Multi-Service Radio Resource Management for 5G Networks

Karimidehkordi, Ali

Publication date:
2019

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

MULTI-SERVICE RADIO RESOURCE MANAGEMENT FOR 5G NETWORKS

BY
ALI KARIMIDEHKORDI

DISSERTATION SUBMITTED 2019

AALBORG UNIVERSITY
DENMARK

Multi-Service Radio Resource Management for 5G Networks

Ph.D. Dissertation

Ali Karimidehkordi

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

Dissertation submitted: October 2019

PhD supervisor: Prof. Preben Mogensen
Aalborg University

Assistant PhD supervisor: Prof. Klaus I. Pedersen
Nokia Bell Labs, Aalborg University

PhD committee: Associate Professor Carles N. Manchon (chairman)
Aalborg University

Professor Riku Jäntti
Aalto University

Dr. Honglei Miao
Intel Deutschland GmbH

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

Department: Department of Electronic Systems

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-521-5

Published by:
Aalborg University Press
Langagervej 2
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2019

To my family

Curriculum Vitae

Ali Karimidehkordi

Ali Karimidehkordi received his B.Sc. in Electrical Engineering-Electronics from Isfahan University of Technology and his M.Sc. degree in Electrical Engineering-Communication Systems from University of Tehran. Since 2016, he has been studying towards the PhD degree in Wireless Communication Networks (WCN) section at Department of Electronic Systems, Aalborg University in close collaboration with Nokia Bell-Labs. His main research interests are radio resource management for the fifth-generation New Radio (5G NR), network optimization, and ultra-reliable low-latency communications (URLLC).

Abstract

The fifth generation of wireless networks (5G) provides services for three traffic classes: enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable and low-latency communications (URLLC). For eMBB scenario, peak user data rate, mobility, and network capacity are the most important key performance indicators (KPIs). For mMTC, a large number of connected devices and network energy efficiency are required, whereas for URLLC latency and reliability are vital. Simultaneously addressing these KPIs is challenging. In general, optimizing for some KPIs leads to degraded performance for others. For example, improving the user data rate and reliability increases the latency, while fulfilling the latency requirements typically has the cost of reduction in reliability, user throughput, and spectrum efficiency. Jointly optimizing these KPIs is one of the main challenges involved in designing 5G networks. The broader scope of this PhD study is on proposing radio resource management (RRM) techniques for 5G networks with the capability to efficiently support different service classes and their diverse requirements.

Interference management techniques can improve network performance in terms of cell capacity and coverage, user throughput, and reliability. In the first part of this PhD project, we present an inter-cell interference subspace coordination (ICISC) scheme for multiple-input multiple-output (MIMO) communications. The key idea in this contribution is to employ precoding algorithms at the transmission-end with the objective of jointly projecting transmitted signal over the desired subspace and aligning the interference onto the predefined interference subspaces at the interfered receiver end. Simulation results show a notable 28% throughput enhancements of the proposed solution. It also benefits from low-computational complexity and requires local channel knowledge makes it attractive for industrial applications.

In the second part, we research on centralized radio access network (C-RAN) architecture with multi-cell scheduling algorithms to improve 5G performance. Specially, we address the challenges of supporting URLLC requirements. Dynamic low-complexity algorithms are proposed to achieve the latency and reliability requirement of URLLC. In comparison to the con-

ventional distributed scheduling, we show that the C-RAN architecture can significantly reduce the undesirable queuing delay of URLLC traffic. The impacts of multi-user scheduling and dynamic switching of serving cells have been investigated. Moreover, we study packet scheduling for mixed URLLC and eMBB traffic in 5G multi-service networks. A low-complexity novel resource allocation method is presented that is latency, control channel, hybrid automatic repeat request (HARQ), and radio channel aware in determining the transmission resources for different users. We run simulations with an advanced 5G New Radio compliant system-level simulator with a high degree of realism to evaluate the performance of the proposed schemes. Simulation results show promising gains in terms of 98% latency improvement for URLLC traffic and 12% eMBB end-user throughput enhancement as compared to conventional proportional fair scheduling.

The third part of the project analyses URLLC radio resource management through single/multi-node connectivity. We evaluate the performance of different scheduling schemes and derive analytical expressions for outage reliability and resource usage. Especially, the overhead and error of transmitting control information and miss-detection of feedback signals are considered for performance evaluation. Low-complexity analytical solutions are presented to minimize the number of allocated resources while satisfying URLLC targets. Numerical results show that the proposed algorithms perform close to those of optimal solutions and achieve a remarkable improvement in resource utilization. Finally, the reliability enhancement with packet duplication is derived and compared against the baseline single-connectivity. It is shown that multi-node connectivity offers significant outage probability improvement over single-connectivity. However, such gains are achieved at the cost of additional radio resource usage.

Resumé

Den femte generation af trådløse netværk (5G) leverer tjenester til tre trafikklasser: forbedret mobilt bredbånd (eMBB), massiv maskintype-kommunikation (mMTC) og ultra-pålidelig og kommunikation med lav latens (URLLC). For eMBB- peak user datahastighed, mobilitet og netværkskapacitet de vigtigste performance indikatorer. For mMTC kræves et stort antal tilsluttede enheder og netværks energieffektivitet, hvorimod for URLLC er latenstid og pålidelighed afgørende. Det er udfordrende at adressere disse krav samtidigt. Generelt fører optimering af nogle krav til forringet ydelse for andre. Eksempelvis, ved at forbedre brugernes datahastighed og pålidelighed øges latensen. Men opfyldelse af latenstidskrav typisk har omkostningerne i form af reduktion af pålidelighed, og spektrum effektivitet. Fælles optimering af disse krav er en af de største udfordringer i design af 5G-netværk. Dette ph.d. projekt drejer sig om at foreslå tekniker indenfor styring af radioressourcer (RRM) til 5G-netværk, for effektivt at understøtte forskellige serviceklasser og deres forskellige krav.

Interferenskontrolteknikker kan forbedre netværkets ydelse med hensyn til cellekapacitet og dækning, og pålidelighed. I den første del af dette ph.d. projekt præsenterer vi et intercell interferens sub space koordination (ICISC) koncept til multiple-input multiple-output (MIMO) kommunikation. Hovedideen med dette bidrag er at anvende forkodningsalgoritmer ved transmissionsenden med det formål at projicere transmitterede signaler over det ønsket underrum og justere interferensen på de foruddefineret interferens underrum ved den interfereret modtagerende. Simuleringsresultater viser bemærkelsesværdige forbedringer af 28% højere datahastighed af den foreslåede løsning. Løsningen har lav kompleksitet og kræver kun viden om de lokal kanal, hvilket gør det attraktivt for industrielle applikationer.

I den anden del af studiet undersøger vi udnyttelse af centraliseret radioadgangsnetværk (C-RAN) med multi-celle RRM algoritmer for at forbedre 5G performance. Særligt adresserer vi udfordringerne med at understøtte URLLC-krav. Dynamiske algoritmer med lav kompleksitet foreslås for at opnå latens- og pålideligheds krav til URLLC. I sammenligning med den konventionelle distribueret RRM, viser vi at C-RAN arkitekturen markant

kan reducere uønsket kø forsinkelse af URLLC trafik. Virkningerne af multicell RRM og dynamisk skiftning af serveringsceller er blevet undersøgt. Ydermere studerer vi RRM for blandet URLLC og eMBB-trafik i 5G multiservice netværk. En ny ressourceallokeringsmetode med lav kompleksitet præsenteres, som er latens, kontrolkanal, hybrid automatisk gentagelsesanmodning (HARQ) og radiokanal afhængig ved bestemmelse af transmissionsressourcerne for forskellige brugere. Vi kører simuleringer med en avanceret 5G New Radio kompatibel systemniveau simulator med høj grad af realisme for at evaluere ydeevnen af de foreslåede løsninger. Simuleringsresultaterne viser lovende gevinster med til 98% latensforbedring for URLLC trafik og 12% eMBB datahastighedsforbedring sammenlignet med konventionel PF (proportional fair) radio ressource allokering.

Den tredje del af projektet analyserer URLLC radioressource allokering via enkelt / multi-node-forbindelse. Vi evaluerer performance af sådanne løsninger, og udleder analytiske udtryk for driftsstabilitet og ressourceforbrug. Især betragtes overhead og fejl ved transmission af kontrolinformation og fejldetektering af feedbacksignaler til ydeevneevaluering. De præsenterede løsninger er designet for at minimere antallet af tildelte ressourcer, mens URLLC-mål opfyldes. Numeriske resultater viser, at de foreslåede algoritmer fungerer tæt på dem med optimale løsninger og opnår en bemærkelsesværdig forbedring af ressourceudnyttelsen. Pålidelighedsforbedringen ved at bruge med pakkeduplicering er udledt og sammenlignet med baseline-enkeltforbindelses transmissioner. Det vises, at multi-node-forbindelse giver en betydelig forbedring af pålideligheden i forhold til enkeltforbindelse. Sådanne gevinster opnås dog på bekostning af højere forbrug af radioressourcer.

Contents

List of Abbreviations

1G first generation

2G second generation

3D three dimensional

3D three dimensional

3G third generation

3GPP 3rd Generation Partnership Project

4G fourth generation

5G fifth generation

5G NR fifth generation New Radio

ACK positive acknowledgement

AMPS Advanced Mobile Phone System

BLER block error rate

BS base station

C-RAN centralized radio access network

CC Chase combining

CCDF complementary cumulative distribution function

CCH control channel

CDF cumulative distribution function

CoMP coordinated multipoint

CQI channel quality indicator

CSI channel state information

D2D device to device

DL downlink

DPS dynamic point selection

DSP digital signal processing

E2E end-to-end

EDGE Enhanced Data Rates for GSM Evolution

eICIC enhanced inter-cell interference coordination

eMBB enhanced Mobile Broadband

FDD frequency division duplex

FDMA frequency division multiple access

gNB fifth generation NodeB

GPRS General Packet Radio Service

GPS Global Positioning System

GSM Global System for Mobile Communications

HARQ hybrid automatic repeat request

HetNet heterogeneous network

HSPA high Speed Downlink Packet Access

IA interference alignment

ICIC inter-cell interference coordination

ICISC inter-cell interference sub-space coordination

IMT International Mobile Telecommunications

IMT-2000 International Mobile Telecommunications-2000

IoT internet of things

IP Internet protocol

IRC interference rejection combining

IS-95 Interim Standard 95

ITU International Telecommunication Union

JT joint transmission

JTACS Japan Total Access Communications System

KPI key performance indicator

L2S link-to-system

LA link adaptation

LDPC low density parity check

LLC low latency communication

LTE Long Term Evolution

LTE-A LTE-Advanced

M-LWDF modified largest weighted delay first

M2M machine to machine

MAC medium-access-control layer

MBB mobile broadband

MCC mission-critical communication

MCS modulation and coding scheme

MIMO multiple-input multiple-output

MMIB mean mutual information per coded bit

mMIMO massive MIMO

MMS multimedia messaging service

MMSE minimum mean square error

mMTC massive Machine Type Communication

MRC maximal-ratio combining

MT mobile terminal

MTC machine-type communication

MU multi user

NACK negative acknowledgement

NMT Nordic Mobile Telephone

NR New Radio

OFDM orthogonal frequency-division multiplexing

OFDMA orthogonal frequency-division multiple access

OLLA outer-loop link adaptation

PAPR peak-to-average power ratio

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

PRB physical resource block

QoE quality of experience

QoS quality of service

RAN radio access network

RE resource element

RRH remote radio head

RRM radio resource management

RSRP received signal received power

RTT round-trip time

SC-FDMA single-carrier frequency-division multiple access

SCS sub-carrier spacing

SE-DPS spectrum-efficient dynamic point selection

SINR signal to interference-and-noise ratio

SMS short message service

SNR signal to noise ratio

SU single user

SVD singular value decomposition

TACS Total Access Communications System

TDD time division duplex

TDMA time division multiple access

TTA throughput to average

TTI transmission time interval

UE user equipment

UL uplink

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:	Multi-Service Radio Resource Management for 5G Networks.
PhD Student:	Ali Karimidehkordi.
Supervisors:	Prof. Preben Mogensen. Aalborg University. Prof. Klaus I. Pedersen. Nokia Bell Labs, 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 close collaboration with Nokia Bell Labs. The work was carried out in parallel with fulfilment of mandatory courses required to obtain the PhD degree.

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

Paper A: A. Karimi, N. H. Mahmood, K. I. Pedersen and P. Mogensen, "Inter-Cell Interference Sub-Space Coordination for 5G Ultra-Dense Networks", *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Toronto, ON, September 2017, pp 1-5.

Paper B: A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Pocovi, and P. Mogensen, "Efficient Low-Complexity Packet Scheduling Algorithm for Mixed URLLC and eMBB Traffic in 5G", *2019 IEEE 89th Vehicular Technology Conference - VTC2019 Spring*, Kuala lumpur, May 2019, pp 1-6.

Paper C: A. Karimi, K. I. Pedersen, and P. Mogensen, "5G URLLC Performance Analysis of Dynamic-Point Selection Multi-User Resource Allocation", *16th International Symposium on Wireless Communications Systems (ISWCS)*, Oulu, August 2019, pp 1-6.

Paper D: A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner and P. Mogensen, "Centralized Joint Cell Selection and Scheduling for Improved URLLC Performance", *2018 IEEE 29th Annual International*

Symposium on Personal, Indoor and Mobile Radio Commu-nications (PIMRC), Bologna, September 2018, pp. 1-6.

Paper E: A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner and P. Mogensen, "5G Centralized Multi-Cell Scheduling for URLLC: Algorithms and System-Level Performance", *IEEE Access*, vol. 6, pp. 72 253–72 262, 2018.

Paper F: A. Karimi, K. I. Pedersen, and P. Mogensen, "Low-Complexity Centralized Multi-Cell Radio Resource Allocation for 5G URLLC", *IEEE Wireless Communications and Networking Conference (WCNC)*, 2020. **Submitted for publication**.

Paper G: A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Berardinelli, and P. Mogensen, "On the Multiplexing of Data and Metadata for Ultra-Reliable Low-Latency Communications in 5G", *IEEE Transactions on Vehicular Technology*, 2019. **Submitted for publication**.

Paper H: N. H. Mahmood, A. Karimi, G. Berardinelli, D. Laselva, and K. I. Pedersen, "On the Resource Utilization of Multi-Connectivity Transmission for URLLC Services in 5G new radio", *in Proc. 2019 IEEE Wireless Communications and Networking Conference (WCNC) Workshops*, Morocco, April 2019, pp 1-6.

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

Praise to the one who no eye can see him, but he observes every eye.

First, let me express my deepest gratitude to my supervisors Klaus I. Pedersen, Preben Mogensen and Nurul Huda Mahmood, whose consistent support and advice helped me to finish the PhD with success. I am really proud to join Aalborg University and have this unique opportunity to work and learn from you. Also, thanks a lot to Gilberto Berardinelli for the great discussions and guidance throughout the research. Special thanks to Carles Manchon, Riku Jäntti, and Honglei Miao, for accepting to be my assessment committee and providing me valuable comments and insightful questions.

