovi Gerardino

Aalborg University
Department of Electronic Systems
Fredrik Bajers Vej 7
DK - 9220 Aalborg

Dissertation submitted: June 2017

PhD supervisor: Prof. Preben Mogensen
Aalborg University

Assistant PhD supervisor: Prof. Klaus I. Pedersen
Aalborg University

PhD committee: Professor Petar Popovski (chairman)
Aalborg University

Associate Professor Gerhard Wunder
Freie Universität Berlin

Principal Researcher Stefan Parkvall
Ericsson

PhD Series: Technical Faculty of IT and Design, Aalborg University

ISSN (online): 2446-1628
ISBN (online): 978-87-7112-981-6

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2017

Curriculum Vitae

Guillermo Andrés Pocovi Gerardino

Guillermo Andrés Pocovi Gerardino obtained his M.Sc. degree in Telecommunications Engineering from Universitat Politècnica de Catalunya (UPC), Spain in 2014. Since September 2014, he has been pursuing the PhD degree in Wireless Communications in the Wireless Communication Networks (WCN) section at Aalborg University in collaboration with Nokia Bell-Labs. His main research interests are related to ultra-reliable communications for upcoming 5G wireless systems.

Abstract

The fifth generation (5G) New Radio (NR) will open the door to a wide range of novel use cases. Besides enhanced Mobile Broadband (eMBB), 5G is expected to be an enabler of Ultra-Reliable Low-Latency Communications (URLLC), where small packets must be correctly transmitted and received in a very short time (up to 1 ms) with a success probability of 99.999%.

The challenging reliability and latency requirements of URLLC, not achievable with the current cellular technologies, call for significant enhancements in the air interface. The first part of the thesis addresses the harmful effects of the wireless channel. Spatial diversity and interference cancellation techniques are shown to play a vital role, decreasing the probability of experiencing large desired-signal fades and reducing the co-channel interference. Particularly, a 4x4 microscopic diversity scheme with second order macroscopic diversity is proposed as a way to fulfil the signal quality outage requirements for virtually-error-free one-shot downlink transmissions in a fully-loaded macro network. Such solutions also offer robustness and resilience against network infrastructure failures, which are shown to be a critical issue especially for the cases where such failures are geographically correlated.

The second part of the study focuses on the time-domain aspects. The different components that contribute to the communication latency are studied in a simplified multi-user setting under different load conditions, showing the tradeoffs between the transmission time interval (TTI) duration, the spectral efficiency (short TTIs have a larger relative control overhead), and the queuing delay experienced at the cells. At low offered loads, it is shown that a short TTI size of 0.25 ms is an attractive solution to achieve low latency. However, as the load increases, longer TTI sizes e.g. 0.5 ms or 1 ms, provide better latency performance as these configurations can better cope with the non-negligible queuing delay.

Based on these learnings, the third part of the study proposes novel link adaptation and scheduling techniques in order to fulfil stringent URLLC requirements. The presented solutions are evaluated by means of highly-detailed system-level simulations of a multi-cell macro environment, where

multiple URLLC users are served in the downlink direction. Among others, a short TTI with sufficiently-fast hybrid automatic repeat request (HARQ) round-trip time is shown to be essential to achieve the 1 ms latency targets in a spectral-efficient manner. An attractive channel quality indicator (CQI) measuring procedure is also studied to reduce the link adaptation mismatches due to the very intermittent URLLC traffic. The proposed techniques significantly improve the URLLC performance, allowing to fulfil the requirements for relatively high loads of URLLC traffic (30-40% resource utilization).

Motivated by the ambitious multi-service capabilities of 5G, the challenge of multiplexing URLLC with eMBB traffic on a shared channel is also addressed. Dynamic resource allocation techniques are presented showing that it is possible to fulfil the stringent URLLC requirements, while still providing acceptable eMBB throughput performance. The benefits of punctured scheduling are also studied, where longer ongoing eMBB transmissions are partly overwritten by the sporadically-arriving URLLC traffic. In this respect, various recovery mechanisms are proposed to reduce the impact on the eMBB users that are punctured. Based on the presented findings, recommendations of the features that must be included in the upcoming 5G to support the large variety of URLLC use cases are provided.

Resumé

Den globale femte generations (5G) New Radio (NR) standard åbner døren for en lang række af nye applikationer. Udover forbedret mobil bredbånd (eMBB) forventes 5G at understøtte Ultra-Reliable Low-Latency Communications (URLLC), hvor små pakker sendes og modtages i løbet af meget kort tid (under 1 ms) med en success rate på 99,999%.

De udfordrende pålideligheds og latens-krav for URLLC, der ikke kan opnås med de nuværende cellulære teknologier, kræver betydelige forbedringer af radio implementationen. Den første del af afhandlingen omhandler de udfordrende påvirkninger som den trådløse radio kanal har på tranmissionen. Specielt påvises at brugen af diversitet fra multiple antenner samt teknikker til udslukning af interferens har en afgørende rolle i at reducere sandsynligheden for at opleve store fald i det ønskede signal-støj forhold. I særdeleshed foreslås 4x4 mikroskopisk antenne diversitet kombineret med makroskopiske transmissioner fra flere celler, som en effektiv metode til at opfylde kravene til signalkvalitet og fejlfrie transmissioner til enheder i et fuldt belastet makronetværk. Denne løsning tilbyder også robust overfor fejl i netværks infrastrukturen som viser sig at være specielt kritiske når de er geografisk korrelerede.

Den anden del af undersøgelsen fokuserer på tidsdomæne aspekter. De forskellige komponenter, der bidrager til kommunikationsforsinkelsen, studeres først i en forenklet system model under forskellige belastningsforhold, der viser kompromiser mellem transmissions tids interval (TTI), spektral effektivitet (kortere TTI'er udviser større relativt spild til kontrol information), og pakke forsinkelser. Ved lave data trafik belastninger påvises at en kort TTI-størrelse på 0,25 ms er en attraktiv løsning for at opnå en lav forsinkelse. Men når belastningen stiger, vil længere TTI størrelser, f.eks. 0,5 ms eller 1 ms, give en lavere forsinkelse, da disse konfigurationer bedre kan klare den ikke-ubetydelige forsinkelse i pakke køen.

På baggrund af disse erfaringer, studeres der yderligere link tilpasnings- og radio ressource allokerings teknikker for at opfylde de strenge URLLC-krav. De præsenterede løsninger analyseres ved hjælp af meget detaljerede systemniveau simuleringer i et multi-celle makro miljø, hvor flere URLLC

brugere serviceres i nedgående retning. Blandt andet er en kort TTI med tilstrækkelig hurtig fysisk retransmissions tid (Hybrid Automatic Repeat Request - HARQ) afgørende for at opnå målet for 1 ms forsinkelse på en spektraleffektiv måde. En ny og effektiv måleprocedure for kanal kvalitet og rapportering (Channel Quality Indicator - CQI) undersøges for dens evne til at til at reducere adaptions-fejl til linket der skyldes den irregulære URLLC-trafik. De foreslåede teknikker forbedrer betydeligt URLLC-ydeevnen, hvilket giver mulighed for at opfylde kravene op til et relativt stort URLLC-trafik niveau (30-40% ressourceudnyttelse).

Motiveret af de ambitiøse krav om at kunne supportere vidt forskellige tjenester i 5G behandles også udfordringen med at multiplekse URLLC og eMBB-trafik på en delt kanal. Teknikker for dynamisk ressource allokering præsenteres der er i stand til at at opfylde de strenge krav til URLLC, samtidig med at der kan opnås en acceptabel ydelse for eMBB tjenester. Fordelene ved en speciel form for "punkteret" ressource allokering dokumenteres, hvor længere igangværende eMBB-transmissioner delvist overskrives af den sporadisk ankommende URLLC-trafik. I den henseende foreslås forskellige mekanismer for at reducere den potentielle negative indvirkningen på de eMBB-brugere, der bliver punkteret. Baseret på de fremlagte resultater anbefales det hvilke funktioner der bør indgå i den kommende 5G NR standard, for effektivt at understøtte URLLC-tjenester.

Contents

List of Abbreviations

1G first generation

2G second generation

3D three dimensional

3G third generation

3GPP 3rd Generation Partnership Project

4G fourth generation

5G fifth generation

5QI 5G QoS class indices

ABS almost blank subframes

ACK positive acknowledgement

AS access stratum

BLEP block error probability

BLER block error rate

BR best resources

BS base station

CAM Cooperative Awareness Message

CB code block

CC Chase combining

CCDF complementary cumulative distribution function

CCH control channel

CDF cumulative distribution function

CL closed loop

CN core network

CoMP coordinated multipoint

CQI channel quality indicator

CSI channel state information

DENM Decentralized Environmental Notification Message

DI dominant interferer

DIR dominant interferer ratio

DPM Dominant Path Model

DRB data radio bearers

E2E end-to-end

eICIC enhanced inter-cell interference coordination

eMBB enhanced Mobile Broadband

FCFS first-come first-served

FDD frequency division duplex

FEU front end unit

FF forgetting factor

GBR guaranteed bit rate

gNB fifth generation NodeB

GoB grid of beams

HARQ hybrid automatic repeat request

HeU Highest eMBB user

IIR infinite impulse response

IMT-2020 International Mobile Telecommunications for 2020 and beyond

IRC interference rejection combining

JT joint transmission

KPI key performance indicator

L2S link-to-system

LA link adaptation

LeU lowest eMBB user

LLC low latency communication

LTE Long Term Evolution

LTE-A LTE-Advanced

MAC medium-access-control layer

MBB mobile broadband

MCC mission-critical communication

MCS modulation and coding scheme

METIS Mobile and wireless communications Enablers for Twenty-twenty (2020) Information Society

MIMO multiple-input multiple-output

MMIB mean mutual information per coded bit

mMIMO massive MIMO

MMSE minimum mean square error

mMTC massive Machine Type Communication

MRC maximal-ratio combining

MT mobile terminal

MTBF mean time between failure

MTC machine-type communication

MTTR mean time to repair

MU multi user

NACK negative acknowledgement

NAS non-access stratum

NC-JT non-coherent joint transmission

NR New Radio

OFDM orthogonal frequency-division multiplexing

OFDMA orthogonal frequency-division multiple access

OL open loop

OLLA outer-loop link adaptation

OSI Open System Interconnection

PDCCH physical downlink control channel

PDCP packet data convergence protocol

PDSCH physical downlink shared channel

PDU packet data units

PF proportional fair

PHY physical layer

PMI precoding matrix indicator

PRB physical resource block

QoE quality of experience

QoS quality of service

RAN radio access network

RE resource element

RLC automatic repeat request

RLC radio link control

ROCA root cause aware

RRH remote radio head

RRM radio resource management

RS reference signal

RTT round-trip time

RU resource utilization

SCS sub-carrier spacing

SDU service data unit

SFBC space-frequency block coding

SFN single-frequency network

SINR signal to interference-and-noise ratio

SPS semi-persistent scheduling

SU single user

SVD singular value decomposition

TB transport block

TCP transmission control protocol

TDD time division duplex

TM transmission mode

TTI transmission time interval

UE user equipment

UMTS Universal Mobile Telecommunications System

URLLC Ultra-Reliable Low-Latency Communications

V2X vehicular to anything

WCDMA Wideband Code Division Multiple Access

Thesis Details

Thesis Title:	Radio Resource Management for Ultra-Reliable Low-Latency Communications in 5G.
PhD Student:	Guillermo Andrés Pocovi Gerardino.
Supervisors:	Prof. Preben Mogensen. Aalborg University. Prof. Klaus I. Pedersen. Aalborg University.

This PhD thesis is the result of three years of research at the Wireless Communication Networks (WCN) section (Department of Electronic Systems, Aalborg University, Denmark) in collaboration with Nokia - Bell Labs. The work was carried out in parallel with mandatory courses required to obtain the PhD degree.

The main body of the thesis consists of the following articles:

Paper A: G. Pocovi, M. Lauridsen, B. Soret, K. I. Pedersen and P. Mogensen, "Signal Quality Outage Analysis for Ultra-Reliable Communications in Cellular Networks", *IEEE Globecom Workshops (GC Wkshps)*, December 2015, pp 1-6.

Paper B: G. Pocovi, M. Lauridsen, B. Soret, K. I. Pedersen and P. Mogensen, "Ultra-Reliable Communications in Failure-Prone Realistic Networks", *International Symposium on Wireless Communication Systems (ISWCS)*, September 2016, pp 1-5.

Paper C: G. Pocovi, K. I. Pedersen and B. Soret, "On the Impact of Precoding Errors on Ultra-Reliable Communications", *International Workshop on Multiple Access Communications (MACOM)*, November 2016, pp 45-54.

Paper D: G. Pocovi, K. I. Pedersen, B. Soret, M. Lauridsen and P. Mogensen, "On the Impact of Multi-User Traffic Dynamics on Low Latency Communications", *International Symposium on Wireless Communication Systems (ISWCS)*, September 2016, pp 1-5.

Paper E: G. Pocovi, B. Soret, K. I. Pedersen and P. Mogensen, "MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks", *IEEE International Conference on Communications Workshops (ICC)*, May 2017, pp 1-6.

Paper F: G. Pocovi, K. I. Pedersen, P. Mogensen and Beatriz Soret, "Radio Resource Management for 5G Ultra-Reliable Low-Latency Communications", *IEEE Transactions on Vehicular Technology*. Submitted for publication. 2017.

Paper G: K. I. Pedersen, G. Pocovi, J. Steiner and Saeed R. Khosravirad, "Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband", *IEEE Vehicular Technology Conference*, September 2017, pp 1-6.

Paper H: K. I. Pedersen, G. Pocovi, J. Steiner and A. Maeder, "Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network Implementations", *IEEE Communications Magazine*. Submitted for publication. 2017.

This thesis has been submitted for assessment in partial fulfilment of the PhD Degree. The thesis is based on the submitted or published papers that are listed above. Parts of the papers are used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and also available at the Faculty.

Acknowledgements

The completion of this project is not only the result of three years of hard work but is also the fruit of proper support, advice, and encouragement from colleagues, family and friends. First of all, I want to thank my supervisors, Preben Mogensen and Klaus Pedersen, for sharing their expertise and providing guidance which was essential to keep me on the right track; not to mention their vital efforts in reviewing this thesis and the published papers.

Thanks to Beatriz Soret and Mads Lauridsen, who sat next to me during these three years, and also provided plenty of support and valuable brainstorming sessions. Thanks also to Jens Steiner whose software development skills and technical expertise considerably improved this work. In addition, I want to thank each of my colleagues at Nokia Bells Labs and Aalborg University for contributing to what I believe is a very enriching and pleasant work environment.

Of course, my family, who have always supported me and and encouraged me to excel in everything I do. They are responsible for everything I have accomplished. My gratitude also to Kirstine who has been with me since I started the PhD journey, and whose love and affection also played an important part.

Finally, thanks to my friends in Venezuela, Denmark and the rest of the world for the good times spent together.

Guillermo Andrés Pocovi Gerardino
Aalborg University, June 2017

Part I

Introduction

Introduction

Historically, a new generation of mobile communication technologies is introduced approximately every 10 years. It all started back in 1981, when the first generation (1G) mobile systems were commercially launched. The 1G employed wireless analog-based connectivity to base stations to relay telephone calls to the public telephone network. These were later replaced by the second generation (2G) in the early 90s, bringing more efficient use of the spectrum and new services, e.g. SMS, fruit of the migration from analog to digital communications. The 2G systems quickly evolved to also provide data connectivity to the mobile subscribers. However, the mobile broadband (MBB) revolution came in the current century with the introduction of the third generation (3G) and later the fourth generation (4G), aka. Long Term Evolution (LTE). The first release of LTE (called LTE Release 8) was first completed in March 2009 by the 3rd Generation Partnership Project (3GPP), with the purpose of satisfying the ever-increasing demands of MBB traffic. LTE and particularly its evolution, known as LTE-Advanced (LTE-A), can provide peak data rates theoretically as high as 3 Gbps. Currently, LTE is widely deployed with more than 1.683 billion subscriptions (1 for every 4 mobile subscribers) [1].

The LTE-A standard is continuously evolving trying to take full advantage of recent research efforts in the field. However, backwards compatibility, among other technical limitations, have triggered research and standardization activities toward the design of a fifth generation (5G) mobile system, that fulfils the objectives for the International Mobile Telecommunications for 2020 and beyond (IMT-2020) [2]. As it will be explained, the IMT-2020 requirements do not only focus on MBB performance (as in 3G and 4G), but also address emerging use cases with very diverse and unprecedented requirements. In essence, 5G shall provide *when needed* much higher throughput, lower latency, higher reliability, and larger connection density as compared to its predecessors [3].

The focus of this thesis is on the service class *Ultra-Reliable Low-Latency Communications (URLLC)*. The goal is that 5G systems must be able to deliver to a mobile user a small data payload in a very short time (up to 1 ms)

with ultra-high probability of success (99.999%). This dissertation will closely analyse the challenges of achieving reliable communications over wireless, investigate the potential of current state-of-the-art techniques, and design novel mechanisms for efficient support of URLLC in 5G networks. As a starting point, the following sections will further describe 5G, focusing on the URLLC service class, as well as presenting the scope of the thesis and related contributions.

1 5G Overview

The 5G mobile communication systems shall fulfil the increasing demands of MBB traffic and bring support for new use cases. As depicted in Fig. I.1, initial research towards 5G started in late 2012, mainly as a mixture of independent research efforts and publicly-funded projects (e.g. METIS [4]). It was not until mid-2016 when the 5G standardization activities started in 3GPP, and the formal requirements were defined [5]. The standardization of 5G is still ongoing, and has recently achieved an important milestone with the finalization of the first study item in March 2017. Although the 3GPP has adopted the name *New Radio* (NR), we will generally use the term 5G throughout the thesis.

It has been agreed that 5G shall bring support for a large variety of use cases, categorized into the following three broad service classes (Fig. I.2):

- enhanced Mobile Broadband (eMBB) represents the evolution of the traditional human-centric broadband traffic, where peak data rates of 20 Gbps and 10.000 times more capacity as compared to LTE shall be supported. Furthermore, decent throughput performance (50-100 Mbps) should be provided in less favourable scenarios and conditions; e.g. densely-populated areas (e.g. stadiums) and under high mobility.
- massive Machine Type Communication (mMTC) is characterized by a very large number of connected devices (up to 1.000.000 devices/km^2 in urban areas) typically transmitting a relatively low amount of non-

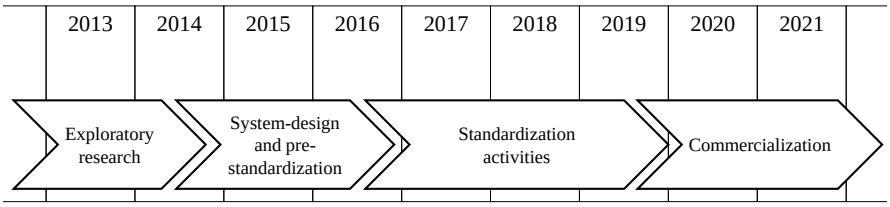


Fig. I.1: Time-line of 5G research, standardization, and commercialization activities [6].

latency critical data. Devices are required to be low cost, and have a battery life above 10 years.

- Ultra-Reliable Low-Latency Communications (URLLC) entail the transmission of sporadic and small packets, with low latency (1 ms) and very high reliability (99.999%).

1.1 Ultra-Reliable Low-Latency Communications

The URLLC key performance indicators (KPIs) and their definitions are presented in Table I.1. In addition to the 1 ms latency with $1 - 10^{-5}$ (99.999%) reliability requirement, a user-plance latency of 0.5 ms shall be fulfilled. The specified 0.5 ms user-plane latency corresponds to a *best-case* value and it is not associated to a specific reliability constraint. The user-plane latency is defined as the *one-way* latency achieved at the layer 2/3 of the radio access network (RAN), in either uplink or downlink direction, and does not account for core network functionalities. Furthermore, URLLC user equipments (UEs) must be capable of performing handovers between cells without interruption of the data connectivity.

It is expected that enabling support for such unprecedented latency and reliability requirements will open the door to novel applications and use cases. Examples of these applications include i) wireless control and automation in industrial environments, addressing the demands of the fourth

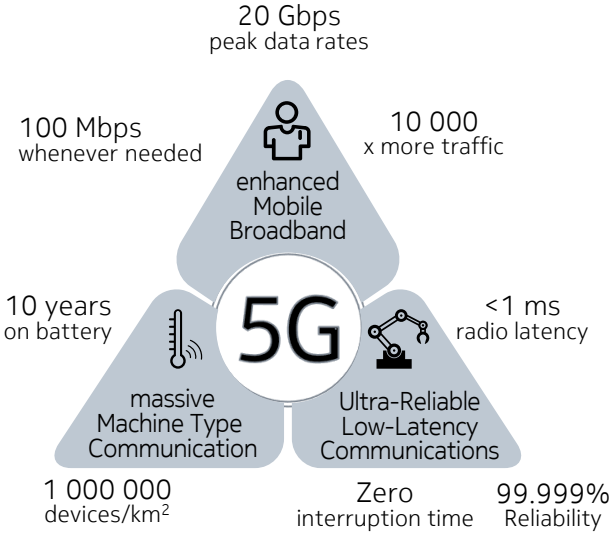


Fig. I.2: 5G service classes and requirements. Source: Nokia.

Table I.1: URLLC KPIs and definitions as defined in [5].

Requirement	Value
User-plane latency The time it takes to successfully deliver an application layer packet/message from the radio protocol layer 2/3 SDU ingress point to the radio protocol layer 2/3 SDU egress point via the radio interface in both uplink and downlink directions	0.5 ms
Reliability The success probability R of transmitting X bits within a certain delay at a certain channel quality	$1 - 10^{-5}$ for 32 Bytes payload with a user plane latency of 1 ms
Mobility interruption time The shortest time duration supported by the system during which a user terminal cannot exchange user plane packets	0 ms

industrial revolution (*Industry 4.0*) [7]. ii) Inter-vehicular communications, where vehicles exchange information in order to increase safety and improve the vehicular traffic flow and efficiency [8]. And iii) real-time tactile internet, allowing wireless control of both real and virtual objects with real-time haptic feedback [9]. In other words, the support for URLLC is envisioned to expand the use of 5G technologies across new markets, creating new business opportunities for mobile operators and telecommunication equipment vendors. In addition, 5G URLLC will positively impact the society, improving our quality of life and saving up time and resources.

Naturally, fulfilling such unprecedented requirements of latency and reliability will require large modifications of the radio interface as compared to what is typically employed in LTE systems. The next section will outline the main components of the lower layers of a communication system. These will be used as a base to identify the most critical aspects (from a reliability and latency point of view) and to motivate the scope of this work.

2 Anatomy of a Communication System

The communication system, which is in charge of delivering the data, plays a very important role in fulfilling the latency and reliability requirements. Fig. I.3 illustrates a generic wireless communication system consisting of one transmitter and multiple receivers. Data received from higher layer applications are stored in user-specific transmission buffers. On each transmission

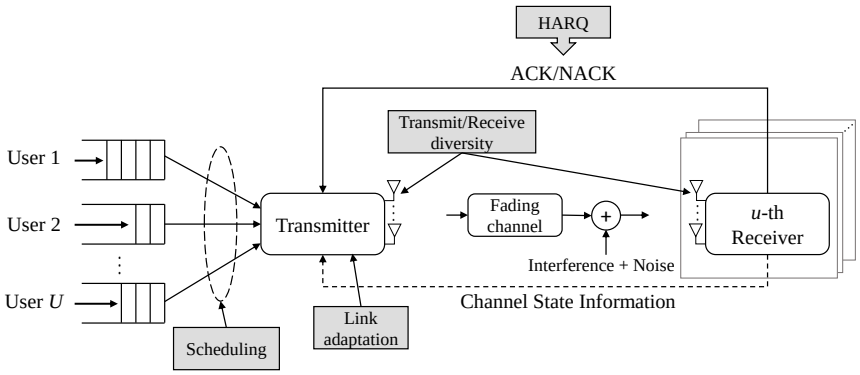


Fig. I.3: Generic communication system. Source: [10].

time interval (TTI), the scheduling entity at the transmitter allocates radio resources to the users, based on their quality of service (QoS) requirements and channel quality as indicated in the channel state information (CSI) reports. Dynamic link adaptation is typically applied to each user's data transmission, trying to adjust the modulation and coding scheme (MCS) to fulfil a certain block error rate (BLER) constraint. The uncertainties in the wireless channel, due to noise and time-variant and frequency-selective fading and interference, can be reduced by exploiting time, frequency and spatial diversity. The latter is typically achieved using multiple-input multiple-output (MIMO) antenna techniques. At the receiver, post-processing of the signal is applied in order to maximize the probability of successfully decoding the data. In case of failed decoding, a negative acknowledgement (NACK) is sent to the transmitter, which typically triggers error-control mechanisms such as hybrid automatic repeat request (HARQ). Besides, mobility of the transmitter and/or receiver could result in handovers with potentially some data interruption time when transferring the communication from one node to another.

3 Latency and Reliability in LTE Networks

The LTE standard was mainly designed to carry human-centric broadband data. Therefore, most of the radio functionalities depicted in Fig. I.3 aim at maximizing the spectral efficiency of the system [11]. As an example, the LTE base station allocates radio resources to the users on a 1 ms TTI resolution[1], which represents the lower bound of the communication latency. The link adaptation is performed to reach a relatively high first-transmission BLER target of 10%-20%. Erroneously decoded packets are retransmitted using

[1]Transmission with shorter TTIs of 0.143 ms is expected to be supported in LTE Release 15, also known as *LTE-Advanced Pro* [12].

HARQ, which adds a minimum of 8 ms to the latency of each retransmitted packet. Most of the LTE commercial deployments employ 2x2 MIMO antenna configurations. While MIMO schemes can be used to provide high order of spatial diversity, spatial multiplexing of parallel data streams is typically preferred for users that experience good channel quality [11]. In case of mobility across cells, LTE employs *break-before-make* handovers, each resulting in typically 40-60 ms data connectivity interruption [13], [14].

We refer to field performance results to give an indication of the latency and reliability performance that LTE systems provide in practice. For example, the work in [14] measures the end-to-end (E2E) latency and handover performance of three commercial LTE networks based on 19,000 km drive tests in Northern Denmark. Fig. I.4 shows the user-plane latency performance, measured as the round-trip time (RTT) of ping messages to a server located in Aalborg University. The figure shows median user-plane latencies between 50-120 ms, depending on the operator. In addition, latencies as high as 250 ms can be observed at the tail of the distribution. One key finding is that the operator-specific core network setup compromises the overall RTT latency performance. Similar latency performance is observed from the measurement campaigns reported in [15]. These results also reveal how the latency drastically varies during the day, depending on the experience load in the system. As discussed in [11], [14], the user-plane RTT latency of LTE, excluding the core network, is approximately 19 ms. The one-way radio latency, estimated as $19/2$ ms $= 9.5$ ms, is therefore not sufficient to meet the URLLC latency requirements defined in Table I.1. Regarding the mobility performance, results from [14] show similar handover data interruption

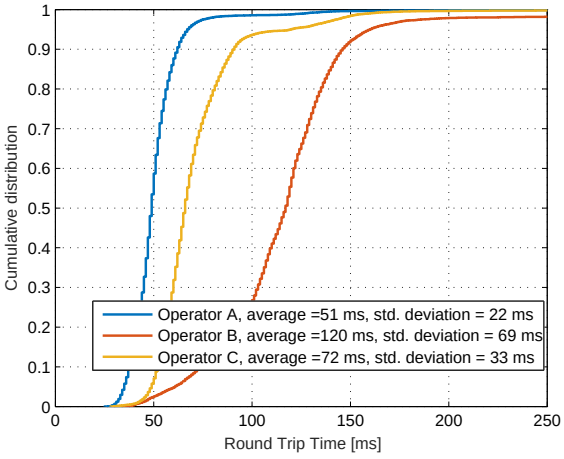


Fig. I.4: Cumulative distribution of the user-plane round-trip time. Source: [14].

times for the three operators: approximately 40 ms in the median, and up to 200 ms in the 99%-percentile. The observed performance is also in line with the reported in other field test campaigns, see e.g. [13].

4 Scope and Objectives of the Thesis

It is therefore evident that LTE has difficulties in fulfilling stringent latency and reliability demands of URLLC. The goal of the thesis is to set the basis for the support of ultra-reliable low-latency communications for the upcoming 5G New Radio, focused primarily on the downlink direction. To accomplish this, we investigate the benefits of current state-of-the-art techniques, and propose novel mechanisms that allow meeting the stringent URLLC requirements. The proposed solutions are evaluated in a dynamic multi-user and multi-cell system-level setting, in order to ensure high degree of realism and practical relevance of the results.

In Fig. I.5, the URLLC requirements are mapped to potential technology enablers, highlighting the specific challenges and solutions that are studied throughout the thesis. A shorter TTI, as compared to LTE, is needed in order to fulfil the 0.5 ms best-case latency required for URLLC. The TTI size cannot be made arbitrarily small, since there is an inherent cost in terms of spectral efficiency [10]. Furthermore, faster processing at both transmitter and receiver is essential to reduce the processing time of data and control, as well as to shorten the HARQ RTT in case of retransmissions. These aspects, namely frame structure, TTI duration, and HARQ recovery mechanisms are an important part of this study.

The reliability can be improved in multiple ways. Naturally, link adaptation should be conducted to achieve a relatively low first-transmission BLER target, e.g. $\leq 1\%$. In this regard, it is important that the receiver accurately estimates the experienced channel quality, and reliably transmits the CSI reports (among other control information) to the transmitter. Particularly, the following challenges are addressed: how to perform efficient link adaptation for URLLC transmissions, how to accurately estimate the channel quality experienced at the receiver, and what is the impact of CSI feedback errors on the communication reliability.

Diversity and interference management techniques are also essential. The goal is to improve the signal quality outage at the receiver, as well as to increase robustness and resilience against failures of the network infrastructure. Microscopic and macroscopic spatial diversity techniques are studied. The latter consists of multiple cells jointly transmitting data to the mobile users. Macroscopic diversity will also help transition towards *make-before-break* handovers with virtually zero interruption time. However, given the time constraint of the study, aspects specific to mobility have been left out of

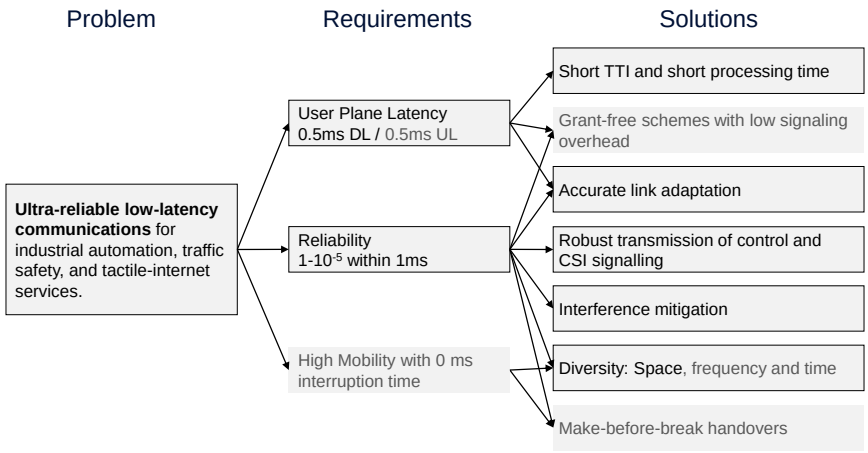


Fig. I.5: Scope of the thesis illustrated in terms of the problem that is addressed, and the potential solutions.

the scope. Grant-free schemes such as semi-persistent scheduling (SPS) are also not considered. The reasoning behind is that these techniques are mainly relevant for the uplink direction, as these avoid the need of a time-consuming scheduling request.

Another important requirement, not shown in Fig. I.5, is to efficiently multiplex URLLC together with other services (e.g. eMBB and mMTC) in the same radio interface. This problem is tackled from two angles. On the one hand, we investigate the functionalities that the 5G packet scheduler must include in order to serve the different users in accordance with their individual QoS requirements. On the other hand, we evaluate how the proposed URLLC enhancements coexist with system aspects relevant to other services, trying to minimize impact on peak data rates, system capacity, among other conventional performance metrics.

To summarize, this thesis tackles the problem of achieving reliable communications focusing on a large set of physical (PHY) and medium-access-control (MAC) layer procedures and functionalities, typically covered under the term *Radio Resource Management* (RRM). The main research questions (Q) and hypotheses (H) that are addressed in the study are listed below:

Q1 How to deal with the stochastic nature of the wireless channel and interference effects?

H1 Spatial diversity and interference management techniques have shown great potential to combat these threats. In order to achieve ultra-reliable communications, large order of macroscopic and microscopic diversity and interference cancellation will be required.

Q2 What is the best transmission strategy for ultra-reliable communications?

H2 A short TTI, fast processing, and short HARQ RTT will be required in order to fulfil the 1 ms latency requirement. Lowering the BLER target of URLLC transmissions will also improve the reliability. However, the TTI length and BLER target cannot be arbitrarily small as some trade-offs between queuing delay and spectral efficiency might exist.

Q3 How to efficiently multiplex URLLC with eMBB traffic on a shared channel?

H3 Service-specific link adaptation and smart resource allocation techniques will be required in order to serve the URLLC traffic per their low latency requirements, while still providing eMBB users with high data rates. Punctured or preemptive scheduling disciplines are among the potential solutions to fulfil this target.

5 Research Methodology

To accomplish our objectives, a classical scientific approach is followed for each part of the study. The overall working methodology is summarized as follows:

1. **Identification the problem:** We investigate the envisioned URLLC use cases in order to identify their respective traffic characteristics, latency and reliability constraints, etc. Next, the open literature is examined in order to determine how today's cellular systems perform in practice, in respect to the previously identified requirements. The communication system is decomposed into several subcomponents, which allows to better identify critical problems and formulate research questions.

2. **Formulation of hypotheses, and potential solutions:** The open literature is visited again in order to find existing techniques that could be applied to solve the outlined problems. For problems where no feasible solution is found, we design our own techniques and formulate the respective hypotheses. When possible, the designed techniques are intended to i) be feasible and applicable to current and future cellular systems, and ii) provide good coexistence with system aspects and functionalities relevant to other services (e.g. eMBB).

3. **Validation of the hypotheses:** The majority of the proposed solutions are evaluated by means of system-level Monte Carlo simulations [16]. This methodology allows us to obtain performance results with high

degree of realism, which would be very difficult to achieve using purely theoretical tools. The employed simulation tools are based on underlying mathematical models and account for most of the elements naturally influencing the performance, e.g. the multi-user and multi-cell dynamics, commonly-accepted radio propagation models, etc.

4. **Analysis of the results:** The formulated hypotheses are tested against the simulation results. In various occasions, the analysis of the simulator output allows us to identify additional weaknesses and problems not previously considered while formulating the initial research questions and hypotheses.

5. **Dissemination of the results:** The proposed solutions and the respective performance results are presented in form of scientific publications. Some novel part of the work is also disclosed via patent applications.

6 Contributions

The main contributions of this study are summarized as follows:

1. **Determining the required level of diversity and interference cancellation to achieve ultra-reliable communications.**

   The low percentiles of the signal to interference-and-noise ratio (SINR) distribution are studied for different combinations of microscopic and macroscopic spatial diversity techniques. The most feasible solutions to achieve very low SINR outage probability as required for ultra-reliable communications are identified. The analysis is conducted for two different network layouts, and taking into account various sources of imperfections, namely failures of the network infrastructure and errors in the uplink feedback channel containing the preferred precoding vector.

2. **Identifying the challenges of achieving low latency in dynamic multi-user systems.**

   The achievable latency in multi-user systems depends not only on the TTI duration, but also on the queuing delays, and the associated traffic model. Tradeoffs between the queuing delays at the cells and the TTI size (and its respective spectral efficiency) are determined. This analysis leads to the conclusion that the optimal transmission strategy of URLLC payloads depends on the load experienced in the system.

3. **Presenting link adaptation and scheduling enhancements for achieving ultra-reliable low-latency communications in cellular networks.**

   This includes a short TTI with fast HARQ round-trip time in order to fit one retransmission within the 1 ms latency budget. Furthermore,

an attractive channel quality indicator (CQI) measuring procedure is proposed which significantly reduces the link adaptation mismatches consequence of the very sporadic interference.

4. **Introducing of a joint link adaptation and resource allocation technique to efficiently schedule URLLC traffic.**

   The proposed enhancement allows for dynamic adjustment of the block error rate (BLER) target in accordance with the instantaneous load experienced at each cell. Furthermore, in cases where URLLC is multiplexed with eMBB traffic, it provides a simple method to determine how the radio resources should be distributed among the two service classes.

5. **Introducing scheduling solutions for dynamic multiplexing of URLLC and eMBB traffic on a shared channel.**

   Two resource allocation approaches for multiplexing URLLC and eMBB are studied: traditional scheduling based and puncturing based. In the latter, eMBB users are served with a 1 ms TTI size, which are partly overwritten by the incoming URLLC traffic. Various recovery mechanisms are proposed to reduce the performance degradation of the eMBB users that are punctured; including a more efficient HARQ retransmission scheme where only the damaged part of punctured transmissions is retransmitted.

6. **Performing a sensitivity analysis of the URLLC and eMBB performance under different settings.**

   The URLLC and eMBB performance is analysed for different conditions of URLLC offered load and payload size, and link adaptation and scheduling strategies. The presented results give valuable insights into the feasible load regions for fulfilling the URLLC requirements, as well as what conditions are more appropriate for dynamic multiplexing of URLLC and eMBB traffic in the upcoming 5G systems.

The thesis is composed of a collection of papers. Therefore, the main findings and contributions are presented in the following publications:

Paper A: G. Pocovi, M. Lauridsen, B. Soret, K. I. Pedersen and P. Mogensen, "Signal Quality Outage Analysis for Ultra-Reliable Communications in Cellular Networks", *IEEE Globecom Workshops (GC Wkshps)*, December 2015, pp 1-6.

Paper B: G. Pocovi, M. Lauridsen, B. Soret, K. I. Pedersen and P. Mogensen, "Ultra-Reliable Communications in Failure-Prone Realistic Networks", *International Symposium on Wireless Communication Systems (ISWCS)*, September 2016, pp 1-5.

Paper C: G. Pocovi, K. I. Pedersen and B. Soret, "On the Impact of Precoding Errors on Ultra-Reliable Communications", *International Workshop on Multiple Access Communications (MACOM)*, November 2016, pp 45-54.

Paper D: G. Pocovi, K. I. Pedersen, B. Soret, M. Lauridsen and P. Mogensen, "On the Impact of Multi-User Traffic Dynamics on Low Latency Communications", *International Symposium on Wireless Communication Systems (ISWCS)*, September 2016, pp 1-5.

Paper E: G. Pocovi, B. Soret, K. I. Pedersen and P. Mogensen, "MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks", *IEEE International Conference on Communications Workshops (ICC)*, May 2017, pp 1-6.

Paper F: G. Pocovi, K. I. Pedersen, P. Mogensen and Beatriz Soret, "Radio Resource Management for 5G Ultra-Reliable Low-Latency Communications", *IEEE Transactions on Vehicular Technology*, 2017. **Submitted for publication**.

Paper G: K. I. Pedersen, G. Pocovi, J. Steiner and Saeed R. Khosravirad, "Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband", Accepted for publication in *IEEE Vehicular Technology Conference*, September 2017, pp 1-6.

Paper H: K. I. Pedersen, G. Pocovi, J. Steiner and Andreas Maeder, "Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network Implementations", *IEEE Communications Magazine*, 2017. **Submitted for publication**.

Paper I: G. Pocovi, M. Lauridsen, B. Soret, K. I. Pedersen and P. Mogensen, "Automation for On-road Vehicles: Use Cases and Requirements for Radio Design", *IEEE Vehicular Technology Conference*, May 2015, pp 1-5.

Paper J: B. Soret, G. Pocovi, K. I. Pedersen and P. Mogensen, "Increasing Reliability by Means of Root Cause Aware HARQ and Interference Coordination", *IEEE Vehicular Technology Conference*, May 2015, pp 1-5.

Papers A-H represent the main core of the thesis, whereas Papers I-J are included in an appendix. In addition to this, four patent applications have been filled:

Patent Application 1: Enhanced Channel Quality Indicator (CQI) measurement procedure for URLLC.

Patent Application 2: Puncturing with low impact on eMBB.

Patent Application 3: Scheduling Mechanism for Ultra-Reliable Low-Latency Communication Data Transmissions.

Patent Application 4: Inter-eNB load report for interference coordination.

In addition, a large part of the time was dedicated to system-level simulator development. Different simulation tools were used during the project, depending on the nature of the problem that was addressed. Spatial diversity and interference management studies (Papers A-C) were conducted using a snapshot-based simulator especially developed for this purpose. The main contributions to this simulator are the following:

- Derivation of the mathematical expressions of the signal model when applying a wide range of techniques, including micro- and macroscopic diversity, interference cancellation and frequency reuse (Paper A).
- These expressions were used to build a Monte Carlo simulator for an in-depth system-level evaluation. The simulator allows us to estimate very low percentiles of the SINR outage distribution when applying the different techniques.
- The simulator includes the effects of multiple cells and users, as well as some of the limitations in real systems, e.g. codebook-based quantized precoding vectors as a way to cope with the limited feedback capacity.
- The simulator was later extended to account for network infrastructure failures (Paper B) and precoding imperfections due to errors in the uplink feedback channel (Paper C).

The rest of the studies (Papers E-H) were mainly conducted in a Nokia - Bell Labs proprietary system-level simulator. The object-oriented C++ simulator is used to generate a large variety of LTE and 5G NR performance results and has been calibrated with system-level simulators from several 3GPP member companies. The simulator includes detailed modelling of the majority of RRM functionalities such as packet scheduling, link adaptation, and HARQ. The main contributions to the development of this simulator are as follows:

- **Improved QoS Management:** the functionality of assigning different QoS classes to the different mobile users was implemented. The required QoS is known at the different layers of the Open System Interconnection (OSI) protocol stack, hence allowing service-specific treatment of data at the different layers.

- **QoS-specific Link Adaptation:** Building on the previous point, the functionality of configuring different link adaptation settings (e.g. BLER target) per service class was implemented.
- **QoS-specific Scheduling:** Similarly, URLLC-specific scheduling enhancements were developed. This includes prioritization of URLLC traffic, as well as novel scheduling policies for multiplexing URLLC with eMBB traffic.
- **Puncturing scheduler:** For cases where critical URLLC traffic overwrites ongoing eMBB transmissions (Papers G-H), different criteria were introduced in order to determine which eMBB transmissions should be punctured.
- **Execution time improvements:** A very large amount of samples are required in order to measure reliability levels of 99.999%. Therefore, some effort was put on analysing the execution time of different instructions of the simulator in order to identify potential sections of the code where the execution time could be reduced.

7 Thesis Outline

The thesis is structured into five main parts and one appendix. The main articles are presented in Parts II, III and IV. Each part includes a short summary of the motivation and main findings of the articles, in order to help the reader understand how the papers relate to each other. The outline of the thesis is as follows:

- **Part I: Introduction -** This part corresponds to the present chapter, where we motivate the work, outline the objectives and contributions, and present the structure of the thesis.
- **Part II: Spatial Diversity as an Enabler of Ultra-Reliability -** This part addresses research question Q1 and is composed of Papers A, B, and C. We study spatial diversity and interference management techniques as a way to deal with the different threats present in the wireless channel, namely large- and small-scale fading effects and the inter-cell interference. Specifically, we study the SINR outage distribution with different settings of these techniques, in order to identify the required diversity order for achieving reliable communications. We present system-level performance results for two different cellular network layouts. We also evaluate the impact of network infrastructure failures, and precoding imperfections due to errors in the uplink feedback channel.

- **Part III: Achieving Low Latency in Multi-User Cellular Networks -** This part addresses research question Q2 and encompasses Papers D and E. The focus is on time-domain aspects, where we analyse the different components that contribute to the communication latency (queuing delay, frame alignment, and transmission delay), and identify the tradeoffs among these components. Paper D presents results for a simplified network setup, whereas Paper E includes highly-detailed system-level simulations. In the latter, we present link adaptation solutions to deal with various challenges.
- **Part IV: Dynamic Multiplexing of URLLC and eMBB -** This part addresses research question Q3 and is composed of Papers F, G, and H. We present solutions for efficient multiplexing of URLLC and eMBB on a shared channel. Two different scheduling approaches are considered, i) URLLC and eMBB are scheduled with a short TTI, and ii) URLLC is scheduled with a short TTI puncturing the longer ongoing eMBB transmissions. We present extensive simulation results showing the performance of these approaches for different conditions of offered load, URLLC payload sizes, etc.
- **Part V: Conclusions -** In this part we summarize the main findings, and provide recommendations and pointers to topics that should be addressed in future studies.
- **Part VI: Appendix -** This part includes additional articles (Papers I-J) authored and co-authored throughout the thesis.

References

[1] Global mobile Suppliers Association (GSA) Press Release, "GSA confirms LTE connects almost 1 in 4 mobile subscribers worldwide: Q4 2016," https://gsacom.com/press-release/gsa-confirms-lte-connects-almost-1-4-mobile-subscribers-worldwide-q4-2016/, Jan. 2017.

[2] ITU-R M.2083-0, "IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond," Sept. 2015.

[3] NGMN Alliance, "5G White Paper," Feb. 2015.

[4] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka *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, 2014.

[5] 3GPP TR 38.913 v14.1.0,, "Study on scenarios and requirements for next generation access technologies," Jan. 2017.

[6] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, Y. Selén, and J. Sköld, "5G wireless access: requirements and realization," *IEEE Communications Magazine*, vol. 52, no. 12, pp. 42–47, 2014.

[7] A. Frotzscher, U. Wetzker, M. Bauer, M. Rentschler, M. Beyer, S. Elspass, and H. Klessig, "Requirements and current solutions of wireless communication in industrial automation," in *Communications Workshops (ICC), 2014 IEEE International Conference on*. IEEE, 2014, pp. 67–72.

[8] G. Pocovi, M. Lauridsen, B. Soret, K. I. Pedersen, and P. Mogensen, "Automation for on-road vehicles: use cases and requirements for radio design," in *Vehicular Technology Conference (VTC Fall), 2015 IEEE 82nd*. IEEE, 2015, pp. 1–5.

[9] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, 2016.

[10] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *Globecom Workshops (GC Wkshps), 2014*. IEEE, 2014, pp. 1391–1396.

[11] H. Holma and A. Toskala, *LTE for UMTS: Evolution to LTE-Advanced*, 2nd ed. Wiley Publishing, 2011.

[12] E. Dahlman, S. Parkvall, and J. Skold, *4G, LTE-Advanced Pro and The Road to 5G*. Elsevier Science, 2016.

[13] A. Elnashar and M. A. El-Saidny, "Looking at LTE in practice: A performance analysis of the LTE system based on field test results," *IEEE Vehicular Technology Magazine*, vol. 8, no. 3, pp. 81–92, 2013.

[14] M. Lauridsen, L. C. Gimenez, I. Rodriguez, T. B. Sorensen, and P. Mogensen, "From LTE to 5G for connected mobility," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 156–162, 2017.

[15] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel *et al.*, "Latency critical IoT applications in 5G: perspective on the design of radio interface and network architecture," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, 2017.

[16] R. F. Coates, G. J. Janacek, and K. V. Lever, "Monte Carlo simulation and random number generation," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 1, pp. 58–66, 1988.

Part II

Spatial Diversity as an Enabler of Ultra-Reliability

Spatial Diversity as an Enabler of Ultra-Reliability

This part addresses the harmful effects of the wireless channel by revisiting well known techniques in order to meet the stringent reliability requirements.

1 Problem Description

Achieving reliable communication over wireless is a difficult task. One major challenge is the stochastic nature of the wireless channel, consequence of the inherent time-varying fading and interference effects. Spatial diversity and interference management techniques have shown great potential for combating these threats, see e.g. [1] and [2]. The goal of this part is to revisit these principles and examine their potential to achieve ultra-reliable communications. To accomplish this, the very low percentiles of the SINR distribution are analysed for different combinations of the following techniques: i) microscopic diversity in the form of closed-loop single-stream MIMO transmission; ii) non-coherent macroscopic diversity, where the received signal from each cooperating cell is independently detected and combined at the UE; and iii) interference management, specifically ideal receiver-based interference cancellation and fractional frequency reuse schemes. In line with the URLLC requirements, special focus is put on the 10^{-5}-th percentile of the distribution, where a SINR value of 0 dB is identified as an appropriate target for *one-shot* virtually error-free transmissions.

The study is conducted in a multi-user multi-cell cellular network, assuming full load conditions. For this purpose, the 3GPP has standardized commonly accepted scenarios and models for system-level performance evaluation [3]. The SINR outage performance is first studied for a traditional hexagonal urban macro cellular network. However, in order to provide higher degree of realism in the presented results, an operational LTE network is also modelled in the simulation tools, and used to examine the performance and required improvements in a realistic deployment.

Due to the unprecedented reliability requirements, the sensitivity of the SINR outage performance to other sources of instability is also analysed. These include errors in the uplink transmission of the preferred precoding matrix indicator (PMI), which can reduce the gain of the considered microscopic diversity scheme; as well as malfunction or failure of the network components and infrastructure.

2 Objectives

The goals of this part of the thesis are the following:

- Investigate the potential of spatial diversity and interference management techniques for achieving ultra-reliable communications.
- Identify the most promising options for achieving the required signal quality outage performance in a practical setting.
- Evaluate the performance in both generic 3GPP-defined scenarios and a realistic network deployment, and identify the most relevant differences in terms of performance.
- Study the impact of other sources of imperfections on the SINR outage performance. Specifically, i) malfunction or failures of the network infrastructure, and ii) errors in the uplink CSI feedback containing the preferred PMI for closed-loop downlink transmissions.

3 Included Articles

The main findings of this part are included in the following articles:

Paper A. Signal Quality Outage Analysis for Ultra-Reliable Communications in Cellular Networks

This article studies the downlink SINR outage performance for different configurations of the considered spatial diversity and interference management techniques. The mathematical expressions for the user-experienced SINR when applying the different techniques are presented, and used to conduct a Monte Carlo system-level evaluation following the 3GPP-defined assumptions for a LTE urban macro cellular network. Based on the obtained performance results, it is discussed which of the options fulfilling the ultra-reliable criteria are most promising in a practical setting.

Paper B. Ultra-Reliable Communications in Failure-Prone Realistic Networks

In order to provide higher degree of realism and practical relevance of the results, the downlink SINR outage performance is analysed for a site-specific network corresponding to a realistic deployment in an European capital. The network model includes real base station positions and parameters (downtilts, antenna patterns, etc), as well as three-dimensional data of buildings, streets, among other elements of the network, which are used for a realistic estimation of the radio propagation characteristics and mobile users locations. In addition, a stochastic model for geographically correlated and uncorrelated equipment failures is presented. The model is used for analysing the SINR outage performance for different failure probabilities and geographical dimensions.