In addition, a thank you to Drothe Sparre and Linda Villadsen for administrative support during the PhD period. I had this chance to seat at Nokia Bell Labs and collaborate with Nokia scientists. I would like to express my gratitude to all my colleagues at Aalborg University and Nokia research centers in Aalborg, Espoo, Bangalore, Paris, and Stuttgart for three years of great collaboration, advises, and technical discussions. Thank you, Jens Steiner and Mads Brix for helping me with software development. Special thanks to Renato, Guillermo, Roberto, and Ali E. for your feedback on the papers, and simulator development.

I sincerely appreciate my parents, my sisters, and my brother for the endless love and support through all the time. The words can not express how much I am proud and blessed for having you in my life.

Finally, thanks to all my friends in Iran, Denmark, and the rest of the world for creating unforgettable memories together.

Ali Karimidehkordi
Aalborg University, October 2019

Part I

Introduction

Introduction

Now, it has been four decades since the first generation (1G) of commercial cellular networks has been deployed. As it is expected, nowadays the fifth generation New Radio (5G NR) is surprising the world with its ubiquitous and diverse services. Back to late 1970s and 1980s, the 1G of cellular networks was initiated by Nordic Mobile Telephone (NMT) (mainly in Nordic countries), Advanced Mobile Phone System (AMPS) in USA, and Japan Total Access Communications System (JTACS) to enable voice communications for mobiles. 1G was developed on an analog basis with frequency division multiple access (FDMA) technology that could provide a few kbps data rate on 30 kHz carrier frequency. As a pioneer for commercial cellular networks, 1G suffered from several technical issues such as limited capacity, poor connection quality, inefficient spectrum occupancy, large phone size, and finally a nationwide implementation.

Transition from analog to digital technology in cellular networks started by launching the second generation (2G) in the 1990s. 2G offers enhanced number of cell supported users, increased data rate and service quality, and improved security because of the digital encryption. It also made several new features commercially feasible including short message service (SMS), multimedia messaging service (MMS).

The Global System for Mobile Communications (GSM) and Interim Standard 95 (IS-95) are the most well-known implementations of 2G networks that are deployed based on time division multiple access (TDMA) and wideband code division multiple Access (WCDMA) technologies. Later, further enhancements led to the development of General Packet Radio Service (GPRS) referred as 2.5G and Enhanced Data Rates for GSM Evolution (EDGE) sometimes specified as 2.75G to further boost service performance. In the early 2000s, the 2G family became widespread and popular in more than 200 countries with more than 80% share of the market [1]. The main goal of designing the first and second generation of mobile networks was to enable (enhance) voice communications for subscribers. However with growth in data-based applications, it became demanding to increase the user experienced rate and improve the support for data-related applications and access to real Internet.

Besides that, there were also some technical issues such as low service quality in less urban areas, and sudden call failure due to the reduction in signal strength.

Key characteristics for the air interface and service requirements of the third generation (3G) mobile communication standards were specified by the International Telecommunication Union (ITU) under International Mobile Telecommunications (IMT)-2000 program [2]. The research for 3G in Europe began during the 1990s with Universal Mobile Telecommunications System (UMTS) project. Later and in 1998, the 3rd Generation Partnership Project (3GPP) formed to align parallel works and develop global specifications for the 3G networks by focusing on upgrading GSM infrastructures. That resulted in WCDMA technology. There were also developments based on IS-95 standard carried by 3GPP2 that led to CDMA2000 mobile air interface standards. Further improvements of 3G networks introduced the high Speed Downlink Packet Access (HSPA) and HSPA+ technologies. 3G supports both time division duplex (TDD) and frequency division duplex (FDD) variants. It operates with 1.25 and 5 MHz channels for CDMA2000 and WCDMA standards, respectively. Several new techniques are introduced to achieve higher speed, and lower delay such as employing higher modulation and coding scheme (MCS), the use of multiple-input multiple-output (MIMO) technology, utilizing an Internet protocol (IP) based core network with enhanced security, etc. 3G offers data rates in order of several Mbps, improved quality of service (QoS) for both voice and data based applications, and enables the use of multi-media communication services such as video conferencing and mobile TV.

In 2008, the ITU announced IMT-Advance as the required specifications for the fourth generation (4G) of cellular networks. Those include: enhanced data rate up to 1 Gbps, reduced the cost per data bit, flexible spectrum access, improved system spectral efficiency up to 3 bit/Hz/cell, support high mobility with smooth handover, and high QoS for multimedia applications [3]. In 3GPP, the early phase study of 4G has been done as the long-term evolution of UMTS project which is often known as Long Term Evolution (LTE). Later, the requirements and key features specifications were identified in Release 8 [4].

LTE benefits from new access technology based on orthogonal frequency-division multiple access (OFDMA) in downlink (DL) transmission. Thanks to the recent advances in and digital signal processing (DSP) technologies, OFDMA offers higher performance in terms of spectral efficiency and multi-user multiplexing gain. It is more robust against time dispersion of wireless channels and benefits from less-complex baseband receivers [5, 6]. In uplink (UL) direction, transmission based on single-carrier frequency-division multiple access (SC-FDMA) is adopted to reduce peak-to-average power ratio (PAPR) [7]. LTE operates in both TDD and FDD modes, at a wide range

of bandwidths (from 1.4-20 MHz) and different carrier frequencies. Utilizing MIMO transmission, which is supported in LTE from the first release, the DL peak data rate of 300 Mbps and 75 Mbps on the UL transmission is provided [6].

LTE is also equipped with a simplified all-IP based architecture that is optimized to reduce the user-plane latency to less than 5 msec [5]. The use of advanced minimum mean square error (MMSE) receivers along with UL power control and dynamic inter-cell interference coordination (ICIC) techniques are supported in LTE to cancel/reduce inter-cell interference and improve the desired QoS for different applications [5].

The evolution to fulfil IMT-Advanced requirements continued in 3GPP by introducing new releases with enhanced capabilities toward faster, more efficient, and reliable 4G networks. The 3GPP-Release 10 and subsequent versions, often called as LTE-Advanced, exploit transmission on extended bandwidth (up to 100 MHz) by user scheduling over multiple carrier components so-called as carrier aggregation. Network implementation with increased number of MIMO antennas along with expanded spatial multiplexing, enhanced inter-cell interference coordination (eICIC), additional improvements for heterogeneous deployments, and make the use of coordinated multipoint (CoMP) are among the LTE-Advanced features to further boost the 4G performance [8, 9]. 3GPP-Release 13 and 14 (also known as LTE-Advanced Pro) are the last evolutions of 4G. LTE-Advanced Pro incorporates with the completion of technologies from the previous releases and includes features for enabling massive internet of things (IoT) connectivity, private networks, interworking with unlicensed access spectrum, and public safety enhancements.

1 5G Requirements

After successful deployments of the 4G networks, it is now time for the fifth generation (5G) to change the world! During the previous years, the number of mobile subscribers has significantly increased. By 2019, more than 5.1 billion people (around two-third of the population) have at least one mobile connection [10]. Thanks to the recent advances in the software/hardware architecture of mobile devices, smartphones operate as a personal computer providing an appropriate platform for developing various web-based applications. Along with the increasing trend to social networks, modern smartphones offer advanced multi-media equipment such as high-quality microphones/speakers, multi-lens cameras that can capture three dimensional (3D) photos/videos with incredible details, and high-resolution displays. It is predicted that the annual mobile traffic will increase from 138 exabytes in 2017 to 930 exabytes in 2022 [11]. The average data per user per month is expected to reach 13.3 gigabytes in 2022, which is around 6x that of user traffic in

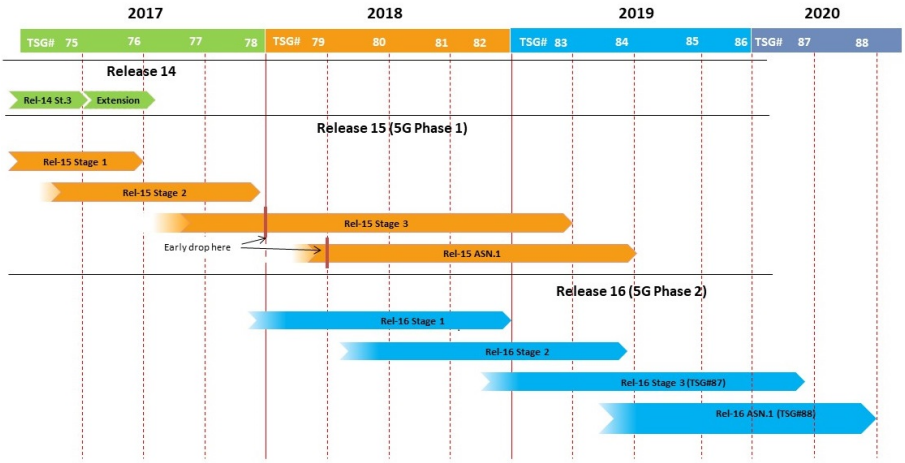

Fig. I.1: Time-line of the 3GPP 5G standardization procedure [3GPP].

2017 [11].

Beside the mobile broadband (MBB) traffic that requires high data rate, there is a fast-growing demand for supporting machine-type communication (MTC) in which a large number of devices and emerging technologies need autonomous communications through wireless technology, and with a large variety of service requirements. Those include access to Internet for a massive number of low-cost low-power smart devices such as sensors, home appliances, health tracking systems, etc. It is forecasted that by 2022, there will be 1.5 billion IoT which are connected to cellular networks [12]. The fourth industrial revolution aims to migrate from wire-based control access to a highly flexible wireless plant. Indeed, there are emerging demands to support various device to device (D2D), machine to machine (M2M), and finally vehicular to anything (V2X) communication with a diverse set of QoS requirements in terms of data rate, power consumption, reliability, and modem cost.

In addition to enhance MBB and MTC, the 5G should be able to support a new service type with extreme requirements for reliability and latency. This is crucial to enable a wide range of use cases such as: factory automation, safety transportation, E-health, autonomous driving, and smart grids.

Research activities towards identifying the future visions of IMT for 2020 and beyond was initiated in 2012. In 2015, the ITU-R specified IMT-2020 as the preliminary descriptions, service perspectives, and the requirements of the next generation mobile communications. As illustrated in Fig. I.1, the 5G standardization procedure started by 3GPP in 2017 and subsequently in 2018, the first full set of 5G standards was completed as 3GPP-Release 15.

2 5G Overview

As envisioned by IMT-2020, 5G shall support three use cases: enhanced Mobile Broadband (eMBB), massive Machine Type Communication (mMTC), and Ultra-Reliable Low-Latency Communications (URLLC). Fig I.2 pictures the key requirements and usage scenarios that the 5G network must be adequately addressed.

- **eMBB:** Increasing demands for mobile broadband services is the main driver behind eMBB. Supporting eMBB needs high data rate, improved quality of services (QoS), increased coverage, and seamless mobility experience. IMT-2020 targets for eMBB are 10 Gbit/s as a peak data rate (100 Mbps in 95% of cases), and support mobility of up to 500 km/h.
- **mMTC:** The main requirements for mMTC are the support high density of connected devices up to 1 million devices per km^2 and 10 years battery life.
- **URLLC:** It offers high service availability and reliable data transmission in a short time budget. Depending on the application different levels of reliability and latency are targeted for URLLC, which one of the most stringent is one-way transmission of 32 bytes packet in maximum one msec and with 99.999% reliability.

To meet the diversity of requirements (and sometimes conflicting) associated with different services, the 5G deploys improved architecture and introduces new technological components.

Utilization of spectrum in higher frequency bands is one of the key solutions to achieve higher data rate with lower latency, enhance network capacity, and provide sufficient services for a large number of subscribers. It is envisioned that future wireless systems will move toward the use of new cm-wave ($3-30$ GHz) and mm-wave ($30-300$ GHz) frequency bands. High propagation loss, atmospheric and rain absorption are among the most common challenges in the utilization of new spectrum, especially in the mm-wave band [13, 14]. However, for the early 5G deployments, the focus is on spectrum below 6 GHz [15].

Together with the move towards higher frequency, massive MIMO deployment and network densification are the main enablers to achieve significant spectrum efficiency and capacity gain. The use of massive MIMO with hybrid beamforming enables simultaneous transmission to multiple users through enhanced spatial dimensions. Besides the enormous benefits, emerging massive MIMO technology imposes several challenges including hardware design issues, pilot contamination, channel estimation, and proper

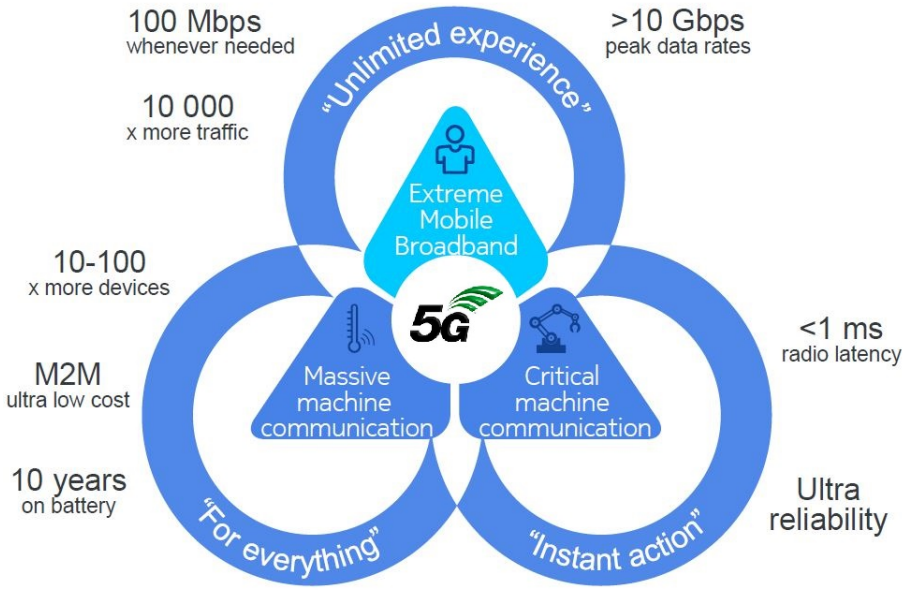

Fig. I.2: 5G use cases and requirements [Nokia].

radio resource management (RRM) that need to be addressed for the efficient deployments.

Cell densification and heterogeneous network (HetNet) deployment are promising ideas to increase network capacity and coverage in 5G [16]. HetNet topology consists of overlaying several BSs with different powers and coverage sizes that are connected to the main macro cell . The objective of adding low power cells is coverage extension, throughput enhancement, and increased energy efficiency [17]. These small cells are connected to the core network via wired/wireless backhaul. HetNet deployment can significantly enhance the frequency reuse factor and take new spectrum into use (mm-wave and cm-wave are suitable to be used at low distance communications [14]). It is also possible to deploy low power cells in small geographical areas with high traffic demands (hot spots). This reduces the load of each BS by distributing users across several nodes. Proper user/cell association methods, interference coordination techniques, and mobility management mechanisms are necessary to guarantee the benefits.

As illustrated in Fig. I.3 and I.4, Release 15 supports two deployment options for 5G NR. Non-stand alone architecture is the first release of the 5G networks, where the NR baseband is connected to the existing LTE infrastructure. S1 interface is used for communication between NR base stations

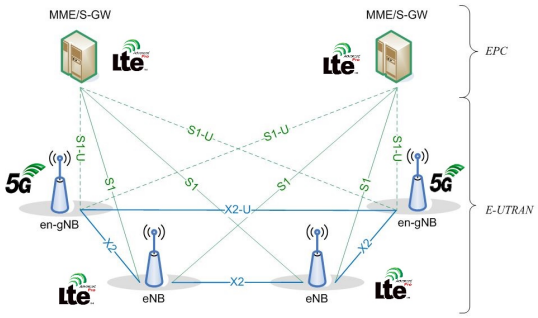

Fig. I.3: Non-stand alone architecture [3GPP].

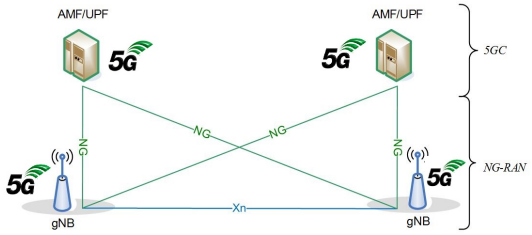

Fig. I.4: Stand alone architecture [3GPP].

(gNB) and the 4G core network. The connection between base stations is established via X2 interface. The mobile terminal is dual-connected to 4G base station (eNB) as a master node and a gNB as the secondary node. Non-stand alone deployment enhances communication speed and reliability by providing access to 5G spectrum.

Stand alone architecture enables the maximum benefits of the 5G to support end-to-end (E2E) services by employing a new 5G core. The NR core uses a cloud-native service-based architecture to develop and manage new applications with different traffic profiles. It supports edge computing and local hosting of network functions close to the mobile terminal access point to reduce E2E latency and improve performance.

The 5G supports E2E network slicing to accommodate with different service classes simultaneously. The concept of network slicing is to create multiple virtual networks over the same infrastructure. The resources are shared and specialized among virtual networks (network slices) so that each slice efficiently serves a particular use case [18]. Depending on the defined business models, a user can be served by multiple slices simultaneously.

centralized radio access network (C-RAN) with powerful computing tools are among the 5G promising solutions to face new challenges. C-RAN provides an appropriate platform for interference management, coordinated beam-

forming, network slice isolation, mobility management, and coverage expansion [17].

The NR adopts a user-centric design. It supports a fully flexible service-specific numerology with scalable bandwidth. To maintain low-latency communications, mini-slot configuration and the use of short transmission time interval (TTI) allocation is adopted in NR.

Similar to 4G, the NR applies OFDMA waveform for DL transmission. In the UL, DFT-S-OFDMA and OFDMA are supported. For frequency bands below 7 GHz, 15 kHz, 30 kHz, and 60 kHz subcarrier spacing are used. 60 kHz and 120 kHz subcarriers are considered for network deployment above 24 GHz. A physical resource block (PRB) of 12 subcarriers is assumed as the minimum allocation size in the frequency domain.

In the time domain, each sub-frame has the length of one msec. Depending on the applied numerology, it includes one or multiple slots of 14 symbols. Moreover, to reduce latency for time-critical applications, mini-slot configuration with length of one to 13 symbols is supported in NR. The NR Release 15 utilizes modulation schemes of QPSK, 16QAM, 64QAM, and 256QAM. Data traffic is encoded using low density parity check (LDPC) codes and for control signalling Polar coding is used.

3 Scope and Objectives of the Thesis

The scope of this PhD dissertation is on the DL RRM for 5G multi-service systems. The majority of the PhD focuses on URLLC as a new service class introduced in 5G NR.

We start by studying interference management as one of the interesting RRM problems and a fundamental element to boost the network performance. Particularly, we focus on interference alignment (IA) inspired techniques as an effective approach to tackle co-channel interference. Efficient adaptation of IA strategies is studied for 5G networks. To harvest its potential benefits for practical deployment, we seek solutions to overcome IA hurdles such as high computational complexity and the need for global channel state information (CSI). This includes adopting a transmission technique that works with local CSI and has low-complexity at transmitter-ends. We favour of simple receiver structure to facilitate the reception procedures. The objective is to provide higher spectrum efficiency for dense deployment scenarios.

In the second part, which forms the main body of this PhD dissertation, we study, design, and evaluate various packet scheduling and link adaptation schemes for URLLC. Complementing the stringent demands and sustaining URLLC are difficult, as there is a fundamental tradeoff between latency, reliability, and spectrum efficiency [19]. Especially, it becomes more

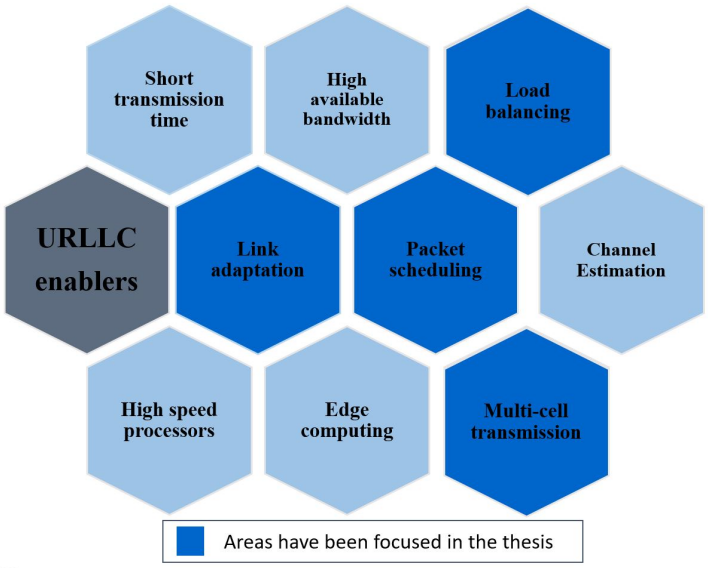

Fig. I.5: Potential URLLC enablers for DL transmission.

challenging when traffic increases. Fig. I.5 shows recent NR enhancements and potential technologies that can be applied to address the challenges and further boost URLLC performance. Transmission on short TTI bases is fundamental for enabling URLLC. From the latency standpoint, this reduces the frame alignment delay, transmission time, and allows performing additional retransmission. Furthermore, having relatively powerful processors that offer low processing times at transceivers is crucial. Assigning higher carrier bandwidth for URLLC is a solution for reducing the transmission time and queuing probability.

Since the spectrum is limited and expensive, resource allocation and link adaptation have fundamental role to address different QoS targets and optimize the network benefit-cost ratio in terms of consumed resources. The PhD covers the problem of downlink packet scheduling for frequency division duplexing. Allocation of time and frequency resources are discussed in line with 3GPP Release 15 specification. The co-existence of URLLC and eMBB traffic is considered. The objective is to make the efficient use of available resources and maximize the network capacity of URLLC and eMBB. The URLLC capacity is defined as the maximum network capability of supporting offered URLLC load while ensuring reliability and latency constraints. For eMBB capacity, achieving the highest throughput with certain level of fairness is desirable. Solutions with reasonable complexity are desirable for potential NR implementation.

Furthermore, centralized multi-cell scheduling schemes are studied. The C-RAN enables developing new RRM techniques, including dynamic load balancing, robust mobility management, interference coordination, and multi-cell transmission. Targeting to increase C-RAN capacity, we first focus on reducing URLLC latency by proper offloading the traffic from congested cells to lightly-loaded neighbouring points. Dynamic solutions compatible with the tight URLLC budget are desirable. In addition, it is essential to manage inter-cell interference and ensure reliability. As the benefits come at the price of higher complexity and energy consumption at both C-RAN and user-ends, it is demanding to have mechanisms for determining multi-cell connectivity and handling the complexity of cell association problem.

Proper system design and performance optimization need an in-depth understanding of communication-theoretic principles of URLLC and studying various components that threaten QoS targets. To address this, a comprehensive study of downlink transmissions is conducted. We investigate a fundamental question on how to multiplex data and control channel. Two approaches to answer this question are separate coding of data and control channel and joint encoding. Reliability and spectrum efficiency of each scheme are analyzed. Through the analyses, a detailed transmission scenario is modelled. Achievable gain from one retransmission, and impairments in feedback channel are taken into account. Performance, cost, and benefits of adopting each approach are evaluated and discussed for URLLC.

In general, successful transmission can be accomplished by setting very conservative MCSs. Although improving the reliability, this solution poses substantial challenges in terms of resource inefficiency and queuing probability. To handle this case, dynamic resource optimization is required. Most of the conventional resource allocation schemes are designed to enhance throughput by assuming relaxed time constraints. Therefore, they are not viable solutions to guarantee high levels of reliability in short time budget. Optimizing the system for URLLC needs to revisit the problem formulation and include new target criteria. The PhD investigates the problem of dynamic link adaptation and packet scheduling in a theoretical framework. The objective is to minimize the average consumed resources (power and spectrum) subject to fulfilling URLLC requirements.

Another potential of C-RAN is its capability to overcome the stochastic nature of wireless channels. Throughout the PhD, we analytically research the reliability enhancement inspired by packet data duplication. This feature involves independent packet transmission from multiple base stations. It is an attractive solution for URLLC. But, it may impose some challenges, including increase in resource usage, interference, and queuing delay. Hence, multi-cell transmission needs to be evaluated and optimized for practical applications.

Following presents the main research questions and hypothesis we ad-

dress in this study:

Q1 How to design a distributed IA solution for 5G?

H1 IA is one of the well-known interference management techniques that may provide significant enhancements in terms of user capacity and coverage. Efficient adaptation of IA based on the real network deployments can be considered as one potential approach to deal with interference in 5G. A distributed solution is required to perform based on local channel knowledge. For practical implantation, it is also desirable to have a low complexity solution.

Q2 How to allocate radio resources to URLLC and eMBB for distributed NR implementation?

H2 A service-specific resource allocation design is required to achieve diverse requirements of URLLC and eMBB. A careful but simple packet scheduling solution is desirable. In addition to service-wise link adaptation, exploiting the gain of frequency selective scheduling, taking latency budget into account, and managing the overhead of control channel may provide considerable performance improvement for both services.

Q3 How to best utilize C-RAN for URLLC in 5G NR?

H3 C-RAN enables fast cell switching on a TTI basis. This might be beneficial for URLLC to mitigate fast fading variations in wireless channels. More importantly, a fully centralized architecture with global knowledge about buffer, load, and channel information offers high degrees of freedom to perform instant load balancing.

Q4 How to multiplex data and control information for URLLC in downlink transmission?

H4 Successful packet transmission is a function of both data and control information. Since URLLC payloads are usually small, the overhead of control information is comparable to the data-block. Joint link adaptation for data and control information significantly improves resource efficiency while ensuring QoS requirements.

Q5 How packet data convergence protocol (PDCP) data duplication impacts URLLC?

H5 It is expected that packet duplication provides significant reliability improvement. There needs further investigations on resource usage and infrastructure expenses.

4 Research Methodology

To approach the research targets, we follow a classical scientific methodology as described below:

1. **Identify the problem, research questions, objectives, and hypotheses:** In the early phase of each project, an extensive literature review along with the study of state of the art relevant papers, patents, and related standards are performed. Regular meetings are organized with industry experts to discuss the existing technologies and discover all parameters, limitations, and issues impacting the performance. Based on the achieved perspective, we identify the problem and objectives, provide relevant research questions, and shape the hypotheses. Whenever possible, the objectives are formulated as optimization problems with different constraints.

2. **Derive solutions:** Open literature is examined to explore the potential solutions. Since most of the formulated problems are combinatorial non-convex optimizations, finding the optimal solutions is not always feasible. In such cases, utilization of the relaxed constraints is explored to propose sub-optimal solutions with reasonable complexity. Although, mathematical modellings give a good view of problem and solution, but for some complex cases such as multi-service multi-node transmission, derivation of mathematical solutions is difficult. In this case, semi analytical heuristic algorithms are developed.

3. **Performance analysis and validation of the proposed solutions:** We verify the hypothesis and analyse the proposed solutions via analytical treatment and Monte-Carlo simulations. Numerical expressions are derived for the performance evaluation and resource utilizations. MATLAB scripts are then developed to compute and compare different solutions. In order to achieve accurate results in complex cases, Monte-Carlo simulations are performed. In this respect, Nokia system-level simulator is utilized to develop, debug, and evaluate the ideas. We align network settings and basic assumptions inline with 3GPP specifications for NR and run the simulations by adopting highly realistic scenarios such as multi-cell multi-service topology modelling, dynamic traffic arrival, and 3D radio propagation. The simulations are conducted with a sufficiently high number of samples to statistically generate reliable results.

4. **Analysis of the results:** The results are analyzed and compared against the hypothesis and initial assumptions. As such, we achieve a better understanding of network behavior in sophisticated scenarios. Further, we revisit the initial hypothesis based on the results.

5. **Dissemination of the results:** Disseminations of this PhD study are published in scientific-related journals, conferences, and project reports. For each study item, we present and discuss the solutions and performance results at project forums, Wireless Communication Networks (WCN) section, and Nokia Bell labs team meetings. The novel ideas are protected via patent applications.

5 Contributions

The key contributions of this PhD study are summarized as follows:

1. **A low-complexity interference coordination scheme for 5G dense network.**

   Applicability of IA techniques for suppressing the interference is studied. A new derivation of IA is presented as the so-called *inter-cell interference sub-space coordination (ICISC)*. The proposed scheme employs a low-complexity precoding algorithm at the transmitter side to jointly project the transmitted stream(s) at the desired sub-space of the intended receiver while aligning the interference at the interference subspace of the interfered nodes. The solutions requires only local channel information and is well-suited for dense network deployments.

2. **Efficient scheduling algorithm for dynamic multiplexing of URLLC and eMBB.**

   An effective algorithm is proposed for co-scheduling of URLLC and eMBB traffic. The solution takes into account several components that impact the performance and spectral efficiency. Those include service type, latency, HARQ, payload size, overhead of control information, and spectrum diversity. Detailed system-level simulations show that the proposed solution offers remarkable gains in terms of reducing the latency of URLLC and enhancing throughput for eMBB traffic. The algorithm has simple architecture that makes it attractive for practical implementation.

3. **Low-complexity centralized multi-cell scheduling algorithms for URLLC in 5G NR.**

   To overcome undesired queuing delay and satisfy URLLC, attractive centralized multi-cell scheduling algorithms are proposed. The solutions provide significant latency improvement and benefit from low-computational complexity in accordance with stringent URLLC processing time requirements. This makes the proposed solution feasible

for real network implementation. Extensive system-level simulation results and discussions are provided to evaluate the performance of the proposed solutions for different network setups and assumptions.

4. **Detailed system-level simulations.**

   Performance evaluation and sensitivity analysis of the solutions are build up by developing 5G NR specific system-level simulations with high degree of realism. This includes multi-cell deployments in line with 3GPP specifications, dynamic traffic model, and flexible user-centric numerology in line with 5G NR proposals for URLLC scheduling. Further, we adopt MIMO transmission with 3D radio propagation, and detailed transmission stack protocols including dynamic link adaptation, control channel adjustment of scheduling grants, modelling HARQ retransmission, etc. Thus, reliable results with a high level of accuracy are generated which can be utilized for 3GPP standardization meetings.

5. **Propose a novel dynamic link adaptation of joint data and metadata for URLLC.**

   A fundamental theoretic study of URLLC for DL transmission is performed. Two multiplexing approaches for data and control namely as *joint encoding of data and control information* and *in-resource control signalling* are investigated. We derive expressions for the resource usage and outage probability. Novel resource allocation and joint link adaptation for data and control information is presented. To this end, a low-complexity solution based on successive convex optimization is proposed. Provided numerical results show considerable resource efficiency gain of the proposed solution.