Paper C. On the Impact of Precoding Errors on Ultra-Reliable Communications

This paper studies the impact of CSI feedback errors on the downlink SINR outage performance. The goal is to quantify the impact of these errors on ultra-reliable communications, and determine what transmission modes (closed-loop or open-loop) are more relevant depending on the feedback error probability. The evaluation is conducted for a traditional urban macro cellular network.

4 Main Findings

Benefits of the studied techniques

Papers A and B show how macroscopic and microscopic diversity techniques can substantially improve the 10^{-5}-th percentile of the SINR distribution. Higher diversity order results in steeper slopes, hence, large performance gains at the very-low percentiles. The benefits of macroscopic diversity come from the i) increased protection against the fast fading effects of the wireless channel, and ii) higher received power (array gain) from precoding the signal at the transmitter and coherent combining at the receiver. On the other hand, macroscopic diversity provides protection against the slow fading effects (a.k.a. shadow fading), and mobility robustness during handovers. Smaller gains are observed from interference cancellation or frequency reuse schemes. The reasoning behind is that interference management techniques do not increase the slope of the SINR distribution since the diversity order remains the same. However, as compared to macroscopic diversity, interference management provides valuable improvements at the median and upper part of the SINR distribution.

Regarding the URLLC requirements, it is shown that traditional MIMO schemes with 2x2 or 4x4 antenna configurations are not sufficient to fulfil the 0 dB SINR target at the 10^{-5} percentile. Those must be complemented with macroscopic diversity and/or interference management techniques in order to ensure the target SINR outage performance. For example, in the 3GPP-defined regular macro network (Paper A), a 4x4 antenna scheme with 2 macroscopic links is presented as the most feasible configuration for achieving the reliability requirements. Similar performance can also be achieved by cancelling the received power from the three strongest interferers. However, ideal interference cancellation of three interferers is harder to achieve in practice.

Performance results in the site-specific network model (Paper B) show generally lower gain from the studied techniques. This is a consequence of the more realistic propagation characteristics and irregular base station deployment. For instance, a 4x4 microscopic scheme with 3 macroscopic links is required in order fulfil the reliability requirements, i.e. one additional macroscopic diversity order as compared to the 3GPP network layout.

Macroscopic diversity consumes radio resources for a single user at multiple cells, potentially reducing the system capacity and increasing the inter-cell interference. To account for these effects, Paper B also studies the impact of the *macroscopic diversity overhead* on the SINR outage performance. It is shown that macroscopic diversity windows between 6 dB and 10 dB (i.e. the maximum allowed received power difference between the strongest cell and the additional cells serving the user) provide a good compromise between diversity gain and resource consumption.

Impact of failures of the cellular infrastructure

Paper B shows that uncorrelated failures do not have a large impact on the SINR outage performance ($<$1 dB degradation for downtime probabilities of 10^{-2}). In this case, there is sufficient overlapping coverage as it is generally the case for dense urban cellular networks. In contrast, settings with spatially correlated failures suffer from large performance degradation since the failures span over large geographical areas in a clustered manner, hence decreasing the probability of having good coverage. For example, the 4x4 configuration with three macroscopic links no longer fulfils the required SINR outage performance if the mean failure probability is higher than 0.5%.

Impact of precoding errors

Paper C shows that errors in the uplink feedback channel have a negative impact on the SINR outage performance. However, even for feedback error probabilities as high as 10^{-2} (i.e. three orders of magnitude larger than

the required reliability), there is a benefit of using closed-loop schemes over open-loop schemes. The explanation behind is the following: even though a non-optimal precoding is applied at the transmitter, the applied PMI is signalled in the downlink scheduling grant, which allows the UE to properly post-process (i.e. coherently combine) the received signal. In other words, the benefits of transmit diversity are degraded but the receive diversity gain is maintained.

Macroscopic diversity provides additional robustness against precoding errors. For instance, a 4x4 MIMO scheme with two orders of macroscopic diversity can achieve the 0 dB SINR outage target at the 10^{-5}-th percentile, even for a precoding error probability of 10^{-2}. With macroscopic diversity, the probability of experiencing feedback errors across the multiple serving cells is reduced. This is a consequence of the applied non-coherent macroscopic scheme, which only requires basic CSI transmission and independent precoding at each cooperating cell.

5 Contributions and Follow-up Studies

The use of *non-coherent joint transmission* (NC-JT), as known in 3GPP, is being considered for 5G NR; see e.g. the following 3GPP contributions: [4], [5]. This is facilitated by the recently-agreed data duplication functionality at the 5G packet data convergence protocol (PDCP) layer, as described in [6]. Apart from the benefits highlighted throughout the presented papers, NC-JT is less sensitive to user's movements and timing mismatch across cooperating cells, and requires lower signalling overhead on the backhaul interface (as compared to coherent schemes) [7].

The analysis has been carried out in a fully-loaded network, and without explicit modelling of the URLLC traffic, resource scheduling, link adaptation mechanisms, etc. However, two follow-up studies evaluate the considered techniques in a more dynamic setting. In [8], on-demand inter-cell coordinated cell muting schemes are studied in a scenario with a mixture of URLLC and eMBB traffic. One of the key points of the presented concept is being *on-demand*, meaning that resources from the interfering cells are only muted when a downlink URLLC transmission takes place at the serving cell. The presented technique shows similar latency reduction as the classical frequency reuse but much lower impact on the eMBB throughput performance.

In addition, [9] analyses the URLLC latency and reliability performance for four different resource allocation methods, namely: i) traditional unicast, ii) single-frequency network (SFN) transmission, iii) sub-band muting at the strongest interfering cell(s), and iv) non-coherent macroscopic diversity. Simulation results in an indoor environment show that muting and macroscopic diversity schemes generally offer the largest improvement in terms of block error rate of URLLC transmissions.

References

[1] S. N. Diggavi, N. Al-Dhahir, A. Stamoulis, and A. Calderbank, "Great expectations: The value of spatial diversity in wireless networks," *Proceedings of the IEEE*, vol. 92, no. 2, pp. 219–270, 2004.

[2] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, "Interference coordination and cancellation for 4G networks," *IEEE Communications Magazine*, vol. 47, no. 4, 2009.

[3] 3GPPP TR 36.814 v9.0.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects," Mar. 2010.

[4] R1-1612841, "Non-coherent Multi-node Transmission for URLLC in the 5G New Radio," *3GPP TSG-RAN WG1 #87*, Nov. 2016.

[5] R1-1703155, "Non-coherent Multi-node Transmission for URLLC in the 5G New Radio," *3GPP TSG-RAN WG1 #88*, Feb. 2017.

[6] 3GPP TR 38.804 v14.0.0, "Study on New Radio Access Technology; Radio Interface Protocol Aspects (Release 14)," Mar. 2017.

[7] M. S. Ali, "On the evolution of coordinated multi-point (CoMP) transmission in LTE-Advanced," *International Journal of Future Generation Communication and Networking*, vol. 7, no. 4, pp. 91–102, 2014.

[8] B. Soret and K. I. Pedersen, "On-demand power boost and cell muting for high reliability and low latency in 5G," in *Vehicular Technology Conference (VTC Spring), 2017 IEEE 85th*. IEEE, 2017, pp. 1–5.

[9] V. Hytönen, Z. Li, B. Soret, and V. Nurmela, "Coordinated multi-cell resource allocation for 5G ultra-reliable low latency communications," in *Networks and Communications (EuCNC), 2017 European Conference on*. IEEE, 2017, pp. 1–5.

Paper A

Signal Quality Outage Analysis for Ultra-Reliable Communications in Cellular Networks

Guillermo Pocovi, Beatriz Soret, Mads Lauridsen, Klaus I. Pedersen, Preben Mogensen

The paper has been published in the
IEEE Globecom Workshops (GC Wkshps), 2015.

Abstract

Ultra-reliable communications over wireless will open the possibility for a wide range of novel use cases and applications. In cellular networks, achieving reliable communication is challenging due to many factors, particularly the fading of the desired signal and the interference. In this regard, we investigate the potential of several techniques to combat these main threats. The analysis shows that traditional microscopic multiple-input multiple-output schemes with 2x2 or 4x4 antenna configurations are not enough to fulfil stringent reliability requirements. It is revealed how such antenna schemes must be complemented with macroscopic diversity as well as interference management techniques in order to ensure the necessary SINR outage performance. Based on the obtained performance results, it is discussed which of the feasible options fulfilling the ultra-reliable criteria are most promising in a practical setting, as well as pointers to supplementary techniques that should be included in future studies.

1 Introduction

Today's wireless communication systems are designed for transporting a wide range of human-centric multimedia and data contents. However, it is expected that future wireless applications will be complemented by a wide range of machine-centered services that will have a big impact on society [1]. Machine-type communications (MTC) with ultra-reliable communication requirements is one of these new use cases attracting interest within the research community. By ultra-reliable, we mean applications requiring the transmission of a certain payload with low latency and high probability of success; e.g. 99.999%. Examples of applications within this category include vehicular communications for safety, tactile internet, industry automation, and energy management [1], [2], [3].

The signal quality outage performance is of key importance to satisfy stringent reliability requirements; in other words, it is essential that the target users perceive a signal quality above a certain value with high probability. In cellular networks, the signal quality is typically measured in terms of desired signal to interference-and-noise ratio (SINR). The SINR is, however, affected by many variable factors that are intrinsic for any wireless systems [4]. For instance, the desired signal power can be highly attenuated as a result of the fading nature of the wireless channel. Also, the received interference can be large due to the typical aggressive reuse of the time-frequency resources for maximizing system capacity in the network.

Studying the importance and potential techniques to combat these threats is the main focus of this paper. Diversity is a well-known technique used to deal with the fading channel. Diversity can be typically achieved in the

space, time or frequency domain. An exhaustive review of the importance of spatial diversity in communication systems is presented in [5]. Appropriate combining of the multiple received signals has substantial importance on the SINR performance, and the most relevant approaches can be found in [6], [7]. Multi-cell cooperation techniques, such as joint transmission and coordinated scheduling, are presented in [8] as a way to improve spectral efficiency and data rates. The SINR outage probability of a joint transmission scheme is analysed in [9]. However, the evaluation in [9] is limited to a specific mobile terminal (MT) at a predefined position in the network. Recent work also evaluates diversity as one of the main enablers for ultra-reliable communication. As an example, [10] investigates utilizing multiple weak links instead of a single powerful link to ensure high availability in a wireless network; whereas [11] evaluates different multiple-input multiple-output (MIMO) antenna configurations to achieve ultra-reliable and low-latency communication. The studies are conducted for a single-user scenario, and hence effects of multi-user and multi-cell interference are not explicitly included in the analysis. Interference management is another complementary approach to improve the signal quality. There is a vast amount of work on interference mitigation and suppression techniques, ranging from static frequency reuse patterns to advanced receivers with interference suppression capabilities [7], [12], [13]. However, their potential to provide ultra-reliable communications requires more studies.

In this paper, we analyse the potential of a wide range of techniques for improving the downlink SINR outage performance in cellular networks. Particularly, spatial diversity techniques at the micro and macroscopic level, as well as interference management approaches, are evaluated. Our objective is to identify the required level of diversity and interference mitigation to achieve very low SINR outage probability as required for ultra-reliable communication. Compared to the studies in [10], [11], our work focuses on a multi-cell/multi-user scenario including diverse system aspects. The chosen evaluation methodology is system-level Monte-Carlo simulations, following the 3GPP-defined LTE simulation assumptions for a traditional macro case.

The rest of the paper is organized as follows: Section 2 provides a brief description of the studied techniques as well as the modelling of the system. The simulation assumptions are outlined in Section 3. Performance results are presented in Section 4. Section 5 discuss the implications of the evaluated techniques, and concluding remarks appear in Section 6.

2 System Model

Fig. A.1 presents a generic cellular network scenario. Base stations (BSs) are strategically deployed to provide wide coverage to a certain area. In a typ-

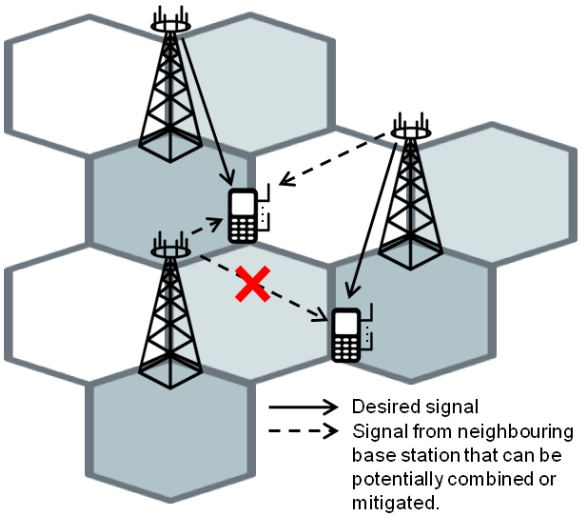


Fig. A.1: Example of typical network layout. The principles of micro and macroscopic diversity techniques, as well as interference mitigation or frequency reuse are also illustrated.

ical scenario, each BS serves multiple MTs within their respective coverage area. From a MT perspective, the desired signal is contaminated with interference generated by neighbouring BSs. Both BS and MT can be equipped with multiple antennas in order to provide increased microscopic diversity. Additionally, various spatially-separated BSs can cooperatively serve the MT to provide higher order of macroscopic diversity and redundancy. Besides, the received interference at the MT can be reduced by applying interference mitigation or frequency reuse schemes. In the following, the models used to evaluate the performance of these different techniques are presented.

2.1 Microscopic Diversity

Microscopic diversity is an effective technique to mitigate the effects of multipath fading. Systems with multiple antennas at the transmitter and/or receiver can provide diversity gains as the multiple spatial transmitter-receiver paths can be coherently combined and decrease the overall probability of experiencing poor channel conditions. We consider a typical MIMO system consisting of T transmit antennas and R receive antennas. A closed-loop MIMO scheme is assumed, where the MT feeds back channel information to the serving BS that is used to determine the antenna transmitter weights (also known as precoding). As our focus in this study is on ultra-reliable communication, and therefore on the lower tails of the SINR distribution, only single-stream transmission cases are considered with maximal-ratio combining (MRC) at the receiver [6]. The R-dimensional received signal $\mathbf{r}_j$ by the

user served by BS j is expressed as

$$\mathbf{r}_j = \mathbf{H}_j \sqrt{\Omega_j} \mathbf{w}_j^t s_j + \sum_{i=1}^{L} \mathbf{H}_i \sqrt{\Omega_i} \mathbf{w}_i^t s_i + \mathbf{n} \ , \tag{A.1}$$

where L is number of interfering signals, $\mathbf{H}_i$ is a RxT matrix whose (m,n)-th element represents the complex channel gain from transmit antenna n at BS i, to receive antenna m; $\mathbf{w}_i^t$ is the T-dimensional precoding vector at the i-th BS; s_i and Ω_i represent the transmitted symbol and the averaged received power from the i-th BS, respectively; for simplicity, $\|s_i\| = 1$; and $\mathbf{n}$ is a $Rx1$ zero mean Gaussian vector with variance σ^2 representing the noise at each receiving antenna. The received signal vector is weighted at the receiver as follows,

$$y = \mathbf{w}_j^r \mathbf{r}_j = \mathbf{w}_j^r \mathbf{H}_j \sqrt{\Omega_j} \mathbf{w}_j^t s_j + \sum_{i=1}^{L} \mathbf{w}_j^r \mathbf{H}_i \sqrt{\Omega_i} \mathbf{w}_i^t s_i + \mathbf{w}_j^r \mathbf{n} \ , \tag{A.2}$$

where $\mathbf{w}_j^r$ is the $1xR$ receive weight vector. For MRC, the optimal weights applied at the transmitter and receiving side are given by [6]

$$\mathbf{w}_j^t = \mathbf{u} \ , \tag{A.3}$$

$$\mathbf{w}_j^r = \mathbf{u}^H \mathbf{H}_j^H \ , \tag{A.4}$$

where $\mathbf{u}$ is the unitary eigenvector corresponding to the largest eigenvalue of the $\mathbf{H}_j^H \mathbf{H}_j$ matrix, and $[\cdot]^H$ denotes the Hermitian transpose. In order to emulate the limited feedback capacity of real systems, $\mathbf{u}$ is quantized and restricted to the predefined set of codewords as used in LTE [14]. The closed-loop transmit weight of each interfering link is randomly generated; this allows to decrease the execution time of the simulations with negligible impact on the performance results [15]. The resulting post-detection SINR expression is given by

$$SINR_j = \frac{\Omega_j \|\hat{\mathbf{u}}^H \mathbf{H}_j^H \mathbf{H}_j \hat{\mathbf{u}}\|^2}{\sum_{i=1}^{L} \Omega_i \|\hat{\mathbf{u}}^H \mathbf{H}_j^H \mathbf{H}_i \mathbf{w}_i^t\|^2 + \|\hat{\mathbf{u}}^H \mathbf{H}_j^H \mathbf{n}\|^2} \ , \tag{A.5}$$

where $\hat{\mathbf{u}}$ denotes the quantized version of the eigenvector $\mathbf{u}$.

2.2 Macroscopic Diversity

Macroscopic diversity is another technique for increasing the reliability. The idea is to have multiple BSs transmitting synchronously the same information, which is then combined at the receiver. Macroscopic diversity provides multiple benefits for achieving reliable communication; for example, it helps

to combat shadowing effects, improves the diversity and redundancy of the system, and increases the total received power of the desired signal. However, this comes at the expense of using transmission resources for a single user at multiple BSs, which can have negative impact on the total network capacity. A simple soft-combining approach as known from 3G is assumed, where the received signal from each macroscopic branch is independently detected and combined [16]. The *SINR* after combining M macro branches is expressed as follows,

$$SINR = \sum_{j=1}^{M} SINR_j \,, \tag{A.6}$$

where $SINR_j$ is the SINR calculated according to (A.5), assuming the MT is connected to BS j. In a cochannel scenario, the best SINR performance is obtained by connecting to the M BSs with the highest received power. This is the approach applied in this study.

2.3 Interference Mitigation

The received interference from multiple BSs can be reduced in order to improve the SINR at the MT. Ideal cancellation of the signal received from the C ($1 \leq C \leq L$) strongest interfering BSs is assumed. This BS subset is denoted as A. The resulting SINR expression is as follows,

$$SINR_j = \frac{\Omega_j \|\hat{\mathbf{u}}^H \mathbf{H}_j^H \mathbf{H}_j \hat{\mathbf{u}}\|^2}{\sum_{i=1}^{L} \Omega_i \|\hat{\mathbf{u}}^H \mathbf{H}_j^H \mathbf{H}_i \mathbf{w}_i^t\|^2 \mathbf{1}_{\{i \in A^C\}} + \|\hat{\mathbf{u}}^H \mathbf{H}_j^H \mathbf{n}\|^2}, \tag{A.7}$$

where $\mathbf{1}_{\{i \in A^C\}}$ denotes the indicator function of the set A^C, and $[\cdot]^C$ denotes the complement; i.e. $\mathbf{1}_{\{i \in A^C\}} = 0$ if the received power from BS i is cancelled; otherwise $\mathbf{1}_{\{i \in A^C\}} = 1$.

2.4 Frequency Reuse

With frequency reuse, the frequency resources are strategically distributed among the different BSs to reduce the cochannel interference and increase the SINR. Apart from the full frequency reuse scheme, a hard frequency reuse with 1/3 reuse factor is considered. This principle is illustrated in Fig. A.1, where the resources used to serve MT within each sector are represented with different colours. Eq. (A.7) can also be used to represent the frequency reuse approach, with A corresponding to the set of BSs not utilizing the same frequency resources as the desired signal. Equal amount of resources per sector is assumed.

3 Simulation Assumptions and Confidence Interval Calculation

3.1 Simulation Assumptions

The evaluation is carried out by analysing the downlink SINR distribution with different degrees of micro and macroscopic diversity, interference cancellation and frequency reuse schemes. A dense macro-cellular network composed of three-sector sites with an inter-site distance of 500 m is simulated, where MTs are uniformly distributed. All BSs are transmitting at full power (full load conditions) at a 2 GHz carrier frequency. A omnidirectional antenna pattern is assumed at the MT, whereas a three-dimensional antenna with horizontal and vertical patterns is assumed at the BS [17], including the effect of antenna downtilt. The propagation is modelled according to [17]. This includes distance-dependant pathloss, the effects of log-normal shadowing and fast fading. The fast fading is independent and identically distributed for each transmit-receive antenna pair, following a complex Gaussian distribution (i.e. the envelope is Rayleigh distributed).

For each MT, the models presented in Section 2 to calculate the experienced post-detection SINR for different techniques are applied. Effects of user mobility and handovers are not explicitly included in the simulations. However, the effect of handover hysteresis margins are implicitly model in the serving BS selection algorithm: each MT identifies the strongest received BSs that are within a certain *handover window*, as compared to the strongest BS. The serving BS for the MT is then randomly selected from the BSs within the *handover window*. This is a simple method for modelling the effect where not all MTs are served by their strongest BS due to the use of handover hysteresis margins in reality. Table A.1 summarizes the simulation assumptions.

The generated SINR samples from all the users are used to form empirical cumulative distribution functions (CDF). For ultra-reliable communication, the key performance indicator is the SINR at the very-low percentiles, considering here the 10^{-5} level in line with [1]. Achieving reliable communication requires that the MT is able to correctly receive both the scheduling grant from the BS and the corresponding transport block with the actual data payload (i.e. information bits). Taking LTE as a reference, the scheduling grant is sent on the physical downlink control channel (PDCCH). Referring to the studies in [18], [19], the MT can decode the PDCCH with 10^{-2} error probability at -5 dB if the PDCCH is transmitted with the strongest coding format. However, in order to achieve truly error free PDCCH reception, an SINR of approximately 0 dB (or higher) is required, when including various imperfections. At 0 dB SINR, also the transport block with data on the physical

Table A.1: Simulation assumptions

Parameter		**Value**
Network layout		3GPP macro case 1
BS transmit power		46 dBm
Carrier bandwidth and frequency		10 MHz @23.0 GHz
Pathloss		$128.1 + 37.6\Delta \log10(R[\text{km}])$ dB
Antenna pattern		BS: 3D with 12° downtilt MT: omnidirectional
Antenna gain		BS: 14 dBi. MT: 0 dBi
Shadowing	Distribution	Log-normal with $\sigma = 8$ dB
	Intra-site correlation	1.0
	Inter-site correlation	0.0
Noise	Power spectral density	-174 dBm/Hz
	Noise figure	8 dB
	Total noise power	-96 dBm
Handover window		3 dB
Fast fading		Rayleigh distributed; Uncorrelated among the different diversity branches
SINR outage target		0 dB at the 10^{-5}-th percentile

downlink shared channel (PDSCH) can be correctly decoded if transmitted with QPSK and a conservative encoding rate[1] (e.g. 1/3) [20]. Hence, in this study we consider 0 dB SINR as the minimum threshold for a MT to have error free downlink reception, without relying on hybrid automatic repeat request retransmissions which further impacts the latency and reliability.

3.2 Statistical Confidence Considerations

To guarantee high statistical confidence of the results, the confidence intervals are estimated using the normal approximation of the Binomial proportion [21]. The interval of a certain percentile $\hat{\gamma}$ is approximately $\hat{\gamma} \pm z_{\alpha/2}\sqrt{\hat{\gamma}(1-\hat{\gamma})/N}$, where N is the generated number of independent samples, $z_{\alpha/2}$ is the $100(1-\alpha/2)$-th percentile of the standard normal distribution, and α is the confidence level associated to the confidence intervals of the experiment. As an example, for a target of 95% confidence level in the estimate of $\hat{\gamma}$ within a $\pm 20\%$ interval we have

$$\frac{1.96\sqrt{\hat{\gamma}(1-\hat{\gamma})/N}}{\hat{\gamma}} < 0.2\,. \tag{A.8}$$

[1]This is a fair assumption given the typically low data rates of use cases requiring ultra-reliablity.

For $\hat{\gamma} = 10^{-5}$, the required number of uncorrelated samples is then given by

$$N = \frac{1}{(0.2)^2}(1.96)^2 \frac{1-\hat{\gamma}}{\hat{\gamma}} = 9.604 \cdot 10^6 \,. \tag{A.9}$$

We go beyond that limit and generate at least $N = 16 \cdot 10^6$ uncorrelated SINR samples for each of the simulated configurations.

4 Results

4.1 Microscopic Diversity

Fig. A.2 shows the empirical CDF of the SINR distribution with different configurations of transmit and receive antennas. For each curve, the surrounding grey area represents the 95% confidence interval. At the 10^{-5}-th percentile, the SINR obtained with a single antenna configuration can be as low as -45 dB. This suggest that single transmit-receive antenna schemes are not appropriate for ultra-reliable communications. In contrast, the increased diversity order obtained with 2x2, 4x2 or 4x4 configurations results in steeper slopes and significantly better SINR performance, being the 4x4 antenna scheme only 6 dB away from achieving the 0 dB SINR target. It is worth mentioning that a 2x2 MIMO configuration is the most commonly used scheme in today's cellular systems such as LTE.

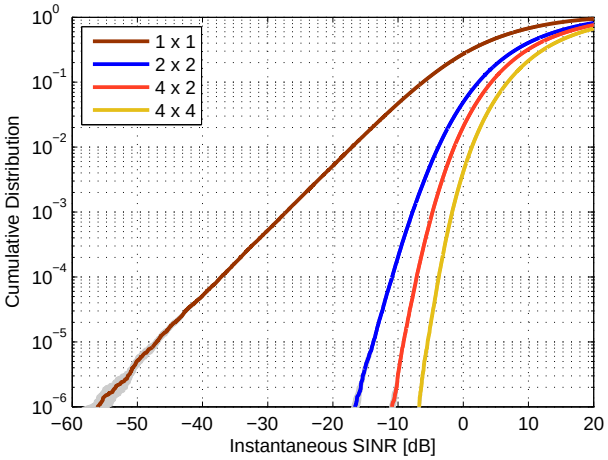


Fig. A.2: SINR performance with different MIMO antenna configurations. The legend indicates Number of transmitting antennas x Number of receiving antennas. For each curve, the surrounding grey area represents the confidence interval at the 95% confidence level.

4.2 Macroscopic Diversity

Fig. A.3 shows the SINR statistics of different macroscopic and microscopic configurations. The confidence intervals are no longer shown; however, the generated number of samples is kept the same, hence the curves have similar confidence as those shown in Fig. A.2. It is observed that macroscopic diversity provides considerable gains in the SINR outage performance. By adding a secondary macroscopic link, more than 6 dB SINR improvement is obtained at the 10^{-5}-th percentile, which allows the 4x4 configuration to fulfil the 0 dB SINR requirement. The macroscopic gain comes mainly from the higher received power as well as the additional protection against shadow fading. The soft-combining is especially relevant for cell-edge MTs, which are likely to receive similar power from the M strongest BSs. Macroscopic diversity also minimizes the negative performance impact of the considered handover hysteresis window, since the probability of not being connected to the strongest BS is reduced.

4.3 Interference Mitigation

Fig. A.4 shows the SINR distribution of 2x2 and 4x4 configurations assuming ideal cancellation of the C strongest interferers. As expected, cancelling the interference provides some SINR gains. However, the slope of the curve is not increased since the diversity order remains the same. Nevertheless, a 4x4 antenna configuration with ideal cancellation of the three strongest interfering links allows to fulfil the 0 dB SINR target with the desired outage probability.

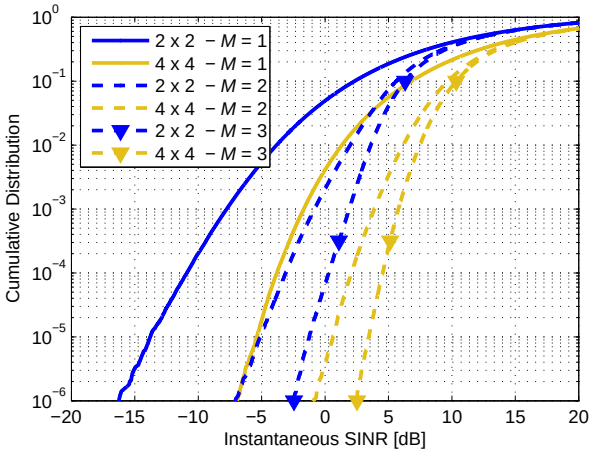


Fig. A.3: SINR performance with different levels of macroscopic diversity applied to a 2x2 and 4x4 MIMO antenna configurations.

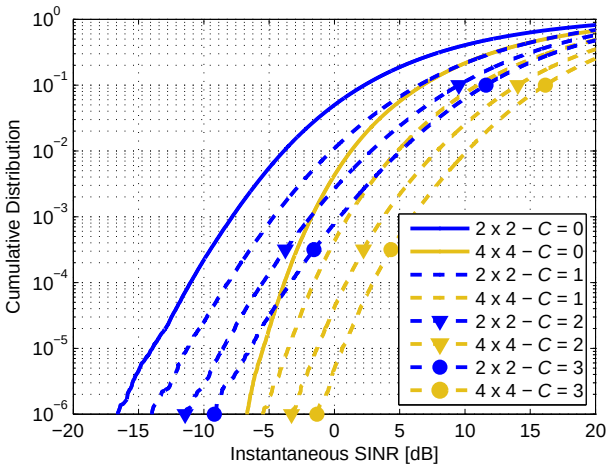


Fig. A.4: SINR performance assuming ideal cancellation of the C strongest interfering signals.

4.4 Frequency Reuse

Fig. A.5 shows the SINR statistics when applying a fixed frequency reuse scheme of $1/3$ and different orders of macroscopic diversity. Similar to the performance observed with interference cancellation, the applied technique does not provide any additional diversity order. But, substantial gains are achieved in the very low percentiles of the SINR distribution due to the drastic reduction of the interference. The 0 dB SINR target can be achieved by using a 2x2 MIMO scheme together with $M = 2$ macroscopic links. Notice that the macroscopic diversity gains are slightly larger than those obtained with the full frequency reuse scheme (see Fig. A.3). In this case, the multiple BSs serving a certain user are likely to be using different frequency resources therefore not interfering with each other.

5 Summary and Discussion

Fig. A.6 presents a summary of the achieved SINR at the 10^{-5}-th percentile for 2x2, 4x2 and 4x4 microscopic antenna schemes together with diverse configurations of macroscopic diversity and interference management techniques. The plotted configurations are selected according to the attained performance and deployment feasibility. The 0 dB SINR target is represented with a horizontal dashed line. Among the various configurations shown in Fig. A.6, a 4x4 antenna scheme with $M = 2$ macroscopic links seems to be the most feasible configuration for achieving the ultra-reliability target. Similar performance can also be achieved by e.g. cancelling the received power

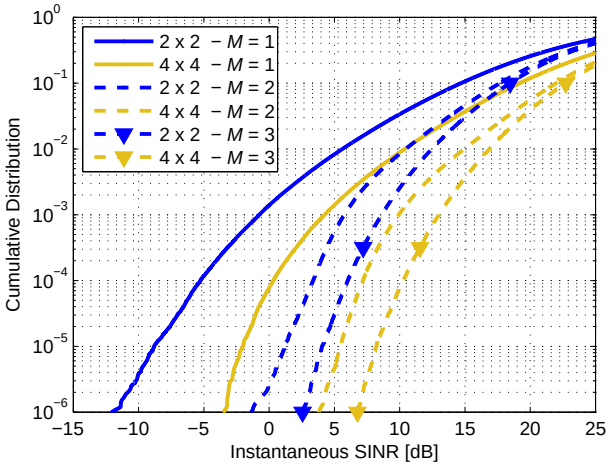


Fig. A.5: SINR performance when applying a frequency reuse scheme of 1/3 with different macroscopic and microscopic diversity configurations.

from the three strongest interferers. However, in practice, full non-linear interference cancellation of three interferers is most likely not possible [22], so the performance obtained with this configuration should be regarded as an upper bound. A 2x2 microscopic diversity scheme on its own is not able to provide the required performance to achieve ultra-reliable communications. The results suggest that such configuration must be complemented with significant reduction of interference by e.g. applying a frequency reuse scheme and having at least two macroscopic links, in order to fulfil the required SINR target. As expected, the 4x2 antenna configuration attains a performance in-between the 2x2 and 4x4 schemes. With 4x2, the 0 dB SINR target can be achieved by adding an additional macroscopic link, compared to what is needed for 4x4 schemes. This configuration is relevant for small devices (e.g. sensors and actuators in wireless-automated industries) where it might be difficult to accommodate a high number of antennas in the MT. The obtained results further motivates relevance of 5G research focusing on cases with at least 4x4 MIMO and various interference suppression techniques [22], [23].

Macroscopic diversity is clearly an important technique to increase the reliability, providing increased diversity against both fast and slow fading, and more transmit power towards the user. Furthermore, macroscopic diversity offers robustness towards BS failures, as well as additional mobility robustness during handovers from one BS to another. However, compared to microscopic diversity, macroscopic diversity consumes resources for a single user at multiple BSs, which can have negative impact on the total network capacity. Secondly, tight coordination and low latency communication between

the BSs involved in macroscopic transmission are needed. Thus, in line with basic communication theory, achieving ultra-reliable communications comes at a cost in terms of reduced average spectral efficiency [4].

The performance results for frequency reuse 1/3 show significant improvements in the SINR performance. In this respect, it is worth noticing, that if applying channel and interference aware frequency domain packet scheduling (as e.g. supported in LTE), then the system converges to an equivalent frequency reuse pattern depending on the offered traffic [20]. Hence, the performance results for frequency reuse 1/3 are equivalent to the performance that would be experienced under fractional load conditions, where each BS only utilize 33% of the available frequency domain resources.

Although the interference mitigation techniques do not improve the distribution (i.e. diversity order) of the desired signal, improvements in the SINR outage performance are still visible. In this study, only ideal non-linear interference cancellation has been considered, although other techniques are also of interest. Among others, it is suggested to conduct further research on other candidate techniques like linear interference-rejection combining (IRC) and more sophisticated proactive or reactive network based interference coordination techniques. As compared to existing studies of network-based interference coordination techniques (see e.g. [13]), the optimization target to have ultra-reliable communication would likely lead to slightly different solutions. Similarly, addition of small cells at strategically chosen locations to enhance reliability is another future research direction [23]. See also [4] for additional pointers to 5G enhancements for ultra-reliable communication.

6 Conclusions

In this study, we have evaluated the potential of diversity and interference-management techniques to achieve ultra-reliability in the 3GPP-defined macro cellular scenario. Micro and macroscopic diversity techniques have been shown to be one of the main enablers of ultra-reliable communications. The evaluated spatial techniques not only provide high diversity to combat the fast fading in the wireless channel, but also increase the robustness of the communication. For a 10^{-5} desired SINR outage, a 4x4 MIMO scheme with second order macroscopic diversity is considered as the most feasible configuration when taking practical implementation considerations into account. Mitigating the interference by either network-based or terminal-based techniques has been identified as a promising complementary solutions to improve the SINR outage performance. Although such techniques do not increase the diversity of the desired signal component, up to 10 dB gain in the outage SINR is achieved by cancelling multiple interferers or applying a 1/3 frequency reuse scheme in the studied scenario.

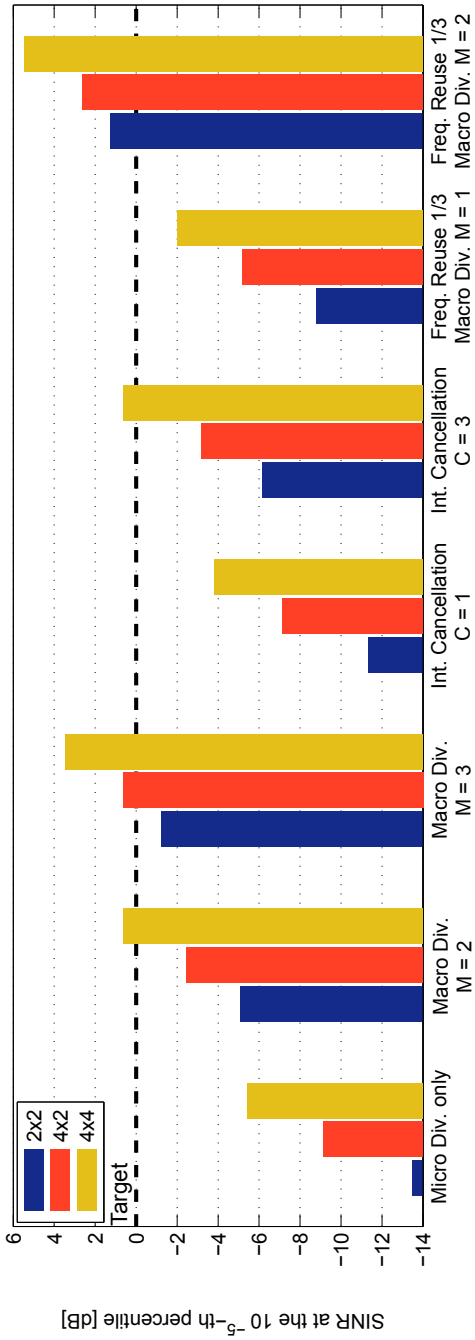


Fig. A.6: Achieved SINR at the 10^{-5}-th percentile for several diversity and interference management configurations.

Future work must further consider the imperfections present in real systems such as non-ideal channel estimation, correlation among the multiple antennas, and more realistic modelling of interference mitigation techniques. Extending the evaluation to include wideband systems with frequency-selective fading channels is also of interest. On a further note, evaluations for other scenarios in addition to the generic 3GPP cases are also recommended. For instance, analysis of scenarios based on data from real network deployments to further assess the degree of reliability that can be supported.

References

[1] A. Osseiran et al., "Scenarios for the 5G Mobile and Wireless Communications: the Vision of the METIS Project", *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26-35, May 2014.

[2] G. P. Fettweis, "The tactile internet: applications and challenges", *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64-70, March 2014.

[3] A. Frotzscher et al., "Requirements and current solutions of wireless communication in industrial automation", *IEEE ICC Workshops*, June 2014.

[4] B. Soret, P. Mogensen, K. I. Pedersen and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks", *IEEE Globecom*, December 2014.

[5] S. N. Diggavi, N. Al-Dhahir, A. Stamoulis and A. R. Calderbank, "Great expectations: the value of spatial diversity in wireless networks", *Proceedings of the IEEE*, vol. 92, no. 2, pp. 219-270, Feb. 2004.

[6] J. M. Romero-Jerez, J.P. Pena-Martin, G. Aguilera and A. J. Goldsmith, "Performance of MIMO MRC systems with co-channel interference", *IEEE International Conference on Communications*, June 2006.

[7] J. H. Winters, "Optimum combining in digital mobile radio with cochannel interference", *IEEE Journal on Selected Areas in Communications*, vol. 2, no. 4, pp. 528-539, July 1984.

[8] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa and M. Tanno, "Coordinated multipoint transmission/reception techniques for LTE-Advanced", *IEEE Wireless Communications*, vol. 17, no. 3, pp. 26-34, June 2010.

[9] D. Ben Cheikh, J.-M. Kelif, M. Coupechoux and P. Godlewski, "Analytical joint processing multi-point cooperation performance in rayleigh

fading", *IEEE Wireless Communications Letters*, vol. 1, no. 4, pp. 272-275, August 2012.

[10] D. Ohmann, M. Simsek and G. P. Fettweis, "Achieving high availability in wireless networks by an optimal number of Rayleigh-fading links", *IEEE Globecom Workshops*, December 2014.

[11] N. A. Johansson, Y. P. Eric Wang, Erik Eriksson and Martin Hesslear, "Radio access for ultra-reliable and low-latency 5G communications", *IEEE ICC Workshops*, June 2015.

[12] G. Boudreau et al., "Interference coordination and cancellation for 4G networks", *IEEE Communications Magazine*, vol. 47, no. 4, pp. 74-81, April 2009.

[13] B. Soret, K. I. Pedersen, N. Jørgensen and V. Fernandez-Lopez, "Interference coordination for dense wireless networks", *IEEE Communications Magazine*, vol. 53, no. 1, pp. 102-109, January 2015.

[14] 3GPPP TS 36.211 v12.5.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation", April 2014.

[15] A. Osseiran and A. Logothetis, "Closed loop transmit diversity in WCDMA HS-DSCH", *IEEE Vehicular Technology Conference*, May 2005.

[16] H. Holma, A. Toskala (editors), "WCDMA for UMTS – Radio Access for Third Generation Mobile Communications", Third edition, Wiley, 2004.

[17] 3GPPP TR 36.814 v9.0.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects", March 2010.

[18] D. L. Villa, C. U. Castellanos, I. Z. Kovacs, F. Frederiksen and K. I. Pedersen, "Performance of downlink UTRAN LTE under control channel constraints", *IEEE Vehicular Technology Conference*, May 2008.

[19] D. Laselva et al., "On the impact of realistic control channel constraints on QoS provisioning in UTRAN LTE", *IEEE Vehicular Technology Conference*, September 2009.

[20] H. Holma and A. Toskala, "LTE Advanced: 3GPP Solution for IMT-Advanced", John Wiley & Sons Ltd, 2011.

[21] L. D. Brown, T. T. Cai and A. Dasgupta, "Confidence intervals for a binomial proportion and asymptotic expansions", *The Annuals of Statistics*, vol. 30, no. 1, pp. 160-201, 2002.

[22] W. Nam, D. Bai, J. Lee and I. Kang, “Advanced interference management for 5G cellular networks", *IEEE Communications Magazine*, vol. 52, no. 5, pp. 52-60, May 2014.

[23] P. Mogensen et al., “5G small cell optimized radio design", *IEEE Globecom Workshops*, December 2013.

Paper B

Ultra-Reliable Communications in Failure-Prone Realistic Networks

Guillermo Pocovi, Mads Lauridsen, Beatriz Soret, Klaus I. Pedersen, Preben Mogensen

The paper has been published in the
International Symposium on Wireless Communication Systems (ISWCS), 2016.

Abstract

We investigate the potential of different diversity and interference management techniques to achieve the required downlink SINR outage probability for ultra-reliable communications. The evaluation is performed in a realistic network deployment based on site-specific data from a European capital. Micro and macroscopic diversity techniques are proved to be important enablers of ultra-reliable communications. Particularly, it is shown how a 4x4 MIMO scheme with three orders of macroscopic diversity can achieve the required SINR outage performance. Smaller gains are obtained from interference cancellation, since this technique does not increase the diversity order of the desired signal. In addition, failures or malfunction of the cellular infrastructure are analysed. Among different types of failures evaluated, results show that failures spanning over large geographical areas can have a significant negative performance impact when attempting to support high reliability use cases.

1 Introduction

Ultra-reliable communications over wireless is an active research topic that will open the possibility of novel applications [1]. For some of the use cases, latencies of a few milliseconds must be guaranteed with reliability levels up to 99.999%. The experienced signal to interference-and-noise ratio (SINR) is a metric closely related to the achievable reliability in cellular systems: the higher the SINR, the more feasible it is to achieve low packet error probability and low communication latency.

Studying the potential of different techniques to achieve the required SINR outage probability for ultra-reliable communications is the objective of this paper. In this context, microscopic and macroscopic spatial diversity techniques have shown promising benefits. As an example, the work in [2] evaluates different multiple-input multiple-output (MIMO) antenna configurations to achieve high reliability in a factory environment, whereas [3] analyses the effectiveness of different transmission methods, including microscopic diversity and hybrid automatic repeat request (HARQ) mechanisms. The work in [4], studies the benefit of combined microscopic and macroscopic diversity schemes in different locations of a regular hexagonal network, however without accounting for the multi-user multi-cell interference, which is typically a performance degrading factor. These effects have been included in our previous contribution, where we have identified the required level of diversity and interference management in a 3GPP hexagonal macro network [5].

So far, fading and interference have been the limiting factors of the end-user reliability performance. However, due to the stringent reliability requirements, other sources of instability and error should also be included

in the analyses. As an example, malfunction or failure of network components could compromise the performance of the network. The causes of network failures can be commonly classified into power, hardware and software, and can occur at the base stations, aggregation nodes, or at the backhaul links [6], [7]. To the best of our knowledge, the system performance impact of these events on ultra-reliable communications has not been quantified.

In this paper, we aim at identifying the required level of diversity and interference management in order to achieve the required SINR outage probability for ultra-reliable communications. We build on the recent study in [5] with the following additional contributions: (i) Analysis in a realistic network deployment based on site-specific data from a European capital. As compared to regular 3GPP scenarios, the use of site-specific models provides higher degree of realism and practical relevance of the results. (ii) System-level evaluation including the effects of multi-user/multi-cell interference and the increased resource consumption as a consequence of the macroscopic diversity overhead. And (iii) impact of the instability and failure-prone characteristics of real cellular deployments, using a stochastic model for geographically correlated and uncorrelated equipment failures. In order to account for these multiple aspects of realism, the chosen evaluation methodology is system-level Monte-Carlo simulations.

The rest of the paper is organized as follows: Section 2 describes the considered network scenario. Section 3 outlines the studied techniques as well as the failure model. The simulation methodology is presented in Section 4. Performance results are shown in Section 5, and concluding remarks are given in Section 6.

2 Network layout

A site-specific network layout of an existing LTE macro deployment in a European capital area is reproduced. A three-dimensional (3D) topography map is used for the considered dense urban area [8]. The map contains 3D data of buildings, streets, open squares, parks, etc. Macro cells are placed following a real deployment in the considered area. The macro site antennas are deployed at different heights, taking the local environment characteristics into account in order to have good wide area coverage. The average macro antenna height is 30 meters, using a few degrees of antenna downtilt. The area includes hundreds of macro-sites with 2 and 3 sectors, with an average inter-site distance of 350 meters, covering a geographical area of around 30 km^2. However, the evaluation is limited to the outdoor locations of the 1.2 km^2 segment depicted in Fig. B.1. The outdoor area is mainly represented by streets, avenues and open squares. Hence, the study is relevant for outdoor high reliability use cases, e.g. vehicular communications (V2X) through cel-

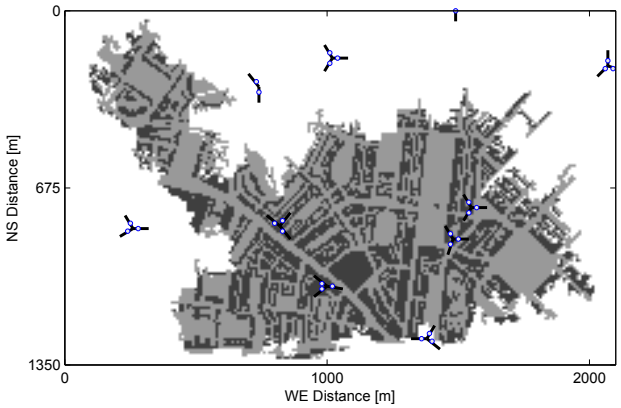


Fig. B.1: Illustration of the network layout. The macro cells are marked with blue circles and a line pointing in the direction of the main lobe of the antenna (a.k.a. broadside). Light and dark grey colour denote outdoor and indoor areas, respectively.

lular infrastructure [9].

The radio propagation characteristics are obtained by using state-of-the-art ray-tracing techniques based on the Dominant Path Model (DPM) [10]. This includes the effects of both distance-dependent attenuation and shadowing. The resulting coverage area for each macro cell varies significantly, as the areas shape according to the streets, buildings, among other objects considered in the network model.

3 System model

A set of cells denoted by $\mathcal{N} = \{1, ..., N\}$ are deployed following the network deployment described in Section 2. The SINR at the mobile terminal (MT) can be improved in multiple ways. For example, both the serving cell and MT can be equipped with multiple antennas in order to provide increased microscopic diversity. Additionally, various spatially-separated cells can cooperatively serve MTs to provide higher order of macroscopic diversity. As an alternative approach, the received interference can be reduced by applying interference cancellation techniques.

3.1 SINR model

In order to evaluate the performance of these techniques, we utilize the signal model described in [5]. Each downlink connection between a MT and a cell is represented by a MIMO system with T transmit antennas and R receive antennas. Closed-loop single-stream transmission modes are considered. In a frequency-flat fading case, the R-dimensional received signal $\mathbf{r}_j$ by a user

served in cell $j \in \mathcal{N}$ is given as follows,

$$\mathbf{r}_j = \mathbf{H}_j \sqrt{\Omega_j} \mathbf{v}_j s_j + \sum_{i \in \mathcal{N} \setminus j} \mathbf{H}_i \sqrt{\Omega_i} \mathbf{v}_i s_i + \mathbf{n} \ , \tag{B.1}$$

where $\mathbf{H}_i$ is a *RxT* matrix whose (*m*,*n*)-th element represents the complex channel gain from transmit antenna *n* at cell *i*, to receive antenna *m*; $\mathbf{v}_i$ is the *T*-dimensional precoding vector used at the *i*-th cell; Ω_i represents the averaged received power from the *i*-th cell; s_i represents the transmitted symbol (for simplicity, $\|s_i\| = 1$); and $\mathbf{n}$ is a *Rx1* zero mean Gaussian vector with variance σ^2 representing the noise power at each receiving antenna. At the receiver, the *R* received signals are combined by applying a weight vector $\mathbf{w}_j$. The transmitter and receiver weights expressions are given by $\mathbf{v}_j = \mathbf{u}$ and $\mathbf{w}_j = \mathbf{H}_j \mathbf{u}$, where $\mathbf{u}$ is the unitary ($\|\mathbf{u}\| = 1$) eigenvector corresponding to the largest eigenvalue of the $\mathbf{H}_j^H \mathbf{H}_j$ matrix, and $[\cdot]^H$ denotes the Hermitian transpose. This combining method corresponds to maximal-ratio combining (MRC), which aims at maximizing the desired signal power at the receiver. The resulting instantaneous post-detection SINR expression is given by,

$$SINR_j = \frac{\Omega_j \|\mathbf{u}^H \mathbf{H}_j^H \mathbf{H}_j \mathbf{u}\|^2}{\sum_{i \in \mathcal{N} \setminus j} \Omega_i \|\mathbf{u}^H \mathbf{H}_j^H \mathbf{H}_i \mathbf{v}_i\|^2 + \sigma^2 \|\mathbf{u}^H \mathbf{H}_j^H\|^2} \ . \tag{B.2}$$

For cases where a MT is served with macroscopic diversity, a simple soft-combining approach as known from Universal Mobile Telecommunications System (UMTS) is assumed. The received signal from each macroscopic branch is independently detected and combined at the MT [11]. The SINR after combining *M* ($1 \leq M \leq N$) macro branches is expressed as follows,

$$SINR = \sum_{j=1}^{M} SINR_j \ , \tag{B.3}$$

where $SINR_j$ is the SINR calculated according to (B.2), assuming the MT is connected to cell *j*. The set of cells serving a certain MT is typically referred to as the active set. A *macroscopic diversity window* is applied for the active set selection procedure; i.e. for each MT, the set of serving cells is limited to those whose received power difference is within a certain window, as compared to the strongest cell. Using an appropriate window size, much of the macroscopic diversity gain can be obtained with a reduced resource consumption and less interference generated as compared to the infinite window case [11].

Finally, cases with interference cancellation are also considered. Ideal cancellation of the signal received from the *C* ($1 \leq C \leq N-1$) strongest interfering cells is assumed. The corresponding SINR expression is similar to (B.2), with the denominator accounting only for the remaining $N - C - 1$ interferers (we refer to [5] for more details).