6. **Studying the potential of data duplication for DL URLLC**

   We study PDCP packet duplication as on of the URLLC enablers. An analytical framework for the resource usage and the reliability enhancements of packet duplication is presented. We discuss the performance and the limitations of data duplication by comparing the results against single-cell connectivity.

The following conference/journal papers are prepared in relation with this PhD study

Paper A: A. Karimi, N. H. Mahmood, K. I. Pedersen and P. Mogensen, "Inter-Cell Interference Sub-Space Coordination for 5G Ultra-Dense Networks", *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Toronto, ON, September 2017, pp 1-5.

Paper B: A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Pocovi, and P. Mogensen, "Efficient Low-Complexity Packet Scheduling Algorithm

for Mixed URLLC and eMBB Traffic in 5G", *2019 IEEE 89th Vehicular Technology Conference - VTC2019 Spring*, Kuala lumpur, May 2019, pp 1-6.

Paper C: A. Karimi, K. I. Pedersen, and P. Mogensen, "5G URLLC Performance Analysis of Dynamic-Point Selection Multi-User Resource Allocation", *16th International Symposium on Wireless Communications Systems (ISWCS)*, Oulu, August 2019, pp 1-6.

Paper D: A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner and P. Mogensen, "Centralized Joint Cell Selection and Scheduling for Improved URLLC Performance", *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Commu-nications (PIMRC)*, Bologna, September 2018, pp. 1-6.

Paper E: A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner and P. Mogensen, "5G Centralized Multi-Cell Scheduling for URLLC: Algorithms and System-Level Performance", *IEEE Access*, vol. 6, pp. 72 253–72 262, 2018.

Paper F: A. Karimi, K. I. Pedersen, and P. Mogensen, "Low-Complexity Centralized Multi-Cell Radio Resource Allocation for 5G URLLC", *IEEE Wireless Communications and Networking Conference (WCNC)*, 2020. **Submitted for publication**.

Paper G: A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Berardinelli, and P. Mogensen, "On the Multiplexing of Data and Metadata for Ultra-Reliable Low-Latency Communications in 5G", *IEEE Transactions on Vehicular Technology*, 2019. **Submitted for publication**.

Paper H: N. H. Mahmood, A. Karimi, G. Berardinelli, D. Laselva, and K. I. Pedersen, "On the Resource Utilization of Multi-Connectivity Transmission for URLLC Services in 5G new radio", *in Proc. 2019 IEEE Wireless Communications and Networking Conference (WCNC) Workshops*, Morocco, April 2019, pp 1-6.

Moreover, three patent applications are filed in cooperation with Nokia Bell Labs as follows

Patent Application 1: Configured Grant Arrangement in Beam-Management Networks.

Patent Application 2: Resource Allocation for IoT Semi-Static Data.

Patent Application 3: Enhanced Data and Control Channel Link Adaptation for NR.

A considerable portion of the PhD period was devoted to simulator development to achieve realistic performance evaluation that can be used for 3GPP standardization perspective. The numerical results of papers A, G, and H are generated using MATALB. Performance evaluations in papers B, C, D, E, and F are conducted by running simulations using Nokia Bell Labs system-level simulator. That is an object-oriented simulator developed on C++ platform well suited for LTE and 5G NR simulations with high degree of realism. The system structure and the applied assumptions are based on complex mathematical models in line with 3GPP guidelines.

Developing a new feature needs a thorough understanding of simulator architecture and identifying detailed modelling of network elements and RRM mechanisms such as channel measurements, cell connectivity, packet scheduling, link adaptation, etc. After successful implementing an idea, an extensive set of evaluations is conducted to test the accuracy and verify the performance of the solution. Finally, the results are validated against other companies. The main contributions and developed features are listed as follows

- **Multi-cell connectivity:** Implementing this feature, the mobile terminal periodically measures cell-specific received signal received power (RSRP) signals. It connects simultaneously to a set of maximum Q cells. CSI is calculated for the connected cells and reported to the network.
- **C-RAN architecture:** It is a platform for centralized scheduling. This enables dynamic time/frequency domain multi-cell multi-user scheduling on a fast TTI basis and allows inter-cell interference coordination.
- **Centralized scheduling algorithms:** Several centralized radio resource allocation schemes were built on the C-RAN platform.
- **Distributed QoS-specific scheduling:** A new multi-service scheduling was developed for URLLC and eMBB traffic.
- **Detailed URLLC statistics:** New features are implemented to collect the performance-related statistics including queuing delay, mobile connectivity, etc.

6 Thesis Outline

This dissertation is outlined in five main chapters. As the thesis has the form of a collection of papers, the detailed research questions, contributions, and performance evaluations are discussed in published/submitted articles that

are presented in Sections II, III, and IV. The chapters begin with a brief introduction and with a description of the objectives and problem formulation. Followings, the main contribution, recommendations, and the respective papers are presented.

The outline of this PhD thesis is described as follows:

- **Part I: Introduction -** This indicates the current chapter. It outlines a brief overview of cellular communications followed by a description 5G NR specifications. Motivations for PhD research are presented following by the relevant state of the art literature review. Next, the hypothesis and research methodology are given. Finally, the main contributions and findings are provided.

- **Part II: Interference Subspace Coordination -** In this part, interference management techniques are studied to address the first research question. It is composed of paper A where we discuss interference management for 5G networks. Especially, a derivative of interference alignment as so called ICISC is investigated. Two variants of ICISC is proposed and the performance results are evaluated for two different network deployments.

- **Part III: Resource Allocation for Distributed and Centralized Deployments -** This chapter answers to the research questions Q2, Q3, Q4 that includes papers B, C, D, E, and F. Several distributed and centralized schemes are discussed for URLLC. An attractive packet scheduling is proposed. The impacts of packet segmentation, the overhead of sending control channel, frequency-selective scheduling, and QoS aware resource allocation are investigated. Variants of multi-cell centralized scheduling are discussed. Detailed system-level simulation results are provided to compare the performance of different scheduling schemes.

- **Part IV: Data and Control Channel Scheduling, PDCP Packet Duplication -** It discusses aspects related to metadata and data allocation to answer the question Q5. To this end, two transmission schemes are investigated. URLLC performance is analysed by taking into account the impacts of sending scheduling grants and error in the feedback channel. A joint link adaptation of data and metadata is proposed to minimize the number of allocated resources while guaranteeing URLLC. The detailed solution is provided in paper G.

  Secondly, we investigate data duplication to address the research question Q5. This includes paper H which analyses the reliability enhancement and resource usage for PDCP packet duplication.

- **Part V: Conclusions -** It summarizes the main findings and contributions of this dissertation. Recommendations and study directions for

future studies are provided.

References

[1] J. Steenbruggen, M. T. Borzacchiello, P. Nijkamp, and H. Scholten, "Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities," *GeoJournal*, vol. 78, no. 2, pp. 223–243, 2013.

[2] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G evolution: HSPA and LTE for mobile broadband*. Academic press, 2010.

[3] ITU-R M.2134, "Requirements related to technical performance for IMT-Advanced radio interface(s)," November 2008.

[4] 3GPP TR 25.913 version 8.0.0 Release 8, "Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)," January 2009.

[5] D. Astely, E. Dahlman, A. Furuskär, Y. Jading, M. Lindström, and S. Parkvall, "LTE: the evolution of mobile broadband," *IEEE Communications Magazine*, vol. 47, no. 4, pp. 44–51, April 2009.

[6] A. Larmo, M. Lindström, M. Meyer, G. Pelletier, J. Torsner, and H. Wiemann, "The LTE link-layer design," *IEEE Communications Magazine*, vol. 47, no. 4, pp. 52–59, April 2009.

[7] N. Abu-Ali, A. M. Taha, M. Salah, and H. Hassanein, "Uplink scheduling in LTE and LTE-Advanced: Tutorial, survey and evaluation framework," *IEEE Communications Surveys Tutorials*, vol. 16, no. 3, pp. 1239–1265, March, 2014.

[8] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "LTE-advanced: next-generation wireless broadband technology," *IEEE Wireless Communications*, vol. 17, no. 3, pp. 10–22, June 2010.

[9] S. Parkvall, A. Furuskar, and E. Dahlman, "Evolution of LTE toward IMT-advanced," *IEEE Communications Magazine*, vol. 49, no. 2, pp. 84–91, February 2011.

[10] GSMA Intelligence 2019, "Requirements for Evolved UTRA (E-UTRA)." [Online]. Available: https://www.gsma.com

[11] CISCO, "VNI mobile forecast highlights," https://www.cisco.com/c/m/en_us/solutions/service-provider/forecast-highlights-mobile.html.

[12] ERICSSON, "Internet of things forecast," https://www.ericsson.com/en/mobility-report/internet-of-things-forecast.

[13] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[14] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: opportunities and challenges," *Wireless Networks*, vol. 21, no. 8, pp. 2657–2676, November 2015.

[15] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, 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.

[16] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[17] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, third-quarter 2016.

[18] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, May 2017.

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

Part II

Interference Subspace Coordination in 5G Networks

Interference Subspace Coordination in 5G Networks

This part of PhD study focuses on interference management for fifth generation (5G) networks by studying distributed interference mitigation schemes inspired by interference alignment (IA).

1 Problem Description

As discussed in the previous section, network densification is one of the promising 5G key solutions to achieve extreme 5G requirements [1, 2]. Increasing the density of base station (BS)s enhances the signal quality for the cell-edge users, fills the holes, expands the coverage, and increases the user experienced data rate [3, 4]. To fully realize the gains of this concept, several issues need to be addressed, including cell association, power control, and most crucial interference management.

There are various technical solutions dealing with interference in different aspects. Time and frequency resource partitioning is one of the main drivers to reduce the user experienced interference [5]. Also, significant performance improvements can be achieved via proper cell/user association and power control mechanisms [6, 7]. In 5G, enhanced coordination between transmitters and utilization of advanced interference rejection receivers are the principal building blocks for interference management [8]. In this respected, advanced interference cancellation/suppression receivers are utilized to estimate the interferer signals and cancel them from the received signal or suppress the interference by means of linear combining of received signals from receiver antennas [9–11].

Recently, growing attention is being driven to IA techniques as a potential solution to cope with interference challenges [12]. IA attempts to align the received interference from multiple sources at a particular receiver subspace by coordinating the transmission across the different terminals [13]. In this way, the interference signal lies in a reduced dimensional subspace at each re-

ceiver, thereby enabling complete interference cancellation. Although highly appealing, IA solutions suffer from several major drawbacks. Namely, the requisite on perfect and global channel state information (CSI), and the need for recursive information exchange between transmitters and receivers [14]. Another limitation of IA is high computation complexity, which grows exponentially with the network size [15, 16]. Therefore, the application of this concept for practical deployments needs further investigation. In particular, it is desirable to have a derivative solution with relatively low-computational complexity that can be implemented in distributed radio access networks.

2 Objectives

This part of PhD study has the following objectives

- Identify the potential and limitations of IA for practical deployments.
- Investigate a low-complexity distributed interference coordination scheme inspired by IA.
- Evaluate the performance the proposed solution for different levels of channel information and network deployment scenarios.

3 Included Articles

Following article includes the main finding of this research part:

Paper A. Inter-Cell Interference Sub-space Coordination for 5G Ultra-Dense Networks

This article studies inter-cell interference sub-space coordination (ICISC) for downlink transmission. We analyse the average user throughput for multiple-input multiple-output (MIMO) communications assuming that the users are equipped with a linear minimum mean square error (MMSE) receiver. An optimization problem is formulated to maximize the sum achievable throughput by designing proper precoder matrices at the transmitter ends. This is a non-convex optimization problem difficult to solve in polynomial time.

A sub-optimal distributed algorithm based on IA is investigated. In particular, the receiver space is divided into two desired and interference subspaces. The goal is to design precoder matrices at the base stations to transmit the signal to the desired subspace of the objective receiver while projecting into the interference subspace of interfered users. To this end, an algorithm is proposed to minimize the weighted squared chordal distance between the

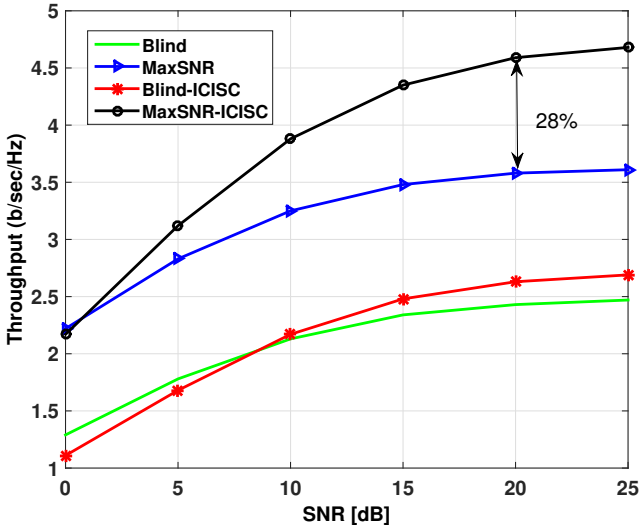

Fig. II.1: Average user throughput versus SNR for Ultra-dense network deployment [17].

transmitted signal and the desired receiver subspace and maximize the distance between the generated interference and desired subspace of other users.

Monte Carlo simulations are provided to evaluate the performance of the proposed algorithm for different user experienced signal to noise ratio (SNR) regimes and network topologies.

4 Main Findings and Recommendations

Fig. II.1 illustrates the average user achievable throughput for two variations of the proposed ICISC, namely as *blind-ICISC* and *MaxSNR-ICISC*. The results are compared against the blind (random) encoding and the well-known *MaxSNR* scheme.

The results show that the proposed ICISC performs well in ultra-dense network deployments where users receive powerful signals and interference. MaxSNR-ICISC achieves up to 28% throughput enhancement in comparison to MaxSNR encoding. It is shown that proper design of desired/interference subspaces has a significant impact on the performance, and improves the achievable throughput by more than 100%. For scenarios with low-density of interference (e.g., urban macro deployment), MMSE receivers efficiently suppress the received interference. Therefore, ICISC offers relatively lower throughput gain. Thus, it is not recommended for such environments.

References

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

[2] S. Chen and J. Zhao, "The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 36–43, May 2014.

[3] B. Basutli, J. M. Chuma, and S. Lambotharan, "Network capacity enhancement in HetNets using incentivized offloading mechanism," *IEEE Access*, vol. 6, pp. 39 307–39 323, 2018.

[4] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, third-quarter 2016.

[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*, vol. 20, no. 3, pp. 120–127, June 2013.

[6] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schober, and L. Hanzo, "User association in 5G networks: A survey and an outlook," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1018–1044, Second-quarter 2016.

[7] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution toward 5g multi-tier cellular wireless networks: An interference management perspective," *IEEE Wireless Communications*, vol. 21, no. 3, pp. 118–127, June 2014.

[8] 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.

[9] V. Fernández-López, K. I. Pedersen, B. Soret, J. Steiner, and P. Mogensen, "Improving dense network performance through centralized scheduling and interference coordination," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4371–4382, May 2017.

[10] C. Wei and W. Liao, "Multi-cell cooperative scheduling for network utility maximization with user equipment side interference cancellation," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 619–635, Jan 2018.

[11] V. Fernández-López, K. I. Pedersen, J. Steiner, B. Soret, and P. Mogensen, "Interference management with successive cancellation for dense small cell networks," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.