3.2 Network failures

A stochastic model is used to study the impact of network failures as an additional source of degradation of the SINR outage performance. Let **P** be a vector composed of N random variables (r.v.) P_1, P_2,..., P_N, where P_n is a Bernoulli distributed r.v. indicating the functional state of cell $n \in \mathcal{N}$. $P_n = 1$ indicates normal operation, and $P_n = 0$ is failure. The probability of failure for cell n is,

$$\mathrm{F}_n = \Pr(P_n = 0) = 1 - \Pr(P_n = 1) \ \ \forall \ \ n \in \mathcal{N}. \tag{B.4}$$

The geographical characteristics of a potential failure is described in terms of the pairwise correlation, i.e.

$$\rho(m,n) = \frac{\mathrm{cov}(P_n, P_m)}{\sigma_{P_n}\sigma_{P_m}} = \frac{\mathrm{E}[P_n P_m] - \mathrm{E}[P_n]\mathrm{E}[P_m]}{\sigma_{P_n}\sigma_{P_m}} \ \forall m,n \in \mathcal{N} \tag{B.5}$$

where $\mathrm{E}[x]$ and σ_x denote, respectively, the expectation and standard deviation of the r.v. x, and $\mathrm{cov}(x,y)$ is the covariance of x and y.

Geographically uncorrelated failures are generally the most common in real networks [6]. Factors such as malfunction of hardware in the cell, buggy software updates, cut of the wired link carrying the data to the core network, etc. can result in localized failures. Correlated failures are less common but not completely absent in real networks. Causes of failure include software upgrades (if performed simultaneously to multiple base stations), hacking of the system (e.g. jamming [12]), natural disasters (storms, earthquakes, etc.), and discharge of batteries after long power outages (especially for small cells with typically small power backup).

We model cases with geographically uncorrelated and correlated failures. The former is among the simplest case as there is independent failure probability at each cell, i.e. $\rho(m,n) = 0$, $m \neq n$. For the case where geographical correlation is present, $\rho(m,n)$ is modelled using an exponential decaying function as follows,

$$\rho(m,n) = \exp\left(\frac{-D_{m,n}}{\mu}\right) \ \ \forall \ \ n,m \in \mathcal{N}, \tag{B.6}$$

where $D_{m,n}$ is the distance (in meters) between cells m and n, and μ is the spatial correlation distance. Given these definitions, we build the covariance matrix $\Sigma(\mathbf{P})$ for the set of cells $\mathcal{N}$ and apply methods such as the ones described in [13], [14] to generate samples of binary r.v. with the specified spatial correlation properties.

4 Simulation methodology

The evaluation is carried out by analysing the downlink SINR distribution with different degrees of micro and macroscopic diversity, and interference management techniques. A snapshot-based simulation approach is applied and the respective assumptions are summarized in Table B.1. Cells are located as described in Section 2 and transmitting at full power (full load conditions) at 2.6 GHz carrier frequency. MTs are uniformly distributed in the corresponding area of interest (see Fig. B.1). The simulation procedure is as follows: Each MT selects its active set of size M (M corresponds to the macroscopic diversity order) from the set of candidate cells according to the average received power. The set of candidate cells corresponds to those cells in normal operation state and is calculated on every simulation snapshot following the stochastic failure model described in Section 3.2. Cells in failing state are assumed to have zero transmit power. Effects of user mobility and handovers are not explicitly included in the simulations. However, the effect of handover hysteresis margin is implicitly modelled in the active set selection algorithm: each MT identifies the strongest received cells that are within a certain *handover window*, as compared to the strongest cell. A serving cell for the MT is then randomly selected from the cells within the handover window. This method models the effect where not all MTs are served by their strongest cell due to the use of handover hysteresis margins in reality.

After the active set selection procedure, the experienced instantaneous post-detection SINR is calculated for each MT following the models in Section 3.1. For each snapshot, the fast fading is independent and identically distributed for each transmit-receive antenna pair, following a complex Gaussian distribution (i.e. the envelope is Rayleigh distributed). Additive noise with a power spectral density of -174 dBm/Hz is considered. It is assumed that MTs are assigned with 10 MHz bandwidth, resulting in a noise power of -96 dBm when including a 8 dB noise figure at the MT. In order to emulate the limited feedback capacity of real systems, precoding vector $\mathbf{u}$ is quantized and restricted to the predefined set of codewords used in Long Term Evolution (LTE) [15].

A large number of snapshots are simulated and the generated SINR samples from all the users are used to form empirical cumulative distribution functions (CDF). In line with [1], the key performance indicator (KPI) is the SINR at the 10^{-5}-th percentile. At this percentile, we consider a 0 dB SINR as an appropriate target to have error-free downlink reception, and therefore fulfil the low latency requirements of ultra-reliability use cases (we refer to [5] for more details).

Table B.1: Simulation assumptions

Parameter	Value
Network layout	Site-specific network
Macro cell transmit power	46 dBm
Carrier frequency	2.6 GHz
Propagation	Dominant path model [10]
Macro cell antenna configuration	Realistic 3D antenna pattern
MT distribution	Uniformly distributed in outdoor locations
MT antenna configuration	Omnidirectional with 0 dBi gain
Noise power spectral density	-174 dBm/Hz
Noise figure	8 dB
Noise power	-96 dBm @10 MHz
Fast fading	Rayleigh distributed; Uncorrelated among the different antenna branches
Handover window	3 dB
μ, ρ	Uncorrelated failures: $\rho = 0$ Correlated failures: $\rho > 0$, $\mu = 415$ m
SINR outage target	0 dB at the 10^{-5}-th percentile

5 Performance results

Fig. B.2 shows the empirical CDF of the SINR distribution with 2x2 and 4x4 microscopic schemes and different orders of macroscopic diversity. For this set of results, all MTs are assumed to have the same macroscopic diversity order equal to M, i.e. no macroscopic diversity window constraint is applied. Normal operation of the network is also assumed. At the 10^{-5}-th percentile, the SINR obtained with 2x2 and 4x4 configurations and no macroscopic diversity is approximately -15 and -7 dB, which is not sufficient for ultra-reliable communications. In contrast, when combined with macroscopic diversity, the diversity order is increased, which results in steeper slopes and significantly better SINR performance. It is observed that $M = 3$ macroscopic links, each with 4x4 MIMO, are required to fulfil the 0 dB SINR requirement. The macroscopic gain comes from the higher received power of the desired signal as well as the additional diversity to combat both fast and slow fading. The soft-combining gain is especially relevant for MTs close to the cell boundaries, which are likely to receive similar power from the M serving cells. Macroscopic diversity also minimizes the negative performance impact of the considered handover hysteresis window, since the probability of not being connected to the strongest cell is reduced.

Fig. B.3 summarizes the SINR performance at the 10^{-5}-th percentile with

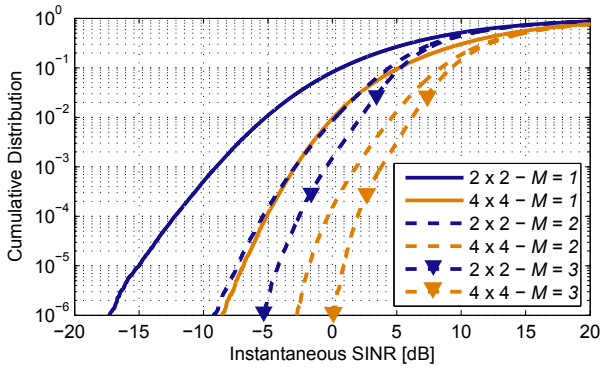


Fig. B.2: SINR performance with different levels of macroscopic diversity applied to 2x2 and 4x4 MIMO configurations. M indicates the macroscopic diversity order.

the different diversity configurations, including cases with ideal interference cancellation of the C strongest interferers (assuming $M = 1$ macroscopic diversity order). The 0 dB SINR target is represented with a horizontal dashed line. As previously mentioned, a 4x4 microscopic scheme with $M = 3$ macroscopic links are required to achieve the necessary SINR outage performance in the considered site-specific network, which is one more macroscopic link compared to our previous evaluation in a standard 3GPP macro scenario [5]. The gain obtained by cancelling up to $C = 3$ interferers is 3.4 dB and 2.5 dB for the 2x2 and 4x4 antenna configurations, respectively, which is not sufficient to fulfil the 0 dB SINR outage target. Compared to the spatial diversity techniques, interference cancellation does not increase the slope of the SINR distribution since the diversity order remains the same. The obtained gains of interference cancellation are smaller than reported in the 3GPP macro network [5], as the considered realistic scenario is less interference-limited.

Increasing the diversity by collocating multiple antennas at the transmitter and/or receiver is a radio-resource efficient way to improve the SINR outage performance. A 2x2 MIMO configuration is the most commonly used scheme in LTE, although 4x4 configurations are also allowed in the standard. Practical implementation of macroscopic diversity schemes is more challenging. Among others, tight coordination and low latency communication between cells are required to support the scheme presented in this study. In addition, macroscopic diversity consumes transmission resources for a single user at multiple cells potentially having impact on the system capacity.

It is reasonable to think that when the performance degradation due to increased resource consumption (and interference) exceeds the diversity gain, macroscopic diversity does not provide any benefit to the system performance. To account for this increased resource usage, the *macroscopic diversity*

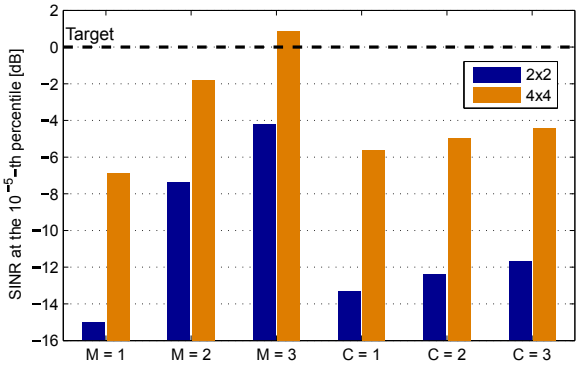


Fig. B.3: Achieved SINR at the 10^{-5}-th percentile with different levels of micro- and macroscopic diversity, and interference cancellation. M and C indicate respectively: level of macroscopic diversity and number of interfering signals cancelled.

overhead (β) is defined as follows [11],

$$\beta = \sum_{m=1}^{M} mR_m \,, \tag{B.7}$$

where R_m is the ratio of MTs with macroscopic diversity order equal to m. The overhead is highest for the case when all the MTs are configured with macroscopic diversity order M, whereas it is lower if this technique is only applied to a fraction of the MTs, e.g. by applying a macroscopic diversity window as explained in Section 3.1. Taking this overhead into account, Fig. B.4 depicts the macroscopic diversity performance with different macroscopic diversity windows. The respective distribution of the macroscopic diversity order for the MTs is written above each group of bars. For each window size, β is calculated according to (B.7) and subtracted (in dB) from the SINR performance obtained via simulations. We refer to this resulting metric as *compensated SINR*. It is worth mentioning that the objective of this metric is to reflect the cost of applying macroscopic diversity, and not to reflect the actual performance. It is observed that the optimal macroscopic diversity window depends on the configuration of interest. For example, a 4x4 MIMO with $M = 3$ achieves the best performance with a 6 dB macroscopic diversity window, whereas a 2x2 scheme with $M = 3$ would require a larger window, e.g. 10 dB .

The last set of results accounts for malfunctions of the cellular infrastructure. Fig. B.5 illustrates the obtained SINR at the 10^{-5}-th percentile when geographically correlated or uncorrelated failures occur with a certain failure probability in the realistic network scenario. For the case of correlated failures, the correlation distance μ is set to 415 meters. This corresponds to a correlation of approximately 0.3 for two cells separated at 500 meter distance. Only the SINR performance with 4x4 and different macroscopic di-

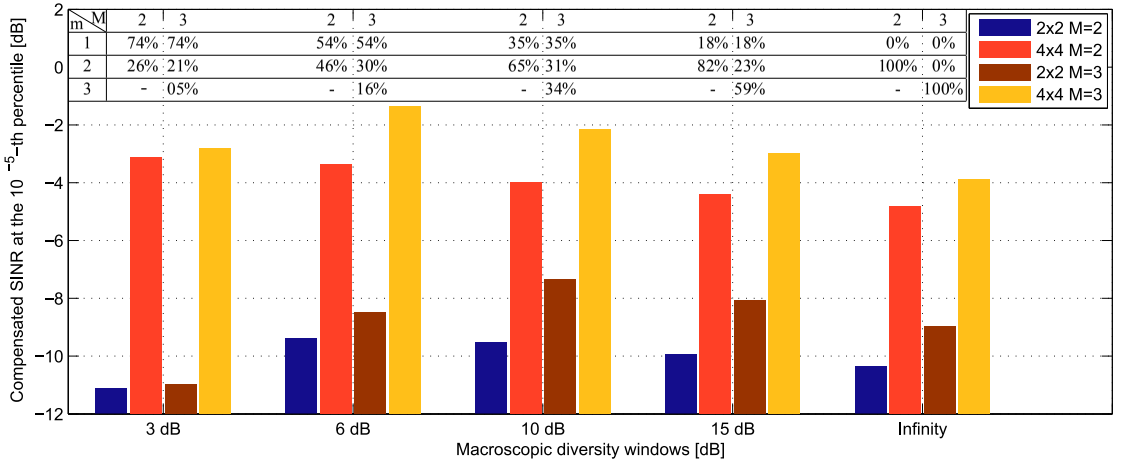


Fig. B.4: Compensated SINR at the 10^{-5}-th percentile for several macroscopic diversity windows accounting for the macroscopic diversity overhead. M refers to the maximum allowed macroscopic diversity order. For each configuration, the percentage of users with macroscopic diversity order m is given above each group of bars.

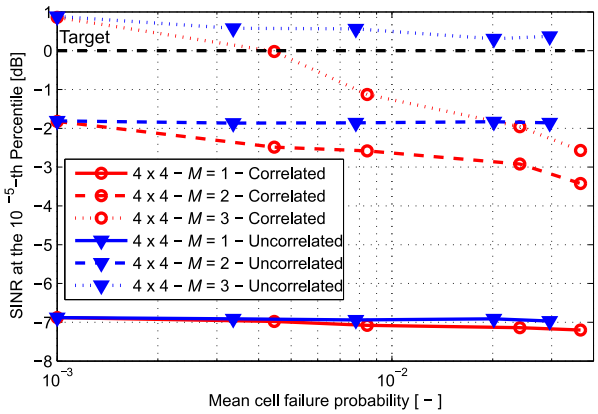


Fig. B.5: Achieved SINR at the 10^{-5}-th percentile for different failure types and probabilities. $\mu = 415$ meters.

versity orders is depicted; however, similar trends are also obtained with 2x2 microscopic diversity schemes. It is observed that uncorrelated failures do not result in significant degradation of the performance. In this case, there is sufficient overlapping coverage as it is generally the case for dense urban cellular networks. In the case of correlated failures, there is relatively high impact for all the evaluated configurations since the failures span over large geographical areas in a clustered manner, hence decreasing the probability of having good coverage. The relative performance degradation is more significant for high macroscopic diversity orders, as the additional (weaker) redundant links would typically experience higher pathloss and therefore are more affected by the noise. For example, the 4x4 configuration with three macroscopic links no longer fulfils the required SINR outage performance if the mean failure probability is higher than 0.5%. It is worth to highlight that the interruption time when switching the connectivity from a failing cell to a fully functional cell could potentially be another source of performance degradation that has not been accounted for in this study. Future work must include the validation of our failure model by e.g. analysing real network mean time between failure (MTBF) and mean time to repair (MTTR) statistics.

6 Conclusion

In this study, we have evaluated the potential of diversity and interference management techniques to achieve very low SINR outage probability as required for ultra-reliable communications. The analysis has been carried out for a realistic site-specific deployment from a big European city. Micro and macroscopic diversity techniques have been shown to be important enablers

of ultra-reliable communications. A 4x4 MIMO scheme with three orders of macroscopic diversity is suggested as a feasible configuration to achieve the required SINR outage performance. Mitigating the interference provides complementary benefit, although it does not increase the diversity order of the desired signal. In addition, failures of the cellular network infrastructure have been considered. Different failure probabilities and geographical dimensions have been evaluated. It has been shown that failures spanning over large geographical areas in a correlated manner can have a significant negative impact when attempting to support high reliability use cases.

References

[1] 3GPP TR 38.913 v0.3.0, “Study on scenarios and requirements for next generation access technologies", March 2016.

[2] N. A. Johansson, Y. P. Eric Wang, Erik Eriksson and Martin Hesslear, “Radio access for ultra-reliable and low-latency 5G communications", *IEEE ICC Workshops*, June 2015.

[3] H. Shariatmadari, S. Iraji and R. Jäntti, “Analysis of transmission methods for ultra-reliable communications", *IEEE PIMRC Workshops*, 2015.

[4] F. Kirsten, D. Ohmann, M. Simsek and G. P. Fettweis, “On the utility of macro- and microdiversity for achieving high availability in wireless networks", *IEEE PIMRC*, Sept. 2015.

[5] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen and P. Mogensen, “Signal quality outage analysis for ultra-reliable communications in cellular networks", *IEEE Globecom Workshops*, December 2015.

[6] A. Azarfar, J. F. Frigon and B. Sanso, “Improving the reliability of wireless networks using cognitive radios", *IEEE Communications Surveys & Tutorials*, vol. 14, no. 2, pp. 338-354, 2nd Quarter 2012.

[7] M. Monemian, P. Khadivi and M. Palhang, “Analytical model of failure in LTE networks", *IEEE Malaysia International Conference on Communications (MICC)*, December 2009.

[8] C. Coletti, L. Hu, H. Nguyen, I. Z. Kovacs, B. Vejlgaard, R. Irmer and N. Scully, “Heterogeneous deployment to meet traffic demand in a realistic LTE urban scenario", *IEEE Vehicular Technology Conference*, September 2012.

[9] G. Pocovi, M. Lauridsen, B. Soret, K. I. Pedersen and Preben Mogensen, “Automation for on-road vehicles: use cases and requirements for radio design", *IEEE Vehicular Technology Conference*, Sept. 2015.

[10] R. Wahl, G. Wölfle, P. Wertz and P. Wildbolz, “Dominant path prediction model for urban scenarios", *14th IST Mobile and Wireless Communications Summit*, June 2005.

[11] H. Holma, A. Toskala (editors), "WCDMA for UMTS - radio access for third generation mobile communications", Third edition, Wiley, 2004.

[12] M. Lichtman, J. H. Reed, T. C. Clancy and M. Norton, “Vulnerability of LTE to hostile interference", *IEEE Global Conference on Signal and Information Processing*, December 2013.

[13] A. D. Lunn and S. J. Davies, “A note on generating correlated binary variables", *Biometrika*, vol. 85, no. 2, pp. 487-490, 1998.

[14] J. H. Macke , P. Berens, A. S. Ecker, A. S. Tolias and M. Bethge, “Generating spike trains with specified correlation coefficients", *Neural Computation*, vol. 21, no. 2, pp. 397-423, February 2009.

[15] 3GPP TS 36.211 v12.5.0, “Evolved universal terrestrial radio access (E-UTRA); physical channels and modulation", April 2014.

Paper C

On the Impact of Precoding Errors on Ultra-Reliable Communications

Guillermo Pocovi, Klaus I. Pedersen, Beatriz Soret

The paper has been published in the
9th International Workshop on Multiple Access Communications (MACOM), 2016.

The layout has been revised.

Abstract

Motivated by the stringent reliability required by some of the future cellular use cases, we study the impact of precoding errors on the SINR outage performance for various spatial diversity techniques. The performance evaluation is carried out via system-level simulations, including the effects of multi-user and multi-cell interference, and following the 3GPP-defined simulation assumptions for a traditional macro case. It is shown that, except for feedback error probabilities larger than 1%, closed-loop microscopic diversity schemes are generally preferred over open-loop techniques as a way to achieve the SINR outage performance required for ultra-reliable communications. Macroscopic diversity, where multiple cells jointly serve the UE, provides additional robustness against precoding errors. For example, a 4x4 MIMO scheme with two orders of macroscopic diversity can achieve the 0 dB SINR outage target at the 10^{-5}*-th percentile, even for a precoding error probability of 1%. Based on the obtained results, it is discussed what transmission modes are more relevant depending on the feedback error constraint.*

1 Introduction

Ultra-reliable communications over wireless is an active research topic that will open the possibility of novel applications [1]. For some of the use cases, latencies of a few milliseconds must be guaranteed with reliability levels up to 99.999%. The signal to interference-and-noise ratio (SINR) outage performance is a relevant metric for ultra-reliable communications. In this context, spatial diversity techniques such as microscopic and macroscopic diversity have shown promising potential. For example, the work in [2], [3] shows that the proper combination of macroscopic and microscopic diversity techniques can provide the required SINR outage performance.

Microscopic diversity is typically used in modern cellular systems, such as the Long Term Evolution (LTE), by use of multiple-input multiple-output (MIMO) antenna techniques. In the downlink, the gains provided by microscopic diversity strongly depend on the availability and accuracy of channel state information (CSI) at the eNodeB. If the channel knowledge is precise enough, closed-loop (CL) schemes, which are known to provide the best performance [4], can be applied. However, in cases of absence or inaccurate CSI knowledge due to e.g. imperfect channel estimation, open-loop (OL) schemes are typically more appropriate.

In frequency division duplex (FDD) modes, where channel reciprocity is not applicable, the eNodeB obtains the CSI through an uplink feedback channel. The CSI contains information about the current channel quality, and the preferred precoding matrix to be applied in downlink CL transmissions. Apart from the typically applied quantization in order to cope with

the limited feedback capacity of real systems, the precoding information is prone to errors due to the intrinsic presence of fading and interference in the wireless channel. The impact of CSI feedback errors have been evaluated from a system capacity point of view. For example, the work in [5] evaluates the influence of CSI feedback errors on the throughput performance of multi-user MIMO systems, whereas [6] demonstrates the significant performance degradation when a UE intentionally reports the wrong CSI to the eNodeB. Previous reliability analyses [2], [3] have not considered these types of imperfections. Our hypothesis is that precoding errors could have a significant impact on ultra-reliable communications, which is what we evaluate in this work.

In this paper we study the impact of CSI feedback errors on the achievable downlink SINR performance in a multi-cell multi-user environment. Our focus is on the very-low percentiles of the SINR distribution in order to quantify the impact of feedback errors on ultra-reliable communications, and determine what transmission modes (e.g. OL or CL) are more relevant depending on the feedback error probability. The complexity of our system model prevents a purely analytical evaluation without omitting important aspects influencing the performance. The evaluation is carried out following the 3GPP-defined simulation assumptions for a LTE macro cellular network that relies on commonly accepted models and methodologies. Mathematical expressions for the user-experienced SINR, when applying the different transmission schemes and related imperfections, are presented in this article and used in the simulations. Long simulations are run to ensure statistical reliable performance results with high level of confidence

The rest of the paper is outlined as follows: Section 2 describes our system model. The simulation assumptions are outlined in Section 3. Performance results are presented in Section 4, followed by concluding remarks in Section 5.

2 System Model

The network consists of a set of $\mathcal{N} = \{1, ..., N\}$ cells, each equipped with T transmit antennas, and a set of $\mathcal{K} = \{1, ..., K\}$ UEs with R receive antennas. Each downlink connection between UE $k \in \mathcal{K}$ and its serving cell $j \in \mathcal{N}$ is represented by a TxR CL MIMO system as shown in Fig. C.1. As our focus is on reliable communications, only single-stream transmission cases are considered [7]. First, each UE estimates the RxT-dimensional channel $\mathbf{H}_{jk}$, whose (m,n)-th element represents the complex channel gain from transmit antenna n at cell j, to receive antenna m at UE k. As a second step, the vector $\mathbf{u}_j$ corresponding to the largest eigenvalue of the $\mathbf{H}_{jk}^H\mathbf{H}_{jk}$ matrix is calculated through singular value decomposition (SVD), i.e. $\mathbf{u}_j = \text{EIG}_{max}(\mathbf{H}_{jk}^H\mathbf{H}_{jk})$. Next, an in-

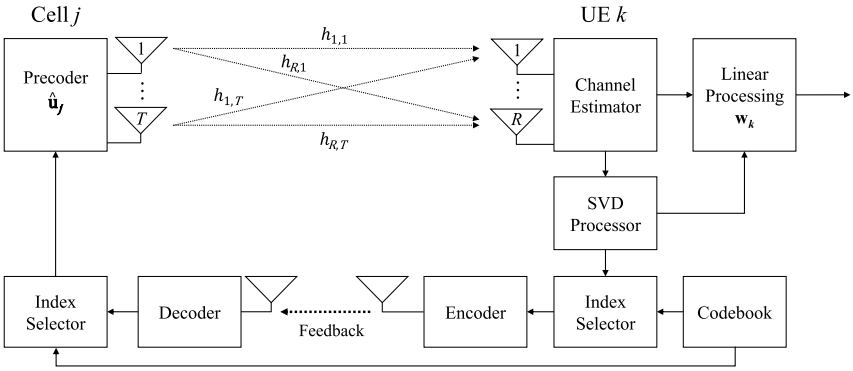


Fig. C.1: Transmitter-Receiver Architecture.

dex selector selects from a pre-defined codebook the precoding vector that matches best with $\mathbf{u}_j$. We refer to this quantized version as $\hat{\mathbf{u}}_j$. The index to the precoder, i.e. precoding matrix indicator (PMI), is transmitted to the cell through the uplink feedback channel.

The cell uses the received PMI to obtain $\hat{\mathbf{u}}_j$, which is then applied in the data transmission. Within each cell, the UEs are served on orthogonal resources, i.e. there is no intra-cell interference as it is also the case for LTE assuming single-stream and single-user MIMO transmission modes [4]. In a frequency-flat fading case, the R-dimensional received signal $\mathbf{r}_j$ by a user (for simplicity, we omit the user-specific index) served in cell $j \in \mathcal{N}$ is given as follows,

$$\mathbf{r}_j = \mathbf{H}_j \sqrt{\Omega_j} \hat{\mathbf{u}}_j s_j + \sum_{i \in \mathcal{N} \setminus j} \mathbf{H}_i \sqrt{\Omega_i} \hat{\mathbf{u}}_i s_i + \mathbf{n} \,, \tag{C.1}$$

where Ω_i represents the averaged received power from the i-th cell, including the effect of the antenna gain and pattern, distance-dependent attenuation and shadowing; s_i represents the transmitted symbol (for simplicity, $\|s_i\| = 1$) and $\mathbf{n}$ is a Rx1 zero mean Gaussian vector with variance σ^2 representing the noise power at each receiving antenna.

In order to maximize the received signal power at the receiver, the R received signals are combined by applying a weight vector $\mathbf{w} = \mathbf{H}_j \hat{\mathbf{u}}_j$. The resulting post-detection SINR expression is given by,

$$SINR_j = \frac{\Omega_j \|\hat{\mathbf{u}}_j^H \mathbf{H}_j^H \mathbf{H}_j \hat{\mathbf{u}}_j\|^2}{\sum_{i \in \mathcal{N} \setminus j} \Omega_i \|\hat{\mathbf{u}}_j^H \mathbf{H}_j^H \mathbf{H}_i \mathbf{u}_i\|^2 + \sigma^2 \|\hat{\mathbf{u}}_j^H \mathbf{H}_j^H\|^2} \,, \tag{C.2}$$

where $[\cdot]^H$ denotes the Hermitian transpose.

The presented microscopic scheme corresponds to transmission mode 6 (TM6) in LTE terminology. TM6 is a special case of CL spatial multiplexing (TM4) where the transmission rank is limited to one. The UE utilizes the

downlink cell-specific reference signals (RS) to perform the channel estimation, and select the preferred PMI (from a common codebook). The eNodeB signals the applied precoding to the UE in the downlink grant [4].

TM6 allows operation with 2 or 4 transmit antennas. For the former case, the LTE Release 8 codebook contains 4 precoding vectors, whereas there are 16 different entries for four transmit antennas [8]. The number of entries have been selected as a tradeoff between the uplink signalling overhead and downlink performance.

In cases where channel information is missing at the eNodeB, spatial diversity gain can be obtained with open-loop transmission modes. LTE transmission mode 3 (TM3) supports OL spatial diversity by use of space-frequency block coding (SFBC) techniques [9], which are based on the space-time block coding initially proposed by Alamouti [10]. SFBC achieves similar diversity order to CL, but with a reduced received power since the transmit beamforming gain is not obtained [10]. The post-detection SINR of a TxR OL MIMO scheme is simply modelled by adding a $10\log_{10} T$ SINR penalty to the performance obtained with a TxR MIMO system assuming full channel knowledge at the transmitter (i.e. without quantized precoding) [10].

As a method to further improve the SINR outage performance, we also consider macroscopic diversity transmissions from M cells to a certain UE [2]. We assume a simple soft-combining approach as known from Universal Mobile Telecommunications System (UMTS), where the received signal from each macroscopic branch is independently detected and combined at the UE [11]. As this scheme rely on non-coherent transmissions, each of the M macroscopic links can be modelled as shown in Fig. C.1. The SINR after combining M ($1 \leq M \leq N$) macroscopic branches is expressed as follows,

$$SINR = \sum_{j=1}^{M} SINR_j \,, \tag{C.3}$$

where $SINR_j$ is the SINR calculated according to (C.2), assuming the UE is connected to cell j.

2.1 Precoding Errors

The gains provided by spatial diversity techniques depends on the accuracy of the CSI at the transmitter [4]. Since the CSI is estimated at the UE and transmitted to the eNodeB through an uplink feedback channel, it is vulnerable to multiple sources of delay and other imperfections. The delays are a consequence of the constrained CSI reporting periodicity and processing time, meaning that the optimal precoding will not be immediately applied at the transmitter. Additionally, errors in the channel estimation could lead to a sub-optimal PMI selection. Another source of degradation is errors in the

uplink transmission of the CSI due to the inevitable presence of fading and interference in the wireless channel.

We focus uniquely on the effect of precoding feedback errors. We assume that errors in the feedback channel can occur with a given error probability P_e. In such cases, the PMI decoded by the eNodeB will be different to the reported by the UE, which will lead to a erroneous precoder selection. The errors in the feedback channel are assumed to be i.i.d for each UE-eNodeB connection.

Since the eNodeB signals the applied precoding in the scheduling grant, the UE can still apply a proper combining weight vector to improve the signal quality at the receiver. In other words, the benefits of transmit diversity are lost but the receive diversity gain is maintained. As a more pessimistic case, errors could alter the applied-PMI related signalling in the downlink grant, resulting in loss of both the transmit and receive diversity gain.

3 Simulation Assumptions

The evaluation is carried out by analysing the downlink SINR distribution for different antenna schemes, transmission methods, and feedback error probabilities. A snapshot-based simulation approach is applied and the respective assumptions are summarized in Table C.1. A large macro-cellular network composed of three-sector sites with inter-site distance of 500 m is assumed, where UEs are uniformly distributed [12]. Cells are transmitting at full power (full load conditions) at a 2 GHz carrier frequency. The simulation procedure is as follows: Each UE selects M serving cells according to the average received power. Effects of user mobility and handovers are not explicitly included in the simulations. However, the effect of handover hysteresis margin is implicitly modelled in the active set selection algorithm: each UE identifies the strongest received cells that are within a certain *handover window*, as compared to the strongest cell. A serving cell for the UE is then randomly selected from the cells within the handover window. This method models the effect where not all UEs are served by their strongest cell due to the use of handover hysteresis margins in reality.

The experienced instantaneous post-detection SINR is calculated for each UE following the models in Section 2. For each snapshot, the fast fading is independent and identically distributed for each transmit-receive antenna pair, following a complex Gaussian distribution (i.e. the envelope is Rayleigh distributed). Additive white Gaussian noise with a power spectral density of -174 dBm/Hz is considered. It is assumed that UEs are scheduled with 10 MHz bandwidth, resulting in a noise power of -96 dBm when including a 8 dB noise figure at the UE.

A large number of snapshots are simulated and the generated SINR sam-

Table C.1: Simulation assumptions

Parameter	Value
Network layout	3GPP Macro case 1
UE distribution	Uniformly distributed in outdoor locations
Macro cell transmit power	46 dBm
Carrier frequency	2.0 GHz
Propagation	$128.1 + 37.6 \log10(R[\text{km}])$ dB
Antenna gain	BS: 14 dBi. UE: 0 dBi
Antenna pattern	BS: 3D with 12° downtilt UE: omnidirectional
Shadowing distribution	Log-normal with $\sigma = 8$ db
Shadowing correlation	Intra-site: 1.0 ; Inter-site: 0.0
Noise power spectral density	-174 dBm/Hz
Noise figure	8 dB
Noise power	-96 dBm @10 MHz
Handover window	3 dB
Fast fading	Rayleigh distributed; Uncorrelated among the different antenna branches
Feedback error probability P_e	10^{-1}, 10^{-2}, 10^{-3}
SINR outage target	0 dB at the 10^{-5}-th percentile

ples are used to form empirical cumulative distribution functions (CDF). Our target is to study the impact of different feedback error probabilities on the SINR outage performance. In line with [1], the key performance indicator (KPI) is the SINR at the 10^{-5}-th percentile. At this percentile, we consider a 0 dB SINR as an appropriate target to have error-free downlink reception, and therefore fulfil the low latency requirements of ultra-reliability use cases (we refer to [2] for more details).

4 Results

The first set of results correspond to the relatively pessimistic case where the PMI applied by the eNodeB is unknown by the UE, thus the UE assumes that the applied precoding is the one that it has previously signalled. Fig. C.2 shows the empirical CDF of the SINR distribution for 2x2 and 4x4 schemes, OL and CL transmission modes, and different feedback error probabilities. Obviously, the 4x4 schemes offer superior performance as compared to 2x2 MIMO schemes. The benefits of CL transmissions over OL schemes are also observable: 4.6 dB and 2.2 dB SINR gain for 2x2 and 4x4 schemes, respec-

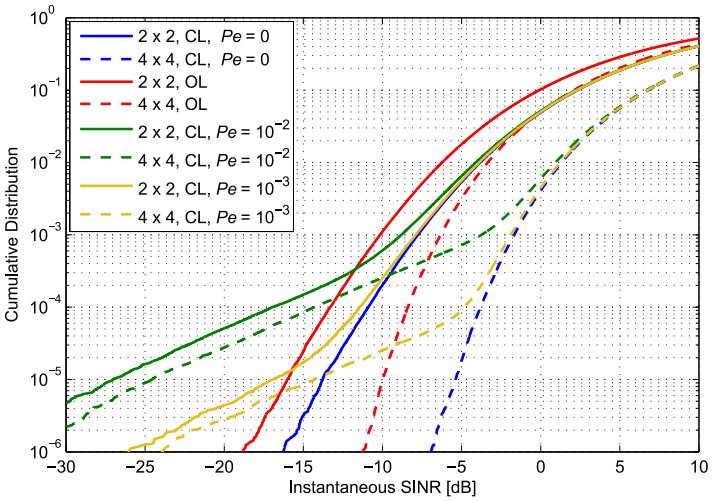


Fig. C.2: SINR outage performance with 2x2 and 4x4 antenna schemes, different transmission modes and precoding error probabilities (P_e). $M = 1$.

tively, at the 10^{-5}-th percentile.

When including the effects of feedback errors, a significant degradation of the performance is observed. For example, even for $P_e = 10^{-3}$, the experienced SINR degradation at the 10^{-5}-th percentile is as high as 8.9 dB and 3.2 dB for 2x2 and 4x4 antenna schemes. The reason is that, when this type of errors occur, the benefits of both transmit and receive diversity are not obtained, i.e. the instantaneously experienced diversity order is equivalent to a 1x1 MIMO system. Under such circumstances, it is shown how OL schemes, which do not require any uplink CSI feedback, offer better performance.

Next, we consider the case where the eNodeB applies an erroneous precoding vector, but the applied PMI is known at the receiver. Fig. C.3 and C.4 shows the SINR distribution with 2x2 and 4x4 antenna schemes, respectively. Cases with second order of macroscopic diversity and $P_e = 10^{-1}$ are also shown. As compared to the performance results in Fig. C.2, errors in the uplink feedback have less impact on the SINR performance. For example, CL configurations with $P_e = 10^{-3}$ and $M = 1$ experience a performance degradation of only 0.3 dB. In this case, the receiver has knowledge of the applied precoding, which allows to fully harvest the receive diversity gain. It is observed that for $P_e \leq 10^{-2}$, the performance of CL schemes is better than OL.

As also observed in previous studies [2], macroscopic diversity provides additional protection against fast and slow fading hence providing significantly better SINR performance. Even for $P_e = 10^{-1}$, only a 1.3 dB perfor-

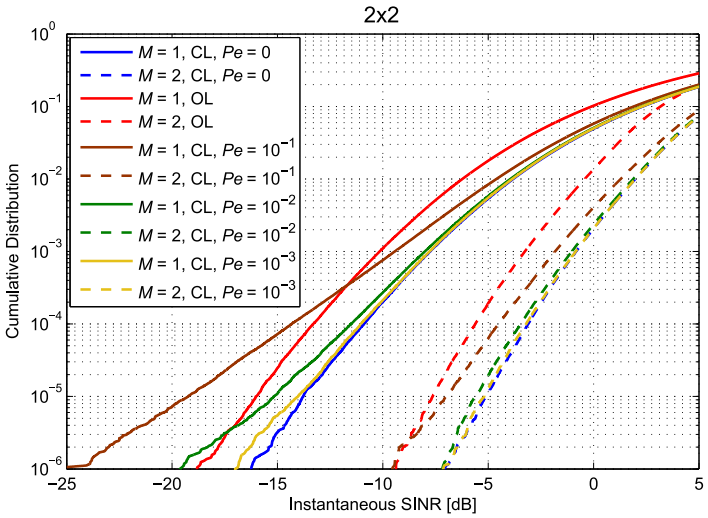


Fig. C.3: SINR outage performance with a 2x2 antenna scheme, different transmission modes and precoding error probabilities (P_e). It is assumed that the applied PMI is known at the UE.

mance degradation is observed for 2x2 and 4x4 MIMO schemes. With macroscopic diversity, the probability of experiencing feedback error across the M links is reduced. Note that compared to the intra-cell MIMO schemes, the considered macroscopic diversity technique relies on non-coherent transmissions and soft-combining of the multiple received signals at the UE, therefore it is only required to report traditional CSI feedback to each of the M eNodeBs.

Fig. C.5 summarizes the achieved 10^{-5}-th percentile SINR performance under different transmission schemes and feedback error probabilities. The 0 dB SINR target is represented with a horizontal dashed line. As also concluded in [2], a 4x4 CL MIMO scheme with $M = 2$ allows to fulfil the 0 dB SINR target. However, it is observed that this is only achievable under certain feedback error probabilities. For instance, if the feedback error probability is $P_e \geq 10^{-1}$, 4x4 MIMO with $M = 2$ no longer fulfils the 0 dB SINR target. The SINR degradation due to feedback errors is much more severe for configurations with low diversity order. For example, 4x4 CL MIMO with $M = 1$ achieves similar performance as 4x4 OL for $P_e = 10^{-1}$; whereas, under the same error probability, 2x2 CL with $M = 1$ is 3.2 dB worse than OL.

5 Conclusions

In this paper we have evaluated the SINR outage performance under different CSI feedback error constraints in order to quantify its impact on ultra-

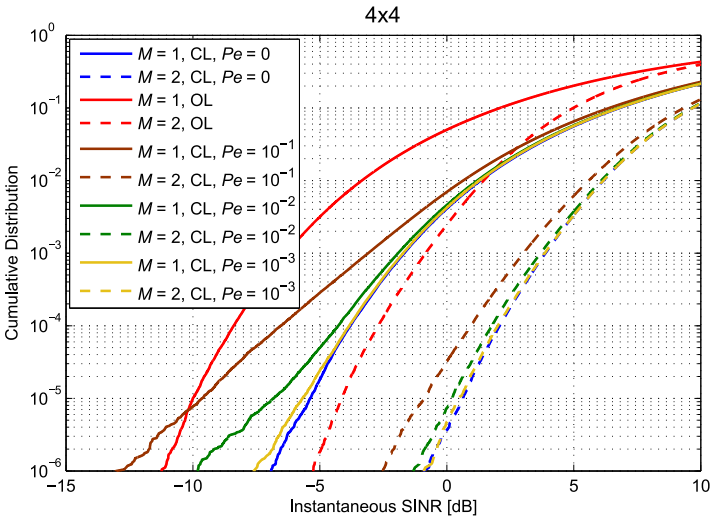


Fig. C.4: SINR outage performance with a 4x4 antenna scheme, different transmission modes and precoding error probabilities (P_e). It is assumed that the applied PMI is known at the UE.

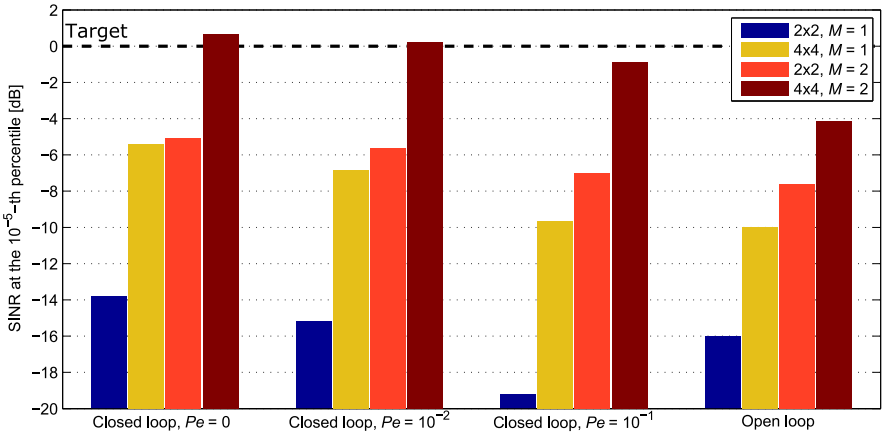


Fig. C.5: Achieved SINR at the 10^{-5}-th percentile for several transmission schemes and precoding error probabilities (P_e).

reliable communications. It have been shown that even for feedback error probabilities as high as 10^{-2} (i.e. three orders of magnitude larger than the required reliability), there is a benefit of using closed-loop MIMO schemes over open-loop schemes. The performance degradation due to errors in the feedback can be reduced by applying macroscopic diversity, as the considered scheme relies on non-coherent independent transmissions from the different macroscopic branches. For instance, a 4x4 MIMO scheme with two orders of

macroscopic diversity can achieve the 0 dB SINR outage target at the 10^{-5}-th percentile, even with a 1% error probability in the CSI feedback. For configurations with low diversity order, a larger performance impact has been observed. For example, closed-loop 2x2 MIMO without macroscopic diversity, performs 3.2 dB worse than open-loop transmissions for a 10% feedback error probability. Future work must also consider other sources of imperfections in the channel information. For instance, as a consequence of non-ideal channel estimation at the UE, or due to delays in the CSI report. This will allow to fully assess the reliability performance in a practical setting.

References

[1] 3GPP TR 38.913 v0.3.0, “Study on scenarios and requirements for next generation access technologies", March 2016.

[2] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen and P. Mogensen, “Signal quality outage analysis for ultra-reliable communications in cellular networks", *IEEE Globecom Workshops*, December 2015.

[3] F. Kirsten, D. Ohmann, M. Simsek and G. P. Fettweis, “On the utility of macro- and microdiversity for achieving high availability in wireless networks", *IEEE PIMRC*, September 2015.

[4] H. Holma and A. Toskala, “LTE for UMTS: Evolution to LTE-Advanced", John Wiley & Sons Ltd, March 2011.

[5] B. Mielczarek, and W.A. Krzymień, “Influence of CSI feedback errors on capacity of linear multi-user MIMO systems", *IEEE Vehicular Technology Conference*, April 2007.

[6] A. Mukherjee and A.L. Swindlehurst, “Poisoned feedback: The impact of malicious users in closed-loop multiuser MIMO systems", *IEEE International Conference on Acoustics Speech and Signal Processing*, March 2000.

[7] D. N. C. Tse, P. Viswanath and L. Zheng, “Diversity-multiplexing tradeoff in multiple-access channels", *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 1859-1874, September 2004.

[8] 3GPPP TS 36.211 v12.5.0, “Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation", April 2014.

[9] M. J. Dehghani, R. Aravind, S. Jam and K. M. Prabhu, “Space-frequency block coding in OFDM systems", *IEEE TENCON*, November 2004.

[10] S. Alamouti, “A simple transmit diversity technique for wireless communications", *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451-1458, October 1998.

[11] H. Holma, A. Toskala (editors), "WCDMA for UMTS - radio access for third generation mobile communications", Third edition, Wiley, 2004.

[12] 3GPPP TR 36.814 v9.0.0, “Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects", March 2010.

Part III

Achieving Low Latency in Multi-user Cellular Systems

Achieving Low Latency in Multi-user Cellular Systems

The achievable latency in a communication system depends not only on the TTI duration, but also on queuing delays, the applied traffic model (and associated payload size), etc. Therefore, a multi-user system-level evaluation with dynamic traffic arrival is desired in order to capture the impact of these multiple components on the metrics of interest.

1 Problem Description

Previously, we have determined the required level of diversity and interference management, such that a URLLC packet is successfully transmitted to a mobile user, with a single virtually error-free transmission, and in a fully-loaded network. This is assuming that sufficient radio resources are available at the serving cell(s), such that each URLLC transmission is immediately scheduled with negligible queuing delay.

The next step is to account for time-domain aspects; more specifically, the multi-user traffic dynamics typically present in cellular systems. However, before moving to a highly-complex system for a holistic performance assessment, it is a prerequisite to understand: what are the main components that contribute to the communications latency, how these components relate to each other, and what is their impact on URLLC.

To answer these questions, the main contributors to the one-way downlink radio latency are first identified, and studied in a simplified system-level setting with dynamic user arrivals. Next, the obtained learnings are used for conducting a more realistic evaluation of the URLLC latency and reliability performance, following the 5G NR evaluation methodology agreed in 3GPP [1]. The used simulation tool accounts for the majority of RRM functionalities, link-to-system mapping for determining the error probability of each data transmission, time-varying traffic and interference, etc. In such a dynamic setting, challenges in the context of link adaptation are identified,

and addressed with various proposed enhancements. The presented performance results allow to determine the maximum URLLC traffic load that can be supported in the network, while still satisfying the stringent latency and reliability requirements.

2 Objectives

The goals of this part of the thesis are the following:

- Identify the main components that influence the communication latency in dynamic multi-user systems.
- Understand how these components are related to each other, and identify potential tradeoffs that must be accounted when formulating novel techniques and enhancements for URLLC.
- Propose solutions for achieving the 1 ms latency and 99.999% reliability required for URLLC.
- Evaluate the URLLC performance in a system-level setting and determine the feasible load regions for fulfilling the URLLC requirements.

3 Included Articles

The main findings of this part are included in the following articles:

Paper D. On the Impact of Multi-User Traffic Dynamics on Low Latency Communications

This article studies the downlink latency performance under different TTI durations (0.25, 0.5, 1 and 2 ms) and load conditions, assuming a mixture of high-priority low latency communication (LLC) traffic and best-effort background eMBB load. The effect of varying relative control channel overhead depending on the TTI size is explicitly taken into account. Some simplifications at the physical layer such as error-free transmissions and omission of fast fading are assumed. The different delay components that contribute to the overall communication latency are described and thoroughly analysed, with special focus on the tradeoffs between the TTI duration, their respective spectral efficiency, and the queuing delays experienced at the cell. Moreover, eMBB throughput performance results under different system configurations are presented.

Paper E. MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks

This paper presents various enhancements for improving the latency and reliability in cellular networks, namely i) a short TTI duration and fast HARQ RTT in order to reduce the transmission and processing time of the URLLC payloads, ii) low-pass infinite impulse response (IIR) filtering of the CQI measurements at the UE for improved link adaptation accuracy, and iii) adjustment of the BLER target of URLLC transmissions in accordance with the average load that is experienced in the system. The benefit of the proposed enhancements is analysed in a fully-dynamic system-level setting, following the evaluation methodology agreed in 3GPP. A standard macro cellular network with 21 cells is assumed, where 210 URLLC UEs are uniformly distributed (average of 10 UEs per cell). The URLLC traffic is modelled as small payloads of 200 Bytes that arrive for each user in the downlink direction following a Poisson arrival process.

4 Main Findings

Tradeoffs between TTI duration, spectral efficiency, and queuing delay.

The analysis of the downlink latency performance in Paper D shows the benefit of adjusting the TTI size depending on the experienced load in the system. At the 99% percentile of the latency distribution, a short TTI size of 0.25 ms provides the best latency performance for low offered loads of LLC traffic. The main benefits come from the low over-the-air delay when transmitting the small LLC payloads. However, at high offered load, the lowest latency is obtained with a 0.5 ms TTI duration. The main reason for this behaviour is the tradeoff between the spectral efficiency and queuing delay: as the load in the system increases, the queuing delay becomes the most dominant component in the total latency. Therefore, it is beneficial to increase the spectral efficiency (by using a longer TTI length with lower relative CCH overhead) in order to reduce the experienced queuing delay at the cell's buffers.

The optimal TTI duration depends not only on the offered load but also on the percentile of interest. For instance, even at low load, there is a benefit of using a 0.5 ms TTI size over a 0.25 ms TTI, if the percentile of interest is 99.9% or above. Furthermore, at very high load, configurations with a 1 ms TTI duration provide the best performance for percentiles beyond 99.9%.

Benefits of the proposed URLLC enhancements

Paper E shows that reducing the HARQ round-trip time from 8 to 4 TTIs signicantly improves the 99.999% percentile of the URLLC latency distribu-

tion. More importantly, it allows to fit one HARQ retransmission within the 1 ms latency budget (assuming a TTI duration of 0.143 ms), hence relaxing the BLER constraint that the URLLC transmissions need to fulfil (e.g. 10^{-2} in the first transmission, and $\leq 10^{-5}$ residual error in the second transmission). The advantage of using two transmission attempts, as compared to a single (very conservative) transmission, is a reduction of the average amount of radio resources required to transmit the small data payloads.

However, in order to perform this in practice, it is essential that the serving cell has accurate knowledge of the channel quality experienced at the UE, such that a proper MCS (fulfilling the BLER constraint) is selected. As shown in Paper E, this is a challenging task given the rapid load (and interference) fluctuations due to the very sporadic URLLC traffic. In this respect, the proposed CQI measurement procedure provides significant gains, as it reduces the mismatch between the channel quality estimation at the cell (based on the periodically-reported CQI) and the actual SINR during the downlink transmission.

In line with the previously identified tradeoffs between spectral efficiency and queuing delay, Paper E shows the importance of adjusting the initial BLER target of the URLLC transmissions in accordance with the experienced load in the system. At low load, a conservative BLER target of 0.1% provides the best latency performance. However, as the load increases, the optimal latency performance is obtained by performing a more aggressive link adaptation, e.g. 1%-10% BLER target. In this case, higher BLER target increases the spectral efficiency of the system, hence reducing the queuing delay experienced at the cells.

URLLC performance in a realistic setting

The URLLC latency at the 99.999% percentile, when applying the proposed enhancements, is summarized in Fig. III.1 (based on results presented in Paper E). As a reference, the resource utilization (i.e. average number of transmitted physical resource blocks (PRBs)) is also included. The URLLC requirements are fulfilled for loads up to 2 Mbps (latency of 0.98 ms), but cases with 4 or 6 Mbps offered load also experience decent performance. As described in Paper E, the gain provided by each enhancement depends on the load conditions. For the sake of brevity, the main findings are summarized as follows:

- At low load, the processing time at the transmitter and receiver is the dominant component in the achievable latency; therefore, significant benefit is obtained by reducing the HARQ RTT to 4 TTIs.
- Low-pass IIR filtering of the CQI measurements at the UE provides the largest gains at high load, i.e when the cell activity is higher and more

sporadic interference is experienced in the network.

- Increasing the BLER target in accordance with the average load in the system provides valuable benefits. The optimal first-transmission BLER target (obtained heuristically via simulations) gradually increases from 0.1% at 1 and 2 Mbps offered load, up to 10% at 8 Mbps load.

Furthermore, the latency performance is also evaluated on a per-user basis. The percentage of users not satisfying the 1 ms latency requirement drastically increases with the load, e.g. 20% and 60% for 4 and 6 Mbps offered load, respectively. As shown in Fig. E.7 (Paper E), there is a strong correlation between the user latency performance and its experienced channel quality.

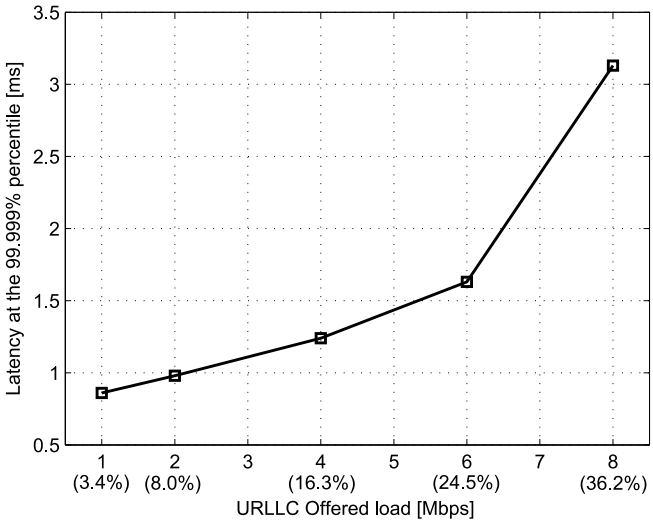


Fig. III.1: URLLC latency at the 99.999% percentile for different offered loads of URLLC traffic. The resource utilization is shown in the x-axis for each evaluated load condition. Carrier bandwidth: 10 MHz

Relation with Part II

Although not discussed in Paper E, the considered closed-loop 2x2 single-stream MIMO with interference rejection combining (MMSE-IRC) receiver was sufficient to fulfil the URLLC requirements for low offered loads of URLLC traffic (as observed in Fig. III.1). The reasons are the following: i) Paper E assumes a time-varying load, meaning that lower inter-cell interference (hence, higher SINR) is generally experienced as compared to the full-load considerations in Part II. ii) Rather than a single-shot transmission, up to two transmissions opportunities within the 1 ms latency budget are considered

in this part of the study. The use of HARQ provides a power gain (from Chase-combining the retransmitted packets) and diversity gain. The latter is mainly due to the likely-different interference conditions experienced per URLLC payload transmission, rather than diversity against fast fading (due to the short time between retransmissions). And iii) the UE deployment in Paper E is limited to users with Geometry factor > -3.5 dB (see [2] for motivation and related discussions). The Geometry factor is equivalent to the average pre-processing SINR experienced by the UE under full-load conditions. Note that, in the standard 3GPP urban macro network, the probability of experiencing a Geometry factor below -3.5 dB is roughly 2% [3].

Nevertheless, the findings in Part II are still valid. For instance, higher order of microscopic and macroscopic diversity would be required in a fully-loaded network, and in cases where timing constraints do not allow a HARQ retransmissions. Macroscopic diversity will also be essential when including the effects of user mobility and failures of the network infrastructure, which have been left out in this part of the study.

References

[1] 3GPP TR 38.802 v14.0.0,, "Study on New Radio Access Technology; Physical Layer Aspects," Mar. 2017.

[2] R1-166398, "URLLC system level simulation assumptions," *3GPP TSG-RAN WG1 #86*, Aug. 2016.

[3] H. Holma and A. Toskala, *LTE for UMTS: Evolution to LTE-Advanced*, 2nd ed. Wiley Publishing, 2011.

Paper D

On the Impact of Multi-User Traffic Dynamics on Low Latency Communications

Guillermo Pocovi, Klaus I. Pedersen, Beatriz Soret, Mads Lauridsen, Preben Mogensen

The paper has been published in the
International Symposium on Wireless Communication Systems (ISWCS), 2016.

Abstract

In this paper we study the downlink latency performance in a multi-user cellular network. We use a flexible 5G radio frame structure, where the TTI size is configurable on a per-user basis according to their specific service requirements. Results show that at low system loads using a short TTI (e.g. 0.25 ms) is an attractive solution to achieve low latency communications (LLC). The main benefits come from the low transmission delay required to transmit the payloads. However, as the load increases, longer TTI configurations with lower relative control overhead (and therefore higher spectral efficiency) provide better performance as these better cope with the non-negligible queuing delay. The presented results allow to conclude that support for scheduling with different TTI sizes is important for LLC and should be included in the future 5G.

1 Introduction

Fifth generation (5G) cellular technologies are expected to bring support for a wide range of use cases [1]- [3]. 5G is foreseen not only to cope with the continuously increasing mobile broadband (MBB) traffic demands, but also to enable novel communication paradigms such as ultra-reliable low-latency communications (URLLC) [2]- [4].

The downlink latency performance in a multi-user cellular network is the focus of this paper. Achieving low latency communication (LLC) is very challenging as it requires the optimization of the multiple components that contribute to the latency budget [5]. The queuing delay at base station nodes is a particularly important component. This is a function of the offered load, traffic dynamics, scheduling strategy and also aspects related to the air interface, e.g. frame structure and transmission time interval (TTI). Examples of studies investigating the queuing delay (and related system aspects) include the work in [6], where the tail distribution of the delay is estimated with different scheduling strategies over a time-slotted fading channel. In the context of cellular networks, the work in [7] analyses the delay performance of various multiple-access schemes with multiple priority classes. In [8], a discrete queuing model is applied to study the downlink throughput and delay performance of a orthogonal frequency division multiple access (OFDMA)-based system. More recently, the work in [9] proposes a flexible frame structure for dynamic scheduling of users with different TTI sizes in accordance to each user requirements. Although short TTI (e.g. 0.25 ms) is beneficial to reduce the over-the-air transmission time, it has a cost in terms of higher signalling overhead and therefore lower spectral efficiency [10]. There is therefore a compromise between the benefits of having short TTI durations, and the experienced queuing delay as a result of the reduced spectral efficiency.

In this work we go a step forward and analyse the tradeoffs between queuing delay and TTI size on a system level. Our main focus is on the achievable latency under different TTI durations and system loads; but we also present relevant results about the spectral efficiency and throughput performance under the different system configurations. We build on the recent study in [9], that proposes dynamic adjustment of the TTI on a per-user basis. The evaluation methodology is dynamic system-level Monte Carlo simulations with bursty traffic, where we consider the effects of different radio channel conditions per user, and varying relative control overhead depending on the TTI size. Despite some simplifications at the physical layer such as error-free transmissions, our simulation framework allows us to draw initial conclusions on the impact of different elements on the total latency, and relevant tradeoffs between spectral efficiency and latency. In a nutshell, our results reveal that as the load increases, the system must gradually increase the TTI size (and consequently the spectral efficiency) in order to cope with the non-negligible queuing delay.

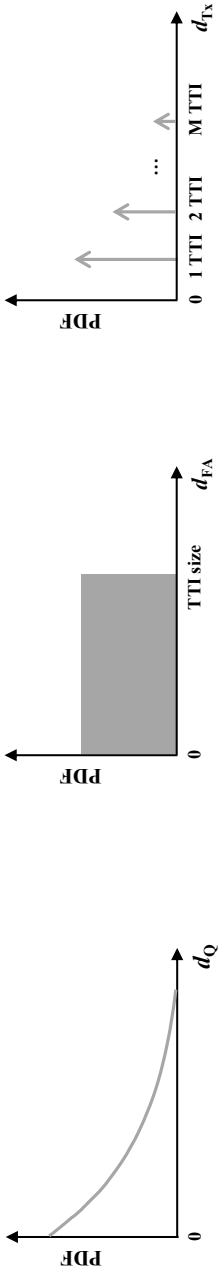
The rest of the paper is organized as follows: Section 2 describes the multiple elements accounting for the user latency. Section 3 presents an overview of the considered frame structure, including multiplexing of users and scheduling format considerations. Section 4 explains the methodology and considered assumptions. Performance results are presented in Section 5, followed by a discussion in Section 6. Finally, conclusions are summarized in Section 7.

2 Latency Composition and Related Definitions

We first describe the various sources that contribute to the downlink latency in a cellular system. A traffic source generates data that are transmitted to a traffic sink via the cellular system. First, the data from higher layers are received at the base station node and are stored in the transmission buffers. Some time is typically required at the base station to process the data and perform the scheduling decision. When the payload is ready to be scheduled, the system must wait to the beginning of the next TTI to transmit the data, assuming a time-slotted system. The data is placed in the radio frame and transmitted to the mobile terminal, where it is subject to a certain processing delay before it is successfully decoded and forwarded to the traffic sink at higher layers. The user-plane one-way latency L for a user scheduled in the downlink can therefore be expressed as [5],

$$L = d_{\mathrm{Q}} + d_{\mathrm{bsp}} + d_{\mathrm{FA}} + d_{\mathrm{Tx}} + d_{\mathrm{mtp}} \quad [\mathrm{s}], \tag{D.1}$$

where d_{Q}, d_{FA} and d_{Tx} represent the queuing, frame alignment, and transmission delay, respectively; and d_{bsp} and d_{mtp} represent the processing delay



(a) Queuing delay: Time it takes for the packet to arrive to the transmission buffer

(b) Frame alignment delay: Time from the moment the packet arrives to the transmission buffer to the beginning of next TTI

(c) Transmission delay: Number of TTIs needed to transmit the packet

Fig. D.1: Sketch of distribution of the different delay components.

at the base station and mobile terminal. Note that we refer to *delay* as the separate contribution of the various components, and *latency* to the sum of all components. Some of these components are described in Fig. D.1. The queuing delay depends on the amount of users that are multiplexed on the same radio resources. Given the random behaviour of packet arrivals, even at relatively low load, there is a probability of experiencing queuing delay due to the instantaneous variation of the incoming traffic. The frame alignment delay depends on the frame structure and duplexing mode. For frequency division duplex (FDD) modes, such as considered in this work, the frame alignment delay is bounded between 0 and the TTI duration, depending on whether the packet reaches the buffer right before or after a TTI begins. The transmission of the payload takes at least one TTI but it can take multiple TTIs depending on the available resources, payload size, radio channel conditions, transmission errors and the respective retransmissions, etc. The processing delay at both base station and mobile terminal depends on their processing capabilities and is typically on the order of a few milliseconds in LTE for each downlink data payload [5]. Shorter processing delay is expected for 5G in order to allow support for lower latency [2].

3 Overview of 5G Flexible Frame Structure

The OFDMA-based frame structure presented in [9] is adopted. Users are flexibly multiplexed on a grid of orthogonal time-frequency tiles, as shown in Fig. D.2. Each tile corresponds to the minimal resource allocation for a user, composed of one subframe in the time domain and a physical resource block (PRB) in the frequency domain. On each scheduling opportunity, an arbitrary number of tiles can be assigned to each user providing therefore high flexibility in terms of TTI length and bandwidth allocation. The control channel (CCH), marked as dark blue in Fig. D.2, is accommodated within the resources assigned to each user (i.e. in-resource CCH). The CCH contains the scheduling grant indicating the specific time-frequency resource allocation for each user, among other relevant link adaptation parameters required to decode the data. The actual resource allocation is performed in accordance with the user-specific service requirements. Using a short TTI (e.g. 0.25 ms) allows to achieve low frame alignment delay and shorter transmission time, at the expense of large CCH overhead. In contrast, the use of long TTIs results in lower CCH overhead, among other benefits that increase the spectral efficiency of the system [9].

3.1 Scheduling format and frame numerology

The CCH and data are multiplexed within the assigned resources per user. This user-specific approach allows to dynamically vary the coding rate of the

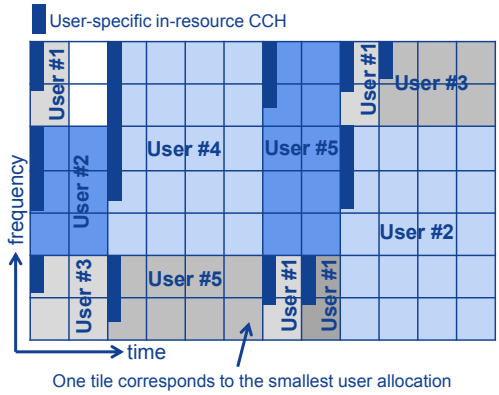


Fig. D.2: User multiplexing example on 5G flexible frame structure.

CCH overhead in order to match the channel conditions of each user (note the difference in size of the user-specific CCH depicted in Fig. D.2). Taking the LTE physical downlink control channel (PDCCH) link-performance as a reference, a minimum of 36 resource elements (REs) are required to transmit the CCH with a block-error rate (BLER) of 1% or less for users experiencing relatively good channel conditions [11]. One RE corresponds to one OFDM subcarrier symbol. Additional robustness is obtained by using higher aggregation levels (i.e. repetition encoding rate) of 2, 4, or 8. Table D.1 summarizes the required number of REs for the CCH depending on the user-specific signal to interference and noise ratio (SINR) [11].

We adopt one of the physical layer numerology options proposed for 5G in [12]. It consists of 16 OFDM symbols per 1 ms, 17.143 kHz subcarrier spacing, and a PRB size of 12 subcarriers. We consider TTI durations of 0.25, 0.5, 1 and 2 ms (4, 8, 16 and 24 OFDM symbols, respectively). On every scheduling opportunity, the resource allocation to a user must be sufficiently large to accommodate the in-resource CCH as well as a reasonable data payload and reference symbols. This sets a constraint on the minimum allocatable resources to a user. As an example, for a TTI of 0.25 ms (4 OFDM symbols and 4x12 = 48 REs within one PRB), the minimal resource allocation to a user varies from 1, 2, 4, and 7 PRBs depending on its SINR value (see Table D.1), when including an additional 10% of reference symbol overhead [11]. For a more exhaustive study on 5G frame numerology options we refer to [12].

4 Simulation Framework

The performance evaluation is based on system-level simulations of a multi-user cellular system. Two types of traffic are simultaneously evaluated. (i) Bursty traffic with a finite payload of B bits per user with random arrivals

Table D.1: Control channel overhead [11]

SINR [dB]	In-resource CCH overhead
$(-\infty, -2.2)$	8x36 = 288 REs
$[-2.2, 0.2)$	4x36 = 144 REs
$[0.2, 4.2)$	2x36 = 72 REs
$[4.2, \infty)$	1x36 = 36 REs

that follow a Poisson process with arrival rate λ (the offered load is $\lambda \cdot B$). We refer to this traffic type as *LLC*. And (ii) full buffer traffic from a single user per cell with infinite payload of downlink data. The latter, referred to as *MBB*, allows us to analyse the impact of different system configurations on the throughput performance. The simulation procedure follows the diagram in Fig. D.3. LLC users arriving to the system are assigned with a SINR randomly chosen from a given distribution. The SINR distribution is taken from a 3GPP regular macro cellular network with 500 m inter-site distance (ISD), where users are uniformly distributed. The SINR distribution captures the effects of distance-dependent attenuation, shadowing and full-load inter-cell interference according to [13]. This approach reproduces the different radio channel conditions depending on the location of the user in a cellular network. Explicit modelling of fast fading is not included. The transmitted data bits on a given time-frequency resource of size (t, bw) are given by,

$$N_{bits} = t \cdot bw \cdot \log_2(1 + SINR) \cdot \eta(t, bw, SINR) \quad [\text{bit}], \tag{D.2}$$

where t corresponds to the TTI duration, and bw is the bandwidth of the allocated resource composed of an integer number of PRBs. The transmission efficiency $\eta(t, bw, SINR)$ represents the relative CCH overhead of the (t, bw)-sized resource. This is calculated as the amount of REs used for the scheduling grant (given in Table D.1 for different SINR values) plus an additional 10% for reference symbols, divided by the total amount of REs in the block of (t, bw) size. LLC users are scheduled with a first-come first-served (FCFS) policy with priority over MBB traffic. Since we are mainly interested in the tradeoffs between the queuing delay and the TTI size, we assume a fixed TTI size per simulation for both MBB and LLC traffic. After the payload of B bits is delivered, the call is terminated. Frequency multiplexing of users can occur for the cases where the transmission of a certain payload occupies less than the available resources in a TTI. Table D.2 summarizes the default simulation assumptions.

Simulations are run with different offered loads and TTI durations, and relevant statistics are obtained for each type of traffic. The main performance indicators for MBB and LLC are, respectively, the downlink experienced throughput and the latency, as defined in Section 2. The processing

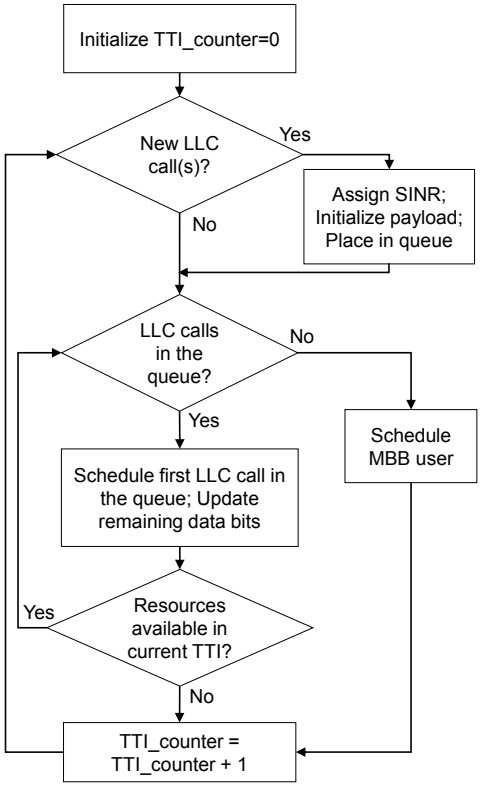


Fig. D.3: Flow diagram of simulation methodology.

delay is assumed to be constant for each call, and is therefore not included in the simulations. The simulation time corresponds to at least 100.000 calls to ensure a reasonable confidence level for the considered performance measures.

5 Results

We start by analysing the impact of different TTI sizes on the MBB throughput performance. Table D.3 summarizes the MBB user throughput for different SINR and offered loads of LLC traffic. An offered load of 4 Mbps corresponds to an average LLC resource utilization of approximately 25%. As expected, the throughput decays when the load increases. It can also be observed how the throughput is affected by the TTI size. Shorter TTIs result in larger CCH overhead and in consequence lower spectral efficiency. For example, the gain from using a 2 ms TTI over a 0.25 ms TTI is, respectively, 17% and 20% for the lowest and highest offered load and a SINR of -3 dB. At +3 dB SINR, the gain from long TTI is reduced to 4% and 8% for 4 Mbps and

Table D.2: Default simulation assumptions

Parameter	Value
SINR distribution	3GPP Macro network with 500 m ISD [13]; Full load conditions
System numerology	16 OFDM symbols per 1 ms; 17.143 kHz subcarrier spacing; 12 subcarriers per PRB [12]
System bandwidth	10 MHz; Effective transmission bandwidth of $\sim$9 MHz (44 PRBs)
TTI size	0.25 ; 0.5 ; 1 ; 2 ms
Scheduling technique	Fixed TTI size for all types of traffic; FCFS scheduling for LLC with priority over MBB
Traffic model	MBB: Single user with full buffer traffic LLC: Poisson arrival process with 1 kB payload
LLC offered load	0.4 - 12 Mbps

12 Mbps offered load, respectively. In general, the largest gains from using a long TTI size are obtained at low SINR. This is mainly due to the larger impact of the CCH overhead for users experiencing poor radio channel conditions.

Fig. D.4 presents the latency at the 50% (median) and 99% percentile for different offered loads and TTI durations. At the median, it is shown that the achievable latency is not significantly impacted by the offered load. In this case, the dominant components of the latency budget are mainly the frame alignment and transmission delay, therefore a 0.25 ms TTI provides the best performance. However, when evaluating the 99% percentile, it is observed that the achieved latency is considerably affected by the load. At low offered load, the optimal TTI size is 0.25 ms. However, as the load increases, the lowest latency is obtained with longer TTI size. Particularly, the 0.5 ms TTI provides equal or better performance for offered loads of 10 Mbps or higher.

The main reason for this behaviour is the queuing delay. As the offered load increases, the queuing delay becomes the most dominant component on the total latency, therefore it is beneficial to increase the spectral efficiency (by using a longer TTI) in order to reduce the experienced delay in the queue. This phenomenon is illustrated in Fig. D.5, where the queuing probability is plotted for different loads and TTI sizes. The queuing probability is defined as the probability that a user is not scheduled in the TTI immediately after arrival. It can be observed that, the shorter the TTI the higher the probability of experiencing queuing. Note that only a few cases experience a queuing probability higher than 50%, which reconfirms the steady behaviour of the

Table D.3: MBB throughput for different TTI sizes, SINR and LLC offered load.

SINR	TTI	4 Mbps off. load		8 Mbps off. load		12 Mbps off. load	
		Throughput	Gain[1]	Throughput	Gain[1]	Throughput	Gain[1]
[dB]	[ms]	[Mbps]	[%]	[Mbps]	[%]	[Mbps]	[%]
-3	0.25	3.29	0	2.13	0	0.98	0
	0.5	3.62	10	2.38	12	1.13	15
	1	3.77	15	2.48	17	1.22	23
	2	3.84	17	2.50	18	1.18	20
0	0.25	6.14	0	4.00	0	1.86	0
	0.5	6.45	5	4.28	7	2.13	14
	1	6.58	7	4.36	9	2.14	15
	2	6.63	8	4.34	9	2.13	15
+3	0.25	10.14	0	6.62	0	3.13	0
	0.5	10.44	3	6.92	5	3.38	8
	1	10.53	4	6.94	5	3.49	11
	2	10.55	4	6.99	6	3.38	8

[1]Gain relative to the 0.25 ms TTI configuration for the respective SINR and offered load parameters.

observed median latency performance. Fig. D.6 shows the distribution of the queuing delay for an offered load of 12 Mbps. It is observed that a 0.25 ms TTI configuration experiences the highest queuing delay (in both mean and tail of the distribution) as a consequence of the lower spectral efficiency. Configurations with 0.5 or 1 ms TTI provide lower queuing delay. At high load, the benefits of lower queuing delay exceed the drawbacks of longer transmission time and frame alignment delay, which results in overall better 99% percentile latency performance (as shown in Fig. D.4).

The tradeoff between the TTI size and the queuing delay is not only evident when increasing the load, but also when analysing the tail of the latency distribution. Fig. D.7 shows the latency distribution for offered loads of 4 Mbps and 12 Mbps. For these two cases, we have run longer simulations such that it allows us to examine with good accuracy up to the 99.99% percentile. Even at relatively low load (4 Mbps, Fig. D.7(a)), there is a gain from using a 0.5 ms TTI over a 0.25 ms TTI if the percentile of interest is above 99%. A similar trend is observed for the high load case (12 Mbps, Fig. D.7(b)). However, in this case the point at which the 0.5 ms TTI becomes better than the 0.25 ms TTI appears much earlier in the distribution. It is also observed that the 1 ms TTI configuration is the best performing solution for percentiles above 99.9%.

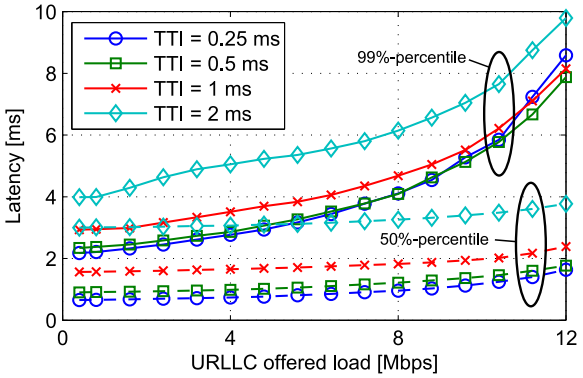


Fig. D.4: LLC latency at the 50% and 99% percentile under different load conditions and TTI sizes.

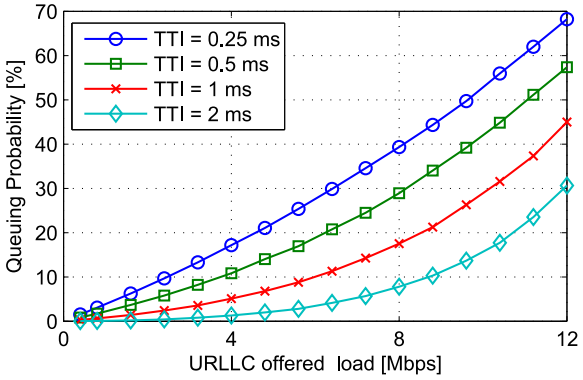


Fig. D.5: LLC queuing probability under different load conditions and TTI sizes.

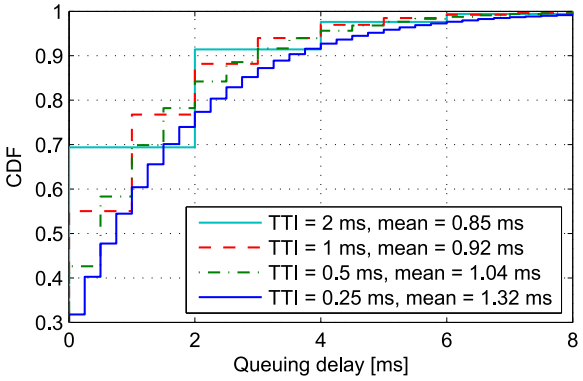


Fig. D.6: Cumulative distribution function (CDF) of LLC queuing delay at 12 Mbps LLC offered load.

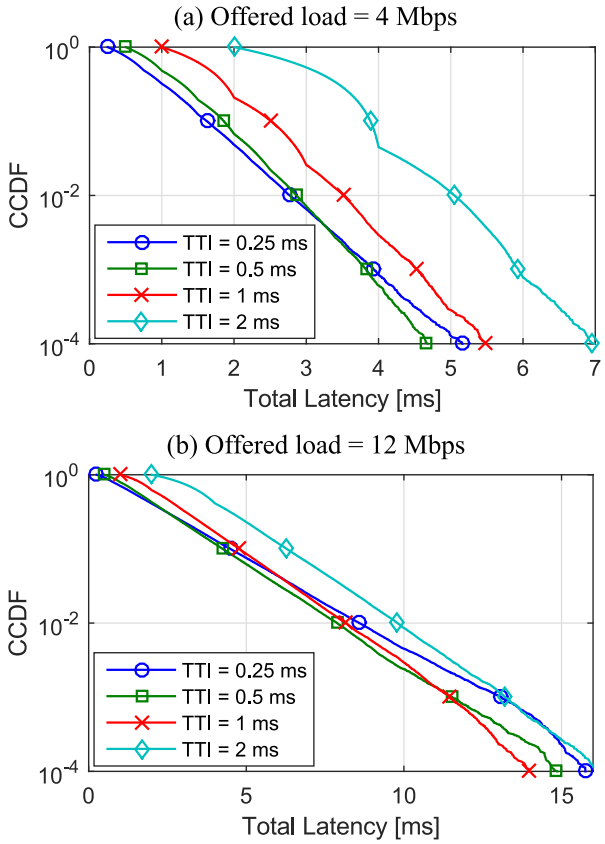


Fig. D.7: Complementary cumulative distribution function (CCDF) of the LLC latency under different load conditions and TTI sizes.

6 Discussion

The presented results show the benefits of using different TTI sizes to achieve low latency, depending on the offered load and the percentile of interest. Particularly, the tail of the latency distribution reveals the importance of using long TTI size (e.g. 0.5 or 1 ms) with higher spectral efficiency in order to reduce the experienced queuing delay. The observed trends are relevant for URLLC use cases, which require latencies of a few milliseconds guaranteed with reliability levels up to 99.999% [2]- [4]. However, the advantages of using different TTI sizes are broader. For example, the TTI duration can be adjusted in accordance to the user-specific radio channel conditions in order to compensate for the control overhead. This benefit has been shown in Table D.3, where the throughput gains of having large TTIs are more significant for users with low SINR. The TTI size can also be selected according

to the individual user's service requirements. Besides URLLC and MBB, another relevant 5G use case is low cost massive machine-type of communication (mMTC) which might only support narrow bandwidth operation and therefore will benefit from long TTIs [2]. Given these manifold benefits, it is expected that a highly flexible scheduling of users, such as illustrated in Fig. D.2, will be of key importance to efficiently support the different use cases and requirements envisioned for 5G.

7 Conclusions

In this paper we have analysed the latency performance with different TTI configurations taking into account the multi-user dynamics of a cellular network. At low offered loads, it is observed how a 0.25 ms TTI is an attractive solution to achieve low latency. The main benefits come from the low frame alignment and transmission delay required to transmit the payloads. However, as the load increases, it has been shown how longer TTI sizes, e.g. 0.5 ms or 1 ms, provide improved performance as these configurations can better cope with the non-negligible queuing delay. The presented results allow to conclude that support for scheduling with different TTI sizes is important to achieve low latency and should be included in future 5G. Our future work will include a more detailed modelling of physical and medium access layer mechanisms including link adaptation, transmission errors and the respective retransmissions. Evaluations with simultaneous use of different TTI size depending on the use case is also of interest.

References

[1] A. Osseiran et al., "Scenarios for the 5G mobile and wireless communications: the vision of the METIS project", *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26-35, May 2014.

[2] 3GPP TR 38.913 v0.3.0, "Study on scenarios and requirements for next generation access technologies", March 2016.

[3] ITU-R M.2083-0, "IMT vision - framework and overall objectives of the future development of IMT for 2020 and beyond", Sept. 2015.

[4] P. Popovski, "Ultra-reliable communication in 5G wireless systems", *International Conference on 5G for Ubiquitous Connectivity*, Nov. 2014.

[5] S. Ahmadi, "LTE-Advanced: A practical systems approach to understanding 3GPP LTE releases 10 and 11 radio access technologies", Academic Press, 2013.

[6] F. Ishizaki and G. U. Hwang, “Queuing delay analysis for packet schedulers with/without multiuser diversity over a fading channel", *IEEE Transactions on Vehicular Technology*, vol. 56, no. 5, pp. 3220-3227, Sept. 2007.

[7] I. Rubin and Z.-H. Tsai, “Message delay analysis of multiclass priority TDMA, FDMA, and discrete-time queueing systems", *IEEE Transactions on Information Theory*, vol. 35, no. 3, pp. 637-647, May 1989.

[8] G. Wunder and C. Zhou, “Queueing analysis for the OFDMA downlink: throughput regions, delay and exponential backlog bounds", *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 871-881, Feb. 2009.

[9] K. I. Pedersen, G. Berardinelli, F. Frederiksen and A. Szufarska, “A flexible 5G frame structure design for frequency-division duplex cases", *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53-59, March 2016.

[10] B. Soret, P. Mogensen, K. I. Pedersen and M. C. Aguayo-Torres, “Fundamental tradeoffs among reliability, latency and throughput in cellular networks", *IEEE Globecom*, Dec. 2014.

[11] D. Laselva et al., “On the impact of realistic control channel constraints on QoS provisioning in UTRAN LTE", *IEEE Vehicular Technology Conference*, Sept. 2009.

[12] G. Berardinelli, K. I. Pedersen, F. Frederiksen and P. Mogensen, “On the design of a radio numerology for 5G wide area", *International Conference on Wireless and Mobile Communications*, Oct. 2015.

[13] 3GPP TR 36.814 v9.0.0, “Evolved universal terrestrial radio access (E-UTRA); Further advancements for E-UTRA physical layer aspects", March 2010.

Paper E

MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks

Guillermo Pocovi, Beatriz Soret, Klaus I. Pedersen, Preben Mogensen

The paper has been published in the
IEEE International Conference on Communications Workshops (ICC), 2017.

Abstract

Ultra-reliable low-latency communications (URLLC) entail the transmission of sporadic and small packets, with low latency and very high reliability. Among many potential areas of optimization for URLLC, the problems of large delays during HARQ retransmissions, and inaccurate link adaptation as a consequence of the rapidly-varying interference conditions are studied. The former is addressed by reducing the TTI length and HARQ round-trip time, as compared to what is used in LTE; whereas including low-pass filtered interference information in the CQI report is also proved to have great potential. Extensive system-level simulations of the downlink performance show that the URLLC requirements, i.e. latencies below 1 ms and 99.999% reliability, are achievable at low load scenarios, whereas some performance degradation (1 - 3 ms latency) is experienced at higher loads due to the increased queuing delay and inter-cell interference.

1 Introduction

Standardization activities towards a fifth generation (5G) New Radio (NR) are gaining big momentum in the 3rd Generation Partnership Project (3GPP) [1]. Ultra-reliable low-latency communication (URLLC) has been agreed as one of the three main use cases, targeting transmission of relatively small payloads with very low latency (<1 ms) and high reliability (99.999%) [2].

The very challenging requirements of URLLC have raised significant attention in academia and industry. For example, the studies in [3], [4] analyse the required improvements on the wireless link. It is shown how a combination of micro- and macroscopic diversity can achieve the signal quality outage probability required for URLLC. In [5], different deployment strategies (e.g. number of cells, frequency reuse pattern) are studied to meet the coverage requirements in a factory automation scenario, while the complementary system-level simulation results are presented in [6]. To achieve low over-the-air transmission delay, reducing the transmission time interval (TTI) is of significant importance [7]. Related to the former, the study in [8] evaluates the downlink latency performance under different TTI durations and load conditions, assuming a mixture of high-priority bursty traffic and best-effort mobile broadband (MBB) traffic. In [9], a link adaptation strategy is presented where the modulation and coding scheme (MCS) is selected according to the target reliability and feedback channel imperfections, assuming that one hybrid automatic repeat request (HARQ) retransmission is allowed.

The applicability of these medium access control (MAC) layer schemes on URLLC is mainly limited by two factors: (i) the relatively large delay that characterizes the HARQ operation, and (ii) inaccurate link adaptation due to the very sporadic URLLC traffic. Motivated by this, we present dif-

ferent enhancements for supporting URLLC in cellular networks. Among others, decreasing the TTI length and HARQ round-trip time (RTT), as compared to LTE, is suggested as a resource-efficient way to improve the latency performance; whereas including time-filtered interference information in the channel quality indicator (CQI) report is proposed to improve the link adaptation accuracy. The benefit of the presented enhancements is evaluated by analysing the downlink latency and reliability performance in a multi-user multi-cell scenario. As it will be shown, the gain provided by each solution depends on the offered traffic load, the inter-cell interference, etc. Given the complexity of the considered problem, the chosen evaluation methodology is dynamic system-level simulations, following the latest URLLC modelling assumptions agreed in 3GPP [10]. Good practice is applied in order to generate trustworthy results.

The rest of the paper is organized as follows: Section 2 outlines the considered network and traffic model, and the performance metrics. Section 3 presents the proposed URLLC enhancements. The simulation assumptions are presented in Section 4. Performance results are shown in Section 5, followed by concluding remarks in Section 6.

2 Network Model and Performance Metrics

System-level evaluations are a powerful tool to analyse the overall behaviour and performance of cellular systems. Particularly, the 3GPP has highlighted the importance of carrying out highly detailed system-level evaluations in order to analyse the impact of time-varying inter-cell interference, and queuing and scheduling delays, on the URLLC performance [10].

2.1 Network Layout & Traffic Model

We follow the modelling and assumptions recently discussed in [10], [11]. A macro-cellular network composed of 7 three-sector sites with 500 meter inter-site distance is assumed. A fixed number of URLLC user equipments (UEs) are uniformly distributed across the network. Unidirectional downlink traffic following the so-called FTP Model 3 (FTP3) is applied. This consists of relatively small packets (typically between 32 and 200 Bytes) that are generated for each UE in the downlink direction following a Poisson arrival process [10].

2.2 Performance Metrics

In line with [1], the key performance indicator (KPI) is the one-way downlink latency that can be achieved with a $1 - 10^{-5}$ probability. The latency is measured from the moment a FTP3 packet arrives at the base station until it is successfully received at the UE. This accounts for the queuing delay in the

cell, defined as the time elapsed between the arrival of the packet at the base station buffers and the execution of the scheduling decision; frame alignment, i.e. time remaining to the beginning of the next TTI; and transmission delay, including the potential HARQ retransmissions that could occur.

3 URLLC Enablers

3.1 Low Latency Frame Structure

We adopt the candidate frequency-division duplex frame structure presented in [12]. It consists of a grid of time-frequency resources where users are dynamically multiplexed via orthogonal frequency division multiple access (OFDMA). The time domain is organized into subframes, each containing a set of physical resource blocks (PRB) in the frequency domain. The physical layer numerology follows the recent agreements in 3GPP: 15 kHz sub-carrier spacing (SCS), 14 OFDM symbols (1 ms) subframe, and a PRB size of 12 sub-carriers (180 kHz) as the baseline configuration; although options with 2^N scaling of the SCS, e.g. 30 kHz or 60 kHz, are also allowed [11]. The 3GPP has also agreed on using different TTI durations depending on the user-specific requirements. Apart from scheduling with a 1 ms subframe resolution, smaller scheduling units composed of e.g. 7 OFDM symbols (0.5 ms 'slot') or 2 OFDM symbols (0.143 ms 'mini-slot'), are also considered for 5G [13].

In line with [12], the control channel (CCH) is accommodated within the resources assigned to each user (i.e. in-resource CCH). The CCH contains the scheduling grant indicating the specific time-frequency resource allocation for each user, among other relevant link adaptation parameters required to decode the data. The coding rate of the user-specific CCH is dynamically adapted to the user's channel conditions, following the link-level performance specified in [14]. Note that although scheduling with a mini-slot provides the lowest latency, it has a cost in terms of higher signalling overhead due to the need of sending more frequently CCH information [12].

3.2 Short HARQ RTT

The HARQ RTT is assumed to scale linearly with the TTI length. Assuming a LTE-alike asynchronous HARQ operation with a minimum RTT of 8 TTIs, even with a TTI duration of 0.143 ms, the HARQ RTT would not satisfy the 1 ms URLLC latency target (i.e. $8 \cdot 0.143 \text{ ms} > 1 \text{ ms}$). Since relying on a single (very conservative) transmission would have a significant cost on the spectral efficiency [9], we study the case where the HARQ RTT is reduced to 4 TTIs in order to allow room for one HARQ retransmission. A diagram of the HARQ procedure with reduced RTT is presented in Fig. E.1. We consider

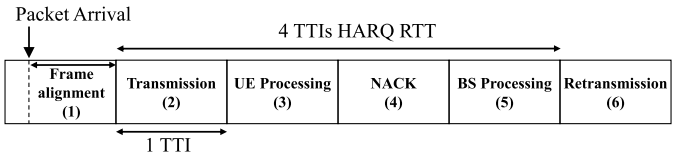


Fig. E.1: Diagram of HARQ operation with 4 TTIs RTT.

a maximum of 1 TTI for frame alignment and for performing the scheduling decision (1). The UE processing time (3) required to decode the initial transmission, and the base station processing (5) of the negative acknowledgement (NACK) are also reduced to 1 TTI[1]. Under such conditions, the maximum latency assuming one HARQ retransmission is reduced to $6 \cdot 0.143$ ms $= 0.86$ ms (excluding the queuing delay). Note that the proposed reduction of the HARQ RTT mainly requires an improvement of the processing capabilities at the UE and base station. A complementary way to relax the processing requirements is by applying the so-called early-feedback techniques [16], which try to predict the outcome of the decoder before the entire decoding finishes.

3.3 Accurate Link Adaptation

The traffic characteristics of URLLC represent a challenge when attempting to perform accurate link adaptation. Due to the relatively small payloads, a URLLC transmission generally occupies a subset of the available radio resources (i.e. PRBs) within a TTI. This fact, together with the sporadic arrival of packets (as specified in Section 2.1), result in a rapidly changing interference pattern. As a result, it becomes difficult to accurately select an appropriate MCS that fulfils a certain block error rate (BLER) constraint. This problem is also well-known from LTE system-level performance analyses in non-fully loaded networks. In such cases, the MCS selection is typically improved by use of outer loop link adaptation (OLLA) mechanisms [17], which "fine tune" the MCS selection according to the received HARQ ACK/NACK feedback messages. These mechanisms are, however, characterized by slow convergence which limits their applicability to URLLC use cases [17].

Our proposal is to modify the UE measurement procedure of the CQI report, by including historical information of the experienced interference. On each TTI n, the UE measures the interference with a certain PRB resolution (a.k.a. sub-band). The interference measurement on the i-th sub-band, $x_i[n]$, is filtered with a low-pass first-order Infinite Impulse Response (IIR) filter,

[1]The processing time would require further reduction if considering uplink-downlink frame misalignment due to the timing advance [15].

resulting in the following smoothed value:

$$s_i[n] = \alpha \cdot x_i[n] + (1 - \alpha) \cdot s_i[n-1] \,, \tag{E.1}$$

where α is the forgetting factor (FF) of the filter ($0 < \alpha < 1$). The CQI, which is periodically reported to the base station, contains the low-pass filtered interference information together with the latest desired-signal fading information. Note that the latter varies in a much lower time scale and, except for very high UE speeds, it is possible to track the channel variations with relatively high accuracy [15]. The FF α determines how much weight is given to the latest measurement as compared to the previous ones. Based on a heuristic analysis using simulations, it has been found that a FF $\alpha = 0.01$ is beneficial for the latency performance.

3.4 BLER Optimization

As presented in [8], there is a tradeoff between spectral efficiency and queuing delay. That is, as the system load increases, it is beneficial from a latency point of view to configure the system for high spectral efficiency (rather than for low latency) in order to cope with the non-negligible queuing delay. This can be achieved by adjusting the BLER target for the link adaptation depending on the system load. At low load, conservative transmissions (e.g. 0.1% BLER target) provide the best latency performance, whereas more aggressive transmissions can be allowed at high load, since the reduction of the queuing and scheduling delay compensates for the larger delay in the air interface (due to the occurrence of a larger amount of retransmissions).

4 Simulation Assumptions

The performance evaluation is based on system-level simulations of a multi-user cellular system. The default simulation assumptions are summarized in Table E.1. The simulator time-resolution is one OFDM symbol, and it includes explicit modelling of the majority of radio resource management functionalities such as packet scheduling and HARQ. Dynamic link adaptation is applied for both data and the in-resource control channel, which results in varying control overhead depending on the user signal quality and TTI duration (see [8], [12]). Additional overhead from reference signals (RS) is also included. The link adaptation for the data transmissions is based on the periodical frequency-selective CQI report from the URLLC users, using standard OLLA to reach a certain BLER target (0.1% as default). Closed-loop 2x2 single-user single-stream MIMO is assumed for each link and the UE receiver type is minimum mean square error with interference rejection combining (MMSE-IRC).

Table E.1: Simulation assumptions

Parameter	Value
Network environment	3GPP Urban Macro (UMa) network with 21 cells and 500 meter inter-site distance [10]
Carrier configuration	10 MHz carrier bandwidth at 2 GHz
PHY numerology	TTI sizes of 0.143 ms, 0.5 ms, and 1 ms; Other numerology settings in line with LTE
Control channel	In-resource control channel scheduling grants with dynamic link adaptation [12]
Data channel MCS	QPSK to 64QAM, with same coding rates as in LTE; 0.1-10% BLER target for first transmissions
RS overhead	4 resource elements per PRB
Antenna configuration	2 x 2 single-user single-stream MIMO with MMSE-IRC receiver
Packet scheduler	Proportional Fair
CSI	LTE-alike CQI and PMI, reported every 5 ms; Interference filtering is applied with FF α
HARQ	Async. HARQ with Chase combining; Max. 6 HARQ retransmissions with 4-8 TTI RTT
RLC	RLC Unacknowledged mode
UE distribution	210 UEs uniformly distributed in the network; 20% indoor, 80% outdoors; 3 km/h UE speed
Traffic model	FTP Model 3 downlink traffic with 200 Bytes payload size
Offered load	1 - 8 Mbps average load per cell

The network layout, UE distribution and traffic follow the description presented in Section 2.1. A set of $N = 210$ URLLC UEs are uniformly distributed in the network, which corresponds to an average of 10 UEs per cell. For each UE, a payload of size $B = 200$ Bytes is generated in the downlink direction following a Poisson distribution with arrival rate λ. The network offered load corresponds therefore to $N \cdot B \cdot \lambda$.

The latency (defined in Section 2.2) of each downloaded FTP3 packet is collected and used to form empirical complementary cumulative distribution functions (CCDF). In line with [1], the main KPI is the achievable latency at the 10^{-5} percentile. The latency is analysed both globally and on a per-user basis. The simulation time corresponds to at least 5.000.000 successfully received packets to ensure a reasonable confidence level for the considered performance metrics.

5 Performance Analysis

Next, we present performance results in order to highlight the benefit obtained from the proposed enhancements.

Impact of the TTI size

Fig. E.2 shows the CCDF of the URLLC latency with TTI sizes corresponding to 2, 7, and 14 OFDM symbols, and an average offered load of 1 Mbps per cell. The system utilization for each configuration, i.e. percentage of PRBs transmitted on average, is shown in the legend. At the 10^{-5} percentile, there is a significant benefit of using short TTI size as it reduces the over-the-air transmission delay, frame alignment, and HARQ retransmission time. The latter particularly impacts the tail of the distribution, since one retransmission typically occurs at the 10^{-5} outage level. This results in a latency of 10 ms for the 1 ms TTI, whereas it is only 3.3 ms for the 0.14 ms TTI. The performance gain of the 0.14 ms TTI as compared to the 1 ms TTI is much smaller (3x shorter latency) than the expected (7x). This is due to the larger control overhead when scheduling with short TTIs, which results in a lower availability of radio resources for URLLC data transmissions. Note that the case with 0.5 ms TTI experience the lowest PRB utilization. Although the signalling overhead is larger as compared to the 1.0 ms TTI case, the 0.5 ms TTI have the advantage of higher resource granularity, which results in more resource-efficient scheduling of the small URLLC data payloads.

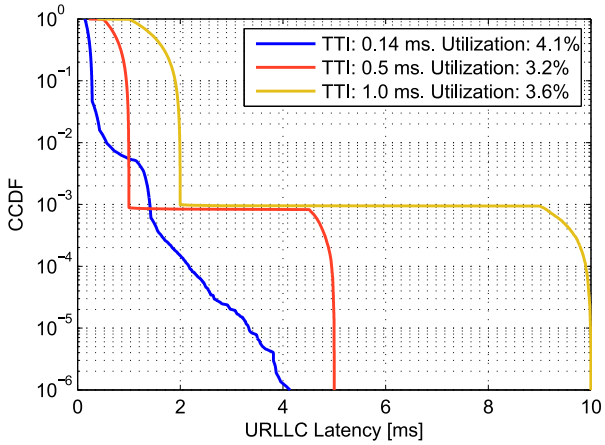


Fig. E.2: Latency distribution with different TTI lengths and 1 Mbps offered load. HARQ RTT: 8 TTIs. BLER: 0.1%. $\alpha = 1$.

Interference filtering

Fig. E.3 shows the latency distribution at different offered loads, and including the benefits of interference filtering (explained in Section 3.3). Only cases with a TTI duration of 0.14 ms are presented, given its large latency-reduction potential. Comparing the 1 Mbps performance in Fig. E.3 with the one shown in Fig. E.2, lower resource utilization and a 1.9 ms latency reduction (from 3.3 ms to 1.4 ms) is obtained at the 10^{-5} percentile, which can be attributed to the more accurate link adaptation. Although not shown, the benefits of interference filtering become even greater at higher offered load, e.g. 2 or 4 Mbps per cell, when the cell activity is higher and more sporadic interference is experienced in the network. For loads up to 4 Mbps, good latency performance (below 2 ms) is obtained despite the relatively high system utilization. As we further increase the load to 6 Mbps, non-negligible queuing starts to occur at the cells' buffers, which deteriorates considerably the achievable latency.

HARQ RTT and BLER optimization

Fig. E.4 show the performance with different configurations of the HARQ RTT and the BLER target, as discussed in Section 3. At 1 Mbps load, the HARQ-related processing delay is the dominant component of the total latency. Hence, by reducing the HARQ RTT to 4 TTIs, the 1 ms latency and 99.999% reliability required for URLLC is fulfilled. The case with 4 Mbps offered load and 0.1% BLER target does not experience significant improvement when reducing the HARQ RTT, since the queuing delay is the dominant component. Instead, it is beneficial to operate at higher BLER target (1%) in

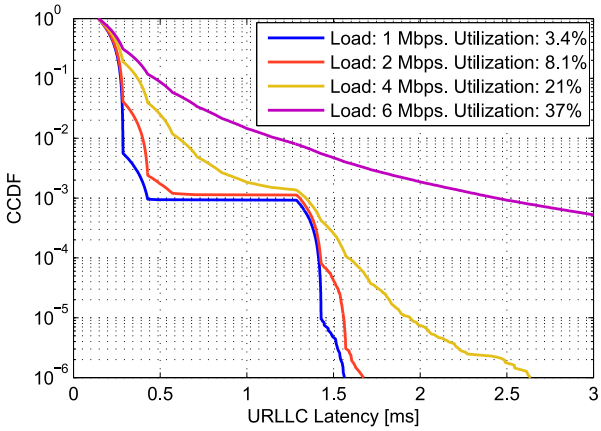


Fig. E.3: Latency distribution at different offered loads. TTI length: 0.14 ms. HARQ RTT: 8 TTIs. BLER: 0.1%. $\alpha = 0.01$.

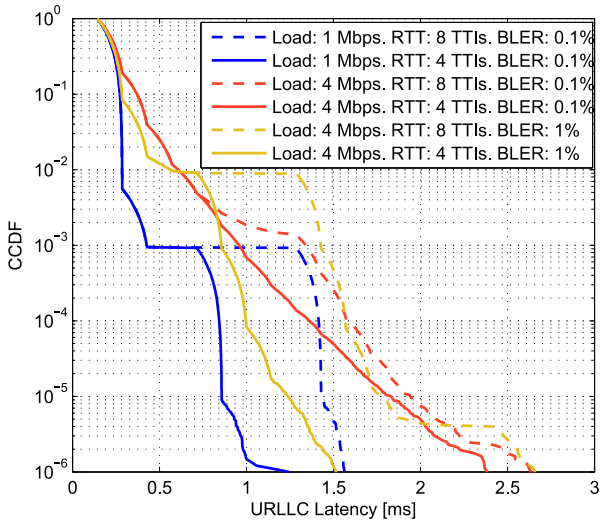


Fig. E.4: Latency distribution with different configurations of the HARQ RTT and BLER target. TTI length: 0.14 ms. $\alpha = 0.01$.

order to increase the spectral efficiency of the system and reduce the queue length. It can be observed that, after reducing the experienced queuing delay, the benefit of short HARQ RTT becomes more evident.