[12] Y. Zhou, L. Liu, H. Du, L. Tian, X. Wang, and J. Shi, "An overview on inter-cell interference management in mobile cellular networks: From 2G to 5G," in *2014 IEEE International Conference on Communication Systems*, Nov 2014, pp. 217–221.

[13] M. A. Maddah-Ali, A. S. Motahari, and A. K. Khandani, "Communication over MIMO X channels: Interference alignment, decomposition, and performance analysis," *IEEE Tr. Inf. Theory*, vol. 54, no. 8, pp. 3457–3470, Aug 2008.

[14] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the K-user interference channel," *IEEE Transactions on Information Theory*, vol. 54, no. 8, pp. 3425–3441, August 2008.

[15] K. Gomadam, V. R. Cadambe, and S. A. Jafar, "Approaching the capacity of wireless networks through distributed interference alignment," in *Proc. IEEE GLOBECOM*, New Orleans, LO, November. 2008, pp. 1–6.

[16] F. Sun and E. de Carvalho, "A leakage-based MMSE beamforming design for a MIMO interference channel," *IEEE Signal Processing Letters*, vol. 19, no. 6, pp. 368–371, June 2012.

[17] A. Karimi, N. H. Mahmood, K. I. Pedersen, and P. Mogensen, "Inter-cell interference sub-space coordination for 5G ultra-dense networks," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, September 2017, pp. 1–5.

Paper A

Inter-Cell Interference Sub-space Coordination for 5G Ultra-Dense Networks

Ali Karimi, Nurul Huda Mahmood, Klaus Ingemann Pedersen, and Preben Mogensen

The paper has been published in the
IEEE 86th Vehicular Technology Conference (VTC-Fall), 2017.

Abstract

In this paper, we present an inter-cell interference subspace coordination scheme for multiple-input multiple-output communications. The method relies on downlink precoding design for distributed multi-cell multi-user networks. In the proposed method, receivers benefit from minimum-mean square error structure. Each receiver separates the received signal space into desired/interference sub-spaces. The key idea in this contribution is to employ precoding algorithms at the transmission-end with the objective of jointly projecting transmitted signal over desired subspace and aligning most of the interference onto predefined interference subspaces at the interfered receiver end. This idea works with only local channel state information available at transmitter-side and benefits from low computational complexity. Simulation results show that the proposed method offers about 28% *throughput enhancements in networks with high dominant interference regimes.*

1 Introduction

The wireless channel is inherently broadcast in nature, where multiple links interfere with each other. Such broadcast nature of the wireless channel has traditionally been viewed as a disadvantage, and was addressed by carefully designing systems to avoid the interference. General interference avoidance techniques involve orthogonalizing transmit resources in time, frequency, space, or codes. Recently, however, there has been a paradigm shift from interference avoidance to interference exploitation, where different interfering sources coordinate their transmission to form an exploitable structure to the interference at the receivers [1].

Interference, unlike noise, can potentially be pre-shaped to give it an exploitable structure. Interference Alignment (IA) is such an "interference-shaping" technique. The main idea in IA is to align the transmission of signals from different transmitters such that all the unwanted interference at a particular interfered receiver overlap in the same signal sub-space. This allows a transmitter-receiver pair to communicate interference free over the dimensions in which the interference signal is not present. Each user in a fully connected K user wireless interference channel can achieve $K/2$ degrees-of-freedom (DoF) by employing IA principles [2].

The fifth generation (5G) cellular network will adopt dense small cells equipped with multiple input multiple output (MIMO) as one of the solutions towards providing flexible wireless connectivity with high spectral efficiency [3]. Efficient interference management in such dense small cell networks is a key research challenge [4], [5]. The natural separation of interference and information signal subspaces through IA lends itself as an effective interference management principle. However, IA techniques suffer

from several major drawbacks. Namely, the requisite on perfect and global channel state information (CSI) [2], and the need for recursive information exchange between transmitters and receivers [6–8]. Other limitations of IA include, high computation complexity which grows exponentially with the network size [9], and the fact that the DoF capacity approximations are often too weak to make accurate predictions about the performance at finite signal-to-noise ratios (SNRs) [10].

This work proposes a derivative of the IA principle specifically targeted for low complexity distributed implementation. Our proposal, titled inter-cell interference subspace coordination (ICISC), involves two steps. First, each receiver designates orthogonal subspaces for the desired signal and the inter-cell interference (ICI) from neighbouring nodes, which are communicated to the transmitters. The transmission at each transmitter is then shaped to project the signal over the desired subspace of its corresponding receiver, while the interference is aligned within the designated 'interference subspace' at the interfered receivers. The designation of the 'desired/interference subspaces' removes the burden of global CSI and/or iterative information exchange. The exponential complexity of the traditional IA solution is also avoided since, the transmit precoder is chosen independently at each transmitter. Furthermore, the freedom to choose the 'interference subspace' allows room for optimizing different performance criteria.

The remainder of this paper is organized as follows: the system model is elaborated in Section 2. Section 3 discusses the problem formulation and details the proposed interference coordination scheme. Simulation results evaluating the performance of the proposed algorithms are presented in Section 4. Finally, Section 5 concludes the paper.

Notations

We use upper-case $\mathbf{H}$ and lower-case $\mathbf{h}$ boldface characters for matrices and column vectors, respectively. Conjugate and conjugate transpose of a matrix $\mathbf{A}$ are denoted by $\mathbf{A}^*$ and $\mathbf{A}^H$, respectively. $\mathbb{C}^{m\times n}$ denotes a $m \times n$ complex matrix . $\mathbf{I}_d$ represents $d \times d$ identity matrix. The trace, determinant, and Frobenius norm operators of a matrix are denoted by $\mathrm{Tr}(.)$, $\det(.)$, and $\| \, . \, \|_F$, respectively. $\mathbb{C}^a$ represents the complex a-dimensional space. The null space is denoted by $\mathcal{N}(.)$. $\mathcal{CN}(\mu, \sigma^2)$ represents the complex Gaussian distribution with mean μ and variance σ^2. and $\mathcal{U}(a, b)$ denotes the uniform distribution with support between a and b, where $(a < b)$. All logarithms are base 2.

2 System Model

Let us consider a downlink multi-cell MIMO interference network of L cells operating at the same resource unit (time-frequency). Each cell consists of

one active pair of transmitter/receiver. The transmitter/receiver in the i^{th} cell are equipped with N_i and M_i antennas, respectively. A set of d_i ($d_i \leq \min(M_i, N_i)$) data symbols is transmitted at each transmission block. The received signal at the i^{th} user ($\mathbf{y}_i \in \mathbb{C}^{M_i \times 1}$) can be expressed as

$$\mathbf{y}_i = \underbrace{\sqrt{\alpha_{ii}\frac{P}{d_i}}\mathbf{H}_{ii}\mathbf{W}_i\mathbf{x}_i}_{\text{Desired Signal}} + \underbrace{\sum_{j\neq i}\sqrt{\alpha_{ij}\frac{P}{d_j}}\mathbf{H}_{ij}\mathbf{W}_j\mathbf{x}_j}_{\text{Interference}} + \underbrace{\mathbf{n}_i}_{\text{Noise}}, \tag{A.1}$$

where P and α_{ij} denote the transmitted power and the path-loss factor between the i^{th} receiver and the j^{th} transmitter, respectively. $\mathbf{H}_{ij} \in \mathbb{C}^{M_i \times N_j}$ and $\mathbf{W}_j \in \mathbb{C}^{N_j \times d_j}$ respectively denote the channel matrix between the i^{th} receiver and the j^{th} transmitter and the j^{th} transmitter precoding matrix with orthonormal columns ($\mathbf{W}_j^H\mathbf{W}_j = \mathbf{I}_{d_j}\ \forall j$). $\mathbf{x}_j \in \mathbb{C}^{d_j \times 1}$ represents signal stream of the j^{th} transmitter, while the additive white Gaussian noise vector at the i^{th} receiver is denoted by $\mathbf{n}_i \in \mathbb{C}^{M_i \times 1}$. We assume that the elements of $\mathbf{H}_{ij}$, $\mathbf{x}_i$, and $\mathbf{n}_i$ are $\sim \mathcal{CN}(0,1)$ with independent and identical distribution (iid). To suppress the noise and received interference, user i exploits a linear minimum-mean square error (MMSE) filter of $\mathbf{F}_i = \mathbf{G}_{ii}^H(\sum_{j=1}^{L}\mathbf{G}_{ij}\mathbf{G}_{ij}^H + \mathbf{I}_{Mi})^{-1}$, where $\mathbf{G}_{ij} = \sqrt{\alpha_{ij}P/d_i}\mathbf{H}_{ij}\mathbf{W}_j$ represents the effective channel between the i^{th} receiver and the j^{th} transmitter [7]. It can be shown that the bandwidth normalized achievable throughput (TP) of the i^{th} receiver can be expressed as [11]:

$$R_i = \log\det\left(\mathbf{I}_{Ni} + \mathbf{G}_{ii}^H\left(\sum_{j\neq i}\mathbf{G}_{ij}\mathbf{G}_{ij}^H + \mathbf{I}_{Mi}\right)^{-1}\mathbf{G}_{ii}\right). \tag{A.2}$$

3 Problem Formulation and the Proposed Solution

3.1 Problem Formulation

The optimum precoding matrices that maximize the network sum TP can be achieved by solving the following optimization problem

$$\max_{\mathbf{W}_i} R_T = \sum_{i=1}^{L} R_i \quad \text{Subject to } \mathbf{W}_i^H\mathbf{W}_i = \mathbf{I}_{d_i}\ \forall i.$$

Note that (A.3) is an non-convex problem and the optimum closed form solution cannot be easily derived [12]. To improve the network TP, we target a decentralized precoding scheme by investigating interference subspace

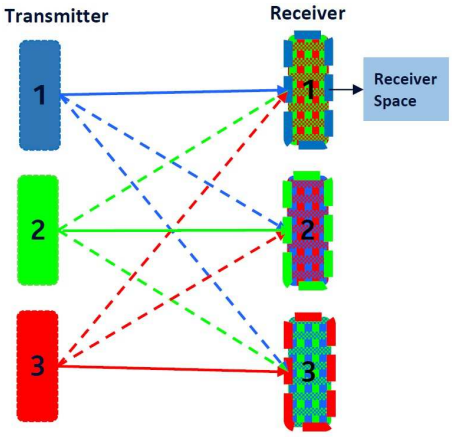

Fig. A.1: Conventional transmission scheme.

coordination between the transmitters and receivers. Fig. A.1 shows the conventional transmission scheme where there is no distinction between the received signal and the interference space. Suppressing multiple interference streams from different sources sharing the same space with the desired signal is challenging for receivers with simple structures (e.g. MMSE) and results in reduced sum network TP.

The main focus of our proposed algorithm is to design a precoding scheme that separates the desired signal subspace from the interference subspace at the receiver-end as conceptualized in IA. The goal is to align the interfering streams in a subspace orthogonal to the desired received signal in order to improve the network TP. It is worth noting that unlike [2] and [6], we are looking for a solution with low complexity, good performance with local CSI, and low recursive signal and information exchange overhead between the transmitters and the receivers.

3.2 Proposed Solution

The proposed solution is conceptually presented in Fig. A.2. The receiver signal space is separated into two orthogonal subspaces: desired ($\mathbf{S}^D$) and interference subspace ($\mathbf{S}^I = \mathcal{N}(\mathbf{S}^D)$). Each receiver broadcasts the desired/interference subspaces to all the transmitters. Each transmitter designs the precoding matrix such that most of the desired signal is projected over $\mathbf{S}^D$ and most of the interference is aligned into $\mathbf{S}^I$ at all interfered nodes.

Due to the lack of global CSI and the limited DoF, perfect separation is not always feasible. We therefore investigate the solution by jointly minimizing the distance between the generated signal/interference at the transmitter side and the defined desired/interference subspace at the receiver side. For ex-

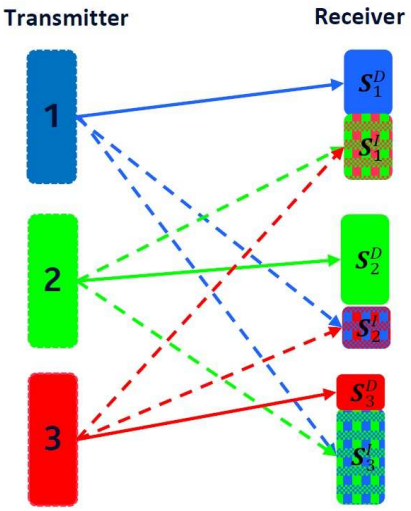

Fig. A.2: Inter-cell interference subspace coordination (ICISC) scheme.

ample, the i^{th} transmitter constructs its precoding matrix to jointly minimize the squared distance between its signal space ($\mathbf{S}_{ii}^{C}$) and $\mathbf{S}_{i}^{D}$ and the distance between the generated interference ($\mathbf{S}_{ji}^{C}$, $j \neq i$) and $\mathbf{S}_{j}^{I}$, $\forall j \neq i$. Exploiting this idea improves the capability of the receiver filters to suppress the interference more efficiently.

The chordal distance for two subspaces of $(\mathbf{S}_a, \mathbf{S}_b)$ with equal dimensions is defined as follow [13]:

$$\begin{aligned} d\left(\mathbf{S}_a, \mathbf{S}_b\right) &= \frac{1}{\sqrt{2}} \parallel \mathbf{V}_a - \mathbf{V}_b \parallel_F \\ &= \left[\frac{1}{2}\mathrm{Tr}\left[\left(\mathbf{V}_a - \mathbf{V}_b\right)\left(\mathbf{V}_a - \mathbf{V}_b\right)^H\right]\right]^{\frac{1}{2}}, \end{aligned} \tag{A.3}$$

where $\mathbf{V}_a$ denotes the orthogonal projector onto $\mathbf{S}_a$. For the subspace of $\mathbf{S}_{ij}^{C}$ originated by the transmission in the channel between the i^{th} receiver and j^{th} transmitter, the orthogonal projector matrix ($\mathbf{V}_{ij}^{C}$) is defined as [14]:

$$\mathbf{V}_{ij}^{C} = \mathbf{H}_{ij}\mathbf{W}_j\left(\mathbf{W}_j^H\mathbf{H}_{ij}^H\mathbf{H}_{ij}\mathbf{W}_j\right)^{-1}\mathbf{W}_j^H\mathbf{H}_{ij}^H. \tag{A.4}$$

A decentralized optimization problem can be formulated at each transmitter-side to minimize the weighted squared distances between the generated signal/interference subspaces and the defined desired/interference subspaces at the receivers. Considering the j^{th} transmitter, the problem can be formulated

as:

$$\min_{\mathbf{W}_j} \text{Cf}(\mathbf{W}_j) \quad \text{Subject to } \mathbf{W}_j^H \mathbf{W}_j = \mathbf{I}_{d_j} \ \ \forall j, \tag{A.5}$$

where $\text{Cf}(\mathbf{W}_j) = \beta_{jj} d(\mathbf{S}_{jj}^C, \mathbf{S}_j^D)^2 - \sum_{i \neq j} \beta_{ij} d(\mathbf{S}_{ij}^C, \mathbf{S}_i^D)^2$. The first term is for minimizing the squared distance between the generated signal and the desired subspace of the j^{th} receiver. The second term is for maximizing the squared distance between the generated interference and the desired subspace of the other receivers, which can be interpreted as projecting the interference onto the interference subspace of interfered nodes. β_{ij} represents the weight coefficient of each distance term in the optimization problem. In this work, we consider the normalized power of the generated signal/interference and define β_{ij} as:

$$\beta_{ij} \triangleq \frac{\alpha_{ij}}{\sum_{i=1,}^{L} \alpha_{ij}}, \quad \forall i. \tag{A.6}$$

In order to solve (A.5) we investigate the following properties of $\mathbf{V}_{ij}^C$ as [14]:

- The column space of $\mathbf{V}_{ij}^C$ is equal to the subspace $\mathbf{S}_{ij}^C$.
- $\text{Tr}(\mathbf{V}_{ij}^C) = d_j$.
- $\mathbf{V}_{ij}^C = (\mathbf{V}_{ij}^C)^H = (\mathbf{V}_{ij}^C)^2$.