Global performance summary

Fig. E.5 shows the 10^{-5}-percentile latency performance at different loads when applying the proposed enhancements. The top of each bar indicates the achieved latency without any of the proposed enhancements, i.e. a fixed BLER target of 0.1% for all loads, $\alpha = 1$, and a HARQ RTT of 8 TTIs. The latency varies from 3.3 ms, at 1 Mbps offered load, to ∼100 ms for a offered load of 8 Mbps. The bottom of each bar indicates the optimized latency performance, and each segment represents the latency improvement provided by a certain enhancement. It is observed that the URLLC requirements are fulfilled for loads up to 2 Mbps (latency of 0.98 ms), but cases with 4 or 6 Mbps offered load also experience decent performance (1.24 ms and 1.63 ms, respectively). The relative gain of the proposed latency and reliability improvements depend on the system load. At low load, the processing time is one of the dominant components of the achievable latency; therefore, significant benefit is obtained by reducing the HARQ RTT. As we increase the load, the cell activity starts to increase resulting in rapidly-varying interference conditions. As a consequence, interference filtering provides large gains, as low-pass information of the experienced interference is implicitly included in the CQI report. On top of this, there is also relevant gain from using a higher

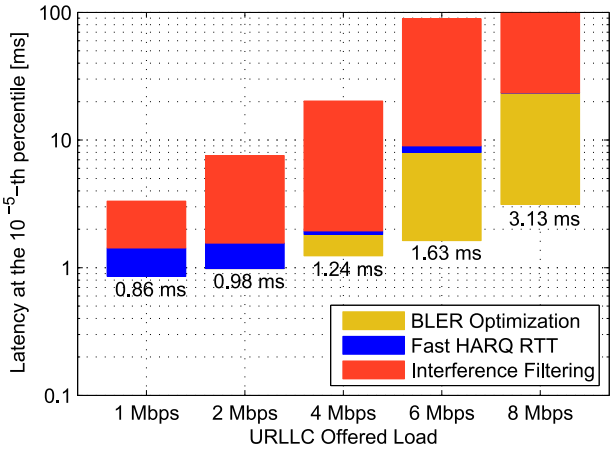


Fig. E.5: Achievable latency at the 10^{-5} percentile under different offered loads. The performance improvement provided by the evaluated techniques is also presented. TTI length: 0.14 ms. $\alpha = 0.01$.

BLER target at high load in order to reduce the queuing delay experienced in the cells (see Fig. E.4). Specifically, the performance at 4, 6 and 8 Mbps offered load is improved by increasing the BLER target from 0.1% to 1%, 5% and 10%, respectively[2].

Per-User performance analysis

The results in Fig. E.2-E.5 show the latency statistics from all the UEs in the network. However, due to the different coverage conditions of the UEs, it is likely to happen that not all the UEs experience the same performance. In order to quantify these effects, we analyse the latency performance for each UE, and determine the ratio of UEs which do not satisfy a certain latency requirement (i.e. outage probability). The latency is analysed at the 10^{-4} percentile, since the amount of samples available per UE is lower as compared to the global analyses. Fig. E.6 shows the UE outage for different offered loads. The HARQ RTT is set to 4 TTIs, whereas the BLER target is configured for each load according to what provides the best latency performance. It is observed that a 1 ms latency with 99.99% reliability can be achieved by all the simulated UEs for loads up to 2 Mbps. As we increase the load, the outage probability drastically increases, e.g. 60% outage probability at 6 Mbps offered load. For a more relaxed latency constraint of 2 ms, the outage is significantly reduced, being no larger than 20% in any of the evaluated load conditions.

Intuitively, there is some correlation between the UE latency performance

[2]The optimal BLER target for each load is obtained heuristically using simulations.

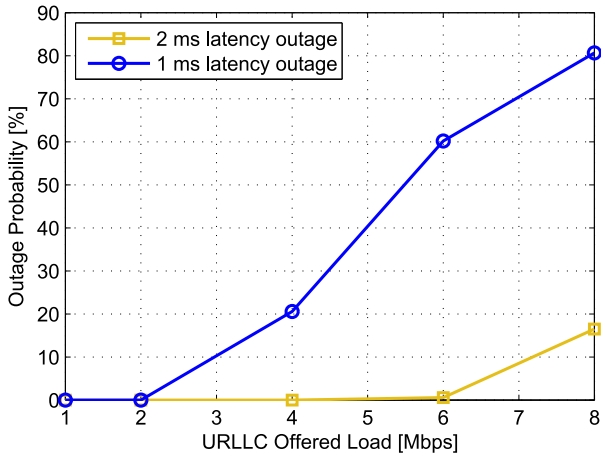


Fig. E.6: Percentage of UEs not satisfying a certain latency requirement at the 10^{-4} percentile. TTI length: 0.14 ms. $\alpha = 0.01$.

and its coverage conditions. Fig. E.7 shows a scatter plot of the per-user latency performance at the 10^{-4}-percentile versus its Geometry factor (Γ). The Geometry factor is equivalent to the average signal to interference-and-noise ratio (SINR) experienced by the UE under full-load conditions; i.e. $\Gamma = P_i/(\sum_{n \neq i} P_n + N_0)$, where P_n is the received signal power from cell n, i denotes the serving cell, and N_0 is the thermal noise power. As expected, users located in cell-edge areas (low Γ) generally experience worse performance as compared to cell-center UEs. At 1 Mbps, the per-user latency performance can be grouped into two groups: (i) users with very good channel quality which do not experience any retransmissions in the evaluated percentile, and therefore achieve a latency of only $\sim$0.4 ms; and (ii) users in less favourable channel conditions, experiencing typically one HARQ retransmission at the 10^{-4} level (resulting in a latency of $\sim$0.86 ms, as discussed in Section 3-B). At 8 Mbps load, the achievable latency is largely affected by the queuing delay. The queuing delay is not constant but naturally varies in accordance to the instantaneous variations of the incoming traffic. This results in a latency performance that varies from user to user but clearly correlated with the UE-specific experienced channel quality. The presented per-user performance statistics justify recent discussions in standardization proposing to perform admission control to discard UEs that are highly likely to not fulfil the required quality of service (due to the UE coverage conditions, system load, etc.) [10].

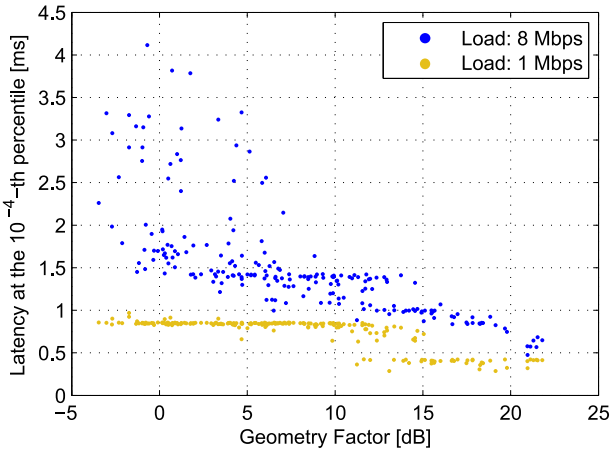


Fig. E.7: Scatter plot of the per-user latency performance at the 10^{-4} percentile versus the Geometry Factor. TTI length: 0.14 ms. $\alpha = 0.01$.

6 Conclusions

Motivated by the traffic characteristics and stringent requirements of URLLC, we have presented different MAC layer enhancements for supporting URLLC in cellular networks. It is shown that the link adaptation imperfections, as a consequence of the very sporadic traffic, can be reduced by including low-pass filtered interference information in the CQI report, whereas short TTI and faster processing at the UE and base station is of significant importance to reduce the delay during HARQ retransmissions. Extensive system-level simulations have been carried out in order to evaluate the benefit of the proposed solutions. It has been shown how latencies below 1 ms with the required 99.999% reliability are achieved at low load scenarios, whereas some performance degradation (1 - 3 ms latency) is experienced at higher loads as a consequence of the higher queuing delay and inter-cell interference. The latency performance has also been evaluated on a per-user basis. The percentage of users not satisfying the 1 ms latency requirement drastically increases with the load, e.g. 20% and 60% for 4 and 6 Mbps offered load, respectively. In this regard, strong correlation is observed between the user latency performance and its experienced channel quality.

Our current work focuses on further enhancements to the link adaptation and scheduling mechanisms, as well as cases with mixed traffic classes, e.g. URLLC and MBB.

Acknowledgment

Part of this work has been performed in the framework of the Horizon 2020 project FANTASTIC-5G (ICT-671660) receiving funds from the European Union. The authors would like to acknowledge the contributions of their colleagues in the project, although the views expressed in this contribution are those of the authors and do not necessarily represent the project.

References

[1] 3GPP TR 38.913 v14.1.0, “Study on scenarios and requirements for next generation access technologies", January 2017.

[2] P. Popovski, “Ultra-reliable communication in 5G wireless systems", *International Conference on 5G for Ubiquitous Connectivity*, Nov. 2014.

[3] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen and P. Mogensen, “Signal quality outage analysis for ultra-reliable communications in cellular networks", *IEEE Globecom Workshops*, Dec. 2015.

[4] F. Kirsten, D. Ohmann, M. Simsek and G. P. Fettweis, “On the utility of macro- and microdiversity for achieving high availability in wireless networks", *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Sept. 2015.

[5] N. Brahmi, et al., “Deployment strategies for ultra-reliable and low-latency communication in factory automation", *IEEE Globecom Workshops*, Dec. 2015.

[6] S. A. Ashraf, I. Aktas, E. Eriksson, K. W. Helmersson and J. Ansari, “Ultra-reliable and low-latency communication for wireless factory automation: From LTE to 5G", *IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Sept. 2016.

[7] 3GPP TR 36.881 v14.0.0, “Study on latency reduction techniques for LTE", Jun. 2016.

[8] G. Pocovi, K. I. Pedersen, B. Soret, Mads Lauridsen and P. Mogensen, “On the impact of multi-user traffic dynamics on low latency communications", *International Symposium on Wireless Communication Systems (ISWCS)*, Sept. 2016.

[9] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji and R. Jäntti, “Link adaptation design for ultra-reliable communications", *IEEE International Conference on Communications*, May 2016.

[10] R1-1609545, “Remaining details of URLLC system level evaluation assumptions", 3GPP TSG-RAN WG1 #86bis, Oct. 2016.

[11] 3GPP TR 38.802 v2.0.0, “Study on new radio access technology physical layer aspects", March 2017.

[12] K. I. Pedersen, G. Berardinelli, F. Frederiksen and A. Szufarska, “A flexible 5G frame structure design for frequency-division duplex cases", *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53-59, March 2016.

[13] R1-1609664, “Comparison of slot and mini-slot based approaches for URLLC", 3GPP TSG-RAN WG1 #86bis, Oct. 2016.

[14] D. Laselva et al., “On the impact of realistic control channel constraints on QoS provisioning in UTRAN LTE", *IEEE Vehicular Technology Conference*, Sept. 2009.

[15] H. Holma and A. Toskala, “LTE Advanced: 3GPP Solution for IMT-Advanced", John Wiley & Sons Ltd, 2011.

[16] G. Berardinelli, S. Khosravirad, K. I. Pedersen, F. Frederiksen and P. Mogensen, “Enabling early HARQ feedback in 5G networks", *IEEE Vehicular Technology Conference*, May 2016.

[17] A. Duran, M. Toril, F. Ruiz and A. Mendo, “Self-Optimization algorithm for outer loop link adaptation in LTE", *IEEE Communications Letters*, vol. 19, no. 11, pp. 2005-2008, Nov. 2015.

Part IV

Dynamic Multiplexing of URLLC and eMBB

Dynamic Multiplexing of URLLC and eMBB

Motivated by the ambitious multi-service capabilities of 5G, this part presents solutions for multiplexing URLLC and eMBB traffic on a downlink shared channel.

1 Problem Description

The previous part showed that the strict URLLC requirements are only fulfilled under certain URLLC offered load conditions. From a cost- and resource efficiency perspective, one promising setup consists on multiplexing the sporadically-arriving URLLC traffic with other service classes, e.g. eMBB. Given the challenge of simultaneously serving a mixture of users with very different QoS requirements [1], numerous decisions have been made in the 3GPP aiming at a highly-flexible physical layer design that provides novel possibilities for multi-user scheduling [2], [3]. Perhaps, the most relevant agreement is the support for different time-domain scheduling resolutions, ranging from 1-3 OFDM symbol *mini-slots*, to one 7-symbol *slot* or aggregation of multiple slots. Naturally, this flexibility offers valuable options for the problem that we try to address, but it also represents a non-trivial challenge of how to utilize those degrees of freedom in an efficient manner.

In this part, solutions for efficient multiplexing of URLLC and eMBB on a downlink shared channel are proposed. Due to the sporadic and unpredictable characteristics of URLLC traffic, the focus is on fully-dynamic scheduling approaches, where the URLLC traffic can be immediately served without the need of pre-reserving radio resources for such traffic. In this respect, the following two resource allocation options are considered [4]: i) *short-TTI based*, where the base station allocates radio resources to both eMBB and URLLC UEs on a mini-slot resolution; and ii) *puncturing based* (also known as preemptive scheduler), where eMBB users can be served with longer TTIs (e.g. 1 ms), which may be partly overwritten or punctured by the urgent URLLC traffic.

Each approach has its pros and cons. For instance, the short-TTI based scheme is simple in the sense that the base station can decide on a TTI basis how to best distribute the resources among the URLLC and eMBB users; being in principle possible to apply existing QoS-aware scheduling algorithms as studied for LTE. This comes at the cost of reduced spectral efficiency from the increased control channel (CCH) overhead, as compared to traditional eMBB operation with 1 ms TTI size. In contrast, the puncturing based approach allows to schedule eMBB users with lower CCH overhead; however, additional mechanisms may be required to make the eMBB users aware of the puncturing event, and to efficiently recover the punctured data.

2 Objectives

The goals of this part of the thesis are the following:

- Identify the challenges of short-TTI based and puncturing based multiplexing of URLLC and eMBB traffic.
- Propose solutions for addressing the identified problems.
- Determine the URLLC capacity of a cellular system, as well as its sensitivity to the URLLC payload size, link adaptation and resource allocation scheme.
- Evaluate the cost in terms of eMBB throughput of satisfying the stringent URLLC requirements, and determine which resource allocation approach (short-TTI based and puncturing based) offers the best eMBB throughput performance.

3 Included Articles

The main findings of this part are included in the following articles:

Paper F. Radio Resource Management for 5G Ultra-Reliable Low-Latency Communications

This article studies the short-TTI based approach for multiplexing URLLC and eMBB traffic on a shared channel. A similar system model as in Paper E is adopted, with the following enhancements: the inclusion of eMBB traffic in the form of 5 additional UEs per cell with full-buffer downlink traffic, a larger carrier bandwidth of 20 MHz, and including cases with smaller URLLC payload sizes of 32 and 50 Bytes. A joint link adaptation and resource allocation technique is proposed, which allows to serve URLLC users in accordance

with their QoS requirements, while still providing acceptable eMBB throughput performance. In addition, the proposed technique provides dynamic adjustment of the block error probability (BLEP) of URLLC transmissions in accordance with the instantaneous load experienced per cell; hence, reducing the need for tedious BLEP adjustment as conducted in Paper E. The benefits of the proposed enhancements are analysed via extensive system-level simulations. In addition, a sensitivity analysis is conducted in order to determine the URLLC performance for different URLLC payload sizes, offered loads, and resource allocation settings.

Paper G. Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband

In this article, eMBB users are scheduled with a 1 ms TTI size, and the sporadically-arriving URLLC traffic is immediately scheduled with a 0.143 ms TTI duration partly overwriting ongoing eMBB transmissions. Small URLLC payloads of 50 Bytes, and 10 MHz bandwidth are assumed in this part of the study. In order to reduce the negative impact when puncturing the eMBB users, a recovery mechanism is presented such that the victim UEs (i.e. those whose part of the transmission has been punctured) are informed of the punctured resources, and disregard the damaged part when decoding the data or performing the HARQ Chase combining. In addition, the paper proposes and evaluates different approaches for deciding which of the radio resources currently used for eMBB transmissions should be punctured. These include *eMBB-unaware* schemes where the resources allocated to URLLC traffic are solely selected based on the channel quality experienced by the URLLC UEs, and *eMBB-aware* approaches which favour puncturing on eMBB users served with either high or low modulation and coding scheme.

Paper H. Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network Implementations

The article describes the novel QoS architecture in 5G, which facilitates ensuring the ambitious end-to-end QoS performance targets. Furthermore, it explains the rationale behind the most recent QoS-related agreements in 3GPP, and gives an overview of the broad set of options for multi-user scheduling in 5G. A more efficient retransmission mechanism for punctured allocations is proposed, where only the damaged part of the initial eMBB transmissions that have been subject to puncturing is retransmitted. Performance results showing the benefits of this technique are presented.

4 Main Findings

URLLC and eMBB performance with short-TTI based scheduling

As shown in Paper F, the introduction of eMBB traffic significantly increases the load and the inter-cell interference experienced in the network. As a consequence, poorer signal quality and larger queuing and transmission delay is experienced for URLLC users, which result in worse performance as compared to the URLLC-only scenario studied in Paper E. For instance, cases with a 200 Byte URLLC payload and conventional resource allocation schemes with fixed BLEP target do not fulfil the URLLC latency and reliability requirements even at low URLLC offered loads.

Due to the tradeoffs between queuing delay and spectral efficiency, the proposed resource allocation technique with dynamic BLEP adjustment provides significant latency improvements. In fact, it allows to fulfil the URLLC requirements for offered loads up to 2 Mbps, although the latency performance at higher loads is still decent (1.4 ms at 8 Mbps load). In addition, the proposed solution is robust and flexible, as it brings valuable latency improvement for a wide range of offered load conditions without requiring fine adjustment of the BLEP target.

The eMBB throughput performance results follow the expected trends: lower throughput as the URLLC offered load increases. It is also observed that settings with the proposed resource allocation technique experience only 3%-10% additional throughput degradation as compared to conventional scheduling schemes. In other words, fulfilling stringent latency and reliability requirements comes at an acceptable cost of reduced spectral efficiency for eMBB users.

Sensitivity to the URLLC payload

Paper F shows that the payload size has a significant impact on the URLLC performance. Particularly, cases with 50 and 32 Byte payload fulfil the URLLC requirements for offered loads as high as 8 Mbps. That is, 4 times higher load as compared to the case with 200 byte payload previously discussed. The improved performance is a consequence of the lower amount of frequency-domain resources (i.e. PRBs) required to transmit smaller URLLC payloads, which increases the probability of scheduling the entire payload immediately after arrival. Due to the larger relative CCH overhead when transmitting smaller URLLC payloads, another relevant finding is the importance of having a fully-flexible and scalable control channel in order to avoid CCH blocking problems, i.e. when radio resources are left unused due to the limited CCH capacity.

URLLC and eMBB performance with puncturing

Papers G and H show that the puncturing based approach fulfils the objective of immediately scheduling the URLLC payloads, without waiting for the eMBB transmissions to be completed. In fact, the puncturing based approach provides similar URLLC latency and reliability performance than the obtained with the short-TTI based approach (for equal assumptions of offered load, payload size, and carrier bandwidth).

The most interesting findings are related to the eMBB performance. Paper G shows that making the eMBB UEs aware of the puncturing provides valuable improvement of the decoding probability of punctured transmissions, e.g. from 21% to 26% for cases where 1/7 of the resource elements are punctured. Moreover, 5% additional throughput gain is obtained if the scheduler prioritizes puncturing eMBB allocations transmitted with low MCS, as these can better tolerate puncturing. Performance-wise, the most valuable recovery mechanism is introduced in Paper H, where only the damaged part of punctured eMBB allocations is retransmitted. By reducing the HARQ retransmission size of punctured transmissions, more resources are available for data transmissions to other users, which results in up to 20% average cell throughput improvement.

eMBB performance summary

As mentioned, the short-TTI based and puncturing based approaches provide similar URLLC latency and reliability performance. In contrast, the eMBB performance strongly depends on the considered scheduling approach, and the offered load of URLLC traffic. To facilitate the comparison, Fig. IV.1 shows the 5th and 50th percentile (median) of the eMBB throughput for different URLLC offered loads and scheduling configurations. It assumes similar settings as in Paper G: URLLC payload of 50 Bytes, and 10 MHz bandwidth. The figure compares the short-TTI based approach, with three different configurations of the puncturing based scheme: i) *Baseline*, where the eMBB UEs are informed of the puncturing; ii) *Baseline + eMBB awareness*, where the scheduler favours puncturing on eMBB users with low MCS index; and iii) *Baseline + PartialReTx*, where only the damaged part of punctured eMBB allocations is retransmitted.

For low load of URLLC traffic (0.1 Mbps), Fig. IV.1 shows that puncturing based schemes offer the best throughput performance (8-12% gain over the short-TTI based approach). The gain is a consequence of scheduling eMBB users with long TTI size and low CCH overhead. However, as the URLLC load increases, the short-TTI based approach can more efficiently accommodate the increased URLLC traffic; hence outperforming puncturing based resource allocation with full-size retransmissions. The overall best configuration is puncturing based with partial retransmissions, as it provides a

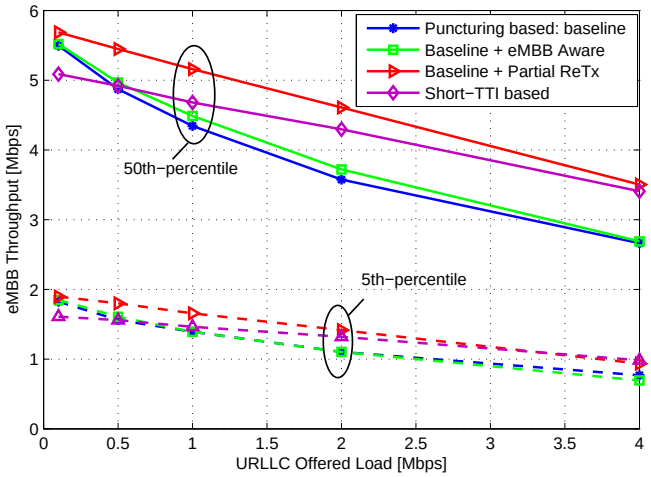


Fig. IV.1: eMBB throughput performance for different scheduling configurations and offered loads of URLLC traffic. The evaluated URLLC offered load conditions are limited to the cases where the URLLC requirements are fulfilled.

spectral-efficient method to recover from cases where only a small part of the initial eMBB transmission is punctured.

References

[1] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *Globecom Workshops (GC Wkshps), 2014*. IEEE, 2014, pp. 1391–1396.

[2] 3GPP TR 38.802 v14.0.0,, "Study on New Radio Access Technology; Physical Layer Aspects," Mar. 2017.

[3] 3GPP TR 38.804 v14.0.0, "Study on New Radio Access Technology; Radio Interface Protocol Aspects (Release 14)," Mar. 2017.

[4] R1-1700374, "Downlink Multiplexing of eMBB and URLLC Transmissions," *3GPP TSG RAN WG1 NR Ad-Hoc Meeting*, Jan. 2017.

Paper F

Radio Resource Management for 5G Ultra-Reliable Low-Latency Communications

Guillermo Pocovi, Klaus I. Pedersen, Preben Mogensen, Beatriz Soret

The paper has been submitted to the
IEEE Transactions on Vehicular Technology, 2017.

Abstract

This paper presents solutions for efficient multiplexing of ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB) traffic on a shared channel. Such scenario presents multiple challenges in the context of radio resource scheduling, link adaptation and inter-cell interference, which are identified and addressed throughout the paper. Among others, we propose a joint link adaptation and resource allocation technique that dynamically adjusts the block error probability (BLEP) of URLLC small payload transmissions in accordance with the instantaneous experienced load per cell. Extensive system-level simulations of the downlink performance show promising gains of this technique, reducing the URLLC latency from 1.3 ms to 1 ms at the 99.999% percentile, with less than 10% degradation of the eMBB throughput performance as compared to conventional scheduling policies. Moreover, an exhaustive sensitivity analysis is conducted to determine the URLLC and eMBB performance under different offered loads, URLLC payload sizes, and link adaptation and scheduling strategies. The presented results give valuable insights on the maximum URLLC offered traffic load that can be tolerated while still satisfying the URLLC requirements, as well as what conditions are more appropriate for dynamic multiplexing of URLLC and eMBB traffic in the upcoming 5G systems.

1 Introduction

The fifth generation (5G) new radio (NR) shall provide support for a wide range of services and applications [1], [2]. Besides enhanced mobile broadband (eMBB), supporting an evolution of today's broadband traffic, 5G will enable ultra-reliable low-latency communications (URLLC), where small payloads must be correctly transmitted and received in a very short time (up to 1 ms) with a success probability of 99.999% [3]. Support for such unprecedented requirements of latency and reliability will open the door to novel use cases, including wireless control and automation in industrial environments [4], inter-vehicular communication for safety [5], smart grids [6], and real-time tactile internet services [7].

The very strict requirements of URLLC, not achievable with the current cellular technologies, call for a broad set of enhancements in the radio interface. The studies in [8]- [10] focus on addressing the harmful effects of the wireless channel, which represent a major challenge to the reliability of the system. It is shown how a combination of micro- and macroscopic spatial diversity techniques plays an important role in dealing with the large- and small-scale fading effects, and the co-channel interference. To achieve low over-the-air transmission delay, the use of short transmission time intervals (TTIs) is of significant importance [11]. The studies in [12] analyse the downlink latency performance with different TTI durations and load conditions

where, in line with [13], the tradeoffs between spectral efficiency, latency and reliability are observed. Under conditions of reliable control channels and feedback, the use of retransmission techniques such as hybrid automatic repeat request (HARQ) can substantially relax the initial block error probability (BLEP) constraint that the URLLC transmissions need to fulfil [14]. The advantage of using multiple transmission attempts, as compared to a single (very conservative) transmission, is a reduction of the average amount of radio resources required to transmit the small data payloads [14]. Building on these studies, we presented in [15] multiple medium access control (MAC) layer enhancements for URLLC, which are corroborated by extensive system-level simulations of the downlink latency performance in a multi-user and multi-cell wide-area environment. In scenarios with only URLLC traffic, it is shown that the link adaptation inaccuracies, as a consequence of the very sporadic traffic and rapid interference variations, represent a major challenge.

Because of the unpredictable nature of URLLC traffic, it is clear that reserving an exclusive set of radio resources for such transmissions would be inefficient from a cost- and resource efficiency perspective. One promising solution is therefore to dynamically multiplex URLLC and eMBB traffic on a shared channel. Here, new challenges arise as compared to the URLLC-only case in [15]; namely (i) how to efficiently distribute the resources between eMBB and URLLC while ensuring their respective quality-of-service (QoS) requirements, and (ii) how to deal with the larger inter-cell interference generated from scheduling eMBB users. The first challenge is typically handled at the packet scheduler functionality. QoS-aware scheduling techniques for cellular systems such as HSPA and LTE have been exhaustively studied in the open literature, see e.g. [16], [17]. It is clear from those studies that users with tight latency constraints should be favoured when allocating radio resources, either by using hard priority or soft priority type-of solutions. Regarding (ii), co-channel inter-cell interference is one of the limiting factors in cellular networks, and has been addressed in many papers, e.g. [18], [19]. In the context of URLLC, [20] and [21] study inter-cell coordinated power boosting and/or cell muting schemes as a way to achieve high reliability and low latency in 5G, whereas [22], [23] analyse different deployment strategies (cell layout and frequency reuse pattern) in order to meet the coverage requirements in a factory automation scenario. Reliable transmission of data under severe interference can also be handled to some extent by selecting a sufficiently robust modulation and coding scheme (MCS) such that low BLEP is achieved [24]. Determining how conservative the BLEP target shall be is however not trivial in dynamic multi-user environments, where multiple URLLC transmissions may need to be accommodated in the same TTI (with limited amount of frequency resources) [15].

This paper presents enhancements for efficient support of URLLC and eMBB in 5G cellular networks, focusing on the downlink. The proposed so-

lutions are derived for a highly-dynamic environment, with multiple users and cells, time-varying traffic and inter-cell interference. Building on the previous work in [15], we consider the case where the network carries only URLLC traffic, and cases where URLLC user equipments (UEs) coexist with traditional best-effort eMBB traffic. The latter case, to the best of our knowledge, has not been previously studied in literature in a dynamic setting. As it will be shown, the challenges for achieving the stringent requirements of URLLC are rather different for those two cases, calling for different solutions. The contributions of this paper can be summarized as follow:

- We present a QoS-aware and radio-channel-aware packet scheduling mechanism, able to efficiently serve the URLLC users in accordance with their QoS requirements, even in the presence of eMBB traffic. Our packet scheduling framework closely interacts with the link adaptation functionality, such that the BLEP of the URLLC transmissions is dynamically adjusted in coherence with the instantaneous load experienced per cell.
- We propose and evaluate an attractive channel quality indicator (CQI) measuring procedure, which significantly improves the URLLC link adaptation accuracy. The proposed procedure applies a low-pass infinite impulse response (IIR) filtering of the measured interference, which provides promising benefits especially in scenarios with large load fluctuations due to the sporadic URLLC traffic.
- An extensive system-level evaluation is carried out to quantify the benefit of the proposed enhancements. The presented results provide a good insight on the maximum offered URLLC traffic load that can be tolerated in the system, as well as its sensitivity to the URLLC-eMBB traffic composition, URLLC payload size, and link adaptation and scheduling setting.
- As there is no such thing as a free lunch, we determine the cost in terms of eMBB throughput for satisfying the stringent URLLC latency and reliability requirements. The proposed solutions allow to fulfil the URLLC requirements, with acceptable impact on the experienced eMBB throughput.

The complexity of our system model prevents a purely analytical evaluation without omitting many important practical aspects. The performance is therefore assessed via highly detailed system-level simulations, following the 5G NR evaluation methodology agreed in the 3rd Generation Partnership Project (3GPP) [2]. The simulator includes explicit and detailed modelling of the majority of radio resource management (RRM) functionalities, and link-to-system mapping for determining the error probability of each data

transmission. When conducting such simulations, good practice in ensuring trustworthy results is applied.

The rest of the paper is organized as follows: Section 2 describes the considered network and traffic model, and the performance metrics. Section 3 outlines the RRM considerations, including the proposed radio resource scheduling and link adaptation enhancements. The simulation assumptions are outlined in Section 4. The performance results are shown in 5, followed by related discussions in Section 6. Finally, concluding remarks appear in Section 7.

2 Setting the Scene

2.1 Network Layout and Traffic Model

We follow the 5G NR modelling assumptions for a wide-area macro cellular scenario as outlined in [2]. This consists of C cells which are deployed in a sectorized manner, with three sectors per site and 500 meter inter-site distance. A set of U UEs are uniformly distributed across the network area. Two different traffic compositions are considered: In case (i), the U UEs are configured with URLLC type-of traffic. This consists of small payload sizes (between 32 and 200 Bytes) that arrive for each URLLC UE in the downlink direction following a Poisson arrival process. This traffic model is known as FTP Model 3 in 3GPP [2]. Case (ii) consists of a mix of URLLC and eMBB UEs. eMBB UEs are modelled with background full buffer best-effort downlink traffic.

2.2 Frame Structure and Numerology

Users are dynamically multiplexed on a time-frequency grid of resources, using orthogonal frequency division multiple access (OFDMA) and frequency-division duplexing (FDD). The physical layer numerology follows the recent agreements in 3GPP: 15 kHz sub-carrier spacing (SCS), 14 OFDM symbols per 1 ms, and a physical resource block (PRB) size of 12 sub-carriers (180 kHz) as the baseline configuration, although options with 2^N scaling of the SCS ($N \in [1, 2, ..., 5]$) are also allowed in the standard [2]. The carrier bandwidth configuration is 20 MHz, corresponding to 100 PRBs. The 3GPP has also agreed on using different TTI durations in accordance with the user-specific requirements. The possible time-domain scheduling resolutions include a *mini-slot*, composed of 1-3 OFDM symbols; a *slot* of 7 OFDM symbols, or aggregation of multiple slots [2]. For the purpose of achieving low latency, we assume that URLLC and eMBB UEs are scheduled on a 2 OFDM symbol (0.143 ms) mini-slot resolution. We refer to [15] and [25] for URLLC and eMBB system-level performance results with different TTI durations.

Each data transmission to a user is indicated with a scheduling grant.

Table F.1: CCH overhead and scheduling format for a 2-symbol (0.143 ms) TTI size [27]

SINR [dB]	In-resource CCH overhead	Frequency-domain minimum allocation size
$(-\infty, -2.2)$	$8 \cdot 36 = 288$ REs	14 PRBs (336 REs)
$[-2.2, 0.2)$	$4 \cdot 36 = 144$ REs	8 PRBs (192 REs)
$[0.2, 4.2)$	$2 \cdot 36 = 72$ REs	5 PRBs (120 REs)
$[4.2, \infty)$	$1 \cdot 36 = 36$ REs	3 PRBs (72 REs)

The scheduling grant contains information on the specific time-frequency resource allocation for each user, the employed MCS, and other transmission parameters required to decode the data. In line with [26], the control channel (CCH) for transmitting the scheduling grant is accommodated within the resources assigned to each user (i.e. in-resource CCH). The coding rate of the in-resource CCH is dynamically adapted in accordance with the user's channel condition, as expressed in the CQI report. In this regard, we assume that the in-resource CCH will carry similar information as the LTE physical downlink control channel (PDCCH), and we therefore use the PDCCH link-level performance [27] as a reference. That is, a minimum of 36 resource elements (REs) in order to transmit the CCH with a BLEP of 1%, with additional repetition encoding in form of aggregation levels 2, 4, 8, depending on the user's channel condition. One RE corresponds to one OFDM subcarrier symbol. Note that the considered in-resource CCH allows for more flexible scaling of the control channel overhead, as compared to LTE where the PDCCH overhead is either 7%, 14%, or 21% [28].

On each scheduling opportunity, the resource allocation to a user must be sufficiently large to accommodate the in-resource CCH as well as a reasonable data payload and reference symbols [25]. Table F.1 summarizes the required number of REs for the CCH depending on the user-specific signal to interference and noise ratio (SINR), as well as the corresponding minimum frequency-domain allocation size assumed in this work.

2.3 Latency Budget

For each UE, data from higher layers are received at the serving cell and stored in a user-specific transmission buffer as illustrated in Fig. F.1. The URLLC latency is measured from the moment a URLLC payload arrives at the serving cell until it is successfully received at the UE. This accounts for various constant and variable components, namely the queuing delay, defined as the time elapsed between the arrival of the payload at the cell's buffers (Fig. F.1) and the execution of the scheduling decision; frame align-

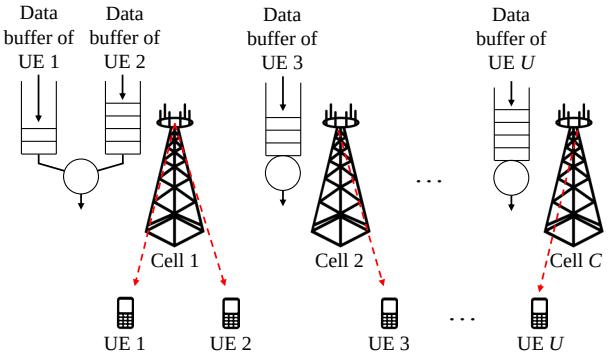


Fig. F.1: System model.

ment, i.e. time remaining to the beginning of the next TTI; and transmission delay. Naturally, the transmission of a URLLC payload takes at least one TTI but may also take multiple TTIs depending on the available resources, payload size, radio channel conditions, transmission errors and the respective HARQ retransmissions that could occur. Fig. F.2 shows an example time-line of the transmission of a URLLC payload. The diagram is simple in the sense that it assumes a single TTI for the initial data transmission of the payload, no queuing delay, and one HARQ retransmission with a 4 TTI round-trip time (RTT) as also considered in [15]. Under these conditions, the maximum latency with a TTI duration 0.143 ms corresponds to $6 \cdot 0.143$ ms = 0.86 ms, hence satisfying the 1 ms latency target.

2.4 Notation

The following notation is used throughout the paper: set of cells and UEs are denoted by $\mathcal{C} = \{1, ..., C\}$ and $\mathcal{U} = \{1, ..., U\}$, respectively. To distinguish the UE type, we define $\mathcal{U}_{urllc} = \{1, ..., U_{urllc}\} \subseteq \mathcal{U}$ as the set of URLLC UEs, and $\mathcal{U} - \mathcal{U}_{urllc}$ as the set of eMBB UEs. We use superscript $\mathcal{U}^c$ to indicate the set of users connected to cell c. The set of PRBs is denoted as $\mathcal{P} = \{1, ..., P\}$.

3 Radio Resource Management Considerations

Needless to say, the packet scheduler and link adaptation functionality play and important role in fulfilling the users' QoS requirements. Dynamic link adaptation is considered for both URLLC and eMBB data transmissions, by setting the MCS based on the user's frequency-selective CQI report [24]. The MCS for the eMBB users is adjusted to reach an average BLEP target of 10%. This is achieved by using the well-known outer loop link adaptation (OLLA) algorithm, where the received CQI values are offset by a certain fac-

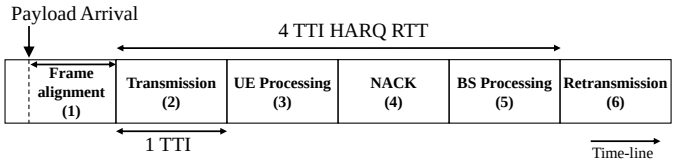


Fig. F.2: Diagram of URLLC downlink data transmissions, assuming one HARQ retransmission.

tor (a.k.a. the OLLA offset) calculated in accordance with the received HARQ ACK/NACK feedback from past transmissions [29].

The link adaptation of URLLC transmissions is more challenging. Naturally, decreasing the BLEP target of the URLLC data transmissions allows to improve the latency and reliability performance [14]. However, in the considered dynamic environment it is not trivial to determine how conservatively/aggressively the MCS should be set for the URLLC users, as scheduling a URLLC UE with extremely low BLEP could result in a lack of radio resources for other URLLC data transmissions [15].

We study joint link adaptation and resource allocation schemes, in the sense that the BLEP target of the URLLC payload transmissions is adjusted based on the amount of PRBs allocated to such transmissions. We first describe the conventional resource allocation approach, where the MCS is adjusted in order to reach a certain fixed BLEP target. Next, we present an improved resource allocation method that dynamically adjusts the BLEP of each individual URLLC transmission in accordance with the instantaneous experienced load per cell.

3.1 Resource Scheduling with Fixed BLEP Target

We assume a packet scheduler where, on each TTI n, each cell $c \in \mathcal{C}$ independently allocates up to P PRBs to its associated users $\mathcal{U}^c$, taking into account the user-specific QoS class and radio-channel conditions [29]. The QoS-awareness is achieved by dividing the scheduling procedure into two stages: a first one where PRBs are allocated to the URLLC users with pending data transmission[1]; and a second one where the remaining PRBs (if any) are allocated to the $\mathcal{U}^c - \mathcal{U}^c_{urllc}$ eMBB UEs (i.e. hard priority type-of scheduling). The actual assignment of PRBs to the users is carried out in a similar way for both traffic classes: a PRB p is assigned to the user u^* that maximizes the well-known *proportional fair* (PF) metric, i.e.,

$$u^* = \arg\max_u \left\{ \frac{r_{u,p}[n]}{T_u[n]} \right\}, \tag{F.1}$$

where n is the discrete time index for the scheduling interval, $r_{u,p}$ is an estimate of the instantaneous supported data rate of user $u \in \mathcal{U}^c$ in the p-th

[1]HARQ retransmissions are prioritized over new transmissions in line with [29].

PRB, and T_u is its average delivered user throughput in the past. The value of $r_{u,p}$ is estimated based on the periodical frequency-selective CQI report sent by each UE, whereas $T_u[n]$ is calculated recursively using a moving average filter. The value of T_u is only updated for users that have data buffered [30]. The cell index has been left out of (F.1) for the sake of simplicity.

On each scheduling interval, the allocation size (i.e. number of PRBs) to a user can be as small as indicated in Table F.1, whereas the maximum allocation size is a function of the available resources, user pending data, and the employed MCS. The MCS is selected in order to reach a fixed BLEP target (10% for eMBB, and 0.1%-1% for URLLC).

3.2 Resource Scheduling with Dynamic BLEP Adjustment

Fig. F.3 describes the operation of the proposed resource allocation scheme with dynamic BLEP adjustment for the URLLC transmissions. The scheduling procedure consists of three steps. Steps (1) and (3) are inherited from Section 3.1. In step (1), each cell c allocates PRBs to its associated URLLC UEs based on their experienced channel quality as expressed in the CQI report. The key point in this step is that each URLLC UE u with pending data transmission receives an allocation of size x_u PRBs ($x_u \in [0, 1, ..., P]$ and $\sum_{i \in \mathcal{U}^c} x_i \leq P$), such that the URLLC payload can be transmitted with a modest initial BLEP target, e.g. 1%. Once the initial $X = \sum_{i \in \mathcal{U}^c} x_i$ PRBs have been allocated, step (2) consists on assigning a proportion Γ ($0 < \Gamma < 1$) of the remaining $P - X$ PRBs to the already allocated URLLC users. The additional resources will allow to transmit the URLLC small payload with a more conservative MCS (i.e. even further reduced BLEP). The question is now: how to select and distribute the $\Gamma \cdot (P - X)$ PRBs among the different URLLC users? Let us assume that each URLLC user u is allocated with γ_u additional PRBs, where γ_u is calculated proportionally to the user's initial allocation size x_u, i.e.,

$$\gamma_u = \frac{\Gamma \cdot (P - X)}{X} \cdot x_u \ . \tag{F.2}$$

The γ_u additional PRBs for each user are selected following the PF rule as described in (F.1).

In step (3), the remaining $(1 - \Gamma) \cdot (P - X)$ PRBs are allocated to the eMBB UEs. Note that for $\Gamma = 1$, eMBB users will only be scheduled on TTIs where no URLLC traffic is present. This setting provides the highest URLLC reliability, at the expense of the largest eMBB throughput degradation; whereas the case with $\Gamma = 0$ corresponds to the baseline scheduling operation (as described in Section 3.1). In essence, the proposed resource allocation technique aims at scheduling URLLC UEs with a modest BLEP target, e.g. 1%, but can be lower depending on the configured Γ, and the experienced URLLC traffic load on each scheduling instant.

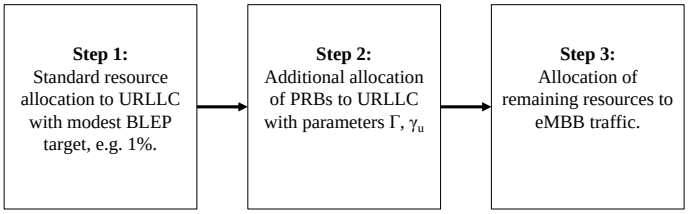


Fig. F.3: Scheduling procedure.

3.3 Accurate CQI Measurements and Reporting

The following enhancement aims at improving the accuracy of the CQI measurements at the UE side, hence giving complementary benefit to the already presented cell-side mechanisms. Accurate link adaptation is challenging especially in the scenario with only URLLC traffic. Due to the relatively small payloads, a URLLC transmission generally occupies a subset of the available PRBs within a TTI. This fact, together with the sporadic nature of URLLC traffic, result in a rapidly changing interference pattern, hence making it difficult to accurately select an appropriate MCS that fulfils the specified BLEP constraint. This problem is also well-known from LTE system-level performance analyses in non-fully loaded networks [31]. However, it is exacerbated in this scenario as drastic variations occur from TTI to TTI and on a PRB basis.

In LTE, the UE determines the CQI based on a finite number of channel quality measurements obtained from a relatively short measuring window [28]. Our proposal is to modify the UE measurement procedure of the CQI report, by including historical information of the experienced interference. On each TTI n, the UE measures the interference with a certain PRB resolution (a.k.a. sub-band). The interference measurement on the i-th sub-band at time instant n, $y_i[n]$, is filtered with a low-pass first-order IIR filter, resulting in the following smoothed value:

$$s_i[n] = \alpha \cdot y_i[n] + (1 - \alpha) \cdot s_i[n-1] \,, \tag{F.3}$$

where α is the forgetting factor (FF) of the filter ($0 < \alpha < 1$). The CQI, which is periodically reported to the serving cell, contains the low-pass filtered interference information $s_i[n]$ together with the latest desired-signal fading information. Note that the latter varies in a much lower time scale and, except for very high UE speeds, it is possible to track the channel variations with relatively high accuracy [28]. The FF α determines how much weight is given to the latest measurement as compared to the previous ones. Following the previous work in [15], we use $\alpha = 0.01$, which provides significantly latency and reliability improvement.

4 Simulation Assumptions

The performance evaluation is based on dynamic system-level simulations following the 3GPP 5G NR methodology [2]. The default simulation assumptions are summarized in Table F.2. The network layout, UE distribution and traffic follow the description presented in Section 2.1. The network is composed of $C = 21$ cells, where $U_{urllc} = 210$ URLLC UEs are uniformly distributed (10 UEs per cell in average). eMBB background traffic is modelled with 105 additional UEs (5 UEs per cell in average) with best-effort full-buffer downlink traffic.

The simulator time-resolution is one OFDM symbol, and it includes explicit modelling of the radio resource management functionalities described in Section 2 and 3. The simulator has been used to generate a large variety of LTE and 5G NR performance results and has been calibrated with system-level simulators from several 3GPP member companies. The basic methodology is outlined in the following: on every TTI, the experienced SINR for each scheduled user is calculated per RE, assuming a minimum mean square error interference rejection combining (MMSE-IRC) receiver [32]. Given the SINR per RE, the effective exponential SINR model [33] is applied for link-to-system-level mapping to determine if the transmission was successfully decoded. Asynchronous adaptive HARQ with Chase Combining is applied in case of failed transmissions, and the SINRs for the different HARQ transmissions are linearly added [34]. Closed-loop single-stream single-user 2x2 MIMO transmission mode is assumed, i.e. benefiting from both transmission and reception diversity against fast fading radio channel fluctuations [8]. Dynamic link adaptation is applied for both data and the in-resource CCH, based on periodical frequency-selective CQI reports from the UEs. The simulator does not consider user mobility; however, the dynamic traffic model and fast fading effects provide significant variability to the channel conditions. Unless otherwise mentioned, we assume low-pass IIR CQI measurements at the UE (Section 3.3) for cases with only URLLC traffic. Each cell independently schedules its users with full priority for URLLC traffic.

For each URLLC UE, payloads of B Bytes are generated in the downlink direction following a Poisson distribution with arrival rate λ. Payload sizes of 32 Bytes, 50 Bytes and 200 Bytes are considered [2]. When presenting the results, we will refer to *URLLC offered load* or simply *offered load*, defined as $L_{urllc} = U_{urllc} \cdot B \cdot \lambda / C$, to indicate the average amount of URLLC traffic that is offered per cell.

The latency (defined in Section 2.3) of each successfully received URLLC payload is collected and used to form empirical complementary cumulative distribution functions (CCDF). For URLLC, the key performance indicator (KPI) is the achievable latency with 99.999% probability, i.e. the 10^{-5} per-

Table F.2: Simulation assumptions

Parameter	Value
Network environment	3GPP Urban Macro (UMa) network with 21 cells and 500 meter inter-site distance [2]
Carrier configuration	20 MHz carrier bandwidth at 2 GHz
PHY numerology	15 kHz subcarrier spacing; 12 subcarriers per PRB; TTI size of 2 OFDM symbols (0.143 ms)
Control channel	Error-free in-resource scheduling grants with dynamic link adaptation [26]
Data channel MCS	QPSK to 64QAM, with same coding rates as in LTE; eMBB: 10% BLEP target for first transmissions URLLC: 0.1-1% BLEP target with different Γ settings
Reference signals overhead	4 resource elements per PRB
Antenna configuration	2 x 2 single-user single-stream MIMO with LTE-alike precoding and MMSE-IRC receiver
Packet scheduler	Proportional Fair with priority for URLLC traffic and different Γ settings
CSI	LTE-alike CQI and PMI, reported every 5 ms; CQI filtering ($\alpha = 0.01$) for cases with only URLLC traffic
HARQ	Async. HARQ with Chase combining and 4 TTI RTT; Max. 6 HARQ retransmissions
RLC	RLC Unacknowledged mode
Traffic composition	Case a) 210 URLLC UEs Case b) 210 URLLC UEs + 105 eMBB UEs
UE distribution	Uniformly distributed; 20% indoor and 80% outdoor locations; 3 km/h UE speed
Traffic model	URLLC: FTP3 downlink traffic with 32, 50 or 200 Byte payload size; eMBB: full buffer
URLLC Offered load	1 - 8 Mbps average load per cell

centile of the CCDF. For eMBB users, we look at the 5th percentile and 50th percentile (or median) of the downlink end-user throughput.

The simulation time corresponds to at least 5.000.000 successfully received URLLC payloads. Assuming that the obtained latency samples are uncorrelated, such sample size allows to estimate the 10^{-5}-th percentile of the latency with an error margin of at most $\pm 5\%$[2], with 95% confidence level [35]. In practice, some correlation is present among the latency samples, slightly increasing the error margin of the results.

[2]The actual error margin depends on the steepness of the distribution at the percentile of interest.

Table F.3: URLLC RU[%] for different URLLC offered loads and URLLC payload sizes.

L_{urllc}	200 Byte Payload		50 Byte Payload		32 Byte Payload	
[Mbps]	w/o eMBB	w/ eMBB	w/o eMBB	w/ eMBB	w/o eMBB	w/ eMBB
1	1.6	3.9	2.2	4.6	2.8	5.3
2	3.3	7.9	4.5	9.3	5.6	10.7
4	6.9	16.3	9.3	18.9	11.7	21.4
6	11.3	24.4	14.4	28.6	17.9	31.8
8	16.6	33.2	20.0	38.2	24.7	43.2

5 Performance Results

The URLLC latency and eMBB throughput performance is evaluated under different offered load conditions, scheduling policies, and URLLC payload sizes. The percentage of PRBs allocated to URLLC traffic (we refer to this as URLLC resource utilization (RU)) is summarized in Table F.3. Here, we assume a fixed BLEP target of 0.1% for URLLC traffic and $\Gamma = 0$. As expected, the URLLC RU increases with the URLLC offered load L_{urllc}. This growth is non-linear for cases without eMBB traffic. Apart from the larger volume of data that needs to be delivered, higher offered load results in larger inter-cell interference and consequently lower signal quality for URLLC users. In contrast, cases with full-buffer eMBB traffic correspond to a fully loaded network. This results in close-to linear increase of the URLLC RU vs L_{urllc}, since the signal quality of the UEs does not change with the offered load.