The cost function $(\text{Cf}(\mathbf{W}_j))$ in (A.5), can be expanded as:

$$\begin{aligned}
\text{Cf}(\mathbf{W}_j) &= \frac{\beta_{jj}}{2} \left\| \mathbf{V}_{jj}^C - \mathbf{V}_j^D \right\|_F^2 - \frac{1}{2} \sum_{i \neq j} \beta_{ij} \left\| \mathbf{V}_{ij}^C - \mathbf{V}_i^D \right\|_F^2 \\
&= \frac{\beta_{jj}}{2} \text{Tr} \left[\left(\mathbf{V}_{jj}^C - \mathbf{V}_j^D \right) \left(\mathbf{V}_{jj}^C - \mathbf{V}_j^D \right)^H \right] \\
&- \frac{1}{2} \sum_{i \neq j} \beta_{ij} \text{Tr} \left[\left(\mathbf{V}_{ij}^C - \mathbf{V}_i^D \right) \left(\mathbf{V}_{ij}^C - \mathbf{V}_i^I \right)^H \right] \\
&= d_j \left(2\beta_{jj} - 1 \right) \\
&- \beta_{jj} \text{Tr} \left[\mathbf{H}_{jj} \mathbf{W}_j \left(\mathbf{W}_j^H \mathbf{H}_{jj}^H \mathbf{H}_{jj} \mathbf{W}_j \right)^{-1} \mathbf{W}_j^H \mathbf{H}_{jj}^H \left(\mathbf{V}_j^D \right)^H \right] \\
&+ \sum_{i \neq j} \beta_{ij} \text{Tr} \left[\mathbf{H}_{ij} \mathbf{W}_j \left(\mathbf{W}_j^H \mathbf{H}_{ij}^H \mathbf{H}_{ij} \mathbf{W}_j \right)^{-1} \mathbf{W}_j^H \mathbf{H}_{ij}^H \left(\mathbf{V}_i^D \right)^H \right].
\end{aligned} \tag{A.7}$$

Since there is no closed form solution for (A.7), numerical algorithms are investigated in this work to find the solution. Similar to the objective functions in [14, 15], one can see that (A.7) is invariant to any rotation from multiplying the precoder matrix by any unitary matrices $(\text{Cf}(\mathbf{W}_j) = \text{Cf}(\mathbf{W}_j \mathbf{U}), \forall \mathbf{U}$:

$\mathbf{U}^H\mathbf{U} = \mathbf{I}_{d_j}$). This implies that the solution can be found over the complex Grassmannian manifold which significantly reduces the computational complexity [16]. For the j^{th} transmitter, the complex Grassmannian manifold $\mathrm{Gr}(N_j, d_j)$ is defined as the set of all d_j-dimensional complex subspaces of $\mathbb{C}^{N_j}$. The precoding matrix of the j^{th} transmitter can be obtained using the steepest descent method as outlined in Algorithm 1.

Algorithm 1

1: Start with an arbitrary precoding matrix $\mathbf{W}_j$ such that $\mathbf{W}_j^H\mathbf{W}_j = \mathbf{I}_{d_j}$ and set the step size $\gamma = 1$.
2: Compute $\frac{\partial \mathrm{Cf}(\mathbf{W}_j)}{\partial \mathbf{W}_j^*}$. The extended form of $\frac{\partial \mathrm{Cf}(\mathbf{W}_j)}{\partial \mathbf{W}_j^*}$ can be found in (A.8) above the next page.
3: Compute the steepest descent direction as:
$\mathbf{Z} = -(\mathbf{I}_{N_j} - \mathbf{W}_j\mathbf{W}_j^H)\frac{\partial \mathrm{Cf}(\mathbf{W}_j)}{\partial \mathbf{W}_j^*}$.
4: **while** $\mathrm{Tr}\left[\mathbf{Z}\mathbf{Z}^H\right] > \varepsilon$ **do**
5: **while** $\mathrm{Cf}(\mathbf{W}_j) - \mathrm{Cf}(\mathbf{W}_j + 2\gamma\mathbf{Z}) \geq \gamma\mathrm{Tr}\left[\mathbf{Z}\mathbf{Z}^H\right]$ **do**
6: $\gamma = 2\gamma$.
7: **end while**
8: **while** $\mathrm{Cf}(\mathbf{W}_j) - \mathrm{Cf}(\mathbf{W}_j)(\mathbf{W}_j + \gamma\mathbf{Z}) < \frac{\gamma}{2}\mathrm{Tr}\left[\mathbf{Z}\mathbf{Z}^H\right]$ **do**
9: $\gamma = \frac{\gamma}{2}$.
10: **end while**
11: Perform the QR decomposition of $\mathbf{W}_j + \gamma\mathbf{Z}$.
$\mathbf{W}_j + \gamma\mathbf{Z} = \mathbf{Q}\mathbf{R}$. Set $\mathbf{W}_j = \mathbf{Q}^{1:d_j}$.
12: Update $\mathbf{Z}$.
13: **end while**
14: **return** $\mathbf{W}_j$.

$$\frac{\partial \mathrm{Cf}(\mathbf{W}_j)}{\partial \mathbf{W}_j^*} = -\beta_{jj}\left(\mathbf{W}_j^H\mathbf{H}_{jj}^H\mathbf{H}_{jj}\mathbf{W}_j\right)^{-1}$$
$$\times\left[\mathbf{H}_{jj}^H\mathbf{V}_j^D\mathbf{H}_{jj}\mathbf{W}_j - \mathbf{H}_{jj}^H\mathbf{H}_{jj}\mathbf{W}_j\left(\mathbf{W}_j^H\mathbf{H}_{jj}^H\mathbf{H}_{jj}\mathbf{W}_j\right)^{-1}\mathbf{W}_j^H\mathbf{H}_{jj}^H\mathbf{V}_j^D\mathbf{H}_{jj}\mathbf{W}_j\right]$$
$$+\sum_{i\neq j}^{L}\beta_{ij}\left[\mathbf{H}_{ij}^H\mathbf{V}_i^I\mathbf{H}_{ij}\mathbf{W}_j\right]\left(\mathbf{W}_j^H\mathbf{H}_{ij}^H\mathbf{H}_{ij}\mathbf{W}_j\right)^{-1}$$
$$-\sum_{i\neq j}^{L}\beta_{ij}\left[\mathbf{H}_{ij}^H\mathbf{H}_{ij}\mathbf{W}_j\left(\mathbf{W}_j^H\mathbf{H}_{ij}^H\mathbf{H}_{ij}\mathbf{W}_j\right)^{-1}\mathbf{W}_j^H\mathbf{H}_{ij}^H\mathbf{V}_i^D\mathbf{H}_{ij}\mathbf{W}_j\right]\left(\mathbf{W}_j^H\mathbf{H}_{ij}^H\mathbf{H}_{ij}\mathbf{W}_j\right)^{-1}. \tag{A.8}$$

One of the main aspects of the proposed algorithm is that it does not need iterative information exchange and updated solutions between transmitters and receivers as in [6, 7]. This makes the proposed method attractive for implementation in practical systems. Moreover, each transmitter can access the local CSI by reciprocity of uplink and downlink channels in time-division-duplex (TDD) transmission. Also, the optimization problem is solved at the transmitters side which usually benefits from powerful processors and imposes less complexity to the receiver-end for the downlink case.

4 Simulation Results

This section provides numerical results in order to evaluate the performance of the proposed ICISC scheme. The performance is compared against two baseline schemes, namely the conventional blind (random) and MaxSNR precoding which is proposed for LTE networks [17]. In blind transmission, the precoding matrix is composed of randomly generated orthonormal vectors. Whereas for MaxSNR, each transmitter selfishly select the precoding matrix to maximize the received SNR at its corresponding receiver. Simulation results are based on at least 1000 independent channel realizations. Block fading channel with Rayleigh distribution is considered. It is assumed that both transmitters and receivers have 4 antennas which is the basic MIMO configuration considered in 5G new radio. Ultra-dense small cell scenario is assumed where receivers suffers from high interference values [3]. For this scenario, the received interference power from each interferer node is modelled as a random variable with the distribution $\alpha_{ij}P(\text{dB}) \sim \mathcal{U}(\text{SNR-5}, \text{SNR})$. Results from two sub-space definitions are presented: blind-ICISC and MaxSNR-ICISC. For blind-ICISC, we consider pre-set $\mathbf{S}_i^D/\mathbf{S}_i^I$. Defined subspaces do not change during transmissions and with channel variations. As an example, for rank 2 transmission the desired and the interference subspaces of the i^{th} receiver can determined as:

$$\mathbf{S}_i^D = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{S}_i^I = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} . \tag{A.9}$$

In MaxSNR-ICISC, the i^{th} user determines the d_i dimensional part of the received space, which has the highest signal strength as the desired subspace. This subspace can be identified as the eigenvectors corresponding to the highest eigenvalues of $\mathbf{H}_{ii}\mathbf{H}_{ii}^H$. The null space of the desired subspace is considered as the interference subspace. The receivers broadcast their own desired subspace to the transmitters. Unlike blind-ICISC, in MaxSNR-ICISC, the defined subspaces are dependent on the channel parameters. Thus, the

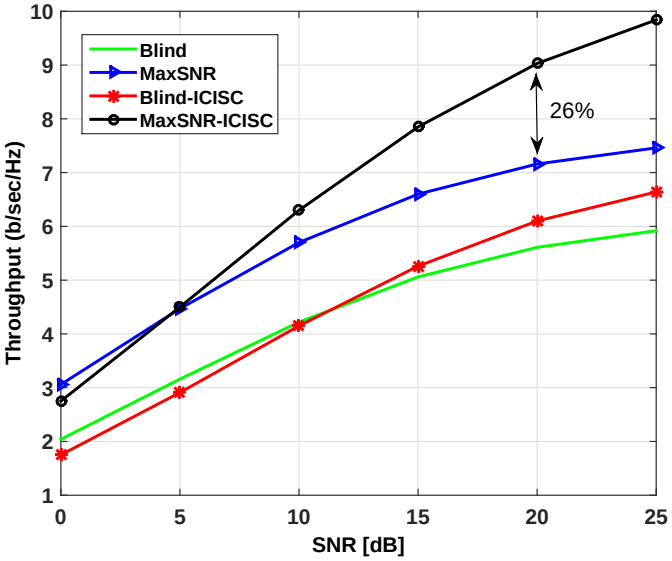

Fig. A.3: Average user TP vs. SNR for for Ultra-dense network of $L = 3$ cells, with the antenna size $M_l = N_l = 4$ and rank $d_l = 2$ transmission.

receiver needs to update the desired subspace and broadcast it to the transmitters as the channel changes.

Fig. A.3 shows the average user TP versus SNR for a network with $L = 3$, $M_l = N_l = 4$, $d_l = 2\ \forall l$ with two-dimensional $\mathbf{S}_i^D$ and $\mathbf{S}_i^I$. It can be seen that proposed precoding schemes outperform the conventional ones specially at high SNR regimes. As an example, at 20dB SNR, the MaxSNR-ICISC achieves 26% higher TP in comparison to the MaxSNR precoding. Blind-ICISC precoding shows 8% improvement over that of blind transmission. In low SNR regimes, where the impact of the interference is limited, MaxSNR precoding shows slightly better performance than MaxSNR-ICISC.

Fig. A.4 depicts the user TP versus SNR for a network with $L = 6$, $M_l = N_l = 4$, $d_l = 1\ \forall l$. For all SNR values, MaxSNR-ICISC can significantly suppress the interference and achieve higher throughput. It can be seen that the proposed scheme works well in ultra-dense small cell scenario where receivers experience high SNR values and suffer strong interference from neighbours. Numerical results show that setting $\varepsilon = 0.05$, the proposed algorithm converges after 50 and 16 iterations in 3 and 6 cell scenarios, respectively. Comparison between the blind-ICISC and the MaxSNR ICISC highlights the importance of appropriate subspace definition in subspace coordination methods.

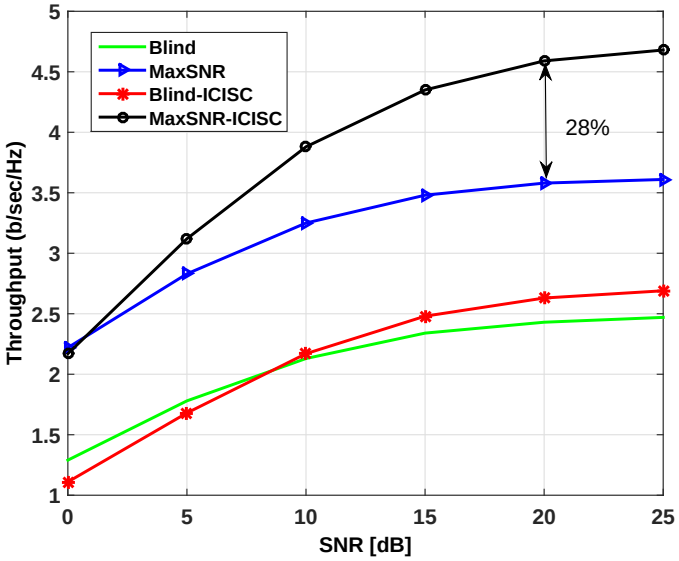

Fig. A.4: Average user TP vs. SNR for Ultra-dense network of $L = 6$ cells, with the antenna size $M_l = N_l = 4$ and rank $d_l = 1$ transmission.

Fig. A.5 evaluate the performance of the proposed scheme for macro-cell scenarios in which the received signal power is considerably stronger than the interference. In these scenarios, it is assumed that the interference has the distribution $\alpha_{ij}P(\text{dB}) \sim \mathcal{U}(\text{SNR-15}, \text{SNR})$. As can be seen, MaxSNR-ICISC proposes a better performance in network with low interference density. For example, it achieves about 12% higher TP in comparison to the MaxSNR scheme in networks with 3 and 6 cells.

5 Conclusion

In this paper, we have proposed a distributed interference management scheme. The proposed inter-cell interference subspace coordination scheme is a low complexity and low overhead derivative form of interference alignment. Each receiver first separates its received signal space into desired/interference subspaces. Then, the transmitters design their precoding matrices with the aim to project their signal to desired subspace of their corresponding receiver and align the interference to the interference subspace of interfered nodes. Simulations results show that the proposed scheme results in up to 28% TP gains over conventional LTE schemes in dense small cell scenarios. Around 10% TP gains are observed for macro scenarios. The proposed scheme is presented

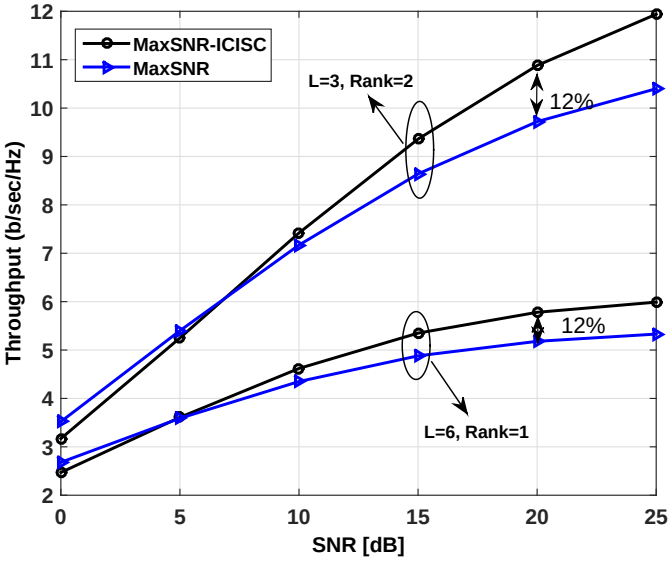

Fig. A.5: Average user TP vs. SNR for macro-cell scenarios of $L = 3, 6$ cells, with the antenna size $M_l = N_l = 4$.

considering broadband services. As part of the future work, we would like to extend the scheme to consider ultra reliable low latency service class as well.

Acknowledgement

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] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[2] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the K-user interference channel," *IEEE Transactions on Information Theory*, vol. 54, no. 8, pp. 3425–3441, Aug. 2008.

[3] N. H. Mahmood, M. Lauridsen, G. Berardinelli, D. Catania, and P. Mogensen, "Radio resource management techniques for eMBB and mMTC services in 5G dense small cell scenarios," in *Proc. 84th IEEE VTC Fall*, Montreal, Canada, September 2016.

[4] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, June 2014.

[5] I. F. Akyildiz, S. Nie, S.-C. Lin, and M. Chandrasekaran, "5G roadmap: 10 key enabling technologies," *Computer Networks*, vol. 106, pp. 17–48, 2016.

[6] K. Gomadam, V. R. Cadambe, and S. A. Jafar, "Approaching the capacity of wireless networks through distributed interference alignment," in *Proc. IEEE GLOBECOM*, New Orleans, LO, Nov. 2008, pp. 1–6.

[7] F. Sun and E. de Carvalho, "A leakage-based MMSE beamforming design for a MIMO interference channel," *IEEE Signal Processing Letters*, vol. 19, no. 6, pp. 368–371, June 2012.

[8] K. R. Kumar and F. Xue, "An iterative algorithm for joint signal and interference alignment," in *Proc. IEEE International Symposium on Information Theory*, Austin, TX, June 2010, pp. 2293–2297.

[9] C. Sun, Y. Yang, and Y. Yuan, "Low complexity interference alignment algorithms for desired signal power maximization problem of MIMO channels," *EURASIP Journal on Advances in Signal Processing*, no. 137, 2012.

[10] U. Niesen and M. A. Maddah-Ali, "Interference alignment: From degrees of freedom to constant-gap capacity approximations," *IEEE Tr. Inf. Theory*, vol. 59, no. 8, pp. 4855–4888, August 2013.

[11] N. H. Mahmood, G. Berardinelli, K. I. Pedersen, and P. Mogensen, "An interference-aware distributed transmission technique for dense small cell networks," in *Proc. IEEE ICCW*, London, UK, June 2015, pp. 119–124.

[12] B. O. Lee, H. W. Je, O. Shin, and K. B. Lee, "A novel uplink MIMO transmission scheme in a multicell environment," *IEEE Transactions on Wireless Communications*, vol. 8, no. 10, pp. 4981–4987, October 2009.

[13] A. Barg and D. Y. Nogin, "Bounds on packings of spheres in the Grassmann manifold," *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2450–2454, September 2002.

[14] S. Bazzi, G. Dietl, and W. Utschick, "Interference alignment via minimizing projector distances of interfering subspaces," in *Proc. 13th IEEE SPAWC*, June 2012, pp. 274–278.

[15] F. Ye, J. Yao, and Y. Li, "An interference alignment based on joint projector interference distances," in *Proc. 11th IEEE WiCOM*, Shanghai, September 2015, pp. 1–5.

[16] J. H. Manton, "Optimization algorithms exploiting unitary constraints," *IEEE Tr. Signal Processing*, vol. 50, no. 3, pp. 635–650, March 2002.

[17] J. Lee, J.-K. Han, and J. Zhang, "MIMO technologies in 3GPP LTE and LTE-Advanced," *EURASIP JWCN*, vol. 2009, no. 1, 2009.

Part III

Resource Allocation for Distributed and Centralized Deployments

Resource Allocation for Distributed and Centralized Deployments

This part of the thesis focuses on the downlink (DL) radio resource allocation (aka scheduling) for distributed and centralized deployments in fifth generation New Radio (5G NR). We follow NR agreement for system modelling and network configuration. Dynamic multi-cell multi-user system-level simulations in line with 5G NR guidelines are developed to obtain results and evaluate the solutions with high degrees of accuracy.

1 Problem Description

Radio resource allocation means dynamic assignment of available resources to users in order to accomplish with different service targets such as throughput, reliability, latency, coverage, etc. It encompasses a large variety of functionalities, including frame structure, time and frequency domain multiplexing, link adaptation, load balancing, etc. Most of the traditional resource allocation schemes are mainly focus on enhancing the spectrum efficiency and user experienced throughput [1, 2]. Therefore, they can not fulfil Ultra-Reliable Low-Latency Communications (URLLC) targets. Utilizing advanced techniques and specialized solutions are essential to maximizing resource efficiency with provisioning URLLC QoS requirements.

Extreme latency budget of URLLC needs adopting a flexible frame structure and user scheduling over short transmission time interval (TTI) lengths of only a few (e.g., two or four) symbols [3]. Employing powerful processors and low-complexity resource allocation schemes are also essential [4, 5]. Due to the sporadic nature of URLLC, co-scheduling URLLC and enhanced Mobile Broadband (eMBB) maximizes the spectrum efficiency [6]. In this respect, the URLLC traffic can be overlaid eMBB via punctured allocation [7, 8]

or, they can be multiplexed when shorter TTIs are adopted [9]. To maximize the benefits and enhance both services, dynamic resource partitioning and scheduling solutions are desirable. Different characteristics of each service, such as payload size, traffic dynamics, and transmission overhead need to be taken into account. The network should exploit the benefits of wireless channel and allocate resources in accordance with specific requirements of each service.

centralized radio access network (C-RAN) architecture is one of the 5G key technologies to overcome some of the hardnesses in radio resource management. Especially, a centralized unit that controls a set of remote radio head (RRH)s via zero-latency optic fiber connections, is an attractive solution for URLLC [10]. While numerous researches have investigated many aspects related to centralized scheduling in wireless networks, most of them focus on eMBB performance targets, such as maximization of the sum throughput, coverage, etc [11, 12]. Besides that, most of the researches are theoretical studies that consider simplified system models, yet they suffer from high computational complexity [13]. It is crucial to revisit the current ideas, update the model based on 5G features such as multiple-input multiple-output (MIMO) transmission and advanced receivers, and propose innovative solutions suitable for URLLC.

This chapter studies resource allocation for URLLC and eMBB through distributed and centralized architectures. The goal is to answer the research question of how to improve URLLC performance in different scenarios. This part of the thesis is conducted in a proposed framework for 5G NR. Therefore, a highly detailed NR compliant system model is adopted. An extensive set of components impacting the performance is taken into account . Those include bursty traffic, periodic channel measurement, MIMO transmission, adopting minimum mean square error (MMSE) receive filter, three dimensional (3D) radio propagation, dynamic link adaptation, HARQ retransmission, etc. We seek for low-complexity solutions that can be used for NR implementations.

2 Objectives

The goals of this part of PhD are summarized as following:

- Design and evaluate a new distributed packet scheduling solution for mixed URLLC and eMBB traffic.
- Study low-complexity centralized multi-cell scheduling schemes to improve URLLC performance for practical deployments.

3 Included Articles

Following papers form the main body of this part:

Paper B. Efficient Low-complexity Packet Scheduling Algorithm for Mixed URLLC and eMBB Traffic in 5G

This paper addresses the problem of downlink packet scheduling for mixed URLLC and eMBB traffic through the distributed architecture. An optimization problem is formulated for the dynamic allocation of time and frequency resources to users. As finding the optimal strategy comes with high computational complexity, a heuristic solution is investigated. The proposed method benefits from low-complexity and is designed by considering components affecting URLLC, such as QoS requirements, payload size, HARQ, and control channel overhead. Extensive system-level simulations following 5G NR guidelines are conducted for performance evaluation. The results are compared against the current state of the art scheduling schemes. In addition, sensitivity analysis of the results for different offered loads is presented, and the impact of queuing delay on the URLLC latency is investigated.

Paper C. 5G URLLC performance Analysis of Dynamic-Point Selection Multi-User Resource Allocation

This paper focuses on the application of C-RAN for URLLC by studying spectrum-efficient dynamic point selection (SE-DPS). SE-DPS is one of the C-RAN solutions that utilizes the potential gain of cell-selection diversity by fast-tracking variations in wireless channels. Here, a mobile terminal connects and periodically performs channel state information (CSI) measurement for a cluster of cells. A cell with the highest instantaneous throughput is reported to the centralized unit for downlink scheduling. This concept requires a limited number of resources for sending CSI in the uplink and has low complexity for cell allocation at the centralized unit. System-level simulations are performed to evaluate the performance of SE-DPS for URLLC. The impacts of offered load and wideband/sub-band channel measurements on the URLLC latency are studied. Moreover, the sensitivity of results versus different multi-cell connectivity parameters is analysed.

Paper D. Centralized Joint Cell Selection and Scheduling for Improved URLLC Performance

This paper studies dynamic load balancing and centralized multi-cell scheduling to further improve URLLC performance in downlink transmission. In this scenario, a mobile terminals connects to a cluster of cells and reports the CSIs of all connected cells. The centralized unit receives information

from all mobile terminals and makes scheduling decisions. In each transmission time interval, the user is scheduled from only one cell (i.e., single-node transmission). The paper focuses on the C-RAN capability of fast load balancing to reduce the queuing delay by offloading traffic from hotspots to less congested points. To this end, an optimization problem is formulated to determine cell allocation and user scheduling. For a set of U active (with buffered data) users, each connected to Q cells, finding the optimum strategy has the complexity of $\mathcal{O}((Q+1)^U)$. A sub-optimal algorithm based on sequential allocation is proposed. The solution has a reduced complexity of $\mathcal{O}(Q \cdot U \log(Q \cdot U))$. It is also aware of the physical downlink control channel (PDCCH) overhead to avoid costly segmentation of multiple URLLC payloads. Results from dynamic highly detailed system-level simulations are presented, and the achieved performance is compared against the baseline distributed scheduling.

Paper E. 5G Centralized Multi-Cell Scheduling for URLLC: Algorithms and System-Level Performance

This is an extended version of the previous study. The paper presents a detailed modelling of C-RAN and multi-cell scheduling. Those include a comprehensive description of network topology, transmission settings, channel measurement, multi-cell connectivity, link adaptation, etc. Some theoretical backgrounds are presented, and several scheduling metrics are evaluated for cell allocation. The impact of packet segmentation is investigated. A solution is proposed to reduce the overhead of control channel information by proper determining of the segmented user. System-level simulations with comprehensive discussions on the performance results are provided. In addition, the impact of different carrier bandwidths is analysed. It is shown that centralized scheduling leads to superior latency performance in comparison to the distributed solution. Related statistics and sensitivity analysis of the results versus different multi-cell connectivity setups are presented.

Paper F. Low-Complexity Centralized Multi-Cell Radio Resource Allocation for 5G URLLC

The paper is built on top of results from the previous studies and proposes improved centralized scheduling for 5G NR. Network settings are updated based on recent 3GPP 5G NR proposals for scheduling URLLC. A subcarrier spacing of 30 kHz is assumed. Based on the QoS requirement and payload size, the TTI length is set to two or four OFDM symbols. Processing times at the network and user-end are reduced to 2.75 and 3.25 OFDM symbols, respectively. In harmony with the previous findings, a new scheduling metric is defined. The latency budget (queuing delay) composes the main body of

the proposed metric to minimize the latency. It also has elements for smart channel and multi-user aware scheduling. In addition, frequency-selective scheduling is added to improve URLLC outage capacity.

The study includes dynamic system-level simulations of a network with 21 macro cells and 2100 mobile terminals to generate statistically reliable results. The centralized URLLC performance is evaluated for different payload sizes and is compared against distributed scheduling and dynamic point selection (DPS).

4 Main Findings and Recommendations

Distributed Packet scheduling for joint URLLC and eMBB traffic.

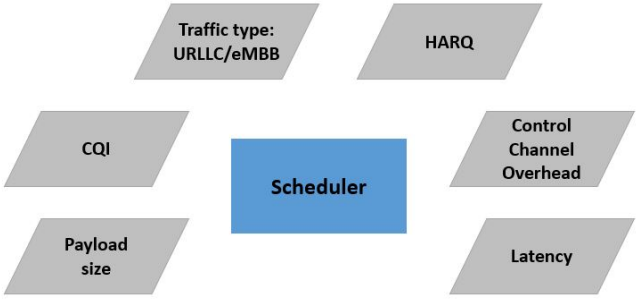

Fig. III.1: Parameters impacting scheduling decision [14].

Paper B presents a simple but efficient resource allocation solution for mixed URLLC and eMBB traffic. As Fig. III.1 illustrates, the solution is built on several essential elements for packet scheduling, including traffic type, channel quality indicator (CQI), hybrid automatic repeat request (HARQ), payload size, control channel overhead, and latency. In line with the requested services, users are prioritized for getting physical resources. Also, different block error rate (BLER) targets and user multiplexing algorithms are adopted to enhance the desired reliability/throughput of the applications.

Table III.1 summarizes the URLLC latency at the outage level of 10^{-5} and, also the average eMBB throughput for different scheduling scenarios [14]. The results show that the proposed scheme dramatically improves latency for URLLC, as well as the throughput of eMBB services. At 15 Mbps URLLC offered load, gains of 99% latency reduction for URLLC, and 17% increase in eMBB throughput are achieved in comparison to proportional fair (PF) scheduling. This is because in the structure of proposed solution, the latency budget is considered as the main factor for user selection in the time domain. Hence, users that are closer to the latency target are prioritized. By doing so,

Table III.1: Network performance for different URLLC offered loads [14].

	Scenario		Offered URLLC load [Mbps]			
			8	**10**	**12**	**15**
URLLC latency [msec]	PF		1.5	2.3	4.5	358
	M-LWDF		1.36	1.63	2.92	22.5
	Proposed		1.24	1.31	1.38	2.27
	Gain to	PF	18 %	57 %	70 %	99.4 %
		M-LWDF	9 %	20 %	53 %	90 %
eMBB throughput [Mbps/cell]	PF		25.3	20.8	16.2	9.2
	M-LWDF		25.3	20.8	16.3	9.3
	Proposed		25.6	21.3	17.07	10.8
	Gain to	PF	1.3 %	2.5 %	5.3 %	17.3 %
		M-LWDF	1.3 %	2.5 %	4.7 %	16.1 %

the latency considerably reduces at low outage level as is the key performance indicator (KPI) for URLLC. The cost of PDCCH is minimized by limiting segmentation of multiple payloads to a maximum one payload with a smaller overhead size. As the last positive point, throughput to average (TTA) is applied for frequency domain multiplexing of URLLC payloads. This helps to increase the reliability of cell-edge users and reduce the total number of resources that are used for transmitting URLLC payloads. It is interesting to note that offering one Mbps URLLC load results in two Mbps decrease in eMBB throughput. This is mainly due to the stringent QoS requirements URLLC.