5.1 URLLC Performance

We first focus on cases with 200 Byte URLLC payloads. Fig. F.4(a) shows the CCDF of the URLLC latency under different offered load conditions, for the scenario without eMBB traffic. At the 10^{-5} percentile, it is observed that the 1 ms latency requirement is fulfilled for all the considered offered loads of URLLC traffic. Specifically, a latency of $\sim$0.86 ms is achieved, matching the latency budget in Fig. F.2. The different components contributing to the URLLC latency are also depicted. The upper part of the distribution (10^{0} – 10^{-3} percentile) represents the case where the URLLC payloads are immediately scheduled and correctly received at the UE. With 10^{-3} probability (equivalent to the URLLC-specific 0.1% BLEP target), the initial URLLC transmissions are not correctly received at the UE side. This triggers a HARQ retransmission, which is immediately scheduled after receiving the HARQ NACK at the serving cell. In addition to this, some temporary queuing delay is experienced at the cells' buffers when operating at a offered load of 8 Mbps, hence degrading the URLLC latency.

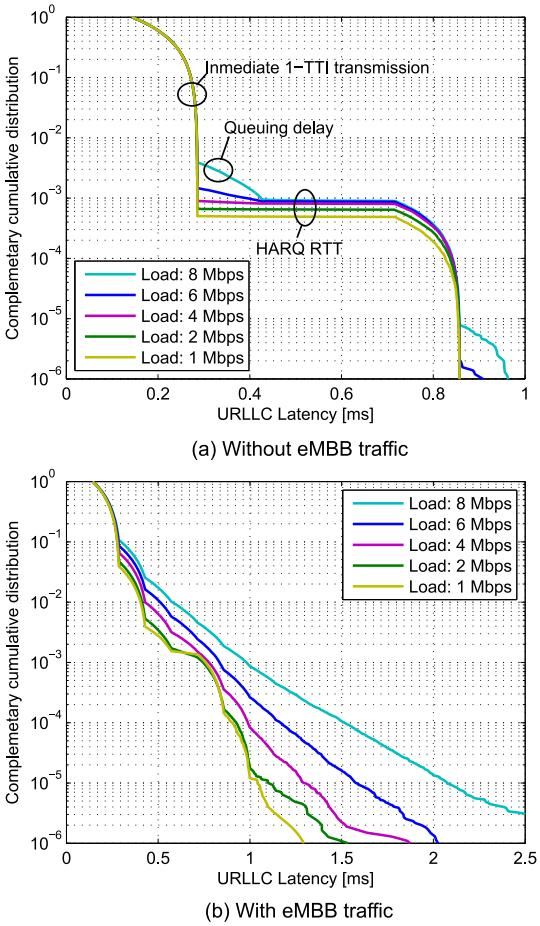


Fig. F.4: URLLC latency distribution for different URLLC offered load conditions, and traffic configurations. URLLC BLEP target: 0.1%; $\Gamma = 0$; 200 Byte payload.

Fig. F.4(b) shows the URLLC latency distribution for cases with eMBB traffic. It is observed that the 1 ms latency with $1 - 10^{-5}$ reliability is not fulfilled, even at low URLLC offered loads. Even though URLLC transmissions are fully prioritized by the packet scheduler, the larger inter-cell interference from scheduling eMBB users significantly degrades the URLLC latency performance. This is highlighted in Fig. F.5, that shows the distribution of the instantaneous post-detection SINR of the URLLC users. Cases without eMBB traffic experience a 7 dB SINR degradation at the median when increasing the offered load from 1 Mbps to 8 Mbps. For the scenario where the network is fully loaded with eMBB traffic, the SINR is independent of the URLLC offered load, and significantly worse than the cases with only

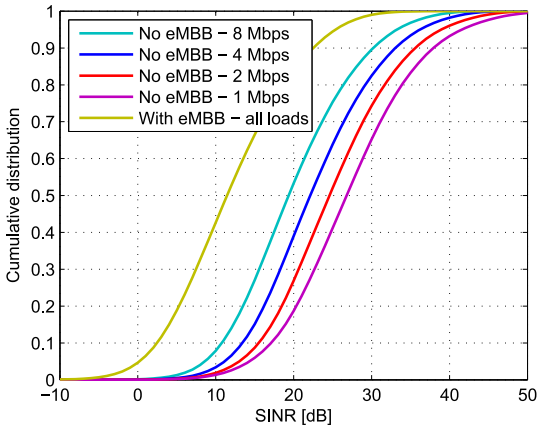


Fig. F.5: Instantaneous per-user per-subcarrier SINR for different URLLC offered load conditions, and traffic configurations. URLLC BLEP target: 0.1%; $\Gamma = 0$; 200 Byte payload.

URLLC traffic. Lower SINR results in lower MCS for data transmissions and higher CCH overhead. As a consequence, larger amount of PRBs (see Table F.3) is required to deliver the URLLC payloads, having a negative impact on the queuing delay at the cell and transmission delay (e.g. a URLLC payload not fitting in a single TTI).

The URLLC latency and reliability performance is tightly related to the scheduling and link adaptation settings. Fig. F.6(a) summarizes the latency performance at the $1 - 10^{-5}$ percentile for different offered traffic loads. For the scenario without eMBB traffic, we show settings with $\alpha = 0.01$ and $\alpha = 1$ of the low-pass CQI filtering enhancement[3]. For the case where eMBB traffic is present, we include different configurations of the scheduling technique ($\Gamma > 0$) presented in Section 3.2. It is observed that the proposed CQI filtering scheme provides large gains, as low-pass information of the experienced interference is implicitly included in the CQI report. These benefits are especially relevant at high load when the cell activity is higher and more sporadic interference is experienced across the network. When looking at cases with eMBB traffic and $\Gamma = 0$, the configured URLLC BLEP target has a large impact on the achievable latency. At low URLLC offered load, it is advantageous to operate with low BLEP target (0.1%) in order to reduce the occurrence of HARQ retransmissions (and the corresponding HARQ processing delay). As the load increases, non-negligible queuing starts to occur at the cells' buffers, which deteriorates considerably the latency performance. Under such circumstances, it is beneficial to operate at a higher BLEP target (1%) in order to increase the spectral efficiency of the system and reduce the

[3]Note that this technique is less relevant in scenarios where eMBB background traffic is present, as stable full-load interference conditions are experienced.

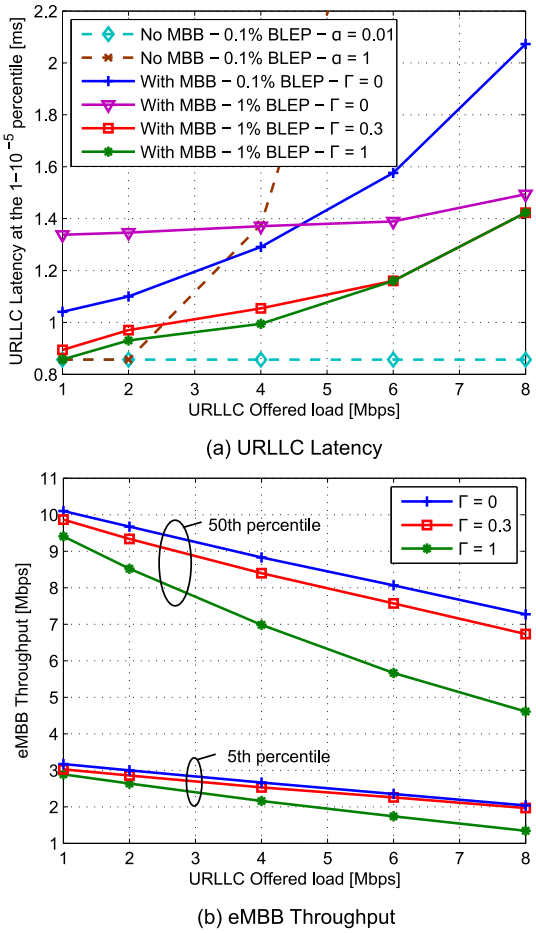


Fig. F.6: Summary of (a) URLLC latency at the $1 - 10^{-5}$-percentile, and (b) eMBB throughput at the 5th- and 50th-percentile. 200 Byte payload.

queue length. Due to these tradeoffs, the proposed resource allocation algorithm ($\Gamma > 0$) provides much better latency performance. Recall from Section 3.2 that the proposed scheduling technique aims at scheduling URLLC UEs with a BLEP target of at most 1%, but can be lower depending on the instantaneous URLLC load at each cell. The 1 ms URLLC latency requirement is fulfilled for offered loads up to 2 Mbps, although the latency performance at higher URLLC offered loads is still decent ($\leq$1.4 ms). Settings with $\Gamma = 1$ provide the best URLLC performance, whereas $\Gamma = 0.3$ still provide relevant latency improvement with only minor impact on the eMBB throughput performance, as will be presented next.

5.2 eMBB Performance

Fig. F.6(b) shows the 5th- and 50th- percentile of the eMBB throughput under different scheduling and traffic settings. As expected, the eMBB throughput decreases as we increase the URLLC offered load. Configurations with $\Gamma = 0$ achieve the highest eMBB throughput, whereas cases with $\Gamma > 0$ experience lower throughput at the expense of reduced URLLC latency, as shown in Fig. F.6(a). Particularly, the setting with $\Gamma = 0.3$ offers significant latency reduction (e.g. from $\sim$1.3 ms down to $\sim$1.05 ms at 4 Mbps offered load) with a small throughput degradation ($\leq$10% for any URLLC offered load condition). Another notable advantage of the proposed technique is that a single setting of parameters (Γ and initial BLEP target) provides significant latency reduction under a wide range of offered load conditions.

5.3 Sensitivity to the URLLC Payload Size

The sensitivity to the URLLC payload size is presented next. As previously observed in Table F.3, settings with smaller payload sizes experience larger RU as compared to the 200 Byte payload case, which is a consequence of the larger CCH overhead. Under these circumstances, the considered in-resource CCH brings relevant benefits, as it allows more flexible scaling of the CCH as compared to LTE. Fig. F.7 shows the URLLC latency at the $1 - 10^{-5}$ percentile for 32 Byte and 50 Byte payload sizes, with and without eMBB traffic. In order to illustrate the benefits of the in-resource CCH, we include cases where the maximum number of scheduled URLLC UEs per TTI is limited to six[4].

[4]Common assumption for LTE performance evaluation, given the limited PDCCH capacity [28].

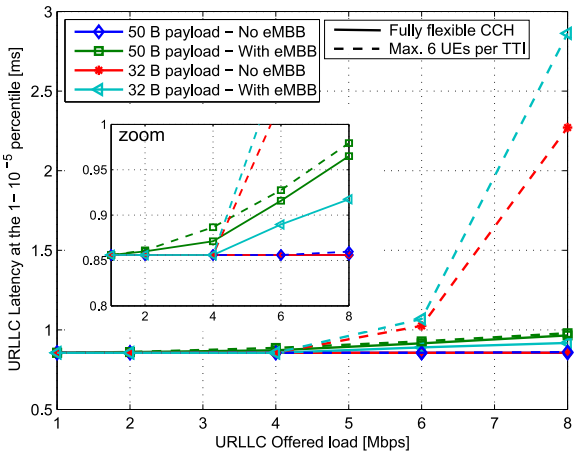


Fig. F.7: URLLC latency performance for different payload sizes and CCH settings. URLLC BLEP target: 0.1%; $\Gamma = 0$.

It is observed that the enforced CCH restriction considerably degrades the URLLC performance. As an example, the setting with 32 Byte payload does not achieve the 1 ms latency requirement for $L_{urllc} \geq 6$ Mbps, even for cases without eMBB traffic. This degradation is mainly due to the enforced CCH restriction, meaning that radio resources are left unused due to the limited CCH capacity. Some performance degradation is also observed for settings with 50 Byte payload size; although the 1 ms latency requirement is still fulfilled. In contrast, cases without the enforced CCH restriction (fully flexible CCH) achieve significantly better performance. The URLLC requirements are fulfilled from low load to high load, also in cases where eMBB traffic is present in the network. Such good performance is mainly due to the lower amount of PRBs required for the transmission of smaller payloads. This is reflected in Fig. F.8, where the empirical distribution of the allocated PRBs per TTI for 1 and 8 Mbps offered load is shown for the scenario with only URLLC traffic. Even though small payload sizes experience larger average RU (Table F.3), the variance of the allocated PRBs per TTI is much larger for cases with a payload size of 200 Bytes. As an example, for cases with 8 Mbps and 200 Byte payload, the 100% of the PRBs are scheduled with a non-negligible 10^{-2} probability. This results in large queuing and transmission delay, meaning that URLLC transmissions are not immediately scheduled upon arrival. This temporary queuing occurs less often for payload sizes of 32 Bytes or 50 Bytes, which explains the much better latency performance (for cases without CCH limitations).

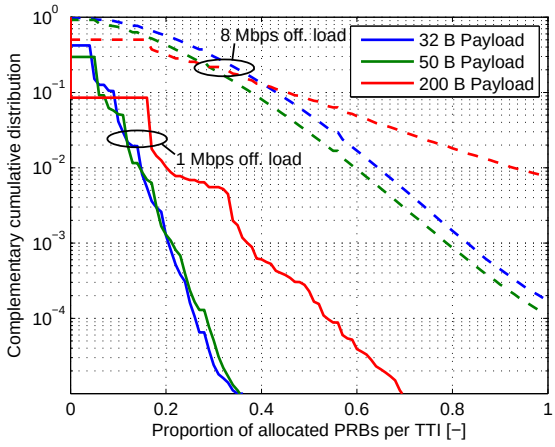


Fig. F.8: Proportion of allocated PRBs per TTI per cell for different URLLC payload sizes and offered loads. No eMBB traffic. URLLC BLEP target: 0.1%; $\Gamma = 0$.

6 Discussion

The URLLC and eMBB performance have been evaluated under different conditions of URLLC offered load, payload sizes and scheduling techniques. The URLLC latency and reliability performance is highly sensitive to the traffic characteristics; particularly, the relation between the available carrier bandwidth and the URLLC payload size. For cases with relatively large payload size (200 Bytes), it is much more challenging to achieve the stringent URLLC latency and reliability requirements, mainly due to the larger transmission and queuing delay. In this regard, the presented scheduling technique with dynamic BLEP adjustment shows promising gains as compared to conventional scheduling algorithms with fixed BLEP target. The proposed solution is highly robust and flexible, as it brings relevant latency improvements over a wide range of offered load conditions without requiring fine adjustment of the link adaptation. Furthermore, it provides to mobile operators a simple method to determine how the radio resources should be distributed among URLLC and eMBB type-of traffic.

Although our focus has been on satisfying the URLLC requirements as defined in 3GPP, it is worth to mention that not all applications are subject to such stringent latency and reliability constraints (see e.g. [5], [6] where use cases requiring 5-15 ms latency with 99.999% reliability are presented). Support for such use cases with more relaxed requirements is expected to have lower impact on the eMBB performance, given the tradeoffs between URLLC latency and eMBB throughput that have been observed in the presented results.

7 Conclusions

In this paper, we have presented solutions for efficient multiplexing of URLLC and eMBB traffic on a shared channel. Specifically, a dynamic resource allocation technique have been proposed which provides a simple, yet effective method to determine how the radio resources should be distributed between the two service classes, in accordance with the well-known tradeoffs between reliability, latency and spectral efficiency. A detailed system-level analysis of the URLLC and eMBB downlink performance shows significant gains of the proposed solution, reducing the URLLC latency from 1.3 ms to 1 ms at the 99.999% percentile, with less than 10% degradation of the eMBB median throughput performance, as compared to conventional scheduling techniques. The main messages brought by this paper are the following: (i) It is possible to multiplex URLLC with eMBB traffic such that the URLLC requirements are fulfilled even under very-high interference from serving the eMBB users. (ii) There is a price to pay in terms of eMBB throughput performance

in order to achieve stringent latency and reliability requirements of URLLC. And (iii) the URLLC performance is highly sensitive to the traffic characteristics, particularly the relation between the available carrier bandwidth, the offered load, and the URLLC payload size. As an example, cases with relatively large URLLC payload size (200 Bytes) fulfil the requirements only for low or medium offered load conditions ($<$ 4 Mbps), whereas settings with smaller payload size (32-50 Bytes) can operate at higher load ($\geq$8 Mbps). In the latter case, we have highlighted the importance of a flexible control channel design in order to avoid problems of control channel blocking, as known from LTE.

Future work must consider a more realistic modelling of eMBB traffic, e.g. finite buffer traffic including the transmission control protocol (TCP) flow control mechanisms. Also, accounting for control channel errors and ACK/NACK misdetections is of relevance to further assess the URLLC latency and reliability performance.

References

[1] 3GPP TR 38.913 v14.1.0, “Study on scenarios and requirements for next generation access technologies", Jan. 2017.

[2] 3GPP TR 38.802 v14.0.0, “Study on new radio access technology physical layer aspects", Mar. 2017.

[3] P. Popovski, “Ultra-reliable communication in 5G wireless systems", *International Conference on 5G for Ubiquitous Connectivity*, Nov. 2014.

[4] A. Frotzscher et al., “Requirements and current solutions of wireless communication in industrial automation", *IEEE International Conference on Communications Workshops (ICC)*, June 2014.

[5] G. Pocovi, M. Lauridsen, B. Soret, K. I. Pedersen and Preben Mogensen, “Automation for on-road vehicles: use cases and requirements for radio design", *IEEE Vehicular Technology Conference*, Sept. 2015.

[6] V. C. Gungor et al., “A survey on smart grid potential applications and communication requirements", *IEEE Transactions on Industrial Informatics*, vol. 9, no. 1, pp. 28-42, Sept. 2012.

[7] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. P. Fettweis, “5G-enabled tactile internet", *IEEE J. Select. Areas Commun.*, vol. 34, no. 3, pp. 460– 473, Mar. 2016.

[8] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen and P. Mogensen, “Signal quality outage analysis for ultra-reliable communications in cellular networks", *IEEE Globecom Workshops*, Dec. 2015.

[9] F. Kirsten, D. Ohmann, M. Simsek and G. P. Fettweis, “On the utility of macro- and microdiversity for achieving high availability in wireless networks", *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Sept. 2015.

[10] D. Ohmann, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, “Diversity Trade-Offs and Joint Coding Schemes for Highly Reliable Wireless Transmissions", *IEEE Vehicular Technology Conference*, Sept. 2016.

[11] 3GPP TR 36.881 v14.0.0, “Study on latency reduction techniques for LTE", Jun. 2016.

[12] G. Pocovi, K. I. Pedersen, B. Soret, Mads Lauridsen and P. Mogensen, “On the impact of multi-user traffic dynamics on low latency communications", *International Symposium on Wireless Communication Systems (ISWCS)*, Sept. 2016.

[13] B. Soret, K. I. Pedersen and P. Mogensen, “Fundamental tradeoffs among reliability, latency and throughput in cellular networks", *IEEE Globecom Workshops*, Dec. 2014.

[14] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji and R. Jäntti, “Link adaptation design for ultra-reliable communications", *IEEE International Conference on Communications*, May 2016.

[15] G. Pocovi, B. Soret, K. I. Pedersen and P. Mogensen, “MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks", *IEEE International Conference on Communications Workshops (ICC)*, May 2017.

[16] G. Song and Y. Li, “Utility-Based Resource Allocation and Scheduling in OFDM-Based Wireless Broadband Networks", *IEEE Communications Magazine*, vol. 47, no. 12, pp. 127-134, Dec. 2005.

[17] T. E. Kolding, “QoS-Aware Proportional Fair Packet Scheduling with Required Activity Detection", *IEEE Vehicular Technology Conference*, Sept. 2006.

[18] G. Boudreau et al., “Interference coordination and cancellation for 4G networks", *IEEE Communications Magazine*, vol. 47, no. 4, pp. 74-81, April 2009.

[19] W. Nam, D. Bai, J. Lee and I. Kang, “Advanced interference management for 5G cellular networks", *IEEE Communications Magazine*, vol. 52, no. 5, pp. 52-60, May 2014.

[20] B. Soret and K. I. Pedersen, “On-demand power boost and cell muting for high reliability and low latency in 5G", *IEEE Vehicular Technology Conference*, June 2017.

[21] B. Soret, G. Pocovi, K. I. Pedersen and P. Mogensen, “Increasing reliability by means of root cause aware HARQ and interference coordination", *IEEE Vehicular Technology Conference*, Sept. 2015.

[22] N. Brahmi, et al., “Deployment strategies for ultra-reliable and low-latency communication in factory automation", *IEEE Globecom Workshops*, Dec. 2015.

[23] S. A. Ashraf, I. Aktas, E. Eriksson, K. W. Helmersson and J. Ansari, “Ultra-reliable and low-latency communication for wireless factory automation: From LTE to 5G", *IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Sept. 2016.

[24] C. H. Yu, A. Hellsten and O. Tirkonnen, “Rate Adaptation of AMC/HARQ Systems with CQI Errors", *IEEE Vehicular Technology Conference*, May 2010.

[25] K. I. Pedersen, et al., “System Level Analysis of Dynamic User-Centric Scheduling for a Flexible 5G Design", *IEEE GLOBECOM*, Dec. 2016.

[26] K. I. Pedersen, G. Berardinelli, F. Frederiksen and A. Szufarska, “A flexible 5G frame structure design for frequency-division duplex cases", *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53-59, Mar. 2016.

[27] D. Laselva et al., “On the impact of realistic control channel constraints on QoS provisioning in UTRAN LTE", *IEEE Vehicular Technology Conference*, Sept. 2009.

[28] H. Holma and A. Toskala, “LTE Advanced: 3GPP Solution for IMT-Advanced", John Wiley & Sons Ltd, 2011.

[29] K. I. Pedersen, et al., “An Overview of Downlink Radio Resource Management for UTRAN Long-Term Evolution", *IEEE Communications Magazine*, vol. 47, no. 7, pp. 86-93, July 2009.

[30] A. Jalali, R. Padovani and R. Pankaj, “Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system", *IEEE Vehicular Technology Conference*, May 2000.

[31] V. Fernández-López, K. I. Pedersen and B. Soret, “Interference characterization and mitigation benefit analysis for LTE-A macro and small cell deployments", *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no.1, April 2015.

[32] M. Lampinen, F. Del Carpio, T. Kuosmanen, T. Koivisto, and M. Enescu, “System-Level Modeling and Evaluation of Interference Suppression Receivers in LTE System", *IEEE Vehicular Technology Conference*, May 2012.

[33] K. Brueninghaus, et al., “Link performance models for system level simulations of broadband radio access systems", *IEEE Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sept. 2005.

[34] D. Chase, “Code combining: A maximum-likelihood decoding approach for combining an arbitrary number of noisy packets", *IEEE Transactions on Communications*, vol. 33, no. 5, pp. 385-393, May 1985.

[35] L. D. Brown, T. T. Cai and A. Dasgupta, “Confidence intervals for a binomial proportion and asymptotic expansions", *The Annuals of Statistics*, vol. 30, no. 1, pp. 160-201, 2002.

Paper G

Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband

Klaus I. Pedersen, Guillermo Pocovi, Jens Steiner, Saeed R. Khosravirad

The paper has been accepted for publication in the
IEEE 86th Vehicular Technology Conference (VTC Fall), 2017.

Abstract

In this paper, we present a punctured scheduling scheme for efficient transmission of low latency communication (LLC) traffic, multiplexed on a downlink shared channel with enhanced mobile broadband traffic (eMBB). Puncturing allows to schedule eMBB traffic on all shared channel resources, without prior reservation of transmission resources for sporadically arriving LLC traffic. When LLC traffic arrives, it is immediately scheduled with a short transmission by puncturing part of the ongoing eMBB transmissions. To have this working efficiently, we propose recovery mechanisms for punctured eMBB transmissions, and a service-specific scheduling policy and link adaptation. Among others, we find that it is advantageous to include an element of eMBB-awareness for the scheduling decisions of the LLC transmissions (i.e. those that puncture ongoing eMBB transmissions), to primarily puncture eMBB transmission(s) that are transmitted with low modulation and coding scheme index. System level simulations are presented to demonstrate the benefits of the proposed solution.

1 Introduction

Research on the 5G New Radio (NR) is gaining further momentum with the closing of the first Study Item on this subject in 3GPP; see especially the following technical reports [1]- [3]. The ambitions for 5G NR are high, aiming for enhanced support for multiplexing of diverse services such as enhanced mobile broadband (eMBB) and low latency communication (LLC) with ultra-reliability constraints [3]- [5]. Simultaneously fulfilling the requirements for a mixture of users with such diverse requirements is a challenging task, given the fundamental tradeoffs known from communication theory [6]. In that respect, the next generation base station (called gNB) scheduler, which orchestrates the allocation of radio resources to different users, plays an important role. The flexible physical layer design [1] - and especially the agile frame structure design [7] - that comes with the 5G NR offers increased degrees of freedom for the scheduler functionality. Facilitating a shift towards a user-centric approach, where the allocation of radio resources for each user is more flexible, and hence can be better optimized in coherence with the users' diverse QoS requirements. Among others, the 5G NR allows to schedule the users with variable transmission time intervals (TTIs) as proposed also in [7]- [9]. Support for variable TTI sizes facilitates matching the radio resource allocations per user in coherence with their radio conditions and QoS requirements. For instance, to schedule LLC users with short TTIs to achieve low latency, accepting the penalty of higher relative control channel overhead; see the recent studies in [10]- [11] on the benefits of variable TTIs for LLC traffic. Similarly, scheduling with variable TTI sizes also provides

advantages for eMBB traffic [9], as it offers a powerful instrument to efficiently adapt to different offered load conditions and the internet transport protocols (TCP) [12] closed-loop flow control mechanisms.

However, despite the benefits of scheduling the users with variable TTI sizes, there are still some non-trivial problems that call for more studies. One of those is how to efficiently multiplex eMBB and LLC on a downlink shared channel, especially for scenarios where the eMBB traffic is primarily scheduled with long TTI sizes, while sporadic arriving LLC traffic must be scheduled immediately with a short TTI size when such payloads arrive at the gNB to fulfill the corresponding latency deadline. In our effort to address this problem, our hypothesis is that a promising solution is to allow punctured scheduling, where a longer ongoing eMBB transmission can be partly replaced (i.e. punctured) by an urgent short TTI transmission to a user with LLC traffic. The fundamental principle of punctured scheduling has some similarities to preemptive scheduling principles as studied extensively for computer networks to accommodate real-time services [13]. However, despite those commonalities, there are several differences and open questions for how to best design punctured scheduling for a 5G NR wireless system. In particular, we study how to minimize the impact on the eMBB users that are harmed (i.e. by overriding part of their transmission). For this purpose, we propose recovery mechanisms for the impacted eMBB users, and suggest custom designed radio resource management (RRM) optimizations to most efficiently multiplex eMBB and LLC traffic, when utilizing punctured scheduling. The proposed methods are evaluated in a dynamic multi-user, multi-cell. Due to the complexity of the 5G NR system and the addressed problems, we rely on advanced system-level simulations for results generation to have high degree of realism. Those simulations are based on commonly accepted underlying models, calibrated with 3GPP 5G NR assumptions [1]- [3], making sure that statistical reliable results are generated.

The rest of the paper is organized as follows: Section 2 further sets the scene for the study by shortly introducing the system model and presenting the problem formulation and related objectives. The proposed punctured scheduling scheme is outlined in Section 3, and the corresponding RRM considerations in Section 4. The performance analysis appears in Section 5, while concluding remarks are presented in Section 6.

2 Setting the scene

2.1 System model

We adopt the 5G NR assumptions as outlined in [1]- [2], focusing primarily on the downlink performance. Users are dynamically multiplexed on a

shared channel, using orthogonal frequency division multiple access (OFDMA). We assume the setting with 15 kHz subcarrier spacing. LLC UEs are scheduled with short TTI of only 2 OFDM symbols, corresponding to a mini-slot of 0.143 ms. eMBB traffic is primarily scheduled with longer TTI sizes of 14 OFDM symbols (1 ms duration), equivalent to two 7-symbol slots (but could also be scheduled with shorter TTI sizes). In the frequency domain, users can be multiplexed on a physical resource block (PRB) resolution of 12 subcarriers. Users are dynamically scheduled, using a user-centric downlink control channel for transmitting the scheduling grant [7]. This includes informing the users on which resources they are scheduled, which modulation and coding scheme (MCS) is used, etc. Asynchronous hybrid automatic repeat request (HARQ) with Chase combining (as also supported for LTE) is assumed. The system is assumed to carry best effort eMBB traffic download, as well as sporadic LLC traffic. The latter is modeled as bursts of small payload size of B bits that arrive for each LLC user in the downlink direction following a uniform Poisson arrival point process with arrival rate λ. Thus, the offered LLC traffic load per cell equals $N \cdot B \cdot \lambda$, where N is the average number of LLC users per cell.

2.2 Problem formulation and objectives

The objective is to serve the eMBB users with high average data rates (i.e. maximizing the spectral efficiency), while serving the LLC users per their low latency requirement with ultra-high reliability. The LLC traffic takes priority over the best effort eMBB data flows, and needs to be immediately scheduled when it arrives at the gNB. The dilemma, however, is that due to the random unpredictable nature of the LCC traffic, the gNB has no solid a priori knowledge of when LLC traffic arrives, and hence when to reserve radio resources for such transmissions. Reserving radio resources for potentially coming LLC transmissions would be inefficient as it results in capacity loss for the eMBB users. On the other hand, when scheduling the eMBB users, the downlink shared channel will in principle be monopolized by such transmissions, causing unnecessary latency to the LLC users that suddenly have data coming. This is the problem addressed in this study.

3 Punctured Scheduling proposal

3.1 Basic principle

The basic principle of the proposed punctured scheduling solution is shown in Fig. G.1. Here, a UE with eMBB traffic is scheduled by the gNB for transmission on the downlink shared radio channel with a long TTI of 1 ms. The former is facilitated by the gNB sending a scheduling grant (transmitted on

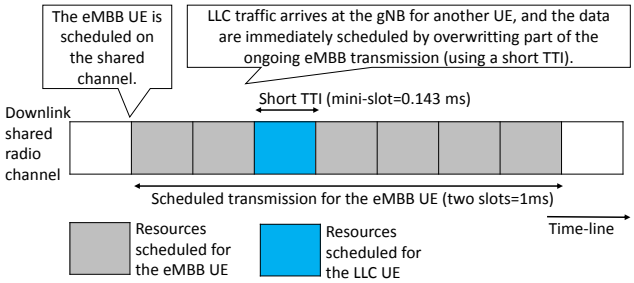


Fig. G.1: Basic principle of downlink punctured scheduling.

the physical layer control channel) followed by the actual transmission of the transport block. During the transmission time of the transport block for the eMBB UE, the shared channel for this transmission is in principle monopolized. However, it may happen that LLC data for another UE arrives at the gNB while the scheduled transmission towards the eMBB UE is ongoing. To avoid waiting for the completion of the transport block transmission to the eMBB UE, we propose to immediately transmit the LLC data by puncturing (i.e. over-riding) part of the ongoing eMBB transmission. The advantage of this solution is that the latency of the LLC data is minimized, at the expense of lower performance of the transmission to the eMBB UE. As some of the resources for the eMBB transmission are corrupted, it essentially results in an error floor, where the performance in terms of block error probability (BLEP) versus SINR for the UE saturates [14]. The impact on the eMBB UE performance from being punctured naturally depends on multiple factors: how many resources have been punctured, whether the eMBB UE is aware of the puncturing, as well as how the information bits for the eMBB transport block (TB) have been encoded, interleaved, and mapped to the physical layer resources [14]. We assume that an eMBB transmission consists of code blocks (CB). The maximum CB size equals Z=6144 bits, and the number of CBs is denoted by C, as for LTE [15]. For the sake of simplicity, we furthermore assume that the CBs are equal size and fully time-frequency interleaved over the assigned resources for the TB.

It should furthermore be noted that the illustration in Fig. G.1 is simple in the sense that the eMBB transmission (in this example) experiences one instance of time-domain puncturing only, by one LLC transmission. However, in a loaded multi-user cellular system, an eMBB transmission may in fact experience puncturing by multiple LLC transmissions, and the LLC transmission(s) may have smaller or larger bandwidth than the eMBB transmission. Aspects of how the gNB select which eMBB transmissions to potentially puncture are further addressed in Section 4 when outlining the assumed scheduling policy.

3.2 Recovery mechanisms

As mentioned, the decoding probability (i.e. 1-BLEP) of the punctured eMBB transmission depends on whether the UE is aware of the puncturing. In line with [14], the performance is improved if the eMBB UE is aware of the puncturing. For an initial eMBB transmission we therefore assume that the UE may be informed of the puncturing. The dilemma here is that the gNB does not know when it schedules the eMBB UE if it will be subject to puncturing later. One option is to allow the gNB to append information of the puncturing (if that happens) to the very last part of the eMBB transmission, i.e. embedded in the last part of the data transmission. However, there is of course the risk that if puncturing does happen, it may happen on the last transmission resources of the eMBB transmission, and hence the indication of the puncturing is lost. If the UE fails to correctly decode a punctured eMBB transmission, a HARQ retransmission is triggered. At this point in time, the gNB knows that the previous transmission was punctured, and hence can inform the UE when scheduling the retransmission by including such information in the downlink scheduling grant. The UE benefits from such information by disregarding the punctured resources of the previous transmission when performing the HARQ soft combining, thereby improving the performance. We assume that HARQ retransmissions consume the same amount of radio resources as the first transmission.

4 Radio Resource Management Algorithms

4.1 Scheduling decisions

For scheduling of the eMBB traffic we assume time-frequency domain radio channel aware Proportional Fair (PF) scheduling, based on periodical frequency selective CQI feedback. The PF scheduling metric $M_{u,p}$ is:

$$M_{u,p}[n] = \frac{r_{u,p}[n]}{R_u[n]}\,, \tag{G.1}$$

where $r_{u,p}$ is an estimate of the instantaneous supported data rate of user u in the p-th PRB, R_u is its average delivered throughput in the past, and n is the discrete time index for the scheduling interval. eMBB users are scheduled with a TTI size of 1 ms. Pending eMBB HARQ retransmissions are prioritized over new eMBB transmissions as also assumed in [16]. By default, the eMBB traffic is scheduled on all available radio resources, assuming there is enough offered eMBB traffic.

When LLC traffic arrives at the gNB, the scheduler aims at immediately scheduling such traffic with a short TTI size of 0.143 ms (corresponding to

2 OFDM symbols). If there are free (unused) radio resources, the LLC traffic is scheduled on those resources. If not, the LLC traffic is scheduled on radio resources currently allocated to eMBB transmissions, i.e. using punctured scheduling. It should be noted that due to the assumed small payload size for the LLC transmissions, only a fraction of the available PRBs are typically needed for each LLC transmission. The question is now which radio resources currently used for eMBB transmissions are the best to puncture? This is a non-trivial question, to which we propose the following punctured-scheduling metric for the LLC users v:

$$M_{v,p}[n] = \frac{r_{v,p}[n]}{R_v[n]} \cdot W_p^{\alpha}[n] \, , \tag{G.2}$$

where W_p is the normalized transport block size of the eMBB user per PRB that is currently scheduled on the PRB p; i.e. the basic PF metric is weighted with a function of the MCS employed for eMBB data transmissions on a given PRB. The exponent α controls how much weight is given to W_p. Based on this scheduling framework, we consider the following three options for punctured scheduling:

- **Best Resources (BR):** $\alpha = 0$. In this case, the pending LLC traffic is scheduled on the PRBs where the LLC users experience the best channel quality as per the CQI feedback. Division of resources among competing LLC users is done following the PF rule.
- **Lowest eMBB user (LeU):** $\alpha = -1$. It is prioritized to schedule the LLC traffic on resources that have been allocated to the eMBB user(s) that use the lowest MCS (among the scheduled eMBB users). The rationale here is that eMBB users with low MCS can better tolerate puncturing.
- **Highest eMBB user (HeU):** $\alpha = 1$. It is prioritized to schedule LLC traffic on resources that have been allocated to eMBB users with highest MCS (among the scheduled eMBB users). The rationale here is to protect the cell-edge eMBB users from experiencing puncturing.

The proposed eMBB-aware scheduling (LeU and HeU) for the LLC transmissions tends to favor puncturing the same eMBB transmission(s) in case several LLC transmissions happens during the same 1 ms TTI interval used for scheduling the eMBB users. Our hypothesis is that the eMBB-aware scheduling options therefore are more attractive.

4.2 Service-specific link adaptation

Dynamic link adaptation (LA) is assumed for both the eMBB and LLC users by setting the MCS for each transmission, based on the users reported CQI.

The MCS for the eMBB users is adjusted to reach an average block error rate (BLER) target of 10%. This is achieved by using the well-known outer loop link adaptation (OLLA) algorithm, where the received CQI values are offset by certain factor (a.k.a. the OLLA offset) calculated in accordance to the received HARQ Ack/Nacks from past transmissions [17]. The OLLA-offset for the eMBB users is only adjusted based on Ack/Nack feedback from eMBB transmissions that have not been punctured; i.e. we only aim at controlling the BLER (10% target) for the eMBB transmissions that do not experience any puncturing. The BLER of the punctured eMBB transmissions will naturally be higher, as the error probability increases with the amount of puncturing.

The LA for the LLC transmissions is conducted to have a BLER target of only 1% to have lower latency. The LA for the LLC users is also conduced based on the users CQI, using standard OLLA to reach the 1% BLER target. Single-stream single-user MIMO transmission is assumed, i.e. benefiting from both transmission and reception diversity against fast fading radio channel fluctuations.

5 Performance Analysis

5.1 Methodology and assumptions

Extensive dynamic system-level simulations are conducted, following the 5G NR methodology in 3GPP [1], [3], assuming a macro-cellular multi-cell scenario. The default simulation assumptions are summarized in Table G.1. All the RRM functionalities described in Section 4 are modeled. Full buffer traffic is used to model the eMBB best effort traffic. A bursty LLC traffic model is used, with 50-byte packets generated following a Poisson arrival process. Different levels of offered LLC traffic load per cell are considered.

Whenever a user is scheduled, the SINR at the receiver is calculated for each subcarrier symbol, assuming a minimum mean square error with interference rejection combining (MMSE-IRC) receiver at the terminal. Inspired by the model in [18]- [19], the SINR values are mapped to the mutual information domain, taking the applied modulation scheme into account. The mean mutual information per coded bit (MMIB) is calculated as the arithmetic mean of the values for the sub-carrier symbols of the transmission [19]. Given the MMIB and the used modulation and coding rate of the transmission, the error probability of a CB is determined from look-up tables that are obtained from extensive link level simulations. For transmissions consisting of more than one CB, we assume identical and independent error performance for all CBs. Thus, the error probability for the transport block is modeled as $P(\mathcal{E}_{TB}) = 1 - (1 - P(\mathcal{E}_{CB}))^{C}$, where $P(\mathcal{E}_{CB})$ is the CB BLEP.

The effect of an eMBB transmission that is punctured is captured as fol-

Table G.1: Summary of default simulation assumptions

Description	Assumption
Environment	3GPP Urban Macro (UMa); 3-sector base stations with 500 meters inter-site distance. 21 cells.
Carrier	10 MHz carrier bandwidth at 2 GHz (FDD)
PHY numerology	15 kHz subcarrier spacing configuration [1].
TTI sizes	0.143 ms for LLC (2-symbol mini-slot). 1 ms for eMBB (two slots of 7-symbols).
MIMO	Single-user 2x2 closed loop MIMO and UE MMSE-IRC receiver.
CSI	Periodic CSI every 5 ms, with 2 ms latency, containing CQI, and PMI.
Data channel modulation and coding	QPSK to 64QAM, with same encoding rates as specified for LTE. Turbo codes.
Link adaptation	Dynamic MCS selection. 1% initial BLER target for LLC 10% initial BLER target for eMBB
HARQ	Asynchronous HARQ with Chase Combining soft combining. The HARQ RTT equals minimum 4 TTIs.
Traffic model	In average 5 full buffer eMBB users per cell. In average 10 LLC users per cell with Poisson arrival of B=50 bytes data bursts.
Scheduling	Proportional fair scheduling of eMBB. Punctured scheduling for LLC traffic following BR, LeU, and HeU.
Link-to-system (L2S) mapping	Based on the mean mutual information per coded bit (MMIB) mapping methodology.

lows: The punctured sub-carrier symbols contain no useful information for the receiver, and hence is modelled as information-less. This effect is included in the calculation of the MMIB and the effective coding rate of the transmission prior to using the look-up tables described above to determine the CB BLEP. In other words, a receiver that is aware of the puncturing incident is assumed to be aware of the exact subcarrier symbols that are punctured. Therefore, the receiver can discard the punctured parts of the physical resources prior to the decoding. Hence, the MMIB for such users is calculated only as the mean from transmission resources that were not punctured and the effective coding rate of the transmission is increased accordingly.

On the other hand, if the UE is unaware of the puncturing the punctured part of the transmission will still be taken as useful signal by the UE thus, used in the decoding process. Therefore, in such scenarios we model the punctured resources as interference only, which decreases the overall MMIB, while keeping the effective coding rate unaffected. The setting in all simu-

lations, except where explicitly mentioned, is that the eMBB UEs are fully aware of the puncturing when it happens.

5.2 Performance results

Fig. G.2 shows the cumulative distribution function (cdf) for the ratio of punctured eMBB resources per user allocation for different offered load conditions of LLC traffic. The ratio of punctured eMBB resources per user allocation is defined as the sum of the sub-carrier symbols allocated to LLC within a given eMBB transmission, divided by the total amount of sub-carrier symbols in the eMBB allocation. As expected, the higher the LLC load, the more eMBB allocations are punctured. At 0.1 Mbps LLC load, only around 10% of the eMBB allocations are punctured, whereas more than 70% puncturing ratio can be observed for a high LLC load of 2 Mbps. The scheduling scheme also impacts the distribution. The LeU and HeU schemes tends to favor puncturing the same eMBB transmission multiple times, in case several LLC transmissions happen during the same 1 ms TTI interval. This results in fewer eMBB allocations being punctured as compared to BR, at the expense of higher puncturing ratio for those transmissions. It is observed that with relatively high probability, the puncturing ratio is either $1/7$ (~ 0.14) or $2/7$ (~ 0.28), which corresponds to the case when LLC allocations puncture the entire frequency sub-band of a certain eMBB allocation with one or two short TTI transmissions.

Fig. G.3 pictures the average decoding probability of eMBB transmissions for different puncturing ratios, including the case where the eMBB UEs

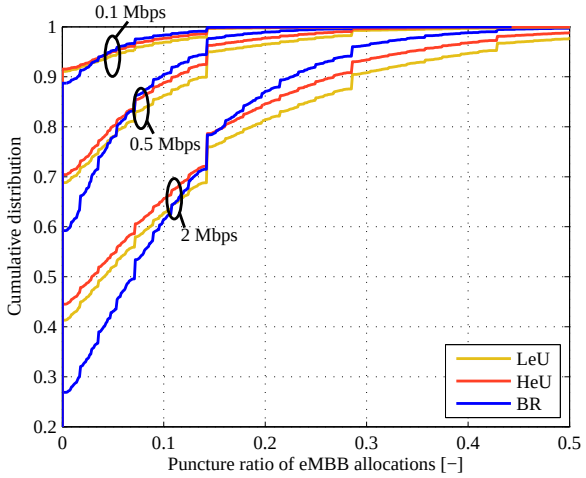


Fig. G.2: Cdf of the ratio of punctured resources per eMBB user allocation.

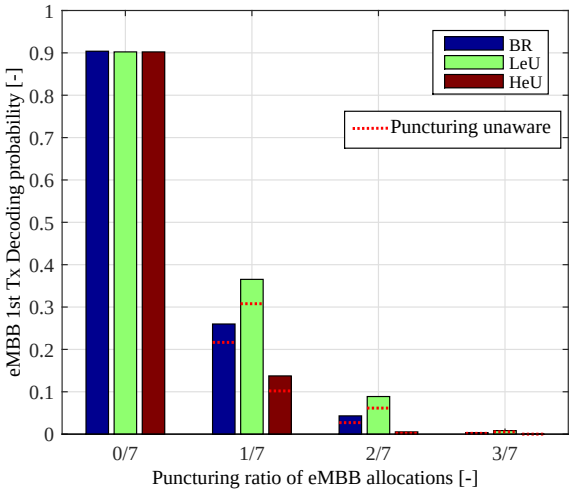


Fig. G.3: Average decoding probability of eMBB transmissions with different puncturing ratio.

are not made aware of the puncturing. eMBB transmissions without any puncturing achieve the 90% decoding probability (or 10% BLER) in line with the service-specific LA setting described in Section IV-B. For cases where an eMBB allocation is punctured on 1/7 of the resources, the decoding probability drastically decreases. It is observed that LeU scheme achieves the best decoding performance, as users with low MCS (typically low coding rate) can better tolerate puncturing. The BR scheme tends to equally affect cell-edge and cell-center eMBB UEs. This results in a decoding probability which is in between what is observed for the other two scheduling schemes. As expected, there is some gain from making the eMBB UE aware of the puncturing (indicated by the dashed line on Fig. G.3); this gain is, however, generally lower than what is reported in [14], although we still observe a clear benefit of making eMBB UEs aware of the puncturing. The reason for observing differences in performance gain of having such puncturing awareness at the eMBB UEs, is expected to be due to the abstract L2S model applied in our system-level study, while findings in [14] are based on more detailed link level simulations. For eMBB users that experience extensive puncturing of 3/7, there is no visible gain by making the eMBB aware of the puncturing. This is because such a large fraction of "lost" resources can anyway not be compensated at the receiver end, and hence is likely to result in failed decoding independent of whether the UE is made aware of it, or not.

The impact on the eMBB performance from the puncturing is shown in Fig. G.4, where a cdf of the eMBB throughput is plotted. It is observed that the eMBB throughput generally declines as LLC traffic is increased, due

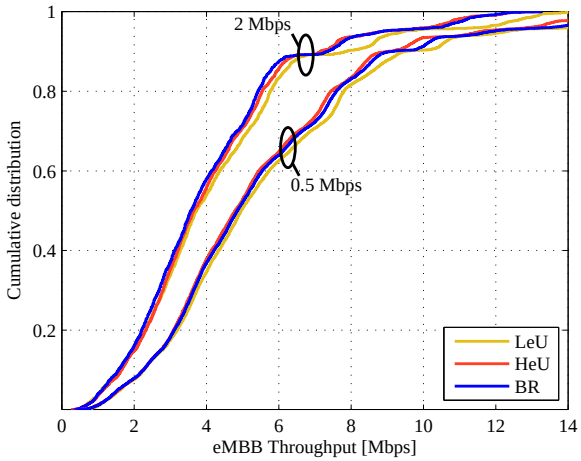


Fig. G.4: Cdf of experienced eMBB user throughput for two different offered loads of LLC traffic.

to more puncturing. The LeU scheme generally offers the best throughput performance. This is due to the larger robustness against puncturing, as also observed in Fig. G.3. The difference in performance between the BR and HeU schemes depend on the offered LLC load. For a LLC offered load of 0.5 Mbps, the performance is as expected: the BR scheme performs better than the HeU, especially in the upper part of the distribution, as HeU favors puncturing of users with high MCS. However, for a higher LLC offered load of 2 Mbps, HeU performs slightly better than BR. This is due to the benefits of concentrating the LLC puncturing in only a few eMBB allocations. Such gain is especially relevant at high LLC load, when the eMBB allocations are more likely to experience puncturing from multiple LLC users.

Fig. G.5 shows the complementary cdf (ccdf) of latency statistics for the LLC traffic for different load conditions. The service-specific 1% BLER target for LLC transmissions is clearly observed in form of a HARQ delay. Looking at the achievable LLC latency at the 10-5 percentile, it is observed that the 1 ms latency requirement for 5G is achieved for both 0.1 Mbps and 2 Mbps load of LLC traffic. Furthermore, no significant difference between the three proposed schedulers is observed at the 10-5 level. One of the reasons is that the proposed puncturing scheduling schemes partly accounts for the experienced channel quality of LLC users. Also, sufficiently good channel quality is experienced across the whole frequency band due to the high diversity from using 2x2 closed-loop single-stream MIMO with MMSE-IRC receiver at the UE.

Finally, Fig. G.6 shows the 50%-ile eMBB throughput (left axis) and the 99.999%-ile latency for the LLC traffic (right axis), for different offered loads

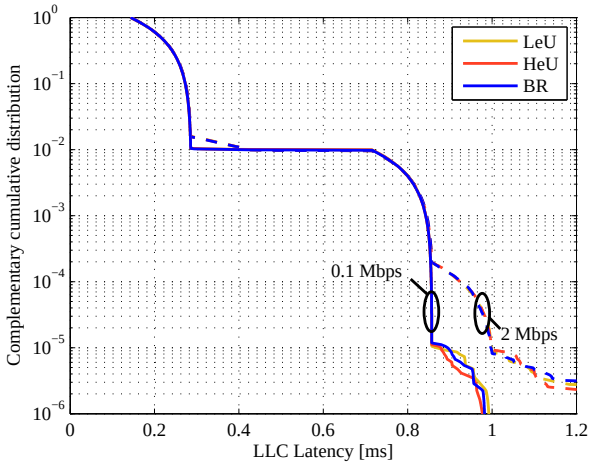


Fig. G.5: Latency distribution (ccdf) of the LLC traffic.

of LLC traffic. Only the LLC latency with the BR scheme is shown as there is marginal difference in performance (as seen in Fig. G.5). The eMBB throughput follows the trends previously described: The LeU scheme offers the best throughput performance due to larger robustness to puncturing. At low LLC offered load, the BR scheme performs better than the HeU, as HeU favors puncturing of users with high MCS (sensitive to puncturing). However, at high LLC offered load, HeU performs slightly better than BR. This is due to the benefits of concentrating the LLC puncturing in only a few eMBB allocations - especially relevant at high LLC load, when the eMBB allocations are more likely to experience puncturing from multiple LLC users.

6 Concluding remarks

In this paper we have presented a punctured scheduling solution, tailored to efficient transmission of urgent LLC traffic on a shared channel with eMBB transmissions. The scheme does not require any pre-reservation of radio resources for transmission of the randomly arriving LLC payloads. Mechanisms to have such solutions perform efficiently are proposed. Those include recovery mechanisms for the eMBB transmissions that experience puncturing, service-specific and puncturing-aware dynamic link adaptation, as well as eMBB-aware scheduling decisions for LLC traffic to minimize the capacity loss for eMBB due to LLC traffic. The presented system-level performance results document the benefits of such solutions, confirming our hypothesis that punctured scheduling (sometimes referred to as preemptive scheduling)

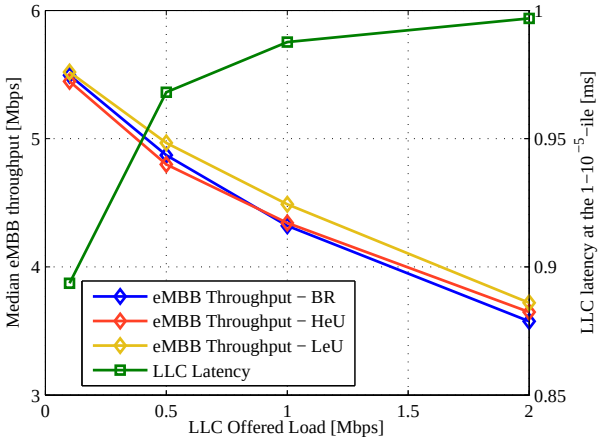


Fig. G.6: 50%-ile MBB throughput (left axis), and 99.999%-ile latency for LLC traffic (right axis).

is attractive and worth pursuing in the design of a 5G multi-service systems.

However, despite of those findings, there are still more options and enhancements for punctured scheduling that are worth studying. Among others, we are currently studying the case where so-called variable block-length HARQ retransmissions are applied, only retransmitting the damaged part of the punctured eMBB transmissions.

Acknowledgment

Part of this work has been performed in the framework of the Horizon 2020 project ONE5G (ICT-760809) receiving funds from the European Union. The authors would like to acknowledge the contributions of their colleagues in the project, although the views expressed in this contribution are those of the authors and do not necessarily represent the project.

References

[1] 3GPP Technical Report 38.802, “Study on New Radio Access Technology Physical Layer Aspects", March 2017.

[2] 3GPP Technical Report 38.801, “Study on New Radio Access Technology: Radio Access Architecture and Interfaces", March 2017.

[3] 3GPP Technical Report 38.913, “Study on Scenarios and Requirements for Next Generation Access Technologies", March 2016.

[4] IMT Vision – “Framework and overall objectives of the future development of IMT for 2020 and beyond", International Telecommunication Union (ITU), Document, Radiocommunication Study Groups, February 2015.

[5] E. Dahlman, et.al., “5G Wireless Access: Requirements and Realization", *IEEE Communications Magazine - Communications Standards Supplement*, December 2014.

[6] B. Soret, et.al., “Fundamental tradeoffs among reliability, latency and throughput in cellular networks", *IEEE Proc. Globecom*, December 2014.

[7] K. I. Pedersen, “A Flexible 5G Frame Structure Design for Frequency-Division Duplex Cases", *IEEE Communications Magazine*, pp. 53-59, March 2016.

[8] Q. Liao, P. Baracca, D. Lopez-Perez, L. G. Giordano, “Resource Scheduling for Mixed Traffic Types with Scalable TTI in Dynamic TDD Systems", *IEEE Proc. Globecom*, Dec. 2016.

[9] K.I. Pedersen, M. Niparko, J. Steiner, J. Oszmianski, L. Mudolo, S.R. Khosravirad, “System Level Analysis of Dynamic User-Centric Scheduling for a Flexible 5G Design", *IEEE Proc. Globecom*, Dec. 2016.

[10] G. Pocovi, B. Soret, K.I. Pedersen, P.E. Mogensen, “MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks", *IEEE Proc. ICC (workshop)*, June 2017.

[11] G. Pocovi, K. I. Pedersen, B. Soret, M. Lauridsen and P. Mogensen, “On the Impact of Multi-User Traffic Dynamics on Low Latency Communications", *Proc. International Symposium on Wireless Communication Systems (ISWCS)*, September 2016.

[12] A. S. Tanenbaum, “Computer networks”, fifth edition, Prentice Hall, 2011.

[13] G. C. Buttazzo, M. Bertogna, G. Yao, “Limited Preemptive Scheduling for Real-Time Systems: A Survey", *IEEE Trans. on Industrial Informatics*, vol. 9, no. 1, pp. 3-15, Feb. 2013.

[14] Technical contribution to 3GPP, Document R1-1700374, “Downlink Multiplexing of eMBB and URLLC Transmissions", Intel Corporation, January 2017.

[15] 3GPP TS 36.212, “Evolved Universal Terrestrial Radio Access (EUTRA); Multiplexing and channel coding", January 2017.

[16] H. Holma and A. Toskala (Editors), “LTE Advanced: 3GPP Solution for IMT-Advanced", John Wiley & Sons Ltd, 2011.

[17] A. Pokhariyal, et.al., “HARQ Aware Frequency Domain Packet Scheduling with Different Degrees of Fairness for UTRAN Long Term Evolution", *IEEE Vehicular Technology Conference (VTC-Spring)*, May 2007.

[18] K. Brueninghaus, et.al, “Link performance models for system level simulations of broadband radio access systems", *IEEE Proc. Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 2306-2311, Sept. 2005.

[19] R. Srinivasan, J. Zhuang, L. Jalloul, R. Novak, and J. Park, “IEEE 802.16m evaluation methodology document (EMD)", in *IEEE 802.16 Broadband Wireless Access Working Group*, Tech. Rep. IEEE 802.16m-08/004r2, `http://ieee802.org/16/tgm/docs/80216m-08_004r2.pdf`, July 2008.

Paper H

Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network Implementations

Klaus I. Pedersen, Guillermo Pocovi, Jens Steiner, Andreas Maeder

The paper has been submitted to the
IEEE Communications Magazine, 2017.

Abstract

In this article, we present a holistic overview of the agile multi-user scheduling functionality in 5G. An end-to-end (E2E) perspective is given, including the enhanced Quality-of-Service (QoS) architecture that comes with 5G, and the large number of scheduling related options from the new access stratum sub-layer, Medium Access Control (MAC), and PHYsical (PHY) layer. A survey of the 5G design agreements from the recently concluded 5G Study Item in 3GPP is presented, and it is explained how to best utilize all these new degrees of freedom to arrive at an agile scheduling design that offers superior E2E performance for a variety of services with highly diverse QoS requirements. Enhancements to ensure efficient implementation of the 5G scheduler for different network architectures are outlined. Finally, state-of-the-art system level performance results are presented, showing the ability to efficiently multiplex services with highly diverse QoS requirements.

1 Introduction

An impressive amount of research related to the upcoming 5G have been published during recent years; as an example, see the survey in [1]. This has formed a solid foundation for progressing also with the 3GPP standardization of 5G, which has recently achieved an important milestone with the completion of the 5G New Radio (NR) Study Item, as captured in the technical reports [2] and [3]. Although 3GPP has adopted the naming NR, we will use the term "5G" throughout this article. The 5G system is set to deliver superior performance for three main service categories: enhanced mobile broadband (eMBB), ultra-reliable low latency communication (URLLC), and massive machine type of communication (mMTC). In achieving the 5G targets, the packet scheduler plays an important role. Both in terms of fulfilling the End-to-End (E2E) Quality-of-Service (QoS) performance targets for each session, as well as in efficiently multiplexing and orchestrating a large number of sessions with highly diverse QoS requirements in one unified system. Specifically, efficient scheduling of URLLC traffic represents a challenging problem, as URLLC is associated with a strict latency target of only 1 ms from the time a packet is delivered to Layer 3/2 in the 5G Radio Access Network (RAN) until it is successfully received, with an outage probability of only 10^{-5}. In addition to the multi-service dimension, the 5G design is also set to scale to a variety of different network implementations.

In this article, we first present a survey of the most important design decisions made in 3GPP that relates to the 5G scheduler design, and in particular to the E2E service delivery capabilities. We explain the rationales behind those design choices, and offer additional insight into how to most efficiently utilize and benefit from the large degrees of freedom that the new 5G de-

sign is set to provide. The new QoS architecture for 5G is presented, highlighting the possibilities of enhanced high-layer scheduling functionality at a new access stratum sub-layer that works in harmony with the advanced Medium Access Control (MAC) layer scheduler, which sits closer to the radio interface; often referred to as the radio scheduler. The 5G radio scheduler comes with many new innovations, especially enabled by the flexible physical layer design. We aim at explaining how the radio scheduler can take advantage of the enhanced 5G physical layer design. Examples of system level performance are presented for two different use cases to illustrate how the 5G scheduler offers performance improvements. To conclude the study, the different enhancements that contribute to the flexible and responsive 5G scheduler design are summarized in Table H.1 at the end of the article. In line with the 3GPP terminology, we refer to terminals as user equipment (UE) and base station as "gNB" (fifth generation Node-B).

2 QoS control and protocol framework

The 5G design includes a new QoS service architecture (as compared to LTE), and several enhancements to the protocol stack [2]. This is illustrated in Fig. H.1, where the QoS architecture is pictured on the left and the user plane protocol stack on the right. The non-access stratum (NAS) filters the data packets in the UE and the 5G core network (CN) to associate the data packets with QoS flows. One or more QoS flows are associated to an E2E session, which is capable of transporting IP, Ethernet, or unstructured datagrams (the latter e.g. for raw machine-type-communication data). For each UE, at least one packet session is established. The access stratum (AS) mapping in the UE and the 5G RAN associates the QoS flows with the data radio bearers (DRBs). This mapping is based on 5G QoS class indices (5QI) in the transport header of the packets, and on corresponding QoS parameters, which are signaled via CN interface when a packet session is established. As illustrated in Fig. H.1, one or more QoS flow(s) can be mapped to a DRB. Hence, the 5G CN and RAN ensures the QoS in harmony by intelligent mapping of QoS flows and DRBs, essentially constituting a two-step mapping of E2E session flows (e.g., IP-flows) to QoS Flows and subsequently to DRBs.

In the 5G RAN at least one default DRB is established for each UE when a new E2E packet session is created. As illustrated in Fig. H.1, an E2E packet session may be mapped to two different QoS flows and DRBs to facilitate cases where the E2E packet session contains data flows with two different sets of QoS requirements; such as e.g. a website with embedded high-definition live streaming video. The 5G RAN may choose to e.g. map a guaranteed bit rate (GBR), or multiple GBR flows to the same DRB. The mapping of E2E session to QoS flows, and DRBs can be updated dynami-

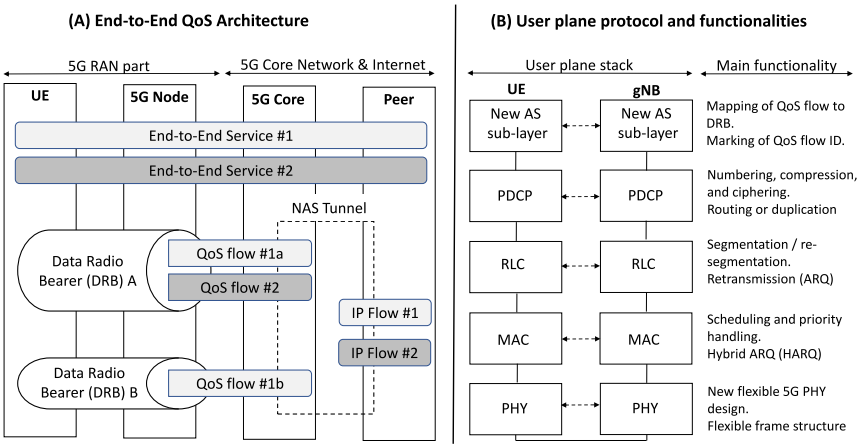


Fig. H.1: QoS architecture, user plane protocol stack, and related functionalities.

cally. This kind of flexibility opens opportunities for applying state-of-the-art higher-layer scheduling policies that differentiate application flows, via the mapping to DRBs, as well as adaptation of DRB requirements for the radio scheduler. The latter mechanisms are also sometimes referred to as higher-layer application-aware scheduling [12], or advanced Quality of Experience (QoE) management [4].

On the terminal side, the concept of reflective QoS eliminates the need to use dedicated flow filters signaled by the network to match traffic to QoS flows. This was one of the main reasons why in LTE, IP traffic was always mapped to default DRBs. In reflective QoS, the terminal derives the mapping of uplink traffic to QoS flows by correlating the corresponding downlink traffic and its attributes, e.g. in Transport Control Protocol (TCP) flows.

On the 5G radio interface, the packet treatment is defined separately for each DRB. Different DRBs may be established for QoS flows requiring different packet forwarding treatment (e.g. associated with different requirements such as latency budget, packet loss rate tolerance, GBR). As will be described in greater detail later, the MAC-level scheduler aims at fulfilling the requirements for the users DRBs, as well as to prioritize accordingly if the system reaches congestion where requirements for all users cannot be simultaneously fulfilled.

The user-plane protocol stack for 5G is illustrated in the right part of Fig. H.1. Here, a new AS sub-layer (with Service Data Application Protocol) is included that is responsible for the aforementioned mapping of QoS flows to DRB and the related marking in uplink. The proposed QoE Manager in [4], can be implemented in this sub-layer. As an example, for the use case of YouTube streaming, the QoE manager in the AS sub-layer may adaptively

monitor and adjust the mapping of QoS flow to DRB, adjusting e.g. the GBR and Latency budget associated with the DRB to guide the lower layer radio scheduler, and ensure a positive end-user experience where the playout of the video starts quickly, and runs smoothly without any re-buffering events (for more details we refer to [4]). For the majority of cases, it is envisioned that all traffic for a UE is mapped to a single (or few) DRB, while the QoE manager at the AS sub-layer takes care of differentiation; e.g. by modifying packet priorities, while only seldomly modifying the QoS parameters of the DRB that the MAC scheduler shall fulfill. Thereby, the QoE manager (aka application-layer scheduler) is operated in harmony with the lower layer radio scheduler to avoid the well-known double responsibility conflict problem from control theory, i.e. avoiding that the higher- and lower-layer scheduler in the worst case make colliding decisions that result in undesirable behaviors.

The packet data convergence protocol (PDCP) layer for 5G inherits the fundamentals from LTE, but also brings valuable enhancements [2]. Among those, PDCP packet duplication is supported as a mean to improve the end-user packet reception reliability, thus being one of the enablers for reaching the reliability part of the 5G URLLC requirement. This means that if a UE is configured with e.g. carrier aggregation or multi-node connectivity [1] [12], the same PDCP packet can be duplicated and sent via different transmission paths, thereby reducing the probability of losing packets. Furthermore, PDCP is responsible for packet re-ordering in case that lower layers do not deliver in-sequence. The Radio Link Control (RLC) includes segmentation and Automatic request repeat (ARQ), while the Medium Access Control (MAC) is the home of the agile radio-layer packet scheduler and the Hybrid ARQ functionality. A large set of PHYsical (PHY) enhancements are coming with 5G [3], which offers significant degrees of freedom for the multi-user, multi-service capable radio scheduler (discussed in more details in the coming sections). On a further note, the concept of network slicing is also supported, where different types of traffic could be handled by separate slices. The network may realize different slices by mapping data to different QoS flows/DRBs based on slice-specific policies, and by scheduling. UEs should be able to aid information related to slice selection, if it has been provided by the NAS (for more information on slicing, see [2] [12]).

3 MAC scheduler overview

A high-level overview of the 5G MAC scheduler functionality is pictured in Fig. H.2. The MAC scheduler is the controlling entity for multi-user radio resource allocations, which is subject to several constraints but also many options for efficiently serving the different terminals. The enlarged number of options for the 5G MAC scheduler, as compared to LTE, naturally offers per-

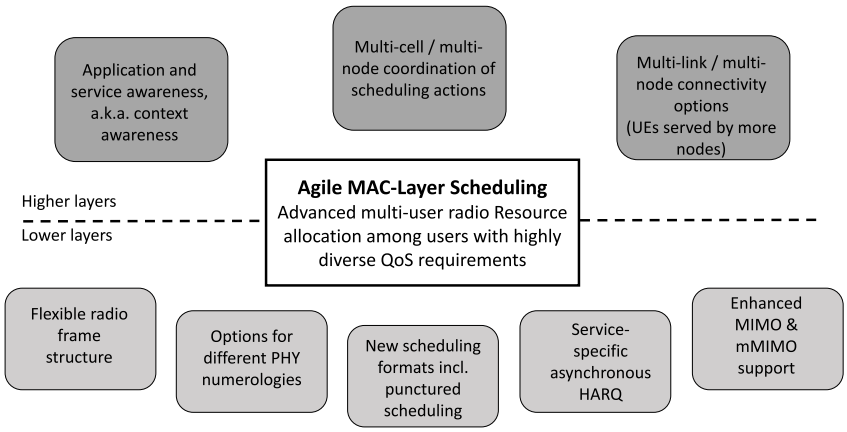


Fig. H.2: High-level overview of MAC dynamic scheduler interfaces and options.

formance improvements, but also presents a non-trivial problem of how to best utilize those degrees of freedom in an efficient manner. The MAC scheduler works by dynamically allocating radio transmission resources (transport blocks) on a per-user basis for downlink and uplink transmissions, separately. The objective of the scheduler is to fulfil the QoS service targets for all the DRBs of the served UEs. This is illustrated in Fig. H.2, where the application and service awareness is provided by the higher layers as discussed in the previous section. Furthermore, the scheduler has to support multi-cell connectivity mode [1] [2], where UEs are configured to be simultaneously served by multiple nodes (and cells). Additionally, there may be other multi-cell coordination constraints; e.g. if enforcing inter-cell interference coordination between neighboring cells where certain radio resources are dynamically muted, and hence not available for dynamic scheduling of users. At the MAC sub-layer, enhanced service-specific HARQ enhancements are included [5]. The 5G PHY layer offers a large set of new options for the MAC scheduler, which enable significant improvements for efficiently multiplexing users with highly diverse service requirements.

Fig. H.3 presents further information on multiplexing of users on the PHY layer and the related MAC sublayer functionality. 5G comes with a new flexible structure, consisting of 10 ms radio frames and 1 ms subframes. The subframes are constructed of slot building blocks of 7 OFDM symbols. For FDD cases, the slots are naturally all downlink (for the downlink band), and all uplink (for the uplink band), while for TDD cases the slots can also be bidirectional (starting with downlink transmission followed by uplink transmission). To support operation in different frequency bands, the PHY numerology is configurable, building on the same base subcarrier spacing (SCS)

of 15 kHz as used in LTE. The SCS can scale from the base value by a factor 2^N, where $N \in [0, 1, 2, 3, 4, 5]$. For 15 kHz SCS ($N = 0$), the slot duration is 0.5 ms, while it equals 0.25 ms for 30 kHz ($N = 1$). Furthermore, mini-slots of 1-3 OFDM symbols are defined as well [3]. The smallest time-domain scheduling resolution for the MAC scheduler is mini-slot, but it is also possible to schedule users on slot resolution, or on resolution of multiple slots (aka slot aggregation). This essentially means that dynamic scheduling with different transmission time interval (TTI) sizes is supported. The later, enables the MAC scheduler to more efficiently match the radio resource allocations for different users in coherence with the radio condition, QoS requirements, and cell load conditions [6-8]. The short TTI size is needed for URLLC use cases [9], but not restricted to such traffic. In the frequency domain, the minimum scheduling resolution is one physical resource block of 12 subcarriers, corresponding to 180 kHz for 15 kHz SCS, 360 kHz for 30 kHz, and so forth [3].

Fig. H.3A shows how different users are multiplexed in the downlink on a FDD carrier (different colors represent transmissions to different users). As can be seen from this example, the majority of the users are multiplexed on slot resolution. Users can be dynamically scheduled with a TTI size of one slot, or multiple slots. For the example in Fig. H.3A, the carrier is configured to allow frequency domain multiplexing of two different PHY numerologies; namely 15 kHz (upper part) and 30 kHz (lower part). The MAC scheduler can freely decide how to schedule its different users on the carriers (i.e. on which PHY numerology, with which TTI sizes, etc.), and it is not visible to the RLC

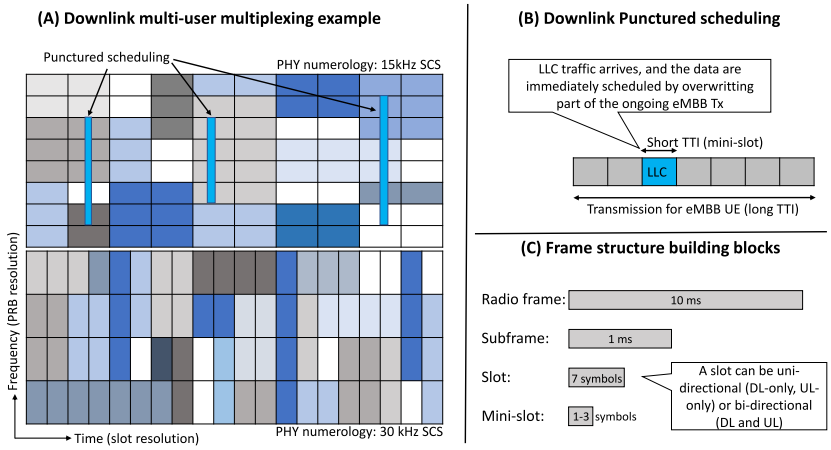


Fig. H.3: Resource allocation framework: (a) downlink multi-user mux, (b) punctured scheduling, and (c) frame structure building blocks.

layer how this is done. It is, however, possible to enforce some restrictions via higher-layer control signaling (radio resource control – RRC) to schedule data from certain DRBs only on a given PHY numerology, and a certain TTI size. Each scheduling allocation (downlink and uplink) is announced to the UE via a PHY downlink control channel carrying the scheduling grant. The downlink control channel is flexibly time-frequency multiplexed with the other downlink PHY channels, and can be mapped contiguously or non-contiguously in the frequency domain. This constitutes a highly flexible design, where the relative downlink control channel overhead can take values from sub-one-percentage values (if e.g. scheduling few users with long TTI size) and up to tens of percentages if scheduling a larger number of users with very short TTI sizes [6] [8]. The design therefore overcomes the control channel blocking problems from LTE (and LTE-Advanced) as reported in [10] [11]. As studied in [6–8], these advantages are achieved by migrating towards a user-centric design with in-resource control channel signalling, as compared to the predominantly cell-centric LTE design. Another advantage brought by the more flexible 5G downlink control channel design is the support of UEs that only operate on a fraction of the carrier bandwidth (e.g. narrowband MTC devices). As is further described in [3], resource allocation for UEs not capable of supporting the full carrier bandwidth is derived based on a two-step frequency-domain assignment process.

Fig. H.3 also illustrates the principle of punctured scheduling for efficient expedition of Low Latency Communication (LLC) traffic. Efficient scheduling of LLC is rather challenging as such traffic is typically bursty (random nature) and requires immediate scheduling with short TTIs to fulfil the corresponding latency budget [9]. Instead of pre-reserving radio resources for LLC traffic bursts (that may, or may not come), it is proposed to use punctured scheduling, which is inherited from the preemptive scheduling ideas known from real-time scheduling in computer networks. The basic principle is as follows [15]: Traffic such as eMBB is scheduled on all the available radio resource (assuming sufficient offered traffic). Once a LLC packet arrives at the gNB, the MAC scheduler immediately transmits it to the designated terminal by overwriting part of an ongoing scheduled transmission, using mini-slot transmission, as illustrated in both Fig. H.3A and H.3B. This has the advantage that the LLC payload is transmitted immediately without waiting for ongoing scheduled transmissions to be completed, and without the need for pre-reserving radio resources for LLC traffic. The price of puncturing is for the user whose parts of its transmission are overwritten. To minimize the impact on the user that experience the puncturing, related recovery mechanisms are introduced [15]. Those include an indication to the victim terminal that part of its transmission has been punctured. This enables the terminal to take this effect into account when decoding the transmission, i.e. it knows that part of the transmission is corrupted. Moreover, options for smart HARQ

retransmission options are considered, where the damaged part of the punctured transmission is first retransmitted. The benefits of those options are further illustrated in the Performance Section.

As illustrated in Fig. H.2, the 5G PHY also offers enhanced antenna techniques – multiple-input-multiple-output (MIMO) schemes [1] [3]. For cases with Single-User (SU) MIMO, this allows to schedule up to eight parallel streams of data to one UE on the same PHY resources. Similarly, enhanced Multi-User (MU) MIMO is supported, where streams towards different users can be scheduled on the same PHY resources. This includes massive MIMO (mMIMO) enhancements, where users can be simultaneous scheduled on different beams; allowing flexible support for implementations with digital beamforming, analog beamforming, and hybrids of those two options. For cases with analog beamforming, the MAC scheduler is typically restricted to only apply time-domain multiplexing between users within each beam, although options for frequency domain multiplexing are not excluded.

4 Flexibility for different network implementations

The 5G scheduler is designed to be applicable for different network implementations [1] [2]. This includes distributed network implementations with separate schedulers implemented in each gNB per cell, as well as more advanced centralized or semi-centralized radio access network solutions [12]. The latter includes cases with a centralized Cloud Edge entity connected via a midhaul interface to a Front End Unit (FEU), which may have RF integrated or may connect to a Remote Radio Head (RRH) via a fronthaul connection [1] [2]. For such advanced network architectures, the implementation of the PDCP and RLC sub-layers is possibly located in the Cloud Edge, while the MAC is distributed over Cloud Edge and FEU, and the PHY is distributed over the FEU and RRH. Given the possible ranges of processing latencies at the different network units, as well as communication latencies over the midhaul and fronthaul interface, the MAC scheduling and the HARQ loop timing of 5G needs to be equally flexible. Thus, in comparison to the strict hardcoded scheduling timing of LTE, 5G offers a much more flexible configuration. The timing between the downlink scheduling and the actual data transmission is indicated as part of the scheduling grant (i.e. on the downlink PHY control channel). The same applies for the timing of uplink data transmissions. The timing relation between the data channel reception, and the time where a corresponding HARQ feedback (positive or negative acknowledgement) shall be sent is also flexibly indicated and configurable. Furthermore, asynchronous HARQ is adopted for both link directions, giving the network full flexibility for deciding when to schedule HARQ retransmissions. See for instance the study in [5] where the HARQ round trip timing is

studied for cases with different fronthaul latencies. The combination of the flexible scheduling timing and asynchronous HARQ is an important enabler that paves the way for supporting cases with decoupled downlink/uplink cell associations, where downlink transmissions are scheduled to the UE from one cell, while uplink transmissions are towards a different cell [13].

Moreover, as compared to the LTE design, the RLC concatenation is replaced with MAC multiplexing, which allows pre-generation and interleaving of PDCP/RLC/MAC headers. This basically means that the time consuming generation of RLC Packet Data Units (PDUs) for each new scheduled transport block (i.e. scheduling instant) as done for LTE is avoided. This makes 5G more efficient and flexible, allowing the RLC and MAC entities to for instance be implemented on different network elements. See more details in [2].

5 Performance results

Performance results from extensive system-level simulations, following the 3GPP 5G simulation guidelines, are presented in the following to illustrate the benefits of some of the 5G scheduling enhancements. Results are presented for a standard three-sector macro scenario, operating at 2GHz with a 10 MHz carrier bandwidth, assuming 2x2 SU-MIMO and the base PHY numerology (15 kHz SCS). We first present downlink eMBB performance results for file download over TCP, using the well-known Reno model [14]. A 2 ms CN delay is assumed from the Client to the 5G RAN. Traffic is arriving according to a homogenous Poisson point process, and users are leaving the system when a download of a 500kB payload is completed. RLC acknowledged mode is assumed. Fig. H.4 shows the performance for short (0.14 ms) and long (1 ms) TTI sizes, considering both the case with low offered traffic and high offered traffic. One of the reported performance metrics is the smoothed round trip time (RTT) of TCP packets in line with the definition in RFC6298. It is observed that the best performance is achieved for the short TTI at the low offered load. This is due to the lower air interface latency that helps to quickly overcome the slow start TCP phase. The higher PHY control channel overhead from operating with short TTIs is not a problem at the low offered load. However, at the high offered load case, the best performance is clearly observed for the case with the long TTI. This is due to the fact that using longer TTIs results in higher average spectral efficiency. If operating with the short TTI size (at high offered load), excessive queuing delays are observed at the gNB due to the lower spectral efficiency because of higher PHY control channel overhead. Thus, the results in Fig. H.4 clearly show the benefit of being able to dynamically adjust the TTI size. See the study in [8] for additional insight.

Next, we present downlink performance for a mixture of eMBB and low

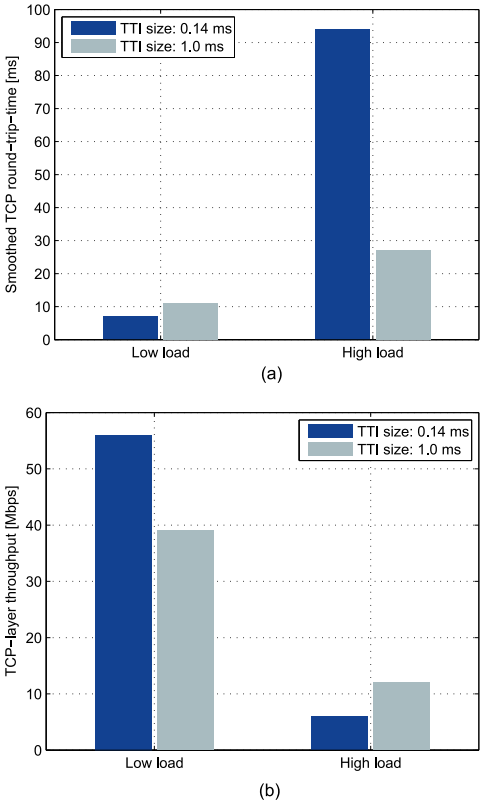


Fig. H.4: Performance of eMBB with different TTI sizes: (a) Smoothed median TCP packet round trip time, (b) median TCP-layer end-user throughput.

latency communication (LLC) type of traffic. In this particular example, there are on average five active eMBB users per macro-cell, performing a download of a 500 kB file size, using TCP. As soon as one of eMBB users finishes its file download, the user is removed and a new one is generated at a random location. In addition, there are on average 10 LLC users per cell, where small latency critical payloads of 50 Bytes are sporadically generated according to a homogeneous Poisson point process, arriving in the gNB. As this scenario corresponds to a fully loaded network, eMBB users are scheduled with a TTI size of 1 ms, using all available PRBs. Hence, no radio resources are reserved for potentially coming LLC traffic. Instead, punctured scheduling is applied whenever LLC payloads appear in the gNB. The LLC payloads are immediately scheduled on arrival with mini-slot resolution (0.14 ms TTI size), overwriting part of the ongoing eMBB scheduled transmissions as also illustrated in Fig. H.3A and H.3B. Due to the urgency of the LLC traffic, we assume RLC transparent mode, and a low initial Block Error Rate (BLER) of

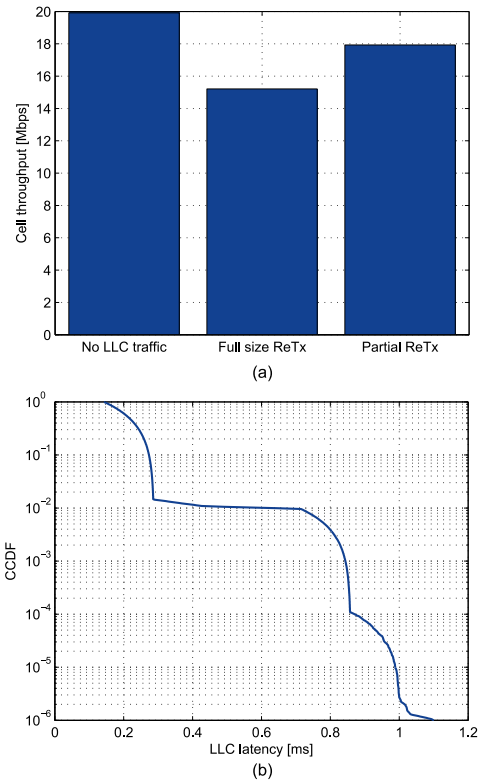


Fig. H.5: Performance of LLC/eMBB with punctured scheduling. (a) average cell throughput (b) ccdf of LLC payload latency.

only 1% for such transmissions to avoid too many HARQ retransmissions. The average cell throughput is illustrated in Fig. H.5A, where the performance is shown for cases with/without LLC traffic. For the cases with LLC traffic, the offered load is such that approximately 12% of radio resources are used for LLC. Two sets of results are shown for the case with LLC traffic: One for the case where the full transport block is retransmitted for failed eMBB HARQ transmissions, and a case where only the damaged part of the eMBB transmission that has been subject to puncturing is retransmitted (labelled as partial retransmission in Fig. H.5A). As observed from Fig. H.5A, the latter option is clearly the most promising solution as fewer radio resources for HARQ retransmissions of eMBB transmissions that have suffered from puncturing are used. However, the cost of using this approach is a slightly larger latency for the eMBB users, as the probability of triggering a second HARQ retransmission is higher, as compared to the case where the first HARQ retransmission includes the full transport block. Fig. H.5B shows the complementary cumulative distribution function (ccdf) of latency of LLC

traffic. The latency is measured from the time where the LLC payload arrives at the gNB until it is correctly received by the UE. The ccdf shows that even under the considered full load conditions, the performance for the LLC traffic fulfills the challenging URLLC target of 1 ms latency with an outage of only 10-5. Hence, the punctured scheduling scheme fulfills its purpose; being able to efficiently schedule the LLC traffic in line with its challenging latency and reliability constraints, while still having efficient scheduling of eMBB traffic without the need for pre-reservation of radio resources for sporadic LLC traffic. Further radio resource management considerations for punctured scheduling are presented in [15].

6 Summary

In this article we have presented an extensive survey of the broad family of packet scheduling related improvements that comes with the new 5G system. Those enhancements and their related benefits are summarized in Table H.1. In short, a new end-to-end QoS architecture is envisioned that offers improved opportunities for application-layer scheduling functionality to ensure satisfactory QoE. The latter works in harmony with the lower-layer agile MAC scheduler. The MAC scheduler comes with a large number of options, primarily offered by the highly flexible PHY design of the 5G new radio; including scheduling with dynamic TTI sizes, flexible timing, different PHY numerologies, new paradigms such a punctured scheduling, etc. In conclusion, the 5G system design, and particularly the scheduler related mechanisms at the different layers, offer opportunities for improved E2E performance, capabilities for more efficiently multiplexing users with highly diverse QoS requirements, and flexibility for different network implementations. System-level performance results confirm that the new scheduling functionalities offer promising benefits.

References

[1] C. Sexton, N. Kaminski, J. Marquez-Barja, N. Marchetti, and L. A. DaSilva, “5G: Adaptable Networks Enabled by Versatile Radio Access Technologies", *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 688-720, 2017.

[2] 3GPP Technical Report (TR) 38.804, “Study on New Radio Access Technology; Radio Interface Protocol Aspects (Release 14)", March 2017.

[3] 3GPP Technical Report (TR) 38.802, “Study on New Radio (NR) Access Technology Physical Layer Aspects", March 2017.

[4] B. Héder, P. Szilágyi, C. Vulkán, “Dynamic and Adaptive QoE Management for OTT Application Sessions in LTE", *IEEE Proc. International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, Sept. 2016.

[5] S. Khosravirad, G. Berardinelli, K.I. Pedersen, F. Frederiksen, “Enhanced HARQ design for 5G wide area technology", *IEEE Proc. VTC-2016-Spring, 5G air interface workshop*, May 2016.

[6] K.I. Pedersen et.al., “A Flexible 5G Frame Structure Design for Frequency-Division Duplex Cases", *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53-59, March 2016.

[7] Q. Liao, P. Baracca, D. Lopez-Perez, L.G. Giordano, “Resource Scheduling for Mixed Traffic Types with Scalable TTI in Dynamic TDD Systems", *IEEE Proc. Globecom*, December 2016.

[8] K.I. Pedersen, M. Niparko, J. Steiner, J. Oszmianski, L. Mudolo, S.R. Khosravirad, “System Level Analysis of Dynamic User-Centric Scheduling for a Flexible 5G Design", *IEEE Proc. Globecom*, December 2016.

[9] G. Pocovi, B. Soret, K.I. Pedersen, P.E. Mogensen, “MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks", *IEEE Proc. ICC (workshop)*, June 2017.

[10] D. Laselva, et.al., “On the impact of realistic control channel constraints on QoS provisioning in UTRAN LTE", *IEEE Proc. VTC 2009 fall*, September 2009.

[11] A.K. Talukdar, “Performance evaulation of the Enhanced Physical Downlink Control Channel in a LTE network", *IEEE Proc. PIMRC*, pp. 987-991, September 2013.

[12] A. Maeder, et.al., “A Scalable and Flexible Radio Access Network Architecture for Fifth Generation Mobile Networks", *IEEE Communications Magazine*, vol. 54, no. 11, pp. 16-23, November 2016.

[13] F. Boccardi, J. Andrews, H. Elshaer, M. Dohler, S. Parkvall, P. Popovski and S. Singh, “Why to decouple the uplink and downlink in cellular networks and how to do it", *IEEE Communications Magazine*, vol. 54, no. 3, pp. 110-117, March 2016.

[14] J. Padhye, et.al., “Modeling TCP Reno Performance: A Simple Model and Its Empirical Validation", *IEEE/ACM Trans. on Networking*, vol. 8, no. 2, pp. 133-145, April 2000.

[15] K.I. Pedersen, G. Pocovi, J. Steiner, “Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband", to appear in *IEEE Proc. Vehicular Technology Conference*, September 2017.

Table H.1: Summary of the agile 5G multi-service scheduling related functionalities.

Functionality	Summary	Benefit
New end-to-end QoS architecture	User data packets are mapped to QoS flows at the UE and CN. UE and RAN maps the QoSs flows to DRBs. DRBs carry QoS flow(s) over the radio interface. QoS differentiation inside NG3 connection is based on packet based QoS flows. Mapping relationship between sessions and DRB is 1 to N and between QoS flows and DRBs N to N. Application scheduler (aka QoE Manager) in the 5G RAN can differentiate application flows, via the recommended mapping to DRBs.	Improved end-to-end QoS control and orchestration.
Packet duplication	Duplication solution for CA and multi-node connectivity cases can use PDCP duplication, so duplicated PDCP packets are sent over different carriers. Supported for both link directions.	Improved RAN reliability.
DRB mapping to PHY	Data from a DRB can be mapped to one or more lower layer PHY numerologies and TTI sizes. It is transparent to the RLC which PHY / TTI is used. The DRB to lower layer PHY mapping can, however, be reconfigured via higher layer RRC reconfigurations.	Full flexibility for optimization of per data flow.
MAC layer concatenation	Replacing RLC concatenation with MAC Multiplexing allows pre-generating and interleaving PDCP/RLC/MAC headers with the respective data blocks. Thereby overcoming the time-consuming on-the-fly generation of RLC packet data units (PDUs) for each new scheduling grant as done for LTE.	Optimized PDU generation, offering higher degrees of freedom for network implementations where e.g. the RLC and MAC is implemented on different hardware units.
Flexible scheduling timing	Timing between DL scheduling grant and corresponding DL data transmission is indicated as part of the scheduling grant (PHY control channel). Timing between UL scheduling assignment and corresponding UL data transmission is indicated as part of the scheduling grant (PHY control channel). Timing between DL data reception and corresponding HARQ ACK/NACK is indicated as part of the scheduling grant (PHY control channel).	Flexible timing for different network implementations, e.g. cloud RAN with different fronthaul latencies and processing time capabilities.
HARQ characteristics	Asynchronous HARQ for both link directions. Support for specific HARQ enhancements such as automatic retransmissions (low latency) and multi-bit HARQ feedback to enable variable block HARQ retransmissions (mainly relevant for large transport block size eMBB transmissions).	Increased timing and scheduling flexibility. Optimized resource efficiency for retransmissions.

Control channel Flexibility	The control channel carrying the scheduling grant (NR-PDCCH) can be flexible time-frequency multiplexed with the other downlink PHY channels. NR-PDCCH can be mapped contiguously or non-contiguously in frequency. Resource allocation for data transmission for a UE not capable of supporting the full carrier bandwidth can be derived based on a two-step frequency-domain assignment process.	Scalable solution, where known problems of control channel blocking from LTE are circumvented.
Variable TTI sizes	Dynamic scheduling with variable TTI sizes is supported. The TTI size can equal one mini-slot, a slot, or multiple slots. The time-duration of mini-slots and slots depends on the chosen PHY numerology. The slot length equals 0.5 ms for 15 kHz SCS, 0.25 for 30 kHz SCS, and so forth.	Reduced latency, and increased flexibility for scheduling in coherence with the users' QoS requirements and RAN conditions. Essential in reaching ultra-reliability targets.
Punctured / preemptive scheduling	Allows to quickly schedule an urgent latency critical payload with a short TTI that overwrites another ongoing downlink scheduling transmission. The concept includes efficient recovery mechanisms, where penalty for the victim UE that experiences overwriting of its transmission is minimized.	Efficient downlink scheduling of sporadic low latency traffic without reserving transmission resources in advance.
PHY numerologies	Configurable PHY numerology with base subcarrier spacing (SCS) of 15 kHz (as in LTE), which can be scaled with 2^N, where $N \in [0, 1, 2, 3, 4, 5]$ for the first 5G NR specs. A cell can be configured to have multiplexing of different PHY numerologies (requires appropriate guard intervals between those).	Scalability to larger frequency ranges and different deployments.
MIMO / Beamforming	SU-MIMO: Support for at least up to eight streams. Can schedule new transmissions and HARQ retransmission on different streams. MU-MIMO: Can schedule up to N users on the same time-frequency resources, either on completely overlapping resources for all N users, or on partly overlapping resources. mMIMO Grid-of-Beams (GoB): Can schedule multiple users on the same time-frequency resources on different beams. For cases with analog beamforming, primarily rely on wideband scheduling of users with time-domain multiplexing per beam.	Improved capacity and coverage.

Part V

Conclusions

Conclusions

1 Summary of the Main Findings

The support of Ultra-Reliable Low-Latency Communications (URLLC) calls for a broad set of enhancements in the different components of the communication system. The methodology adopted in this work consisted on a progressive approach, where the level of complexity and accounted functionalities and effects were gradually increased. In this respect, the study was structured into three main parts. The first part of the thesis focused on investigating the potential of spatial diversity and interference management techniques to combat the harmful effects of the wireless channel. Specifically, the very low percentiles of the signal to interference-and-noise ratio (SINR) distribution were analysed in a multi-user and multi-cell cellular network, assuming full-load conditions, *one-shot* downlink transmissions, and without explicit modelling of the URLLC traffic and scheduling mechanisms.

Microscopic and macroscopic diversity techniques were shown to be one of the main enablers of ultra-reliable communications. Higher diversity order results in steeper slopes of the SINR distribution, hence, substantial SINR gains were observed at the very low percentiles. Interference management in the form of ideal interference cancellation of the strongest(s) interferers or frequency reuse schemes, provided lower but still valuable gains at the 10^{-5}-th percentile of the distribution. Based on the obtained results, a 4x4 MIMO scheme with second order macroscopic diversity was proposed as the most feasible configuration for achieving the 99.999% reliability requirements for a 3GPP urban macro network. The performance was also evaluated in a site-specific network model based on a realistic deployment in an European capital, and including real base station and mobile user positions, and ray-tracing-based propagation maps. For this case, lower gains were generally observed from the studied techniques due to the more irregular deployment and propagation characteristics.

Motivated by the unprecedented reliability requirements, various sources of imperfections and instability were also studied. The considered microscopic diversity schemes require uplink feedback containing the preferred

precoding matrix indicator. In this regard, it was shown that, even for uplink feedback error probabilities as high as 10^{-2} (i.e. three orders larger than the URLLC outage target), there is a benefit of using closed-loop MIMO schemes as compared to feedback-less open-loop transmission modes. In addition, a stochastic model for studying the impact of spatially correlated and uncorrelated base station failures was presented. Results showed that uncorrelated failures are not so critical, whereas geographically correlated failures can have a large performance impact for failure probabilities above 0.5%.

The second part of the study accounted for time-domain aspects. The multiple delay components that contribute to the communication latency were identified and carefully studied in a simplified system-level setting with dynamic user arrivals. The analysis revealed that, even at relatively low loads of low-latency traffic, there is a benefit of optimizing the system for high spectral efficiency, rather than for low latency, in order to minimize the probability of experiencing queuing delays at the cell buffers.

Based on these learnings, various radio resource management techniques were proposed as a way to fulfil the 1 ms and 99.999% reliability requirements of URLLC. In this third part of the study, the proposed solutions were evaluated via highly-detailed system level simulations, which reflected the combined performance of the large set of functionalities typically present in real systems. Among others, the importance of using a short transmission time interval (TTI) of 0.143 ms and sufficiently-short HARQ round-trip time, to allow one retransmission within the 1 ms latency budget, was shown. The latter significantly relaxes the block error rate constraints that the URLLC transmissions need to fulfil, resulting in a much more efficient transmission of the URLLC payloads and lower probability of experiencing queuing delays. Also, an attractive CQI measuring procedure was proposed to reduce the link adaptation inaccuracies as a consequence of the rapid interference variations. This technique significantly reduced the 99.999%-percentile of the URLLC latency, especially at medium to high load when more intermittent interference is experienced in the network. When applying these enhancements, the URLLC requirements were fulfilled for relatively high load of URLLC traffic (up to 40% resource utilization in some of the cases).

Efficient multiplexing the sporadically-arriving URLLC traffic with conventional enhanced Mobile Broadband (eMBB) traffic is another hard problem studied. In this respect, it was observed that eMBB traffic can significantly degrade the URLLC latency and reliability performance. This is a consequence of the larger inter-cell interference, which reduces the signal quality of URLLC users and increases the queuing delay and transmission time of the payloads. To address this challenge, a resource allocation technique was proposed which provides dynamic adjustment of the block error rate (BLER) target of URLLC transmissions in accordance with the instantaneous load experienced at each cell. The proposed solution reduced the 99.999% percentile

of the latency from 1.3 ms to 1 ms, with less than 10% eMBB throughput degradation as compared to conventional scheduling techniques, leading to the conclusion that it is feasible to dynamically multiplex URLLC and eMBB traffic, while still fulfilling each service requirements.

Two resource allocation approaches for efficient multiplexing of URLLC and eMBB were thoroughly studied: traditional scheduling based and puncturing based. For the latter, eMBB users are served with a 1 ms TTI size, which are partly overwritten by the incoming URLLC traffic. Both approaches fulfilled the objective of immediately transmitting the URLLC payloads without needing to reserve radio resources for such traffic. Naturally, some performance degradation is experienced for the eMBB users that are punctured. However, various recovery mechanisms were presented to reduce this damage. In particular, the proposed HARQ retransmission scheme, where only the damaged part of punctured transmissions is retransmitted, provided up to 20% eMBB throughput gain as compared to the case where the entire transport block is retransmitted.

The URLLC latency performance is highly sensitive to the traffic characteristics; particularly, the relation between the available carrier bandwidth and the URLLC payload size. For the 200 Byte payload case, doubling the bandwidth from 10 MHz to 20 MHz, allowed to tolerate an URLLC offered load four times larger, while still satisfying the URLLC requirements. Intuitively, larger bandwidth availability allows to transmit more data within a single TTI, leading to the conclusion that the expected URLLC payload size should be considered when deploying future 5G URLLC networks. Similarly, the presented results provide valuable insights into whether it is feasible to share the resources with eMBB traffic, as well as which scheduling scheme (puncturing based or traditional scheduling based) is more appropriate for URLLC-eMBB multiplexing.

Finally, an end-to-end service delivery and scheduling perspective was given. This included an overview of the broad set of options for multi-service scheduling agreed in 3GPP, where it is shown that the majority of techniques proposed in this thesis are incorporated in the upcoming 5G NR.

2 Recommendations

To answer the research questions formulated in Part I - Section 4, the following recommendations are provided:

- Use a 4x4 closed-loop microscopic diversity scheme with second order macroscopic diversity to fulfil the reliability requirements for one-shot downlink transmissions under full load conditions. For cases where one HARQ retransmission is permitted within the latency budget and lower load is generally experienced, 2x2 diversity schemes are sufficient.

- Use a short transmission time interval (0.1-0.2 ms) with sufficiently-fast processing at the transmitter and receiver in order to fit one HARQ retransmission within the 1 ms latency budget. Adopt the proposed joint link adaptation and resource allocation technique to efficiently serve URLLC traffic under a wide range of offered load conditions, and without requiring fine adjustment of the BLER target.
- Use punctured scheduling for efficient multiplexing of URLLC and eMBB. In order to have this working efficiently, apply the proposed recovery mechanisms such that i) eMBB users are made aware of the puncturing, and ii) only the punctured part of the eMBB transmissions is included in the HARQ retransmission.

3 Future Work

Despite the presented findings and contributions, there are still some aspects and problems that should be addressed in future studies. First of all, although most of the work explicitly accounted for the control channel overhead, errors in the control channel (containing the scheduling grant and ACK/NACK HARQ feedback) were not considered. Determining how the presented results would differ when accounting for these errors is of relevance; see e.g. [1] for insights into the impact of the control channel errors on the reliability performance, and [2] where control channel success probabilities of 99.9% are found sufficient to fulfil the reliability requirements.

The inter-cell interference from serving eMBB users on the same radio channel was dealt by selecting a low modulation and coding scheme such that URLLC transmissions are transmitted with sufficiently low BLER. It would be interesting to complement the proposed URLLC resource allocation enhancements with other interference-reduction techniques known from literature, e.g. inter-cell coordinated cell muting schemes.

Various aspects were left out of the scope due to the limited time of the PhD study. For instance, one interesting research direction is to study the impact of mobility on the URLLC performance, as well as to investigate what kind of enhancements are required to guarantee virtually-zero interruption time of the data connectivity during handovers. Achieving reliable communication in the uplink direction is another important research topic. For this case, there is the challenge of limited transmission power at the user equipment, as well as the need for a scheduling request (for cases with dynamic scheduling) prior to the uplink data transmissions. To overcome the latter, grant-free schemes have shown promising benefits [3]. Finally, performing a similar analysis assuming time division duplexing (TDD) modes is also of relevance. The use of TDD results in additional timing constraints and cross-link interference which might be harmful for ultra-reliable communications.

References

[1] H. Shariatmadari, Z. Li, S. Iraji, and R. Jantti, "Control channel enhancements for ultra-reliable low-latency communications," in *IEEE International Conference on Communications (ICC) Workshop*, May, 2017.

[2] R1-1700265, "Analysis of URLLC reliability in DL HARQ," *3GPP TSG RAN WG1 NR Ad-Hoc Meeting*, Jan. 2017.

[3] R. Abreu, P. Mogensen, and K. I. Pedersen, "Pre-scheduled resources for retransmissions in ultra-reliable and low latency communications," in *Wireless Communications and Networking Conference (WCNC), 2017 IEEE*, 2017, pp. 1–5.

Part VI

Appendix

Paper I

Automation for On-road Vehicles: Use Cases and Requirements for Radio Design

Guillermo Pocovi, Mads Lauridsen, Beatriz Soret, Klaus I. Pedersen, Preben Mogensen

The paper has been published in the
IEEE 82nd Vehicular Technology Conference (VTC Fall), 2015.

Abstract

The support of mission-critical communication (MCC) opens the possibility to implement a broad range of novel applications. V2X communication for traffic safety and automation is, among others, one of these innovative applications expected to bring big benefits to society: accidents are prevented, driving times are reduced, and carbon dioxide is saved. In this regard, we first present a system model and fundamental definitions of reliability, latency and availability. Relying on these definitions, a systematic review of requirements for the huge variety of V2X applications is provided, including insights into the expected evolution towards autonomous driving. The many challenges introduced by V2X use cases are emphasized and compared to today's wireless system capabilities. Finally, we give our vision on the design of future radio technologies for the support of this kind of communications.

1 Introduction

Efficient support for machine-type communication (MTC) over wireless is an active research topic gaining increased attention. Especially MTC use cases with mission-critical communication (MCC) requirements are challenging as such services are subject to much tighter latency and reliability requirements than e.g. is the case for mobile broadband (MBB) services. Hence, how to best accommodate both MBB and MCC in the same wireless system presents several new challenges, given the fundamental trade-offs of optimizing for spectral efficiency, latency, and reliability [1]. Examples of wireless system standards with ongoing MTC/MCC related research include (among others) IEEE 802.11p [2] and 3GPP LTE [3], as well as studies on a future 5th Generation (5G) radio standard [4].

A prerequisite for studying MCC over wireless networks is a solid understanding of the use cases and related definitions. In this study we focus on the class of use cases related to vehicular applications. Vehicular use cases include both communication between vehicles, as well as between vehicles and infrastructure (i.e. base stations / access points) or pedestrians, commonly denoted as V2X communication. Given this starting point, we first aim at presenting a generic, yet simple, system model and the related definitions of latency, reliability, and availability. Secondly, a review of V2X applications and their requirements for road safety, traffic efficiency, and infotainment is provided, based on material from ETSI [5] and the US Department of Transportation (DOT) [6]. Especially the characteristics of these different use cases in terms of message rate, message payload size, latency requirements, distinctiveness and reliability are summarized, and put into a wireless system perspective. Furthermore, the expected evolution towards autonomous vehicles is explored [7], [8]. It is identified how such use cases further pushes

the requirements of V2X communication. The presented information on use cases offers a solid basis for definition of realistic traffic models, and their corresponding Quality of Service (QoS) requirements, that can be applied in wireless system research of MCC. Finally, radio design implications for the identified V2X applications are discussed for different wireless system standards. The latter also includes an outlook towards the challenges and related requirements for the upcoming 5G radio system(s), as identified by the EU funded joint research collaboration project METIS [9] and the International Telecommunications Union (ITU) for International Mobile Telecommunications (IMT) for 2020 and beyond [4].

The rest of the paper is organized as follows: Section 2 presents the generic system model, while the related fundamental definitions for MCC are presented in Section 3. V2X use cases and related requirements and models are presented in Sections 4 and 5. Radio design implications are covered in Section 6, including an outlook towards V2X MCC for 5G. Finally, concluding remarks are summarized in Section 7.

2 System model

Fig. I.1 depicts the generic system model. It contains a traffic source and a traffic sink that represent the application layer. The traffic source generates data that are transmitted to the traffic sink via the communication system. For open loop applications, the traffic source generates data without awaiting any feedback from the traffic sink, and without knowing if the traffic sink correctly receives the data payloads. For closed loop applications, the traffic sink provides feedback to the traffic source, e.g. acknowledgements (ACK) for each of the sent application layer payloads. In the considered V2X use cases, the traffic source and sink may be either at the vehicle or infrastructure.

The communication system in Fig. 1 represents the complete system that carries the generated data payloads from the traffic source to the traffic sink. The communication system includes at least one wireless link, but could also include multiple wireless links. As an example, if the communication is from vehicle A to the infrastructure, and from there to vehicle B, then at least two wireless links are part of the communication system. In addition, the infrastructure may include one or multiple backhaul links as well.

The lower layers of the wireless link(s) involve delivering the payload on the air interface. A generic wireless system includes a transmission buffer where the data received from higher layer applications are stored; a scheduler entity allocating radio resources; the transmitter, that tries to adapt the transmission parameters to the variability in the wireless channel, subject to noise and time-variant and frequency-selective fading and interference; and the receiver, where the signal is equalized (and other post-processing proce-

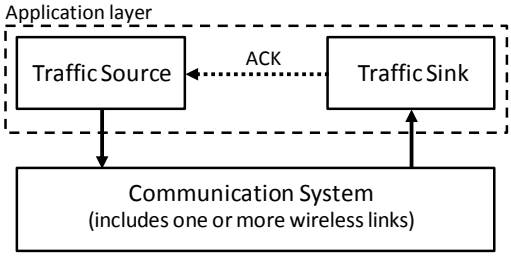


Fig. I.1: System model

dures) in order to maximize the probability of successful decoding. In case of failed decoding, the transmitter can be asked to retransmit. All these elements introduce variability to the transmission, having an impact on the achievable reliability and latency [1].

3 Related definitions

3.1 Reliability definition & service degradation

For an unacknowledged communication, reliability is defined as the probability P that the traffic sink correctly receives a payload of B bits within a maximum latency T [9]. A definition of reliability for acknowledged cases is given by the traffic source correctly receiving an ACK for the payload of B bits within a maximum latency T, with probability P. The acknowledged case is in general more challenging for two reasons: the associated transmission latency of the ACK must be included in the total latency budget, and the ACK message is itself subject to transmission errors.

The definitions above suggest that when the payload of B bits is received after the latency constraint T, it is counted as one error event violating the constraint. Going one step further, one may wonder what happens if the payload is correctly received after $T + \Delta$ (where Δ is small): is the value of the payload completely outdated, or does it still have some value?. The answer is closely related to the nature of the application.

Let us take the example of a typical safety application for V2X, namely the braking warning. Fig. I.2 depicts two vehicles driving in the same direction. The vehicle A in front decides to decelerate for a certain reason, e.g. dangerous road conditions or obstacle in the road. In order to avoid a collision, A informs the following vehicle B (via infrastructure or direct vehicle-to-vehicle) about the braking event. The potential crash-avoidance actions that can be performed by B strongly depend on the delay which this vehicle receives the information. In accordance, we present a more sophisticated definition of errors (Fig. I.3 (a)):

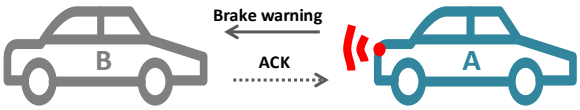


Fig. I.2: V2X application example: braking warning

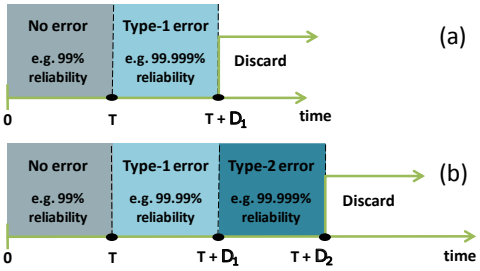


Fig. I.3: Categorization of error cases when service degradation is allowed.

No error

The payload is received within a certain latency constraint T. Vehicle B can e.g. apply the brakes to avoid any potential danger. The associated probability of this event is defined as P_0.

Type-1 error

The payload is received within a latency constraint $T + \Delta_1$. Vehicle B detects that the collision is unavoidable hence deciding to apply collision mitigation measures such as hard braking plus optimal set up of seat-belts and air-bags. The associated probability of this event is defined as P_1, where $P_1 > P_0$.

Intolerable error

The payload is received after the latency constraint $T + \Delta_1$. The collision already occurred therefore there is no reason to deliver the data.

Notice that the presented definition is generic and can –in principle– allow an arbitrary amount of type-n error cases (see Fig. I.3 (b)). For example, one could have allowed e.g. a *no error* case where only soft-braking is applied; a *type-1 error* case where hard-braking and/or evasive manoeuvres are performed; and, finally *type-2 error* where the unavoidable collision measures are executed.

Furthermore, the presented example also works to exemplify the acknowledged reliability definition. Vehicle A can ask B to confirm the reception of the message (i.e. send an ACK). Assuming A knows the (estimate) position of B, vehicle A can apply certain measures to avoid or decrease the impact

of a rear-end crash if the acknowledgement is not received within a certain latency constraint.

3.2 Availability definition

Closely related to reliability, availability is another important metric in V2X communications. Precisely due to their connection, it is difficult to find a consensus in the literature on the definition of these two metrics. We adopt a similar definition as the one in [9].

Space availability/coverage

Space availability is defined as the percentage of area where a required metric by a certain user is achieved, assuming normal operation of the network. Signal strength and signal-to-noise-and-interference ratio are typical parameters for this metric in radio systems. Space availability is a prerequisite for reliable communication.

Time availability/robustness

For a certain service area, time availability is defined as the percentage of time the communication system is capable of providing the required service. Infrastructure equipment failure is one of the events with negative impact on time availability. Notice that when the number of error-types in our definition grows, reliability converges to robustness, since the number of admissible retransmissions increases and the success of the transmission is only limited by the time availability of the vehicular and network infrastructure.

4 Applications and requirements of V2X communications

Table I.1 presents a detailed summary of the most representative use cases enabled by V2X communications. For each use case, information about the communication type, transmission frequency, maximum end-to-end (E2E) latency, distinctive characteristics (if any), and reliability is itemized. The described use cases information is based on the material from ETSI [5] and US DOT [6], [10].

It is observed from Table I.1 a large variation in terms of requirements for the different use cases. The majority of safety-related applications rely on broadcast of small payloads at 10 Hz transmission rate with a 100 ms latency constraint. Among these, we have *pre-crash sensing* and *cooperative platooning* as the use cases that, due to their very critical nature, have the tightest communication requirements requiring up to 50 Hz update rate and

20 ms E2E latency. On the other hand, we have infotainment applications (e.g. internet browsing and media streaming) that require high data rates (some Mbps) with relatively relaxed latency constraints (hundreds of ms).

We take the safety-related use cases as an example to give insights into the modelling of MCC in a wireless system. Active road safety applications are categorized into *cooperative awareness* and *road hazard warning* for periodic and event-triggered transmission of safety messages, respectively [5]. These types of messages differ not only in the role and transmission mode, but also in the dissemination policy. The principle of cooperative awareness applications is that each vehicle periodically broadcasts short messages containing real-time information about their position, speed, travelling direction, etc, enabling vehicles to be mutually-aware of their presence and warn the driver when imminent danger is detected. Cooperative Awareness Message (CAM) [11] is the message format standardized for this type of applications and it is typically delivered to all the neighbouring vehicles.

Road hazard warning applications rely on the broadcast of short messages that are triggered after the detection of a hazardous event e.g. obstacle in the way, slippery road. The message format standardized for this purpose is called Decentralized Environmental Notification Message (DENM) [12] which contains information about the detected event, and is delivered to vehicles potentially affected by such event.

Based on the system model presented in Section 2, the traffic is modelled as follows: CAMs and DENMs are generated by the traffic sink in a periodic and event-triggered fashion, respectively. For the former, an update rate of 10 Hz is typically used in order to support the majority of safety applications. The packet size typically varies between 50 B to 250 B depending on the inclusion of a low-frequency data container which contains static and not highly dynamic information used to support use cases requiring lower update rate (e.g. 1-2 Hz) [11], [13]. For the DENMs, the packet size varies, being above 1000 B if a detailed description of the event is present [12], [13]. It is worth mentioning that the addition of a security overhead can add up to 250 B additional data to the safety message.

The packet is sent to the system where it is modulated and sent over the wireless channel, and eventually delivered to the traffic sink, which must receive the generated payloads within a certain latency as specified in Table I.1.

It is worth mentioning that none of the use cases are attached to a specific reliability constraint; however, the different priority levels emphasized in the “distinctiveness" column in Table I.1 describe the relative importance of a particular use case which can be used to give an insight into the required reliability level. Not by coincidence, the use cases requiring (very) high priority are commonly attached to high update rates. The reason why reliability might not be a crucial requirement is that near-term implementa-

Table I.1: V2X use cases and requirements [5], [6], [10]

Application Class	Use Case	Communication type	Update Rate [Hz]	E2E Latency [ms]	Distinctiveness	Reliability [%]
Active Safety	Emergency electronic brake lights	Event-triggered message broadcast informing about braking event	10	100	High priority	Not Specified
	Emergency Vehicle Warning	Periodic permanent broadcast by emergency vehicle	10 in [5] 1 in [10]	100 in [5] 1000 in [10]	Authentication of the sender	
	Motorcycle warning	Periodic permanent broadcast by the 2-wheel vehicle	2	100	-	
	Pedestrian Warning	Periodic message broadcast by road-side sensing infrastructure or human device	1 in [5] 10 in [10]	100	-	
	Wrong way driver	Even-triggered broadcast by vehicle driving in wrong way	10	100	-	
	Stationary vehicle warning	Even-triggered broadcast by immobilized vehicle	10	100	High priority	
	Road work Warning	Even-triggered periodic broadcast by road-side unit	2 in [5] 1 in [10]	100 in [5] 1000 in [10]	-	
	Overtaking vehicle warning	Broadcast of overtaking state	10	100	Large commun. range > 500 m	
	Do not pass warning	Periodic message broadcast of basic safety information	10	100	High priority	
	Lane change assistance	Point-to-point session for cooperation between involved vehicles	10	100	Relative pos. accuracy: < 2 m	
	Lane change warning	Periodic message broadcast of basic safety information	10	100	-	
	Pre-crash sensing	Periodic broadcast + point-to-point session for cooperation between involved vehicles	10 in [5] 50 in [10]	50 in [5] 20 in [10]	Very high priority	
	Left turn warning	Periodic message broadcast of basic safety information	10	100	-	
	Merging traffic warning	Periodic message broadcast of basic safety information	10	100	-	
	Cooperative merging assistance	Point-to-point session for cooperation between involved vehicles	10	100	Relative pos. accuracy: < 2 m	
	Hazardous location	Event-driven broadcast by vehicles detecting the hazardous location	10 in [5] 2 in [10]	100 in [5] 500 in [10]	-	
	Intersection collision warning	Periodic message broadcast of basic safety information	10	100	High priority	
	Cooperative forward collission warning	Point-to-point two-way communication among vehicles in [5] One way broadcast in [6]	10	100	Relative pos. accuracy: < 1 m; High priority	
Traffic information & efficiency	Traffic light optimal speed	Periodic message broadcasted by road-side infrastructure	2	100	Positioning accuracy: < 2 m	
	Cooperative Platooning	Point-to-point two-way communication among vehicles	2 in [5] 50 in [10]	100 in [5] 20 in [10]	Relative pos. accuracy: < 2 m; High priority	
	In-vehicle signage	Periodic message broadcast by road-side infrastructure	1	500	-	
	Traffic information and recommended itinerary	Periodic traffic information message broadcast by road-side infrastructure	1-10	500	-	
	Map download/update	point-to-point session between vehicle and infrastructure	N/A	500	High data rates	
	Electronic toll collection	Periodic broadcast + point-to-point session between vehicle and infrastructure	1	200	-	
Infotainment	Point-of-interest notification	Periodic message broadcast by road-side infrastructure	2	100	-	
	Browsing, streaming, Download of media	Internet access provided by road-side infrastructure or cellular network	N/A	500	High data rates	
	Instant messaging	Point-to-point session with instant messaging server	N/A	500	-	

tions of V2X technology are expected to solely warn and inform drivers about potential danger instead of taking full control of the vehicle. In fact, 802.11p, which is the de-facto standard to support initial V2X safety applications, uses very simple best-effort transmission approaches making it difficult to ensure high reliability especially at high vehicular density and/or high update rates; see [14] - [16], for example. The open literature, however, claims that a communication reliability[1] above 95% is considered good enough to support the majority of safety applications [14], [17].

Availability is another important performance metric for V2X communications; although not specified, it is worth mentioning that high degree of space and time availability is essential especially for safety use cases. Note also that none of the applications specify the need of acknowledged reliability or tolerance to service degradation. This is due to the fact that most of the safety use cases were designed having in mind the capabilities and limitations of 802.11p, which does not allow a straightforward implementation of the presented definitions.

To summarize, there is a large and diverse amount of applications that can be enabled by V2X communication, both safety-related applications requiring MCC and infotainment applications with typical MBB requirements. Notice that none of the use cases simultaneously demand stringent requirements of throughput, latency and reliability. This fact is very relevant from a radio design perspective and will be analysed later.

5 Autonomous driving vision

The continuing advances in technology are expected to offer solutions in the vehicular field that further depart from current paradigms. In 2014, the SAE organization published the J3016 standard [19] which aims at providing a common terminology and classification levels for driving automation. Five levels of vehicle automation are defined, ranging from level 0: *No Automation* to level 5: *Full Automation*. We focus on the latter in which the vehicle performs all the driving functions without expected assistance from the driver.

It is claimed that fully autonomous vehicles can entirely penetrate the market between 2025-2030 [20], [21]. To fill the technological gap, there are many projects currently dealing with the definition, development and testing of features for autonomous driving. For example, the AdaptIVE consortium [8] or AutoNet2030, which aims at developing autonomous driving technologies for a 2020-2030 deployment horizon.

Despite these many ongoing efforts, there is still no consensus about what autonomous vehicles are nor what will be needed, from a communication

[1]Due to the unacknowledged connection-less nature of 802.11p, communication reliability is measured in terms of packet reception rate [14], [18].

Table I.2: METIS requirements for V2X [9]

Requirement	Value
Payload size (MAC)	1600 Bytes
Latency	< 5 ms
Update Rate	5 - 10 Hz for periodic and event-triggered messages
Reliability	99.999 %
Availability	$\sim$100 %
Device density	Vehicular: 100-1000 dev/km^2; Pedestrians: 150-5000 dev/km^2

system point of view, to support such use cases. Intuitively, autonomous driving applications will demand more stringent requirements compared to those in Table I.1. For instance, the METIS consortium has defined a set of communication requirements at the MAC layer for autonomous driving (see Table I.2) [9]. Notice that the update rates, payload size and traffic type are relatively similar to the presented in Table I.1, however, with much more stringent requirements of latency, reliability and availability. All these elements represent challenges for the wireless system.

6 Radio design implications

6.1 Current communication systems alternatives for V2X

There are various wireless technologies that can match the V2X application requirements more or less effectively [22]. In this section we present a brief overview of the main system alternatives: IEEE 802.11p and 3GPP LTE.

IEEE 802.11p (ITS-G5 in Europe) has been proposed as the standard to support V2X communications. This standard is basically a modified version of the 802.11 specifically designed to deal with the numerous challenges in vehicular environments [2]. The main drawback of 802.11p is its decentralized ad-hoc nature which results in large probability of packet collisions, especially in dense scenarios. Furthermore, the connection-less and unacknowledged mode of communication implies challenges in establishing very reliable communication, becoming even more difficult in two-way communication scenarios [23]. Finally, it is not clear if it can support the high-bandwidth demands of infotainment applications [24].

Using a cellular-based system such as LTE is another approach gaining increased attention [3], [13], [22], [25]. LTE achieves E2E latencies on the order of 20-40 ms. Similar latency numbers have also been observed in vehicular environments (through simulations) [13], [16]. Based on these studies, it is expected that LTE can support the majority of expected initial applications listed in Table I.1. Compared to 802.11p, the planned-infrastructure approach

inherent in LTE (or in cellular-based systems, in general) results in better performance especially in terms of coverage and communication range, and also better support for applications with different QoS requirements [13], [16]. The lack of support for local data exchange implies, however, dependency on the availability (both space and time) of the cellular infrastructure.

6.2 Outlook to 5G

Current communication systems will, in principle, be able to support expected near-term implementations of V2X for safety purposes. However, it is still not clear if they can provide the very high reliability and low latency requirements needed for autonomous driving applications.

The ITU is currently working on defining the overall objectives to be addressed by IMT for 2020 and beyond systems (commonly known as 5G) [4]. Autonomous driving is just one example of the many applications (others include health care, industrial automation, etc [26]) that could benefit by the support of MCC. Motivated by this, the main breakthrough of such 5G system(s) is expected to be the capability to provide flexible and configurable support for multiple applications with very different requirements, ranging from typical MBB services needed for e.g. infotainment applications, to low latency and high reliability for MCC purposes.

As explained in [1], there is a fundamental tradeoff between throughput, latency and reliability; however, the fact that none of the envisioned applications for 5G (V2X field and many others described in [26]) simultaneously demand stringent requirements of these three performance metrics, suggests that it is feasible, although challenging, to design a single wireless system capable of supporting all these services.

7 Conclusions

In this paper we have defined key performance indicators for MCC use cases such as latency, reliability, spatial coverage and temporal availability. A review of today's known V2X applications for active road safety, traffic efficiency, and infotainment is presented. Here it is found that the equivalent application layer traffic typically can be represented by open loop models, where the traffic source generate moderate payload sizes of approximately 50-250 bytes at a rate of 1-10 Hz. For most of the applications, the latency requirement is on the order of $\sim$100 ms, with only few cases demanding 20 ms. Authorities have not specified exact values for reliability, although distinctiveness in terms of relative priorities are listed. But, it is evident that high reliability as well as high degree of spatial coverage and temporal availability are required. Migration towards future autonomous driving use cases

will further tighten the requirements for latency, and especially calling for ultra reliability, as well as correspondingly high spatial coverage and temporal availability. Finally, the feasibility of using wireless standards like IEEE 802.11p and 3GPP LTE have been elaborated, as well as an outlook towards 5G. Among others, 5G is estimated to be capable of meeting the challenging requirements of supporting larger variety of multiple types of services, as compared to what is feasible with today's wireless systems.

References

[1] B. Soret, P. Mogensen, K. I. Pedersen and M. C. Aguayo-Torres, "Fundamental Tradeoffs among Reliability, Latency and Throughput in Cellular Networks", *IEEE GLOBECOM*, Dec. 2014.

[2] IEEE Std 802.11p-2010, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 6: Wireless Access in Vehicular Environments", June 2010.

[3] 3GPP Work Item Description S1-150284, "Study on LTE support for V2X services", Feb. 2015.

[4] ITU Working Document 5D/TEMP/548-E, "IMT Vision — Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond", Feb. 2015.

[5] ETSI TR 102 638 V1.1.1, "Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Definitions", 2009.

[6] DOT HS 811 772, "Development of Performance Requirements for Commercial Vehicle Safety Applications. Final Report", May. 2013.

[7] European Commission Report SMART 2010/64, "Definition of necessary vehicle and infrastructure systems for automated driving", 2011

[8] A. Amditis, "From Advanced Active Safety Systems to Automated Systems: From interactIVe to AdaptIVe and beyond", Feb. 2015.

[9] METIS ICT-317669 Deliverable 1.1, "Scenarios, requirements and KPIs for 5G mobile and wireless system", Apr. 2013.

[10] DOT HS 809 859, "Vehicle Safety Communications Project Task 3 Final Report: Identify Intelligent Vehicle Safety Applications Enabled by DSRC", March 2005.

[11] ETSI EN 302 637-2 V1.3.0 Draft, "Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service", 2013.

[12] ETSI EN 302 637-3 V1.2.0 Draft, “Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Decentralized Environmental Notification Basic Service", 2013.

[13] J. Calabuig, J. F. Monserrat, D. Gozálvez and O. Klemp, “Safety on the Roads: LTE Alternatives for Sending ITS Messages", *IEEE Vehicular Technology Magazine*, vol. 9, no. 4, pp. 61-70, Dec. 2014.

[14] Y. Saleh, F. Mahmood and B. Abderrahim, “Performance of beacon safety message dissemination in Vehicular Ad hoc NETworks (VANETs)", *Journal of Zhejiang University SCIENCE A*, vol. 8, no. 12, pp. 1990-2004, Nov. 2007.

[15] Z. Wang and M. Hassan, “The Throughput-Reliability Tradeoff in 802.11-Based Vehicular Safety Communications", *IEEE Consumer Communications and Networking Conference*, Jan. 2009.

[16] Z. Hameed Mir and F. Filali, “LTE and IEEE 802.11p for vehicular networking: a performance evaluation", *EURASIP Journal on Wireless Communications and Networking*, 2014.

[17] N. An, T. Gaugel and H. Hartenstein, “VANET: Is 95% probability of packet reception safe?", *International Conference on ITS Telecommunications*, Aug. 2011.

[18] F. Bai and H. Krishnan, “Reliability Analysis of DSRC Wireless Communication for Vehicle Safety Applications", *IEEE Intelligent Transportation Systems Conference*, Sept. 2006.

[19] SAE Intl., “SAE On-Road Automated Vehicle Standards Committee Draft Information Report J3016: Taxonomy and Definitions", 2014.

[20] European Technology Platform on Smart Systems Integration (EPoSS), “European Roadmap Smart Systems for Automated Driving", Jan. 2015.

[21] Morgan Stanley Research, “Autonomous Cars: Self-Driving the New Auto Industry Paradigm", Nov. 2013.

[22] G. Araniti, C. Campolo, M. Condoluci, A. Iera and A. Molinaro, “LTE for vehicular networking: a survey", *IEEE Commun. Magazine*, vol. 51, no. 5, pp. 148-157, May 2013.

[23] M. Torrent-Moreno, M. Killat and H. Hartenstein, “The challenges of robust inter-vehicle communications", *IEEE Vehicular Technology Conference*, Sept. 2005.

[24] M. Amadeo, C. Campolo, and A. Molinaro, “Enhancing IEEE 802.11p/WAVE to Provide Infotainment Applications in VANETs”, *Elsevier Ad Hoc Networks*, vol. 10, no. 2, pp. 253-69, Mar. 2012.

[25] ETSI TR 102 962 V1.1.1, “Intelligent Transport Systems (ITS); Framework for Public Mobile Networks in Cooperative ITS (C-ITS)", 2012.

[26] G. P. Fettweis, “The Tactile Internet: Applications and Challenges", *IEEE Vehicular Technology Mag.*, vol. 9, no. 1, pp. 64-70, March 2014.

Paper J

Increasing Reliability by Means of Root Cause Aware HARQ and Interference Coordination

Beatriz Soret, Guillermo Pocovi, Klaus I. Pedersen, Preben Mogensen

The paper has been published in the
IEEE 82nd Vehicular Technology Conference (VTC Fall), 2015.

Abstract

The arrival of mission critical applications in the context of vehicular, medical and industrial wireless communications calls for reliability constraints never seen before in cellular systems. Enhanced Inter-Cell Interference Coordination (eICIC) has been widely investigated in the context of LTE-A Heterogeneous Networks, but always with load balancing and resource partitioning purposes. Given the broad range of new use cases targeting ultra high reliability, we propose the use of on-demand eICIC for reducing the BLER of the retransmissions of critical users while minimizing the impact to the rest of the network. Combined with a ROot Cause Aware HARQ (ROCA-HARQ), which provides additional information when a transmission fails, the joint mechanism is relevant for any LTE/LTE-A deployment and can be easily implemented in a real network. System-level simulations show attractive BLER reductions up to 80% with little impact in throughput performance (loss in user throughput below 6%).

1 Introduction

Mission critical communications (MCC) have become a hot research topic with a broad range of applications, from vehicular to medical and industrial use cases. What makes the topic very challenging from a radio perspective is the required reliability and latency constraints, never seen before in cellular systems [1]. For some of the cases, the reliability requirement can go up to 99.99964% (a.k.a. six-sigma) with latencies of few milliseconds. Perhaps the field attracting more attention from the wireless community is Vehicular Communications, or V2X, where the X can be another vehicle, a pedestrian or the network infrastructure. In this regard, 3GPP has just approved a Study Item for Rel-13 [2] with the goal of investigating the support of V2X communications over LTE/LTE-A[1], involving unicast, multicast and broadcast communication.

The use of HARQ – an error correction mechanism based on retransmissions – in the physical (PHY) and medium access control (MAC) layers of LTE [8] is essential to increase the reliability. The receiver produces either an ACK, for the case of error-free reception, or a NACK if some errors are detected. Upon reception of a NACK message, the desired packet will be sent again. HARQ is used to increase the spectral efficiency and reliability by means of redundant transmissions. The higher the number of transmission attempts, the higher the probability of success. However, some MCC applications may not tolerate several retransmission attempts for fulfilling the

[1]Other examples of wireless system standards with ongoing MCC related research include IEEE 802.11p and future 5G radio systems. Although out of the scope of this paper, the proposed solution may also be relevant for them.

latency constraint. One of the most important limiting factors for a successful retransmission is the variability in the interference, which can be mitigated from the network side through a coordinated inter-cell algorithm.

Enhanced Inter-Cell Interference Coordination (eICIC) was standardized in 3GPP Rel-10, and since then plenty of works have investigated its implementation and performance (see e.g. [4] - [6] for an overview of the feature and [7] for baseline performance results). The main principle of eICIC consists on periodically muting some of the subframes in the aggressor cell in order to reduce the interference to the victim users. eICIC was conceived for load balancing and resource partitioning in co-channel Heterogeneous Networks (HetNets), and until now such approach has been followed by the majority of researchers. With new use cases targeting very high reliability, we see an opportunity to broaden the scope of eICIC, such that interference coordination is applied with the goal of increasing the reliability of critical users.

We propose a combined mechanism with Root Cause-Aware HARQ and eICIC for improving the BLock Error Rate (BLER) performance of the first retransmission, reducing the need for a second or third retransmission. With knowledge of the interference conditions when the transmission has failed, the network can coordinate the transmission of neighbouring cells in the subframe where the retransmission will take place. As discussed along the paper, it is a rather simple but effective solution, in which the reliability of MCC is significantly improved with a small cost in complexity.

The rest of the paper is organized as follows. The problem is outlined in Section 2. In Section 3 we give an overview of the two main techniques in which our proposal relies, HARQ and eICIC, and describe the main principle of the proposed algorithm. The combined solution is further discussed in Section 4, analysing different aspects related to overlapping, link adaptation and signalling exchange. The simulation results are discussed in Section 5, and some concluding remarks close the paper in Section 6.

2 Description of the Problem

We define a communication as reliable if a certain payload is delivered within a given latency requirement and with a given probability of success. In case of HARQ, the NACK indicates that the transmission has failed and a retransmission is needed, with a Chase Combining (CC) gain that in principle goes up to 3 dB in the ideal case and for the first retransmission.

Time-variant signal and interference fluctuations in the network makes things much more challenging. On the one hand, conducting link adaptation (i.e., selection of modulation and coding) and air interface aware packet scheduling at the eNB is not trivial, owing to the inaccuracies and delays of

the channel state information (CSI) feedback from UEs. On the other hand, traffic variations and dense deployments lead to frequent transitions in the interferers from ON to OFF and vice-versa [9] [10], such that the SINR at the retransmission subframe may vary significantly from the value at the transmission.

With proper HARQ and interference coordination, we aim at improving the reliability performance of MCC applications whose latency constraint allows two transmission attempts but not a third one. The same reasoning can of course be applied for the third transmission attempt and so on, and only for the sake of simplicity we focus in the former.

3 Combined ROCA-HARQ and eICIC

3.1 ROot Cause Aware HARQ (ROCA-HARQ)

LTE uses asynchronous HARQ for the downlink, with up to 8 HARQ processes running in parallel. The specific subframe in which the retransmission is scheduled is decided by the transmitter once the NACK has been received.

In traditional HARQ, the confirmation upon reception of a packet is a boolean, ACK/NACK. However, when a failure for a MCC user occurs, it is convenient to add some extra information of the root cause of the error at the corresponding subframe. Although the eNB receives the CSI reports from the UE capturing the channel conditions, this information is not sent instantaneously and per TTI (whereas the channel does change instantaneously), and therefore subject to delays and variability. Thus, a ROot Cause Aware HARQ (ROCA-HARQ) includes updated channel information in the NACK report to inform the transmitter about the real interference conditions at the subframe of the failed packet. Based on it, the transmitter can better adjust the retransmission to increase the probability of success.

3.2 Interference Coordination and Dominant Interferer

eICIC was introduced in LTE-A to reduce the downlink interference to victim users by muting the dominant interferer, which plays the role of aggressor cell. The targeted scenario was a co-channel HetNet with a mixture of macro and small cells transmitting at a lower power. The interference is mitigated by periodic muting of the macro layer – so-called eICIC. The eICIC mechanism is therefore grounded in a benefit-cost balance: on one hand, the victim users benefit from the muted subframes in the aggressor cell, named Almost Blank Subframes (ABS). On the other hand, there is a cost in terms of performance degradation in the said muted layer, which does not schedule users during ABS.

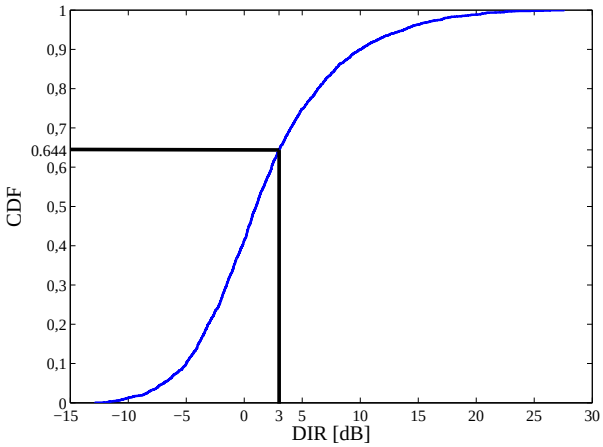


Fig. J.1: CDF of the DIR for the simulated scenario with finite buffer traffic and offered load of 30Mbps.

Beyond the classical 3GPP HetNet scenario assumed in the majority of eICIC studies, the existence of a Dominant Inteferer (DI) is not straightforward in other topologies, like it is the case of more irregular networks, homogeneous networks or dense deployments of small cells, where all the transmitters use the same transmission power [9]. In these cases muting the strongest interferer may result in little SINR improvement for some users, jeopardizing the net balance (same cost, lower benefit). One good indicator of the dominance of the strongest interferer is the Dominant to Interference Ratio (DIR), defined as

$$DIR = \frac{I_{strongest}}{\sum_{i \neq strongest} I_i + N} \tag{J.1}$$

where $I_{strongest}$ is the power received from the DI, I_i is the power received from interferer i and N is the thermal noise power. For example, in the scenario simulated in this paper with 21 clusters of 4 small cells per cluster (see simulation details in Section 5) the CDF of the DIR of the users can be seen in Figure J.1. As expected, not all users have a high value of DIR: 64% of the users experience a DIR below 3dB. We take this information into account in the design of the interference coordination scheme. In particular, the identifier of the strongest interferer is fedback in the NACK report, but only if it is dominant enough, i.e. the DIR at the instant of the fail is above a threshold.

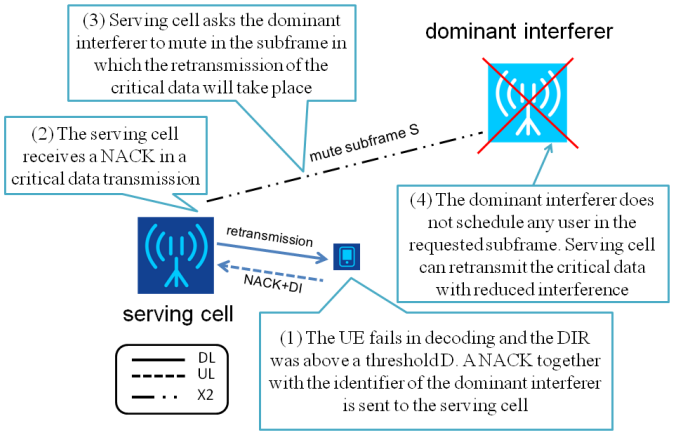


Fig. J.2: Dynamics of the ROCA-HARQ and IC algorithm.

3.3 Proposed Algorithm

The main idea consists of reducing the interference during a retransmission by muting the dominant interferer for the receiver. The benefit of muting the strongest interferer is quantified by checking if the DIR of the victim user is above a threshold D (performed at the UE side). If yes, then the NACK with the DI is reported and the aggressor cell will be asked to mute [2] in the subframe in which the retransmission will take place. The algorithm is illustrated in Figure J.2, and the pseudo code is shown in Algorithm 1. In the eNB side, it is avoided to have situations in which a cell is requested to mute at the same subframe in which a retransmission with reduced interference was planned in said cell. This and other important aspects inferred from the application of the algorithm are discussed in the next section.

4 Implementation Issues

4.1 Muting Overlapping

The proposed muting coordination is especially challenging when one small cell is simultaneously aggressor and serving a victim user. In this case, it is important to coordinate the retransmissions such that they take place in different subframes. The way to solve this situation is rather simple: the

[2]Other options beyond muting the aggressor include reducing the power or doing some rank adaptation together with advanced receivers [9].

eNB will look for a subframe in which it has not been previously asked to mute, postponing the retransmission event when the overlapping happens. Naturally, this results in increased latency for the user that has to postpone his retransmission, but the impact in the global performance is expected to be minor, since such situation happens rarely.

4.2 Inter-cell Signalling Exchange

It is worth to emphasize that the standard eICIC muting is configured periodically, and exchanged among eNBs via X2. In contrast, we are considering sporadic and on-demand single-subframe muting, **used only when needed**. This way, we expect to improve the performance of the retransmission mechanisms without jeopardizing the performance of the rest of the network.

Rel-10 specifications include several enhancements for the X2 application protocol to facilitate collaborative configuration of ABS muting patterns between eNBs. The macro eNB can send a X2 Load Information message to the small cells with ABS information. The ABS information includes information of the currently used ABS muting pattern at the macro-eNB (expressed with a 40-bit word for FDD cases). The ABS information can also be exchanged between macro eNBs to align that neighbouring macro eNBs use the same, or overlapping, ABS muting patterns. The exact definition of the various X2 messages that form the light-weight inter-eNB eICIC coordination protocol can be found in the X2 application protocol specification [11]. The implementation of the proposed scheme in a standard LTE-A network requires small updates in the X2 signalling. Thus, a cell serving a victim user can request (instead of inform) the aggressor cell to mute its transmission in a particular subframe. Notice that the new message could be used for many other purposes, not only reliability.

The impact in the latency of a coordinated muting is also related to the intrinsic X2 signalling delays, such that for delays below X ms the first retransmission can be scheduled only X ms ahead the NACK reception.

4.3 Link Adaptation

The use of interference coordination results in more severe time-variant interference fluctuations in the network. The challenge is solved in Rel. 10 with the introduction of restricted CSI measurements for advanced UEs [5], so that the eNB receives such reports corresponding to normal subframes and ABS, respectively. The Rel. 10 measurement restrictions are suitable for the standard periodic muting, but it is not the case of the solution proposed in this paper. Here we are talking about instantaneous muting, such that it is not feasible for the network to configure the UEs to take dedicated measures during a single or few subframes without knowing in advance when

Algorithm 1 `Combined ROCA-HARQ and eICIC`

```
1: procedure ROCA-HARQ {UE side}
2: fail(1 : end) = false
3: iter = 0
4: while (iter < iter_max) do
5:    if (fail(iter)) then
6:       if (DIR(iter) > D) then
7:          send NACK + DI
8:       else
9:          send NACK
10:      end if
11:   end if
12:   iter++
13: end while
14: end procedure
15:
16: procedure IC {eNB side}
17: muting(1 : end) = false
18: iter = 0
19: while (iter < iter_max) do
20:   if (muting request for subframe iter+X) then
21:      muting(iter + X) = true
22:   end if
23:   if (received NACK+DI) then
24:      while (muting(iter + Y)) do
25:         {The eNB has been requested itself to mute}
26:         Y++
27:      end while
28:      send muting request for subframe iter+Y
29:   end if
30:   iter++
31: end while
32: end procedure
```

and if it will happen. Nevertheless, the amount of muting that the algorithm implies is very small, as it will be seen in the results, and therefore the impact of the reduced accuracy in the link adaptation is minor. In any case, the Outer-Loop Link Adaptation (OLLA) mechanism helps in coping with this additional variability.

5 Validation and Results

5.1 Simulation Methodology

For the simulation assessment we assume an LTE-Advanced system with dense clusters of small cells, being the feature relevant also for other topologies. The main simulation settings are collected in Table J.1. 21 clusters with 4 small cells per cluster and all eNBs operating in the same carrier frequency at 3.5 GHz and with 10 MHz bandwidth are considered. The antenna pat-

Table J.1: Summary of simulation settings.

Parameter	Value
Network Layout	21 clusters with 4 small eNBs per cluster
Transmit power	30 dBm
Bandwith	10 MHz at 3.5 GHz carrier frequency
Subframe duration	1 ms (13 data plus 1 control symbols)
Modulation and coding schemes	QPSK (1/5 to 3/4), 16-QAM (2/5 to 5/6), 64-QAM (3/5 to 9/10)
HARQ modeling	Maximum 4 transmissions. Ideal CC with efficiency = 100% and non-ideal CC with efficiency = 80 %
Transmission mode	2x2 closed loop with rank adaptation
Small cell path loss	ITU-R UMi [12]
Shadow fading	Lognormal, std = 4 dB (LOS) std = 6 dB (NLOS)
eNB packet scheduler	Proportional Fair (PF)
UE capabilities	Interference Rejection Combining

tern is omnidirectional and the transmit power is 30 dBm. The path loss model is ITU-R urban micro-cell (UMi), with different expressions for the LOS and NLOS cases. The system-level simulator is time-based and includes all the major LTE functionalities such as link adaptation, HARQ, and packet scheduling. In every 1 ms subframe, the SINR of each user is calculated per subcarrier according to the chosen receiver type. HARQ with Chase Combining is applied in case of failed transmission. Both ideal and non-ideal CC are considered, the latter with an efficiency of 80% [13]. The link adaptation module decides the modulation and coding scheme for the first transmission based on frequency-selective feedback from the users. The simulator does not consider user mobility; however, the user sessions are generally short and the SINR calculations include the effect of variable fast fading. Closed loop 2x2 single-user MIMO with rank adaptation is assumed. Packet scheduling is performed in the time domain only, with one user per TTI. This allows us to increase the number of OFDMA symbols per TTI from 11 to 13 to improve the data rate. The receiver type at the user equipment is MMSE-IRC. It is also assumed that empty cells do not cause any interference.

Users arrive to the system according to a Poisson distribution. Each user has a payload of 0.5 MB to be transmitted, and leave the system once the transmission is completed. The parameter to be set is the average offered

load per cluster, which varies in the simulation from 20 Mbps to 40 Mbps, corresponding to medium to high load in this scenario. All the users in the network are assumed to have MCC requirements. Therefore, the algorithm is applied to all of them. This is an extreme case since in a real network usually not all supported communications are of this kind.

The combined ROCA-HARQ and IC algorithm is compared to a standard configuration in which no coordination is done for the retransmissions. In both cases, the target BLER of the first transmission is set to 20%. [3]

5.2 Simulation Results

We examine first of all the BLER of the first retransmission (second transmission attempt) with ideal CC and under different load conditions: 20 Mbps (medium load) in Figure J.3 and 40 Mbps (high load) in Figure J.4. The threshold D for the DIR is set to 3 dB, which was shown by extensive simulation campaigns to provide the best tradeoff. As expected, the use of the ROCA-HARQ + IC algorithm (solid line) reduces the error rate as compared to not using any special coordination (dashed line). The reduction is quantified in the Figures showing the relative difference in two points of the CDF, namely 50%-ile and 95%-ile, with values ranging from 31% to 51%.

The benefits of the ROCA-HARQ + IC algorithm are pushed up even further when the CC process is not ideal, with an efficiency of 80% in our simulations. Then, the BLER of the first retransmissions is reduced as shown in Figure J.5, with a reduction that goes up to 80% in 50%-ile and 70% in 95%-ile.

The amount of retransmissions needed per packet with the algorithm is also reduced, and consequently the latency associated to these MCC users. With ideal CC, the percentage of successful first retransmissions increases more than 20% when applying ROCA-HARQ + IC as compared to not using it, and with non-ideal CC the improvement is approximately 45%.

Another KPI of interest is the amount of muting that the ROCA-IC algorithm is introducing. It is actually very minor, with values that range from 1% of the total time with 20 Mbps to 3.6% with 40 Mbps. The impact in the throughput of the aggressor cell is therefore marginal. In general, very little muting is needed with low or medium load, since cells are empty a good part of the time and the system is not limited by interference. As the load increases, the number of active users per cell goes up and the interference becomes a limiting factor.

The user throughput performance with ideal CC is shown in Figure J.6. The bar plot shows the loss in throughput when using the algorithm as compared to not using it, with three points of the throughput CDF, namely per-

[3]Notice that an alternative to the ROCA-HARQ + IC scheme is to set a lower target BLER in the first transmission, but that has a deeper impact on the spectral efficiency.

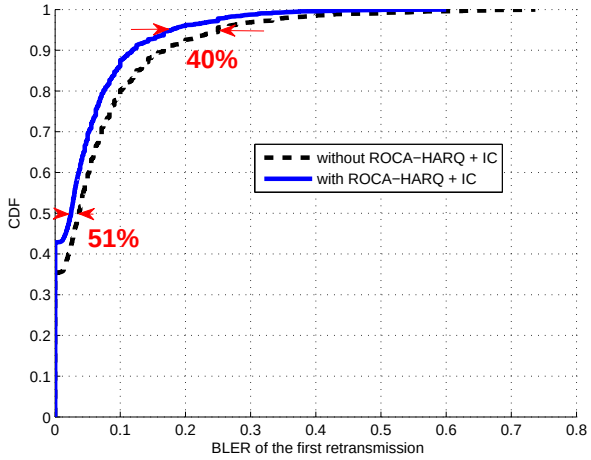


Fig. J.3: BLER of the first retransmission for finite buffer 20Mbps of offered load and ideal CC. The reduction in the 50% and 95%-ile of the CDF is shown.

centiles 5%, 50% and 95% of the user throughput. The ROCA-HARQ+IC algorithm gets very close to the baseline case, with a loss that is below the 6% for ideal CC with 40Mbps and non-ideal CC with 20 Mbps, and insignificant for ideal CC with 20 Mbps. Nevertheless, notice that we are assuming all users to have MCC transmissions, which is an extreme case (i.e. all users in the network are affected by the mechanism).

6 Conclusions

We propose the combined use of Root Cause Aware HARQ and interference coordination for reducing the BLER of the retransmissions of MCC users with stringent reliability and latency constraints. eICIC is used here with reliability purposes instead of the classical load balancing approach. To support it, a modified HARQ with reported root cause information assists the transmitter when an error occurs for coordinating the proper interference conditions in the retransmission. The simulation results quantify the gains of the algorithm, that can go up to 80% with a minor cost in throughput reduction (below 6%) and complexity. Future work includes more sophisticated coordination, for example rank adaptation or transmission power reduction.

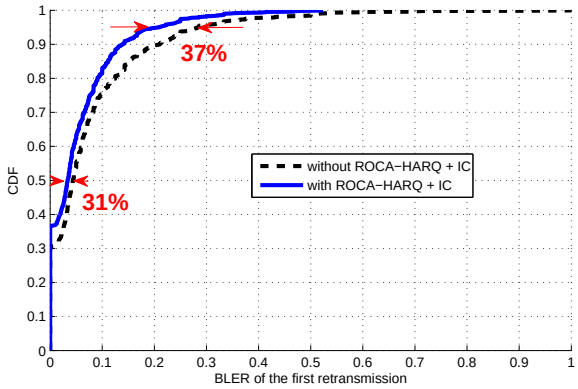


Fig. J.4: BLER of the first retransmission for finite buffer 40Mbps of offered load and ideal CC. The reduction in the 50% and 95%-ile of the CDF is shown.

References

[1] A. Osseiran 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] 3GPP Work Item Description, “Study on LTE support for V2X services, " S1-150284, February 2015.

[3] A. Damnjanonic et al., “A Survey on 3GPP Heterogeneous Networks, " *IEEE Wireless Communications Magazine*, vol. 18, no. 3, pp. 10-21, June 2010.

[4] D. López-Pérez et al., “Enhanced Intercell Interference Coordination Challenges in Heterogeneous Networks," *IEEE Wireless Communications Magazine*, Pages 22-30, Vol. 18, Issue 3, June 2011.

[5] K.I. Pedersen, Y. Wang, S. Strzyz, and F. Frederiksen, “Enhanced Inter-Cell Interference Coordination in Co-Channel Multi-Layer LTE-Advanced Networks," *IEEE Wireless Communications Magazine*, June 2013.

[6] B. Soret, H. Wang, K. I. Pedersen, and C. Rosa, “Multicell Cooperation for LTE-Advanced Heterogeneous Network Scenarios," *IEEE Wireless Communications Magazine*, vol. 20, no. 1, pp. 27-34, Feb. 2013.

[7] Y. Wang, B. Soret, and K. I. Pedersen, “Sensitivity Study of Optimal eICIC Configurations in Different Heterogeneous Network Scenarios," *IEEE Vehicular Technology Conference (VTC)*, Sept. 2012.

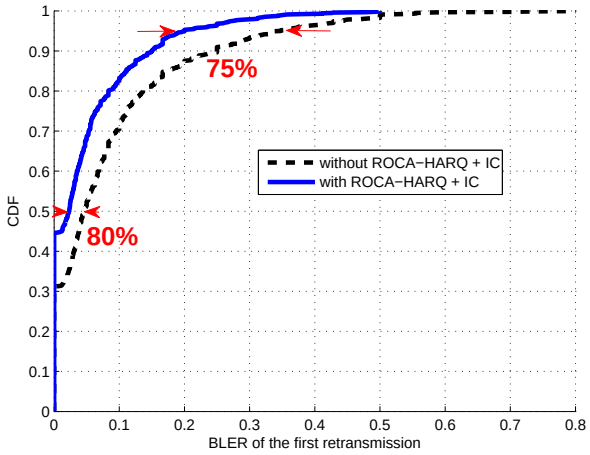


Fig. J.5: BLER of the first retransmission for finite buffer 40Mbps of offered load and non ideal CC (efficiency = 80%). The reduction in the 50% and 95%-ile of the CDF is shown.

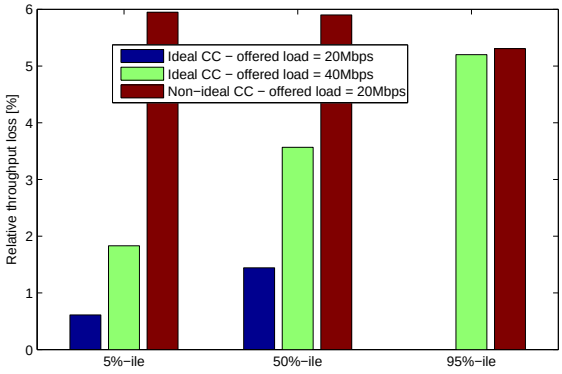


Fig. J.6: Throughput loss for finite buffer with two different offered loads. The reduction in the 5%, 50% and 95%-ile of the CDF is shown.

[8] J. F. Chengi, “Coding Performance of Hhybrid ARQ Schemes," *IEEE Transactions on Communications*, vol. 54, no. 6, pp. 1017-1029, 2006.

[9] B. Soret, K. I. Pedersen, N. Jørgensen, V. Fernandez-Lopez, “Interference Coordination for Dense Wireless Networks," *IEEE Communications Magazine*, Jan. 2015.

[10] V. Fernandez-Lopez, K. I. Pedersen, B. Soret, “Effects of Interference Mitigation and Scheduling on Dense Small Cell Networks," *IEEE Vehicular*

Technology Conference (VTC), 2015.

[11] 3GPP Technical Specification 36.423, “Group Radio Access Network; Evolved Universal Terrestrial Radio Access Network (E-UTRA); X2 application protocol (X2), " April 2011.

[12] 3GPP Technical Report 36.814, “Further Advancements for E-UTRA Physical Layer Aspects," version 9.0.0, March 2010.

[13] F. Frederiksen, and T. E. Kolding, “Performance and Modeling of WCDMA/HSDPA Transmission/H-ARQ Schemes," *IEEE Vehicular Technology Conference (VTC)*, 2002.

ISSN (online): 2446-1628
ISBN (online): 978-87-7112-981-6

AALBORG UNIVERSITY PRESS