Centralized Multi-Cell Scheduling

The impacts of frequency-selective scheduling and SE-DPS on URLLC are discussed in paper C. System level-results show that for lower offered loads, wide-band scheduling and frequency-selective allocation both provide the same latency performance. Similar behavior is observed for SE-DPS and distributed (no-DPS) case. This is because only part of the spectrum is occupied most of the time. In such cases, interference is limited, users experience relatively good connection quality, and queuing probability is not considerable.

Queuing delay becomes more dominant when the offered load increases. Applying frequency-selective scheduling and SE-DPS show clear benefits to reduce the URLLC latency. Appropriate link selection and user multiplexing in the frequency domain achieve the desired diversity gains and allow users to be scheduled with higher levels of modulation and coding scheme (MCS). Therefore, a lower number of resources are required. This results in reducing the queuing delay and improves the latency. Simulation results show that at 15 Mbps load, the URLLC latency at 99.999% percentile equals 12 msec when

distributed, wide-band scheduling is adopted. Applying frequency-selective allocation improves the latency to 2.7 msec. The reliability target is then fulfilled at 1.8 msec with joint SE-DPS and frequency selective scheduling [15]. Sensitivity analysis shows that the major gain of SE-DPS is achieved if mobile terminals connect to maximum of two cells in 3 dB window size.

Papers D, E, and F focus on centralized multi-cell scheduling and dynamic fast load balancing to enhance URLLC performance. The architecture of C-RAN and initial performance results are provided in papers D and E where it is shown that centralized scheduling is promising solution for URLLC. As described in the papers, the provided gain is proportional to the load condition. At low-offered loads, the latency mainly depends on transmission/processing times and HARQ retransmission delay. Hence, the distributed case and centralized solution have similar performance. However, as the load increases the centralized scheduling substantially decreases the queued users by offloading traffic from congested points to neighbouring cells. The results show that despite additional generated interference of load balancing, the proposed low complexity centralized scheduling is beneficial and enhances the URLLC outage capacity.

Several scheduling metrics are analyzed for user allocation. It is observed that including delay as part of the scheduling metric is vital for URLLC. Moreover, segmentation has significant impact on reducing the tail of URLLC latency. The effect of spectrum bandwidth on URLLC is investigated. The simulations results indicate that by expanding the bandwidth from 10 MHz to 20 MHz, the affordable URLLC traffic while maintaining the targets, grows exponentially.

Frequency domain multiplexing is added and the scheduling metric is updated in paper F to further boost the centralized performance. The upgraded scheduling metric benefits from elements of latency budget, channel quality, and multi-cell connectivity that lead to superior resource allocation. Dynamic system-level simulations results show that the proposed centralized solution in paper F significantly improves URLLC latency. As one example, Fig. III.2 depicts the complementary cumulative distribution function (CCDF) of URLLC latency for different scheduling schemes and offered traffic. For 18 Mbps load and at the outage level of 10^{-5}, 97% and 90% latency improvements are achieved in comparison to the distributed scheduling and DPS, receptively [16]. The superiority of centralized solution comes as the results of fast switching between the serving cells to overcome fading, and most importantly performing instant load balancing to reduce the traffic at congested cells.

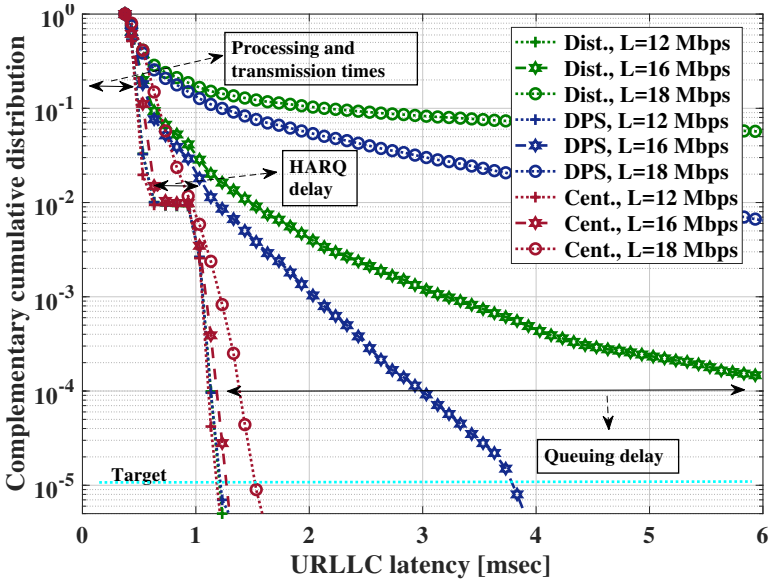

Fig. III.2: URLLC latency for different scheduling methods assuming four OFDM symbols TTI size [16].

Recommendations

Following summarizes the recommendations from the findings of this chapter:

- For optimizing URLLC scheduling, it is essential to take into account packet characteristics such as QoS, payload size, HARQ, latency, and the overhead of control channel information. Multi-user frequency selective scheduling provides substantial diversity gain for meeting URLLC targets. To achieve the maximum benefits, it is essential to employ a mechanism that avoids the segmentation of multiple payloads. If the available resources are not sufficient for one full payload, it is recommended to assign remained PRBs to a user with low PDCCH overhead.

- Centralized multi-cell scheduling is an attractive solution for URLLC. A Low-complexity algorithm (in line with the tight URLLC processing time) for fast load balancing and scheduling can significantly reduce the queuing probability, thus the latency for URLLC. A measurement set size of maximum two cells with received signal received power (RSRP) widow size of 3 dB provides sufficient degrees of freedom for efficient cell allocation.

References

[1] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 2, pp. 678–700, Second 2013.

[2] N. Abu-Ali, A. M. Taha, M. Salah, and H. Hassanein, "Uplink scheduling in LTE and LTE-Advanced: Tutorial, survey and evaluation framework," *IEEE Communications Surveys Tutorials*, vol. 16, no. 3, pp. 1239–1265, Third 2014.

[3] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, 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.

[4] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 127–141, January 2018.

[5] 3GPP Technical Documents R1-1813120, "Discussion on the RAN2 LS on TSN requirements evaluation," November 2018.

[6] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–5.

[7] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38 451–38 463, 2018.

[8] A. Anand, G. D. Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, April 2018, pp. 1970–1978.

[9] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28 912–28 922, 2018.

[10] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, third-quarter 2016.

[11] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schober, and L. Hanzo, "User association in 5G networks: A survey and an outlook," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1018–1044, Second-quarter 2016.

[12] V. Fernández-López, K. I. Pedersen, B. Soret, J. Steiner, and P. Mogensen, "Improving dense network performance through centralized scheduling and interference coordination," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4371–4382, May 2017.

[13] H. Zhou, S. Mao, and P. Agrawal, "Approximation algorithms for cell association and scheduling in femtocell networks," *IEEE Transactions on Emerging Topics in Computing*, vol. 3, no. 3, pp. 432–443, September 2015.

[14] A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Pocovi, and P. Mogensen, "Efficient low complexity packet scheduling algorithm for mixed URLLC and eMBB traffic in 5G," in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*, April 2019, pp. 1–6.

[15] A. Karimi, K. I. Pedersen, and P. Mogensen, "5G URLLC performance analysis of dynamic-point selection multi-user resource allocation," *in Proc. 2019 International Symposium on Wireless Communication Systems (ISWCS)*, August, 2019.

[16] ——, "Low-complexity centralized multi-cell radio resource allocation for 5G URLLC," *Submitted to 2020 IEEE Wireless Communications and Networking Conference (WCNC)*, 2020.

Paper B

Efficient Low Complexity Packet Scheduling Algorithm for Mixed URLLC and eMBB Traffic in 5G

Ali Karimi, Klaus I. Pedersen, Nurul Huda Mahmood,
Guillermo Pocovi, and Preben Mogensen

The paper has been published in the
IEEE 89th Vehicular Technology Conference (VTC2019-Spring), 2019.

Abstract

We address the problem of resource allocation and packet scheduling for a mixture of ultra-reliable low-latency communication (URLLC) and enhanced mobile broadband (eMBB) traffic in a fifth generation New Radio (5G NR) networks. A novel resource allocation method is presented that is latency, control channel, hybrid automatic repeat request (HARQ), and radio channel aware in determining the transmission resources for different users. This is of high importance for the scheduling of URLLC users in order to minimize their latency, avoid unnecessary costly segmentation of URLLC payloads over multiple transmissions, and benefit from radio channel aware multi-user diversity mechanisms. The performance of the proposed algorithm is evaluated with an advanced 5G NR compliant system level simulator with a high degree of realism. Simulation results show promising gains of up to 98% latency improvement for URLLC traffic and 12% eMBB end-user throughput enhancement as compared to conventional proportional fair scheduling.

1 Introduction

The fifth generation New Radio (5G NR) is set to support different services such as ultra-reliable low-latency communications (URLLC) and enhanced mobile broad (eMBB) [1]. For URLLC, various classes with different quality of service (QoS) requirements are defined by 3GPP, where one of the most stringent service target is one millisecond (msec) latency at 99.999% reliability [2]. An overview of communication theoretic principles of URLLC can be found in [3, 4]. A flexible multi-service capable frame structure has been studied in [5]. Several contributions in the literature have also studied various resource allocation techniques to enhance the performance of URLLC in 5G NR. The authors in [6] study the problem of user (UE) selection and scheduling for URLLC, where only one UE is scheduled in each transmission time interval. In [7, 8], the authors formulated a multi-dimensional 0-1 Knapsack problem for low-latency communications to select and drop delayed packets from the network. It has been shown in [9] that wide-band allocation maximizes the outage capacity of URLLC and dynamic multiplexing of URLLC and eMBB significantly improves the spectral efficiency. Dynamic link adaptation and multiplexing of URLLC and eMBB traffic on a shared channel were studied in [10, 11]. Finally, several pre-emptive scheduling schemes for multiplexing of URLLC and eMBB traffic are proposed in [12, 13].

In this paper, we present additional scheduler advancements as compared to earlier published studies. For scheduling of the high-priority UEs, we propose a resource allocation scheme that is payload and control channel aware, and exploits the radio channel time-frequency variations. The payload awareness is incorporated in the scheduler by favouring scheduling of full URLLC

payloads without segmenting those over multiple transmissions. At most one UE per URLLC scheduling interval is subjected to segmentation, limited to the UE with the minimum segmentation cost. Moreover, the buffering time of individual payloads are explicitly taken into account in the scheduling decisions, as compared to the latency target. The overhead from the physical layer control channel to signal the scheduling grant to the UEs is also explicitly incorporated in the presented resource allocation framework. Finally, the proposed scheduler also has an element of radio channel awareness to gain from multi-user diversity.

State-of-the-art 5G NR compliant multi-cell dynamic system level results are presented to demonstrate how the proposed solution performs under different load regimes. The results confirm that the proposed resource allocation algorithm improves the latency performance of URLLC users, and also enhances the end-user throughput for the eMBB users.

The rest of the paper is organized as follows: the system model and problem formulation are elaborated in Section 2. Section 3 discusses the proposed packet scheduling algorithm. Simulation methodology and performance results are presented in Section 4. Finally, the study is concluded in Section 5.

2 Setting the Scene

2.1 Basic System Model

We adopt the 5G NR specifications as outlined in [1, 14], focusing primarily on the downlink (DL) performance for frequency division duplexing (FDD) mode. The network consists of C cells forming a three-sectorized deployment with 500 meters inter-site distance corresponding to the 3GPP urban macro (UMa) deployment [14]. A set of U URLLC and M eMBB UEs are randomly distributed over the entire network area. For each URLLC UE, bursts of small payloads of B bytes arrive at the network according to a Poisson point process with arrival rate of λ [payload/sec]. This traffic model is known as FTP3 in 3GPP [15].

Full buffer traffic with infinite payload size is assumed for eMBB UEs. In the t-th transmission time interval (TTI), the sets of active (with data) URLLC and eMBB UEs connected to cell c are denoted by $\mathbf{U}^{c,t}$ and $\mathbf{M}^{c,t}$, respectively.

Both eMBB and URLLC traffic are dynamically multiplexed on a shared channel, using orthogonal frequency division multiple access (OFDMA) with 30 kHz sub-carrier spacing. A short TTI size of 4 OFDM symbols (0.143 msec) and a physical resource block (PRB) resolution of 12 sub-carriers is assumed as the minimum time and frequency scheduling unit.

The base stations and users are each equipped with two transmit/receive

antennas. UEs exploit linear minimum mean square error interference rejection combining (LMMSE-IRC) receiver to suppress noise and received interference. Each UE periodically measures the channel and interference for each resource element (RE) and reports a frequency-selective channel quality indicator (CQI) per sub-channel of eight PRBs. The reported CQIs are subjected to processing delay before being applied at the network for DL transmission.

User-centric control channel transmission is assumed to indicate scheduling grant of scheduled UEs [16]. Thus, whenever a user is scheduled, both a user-specific scheduling grant on the physical downlink control channel (PDCCH) and the actual transport block (data) on the physical downlink shared channel (PDSCH) are transmitted. The PDCCH size is dynamically adjusted based on the reported wide-band signal to interference plus noise ratio (SINR) value to guarantee low probability of failure. In line with [10, 16], the PDCCH is transmitted with aggregation level 1,2,4, or 8 depending on the experienced SINR at the UE, where the aggregation consumes 36 REs.

Dynamic link adaptation is applied for transmission of the PDSCH. As the CQI is subjected to reporting delay and other imperfections, the well-known outer loop link adaptation (OLLA) is applied to control the block error rate (BLER). In line with [10, 17], the OLLA offset is adjusted to achieve 1% and 10% BLER of the first PDSCH transmission for URLLC and eMBB, respectively. In case of packet failure, the UE will feed back a negative acknowledgement (NACK), and the corresponding hybrid automatic repeat request (HARQ) retransmission is scheduled by the network. Asynchronous HARQ retransmission with Chase combining and a maximum of six retransmissions are assumed [18, 19].

2.2 Latency Components

The one-way URLLC latency (Y) is defined from the time that a URLLC payload arrives at the network, until it is successfully decoded at the UE. If the UE correctly receives the packet in the first transmission, the latency equals the first transmission delay as:

$$Y = d_{fa,q} + d_{bsp} + d_{tx} + d_{uep}, \tag{B.1}$$

where $d_{fa,q}$ denotes the frame alignment and queuing delay. The payload transmission time is denoted by d_{tx}. Processing time at the network and the UE are represented by d_{bsp} and d_{uep}, respectively. The frame alignment delay is a uniformly distributed random variable taking values between zero and one TTI. The queuing delay accounts for the time where the the payload arrives at the base station until is considered for scheduling in the next upcoming TTI. The transmission time is a discrete random variable. Depending on the packet size, channel quality and scheduling strategy, d_{tx} varies from one to multiple TTIs. The processing times at the network (d_{bsp}) and the

UE (d_{uep}) are assumed to be constants, equal to 2.75 and 4.5 OFDM symbols, respectively [20]. In case of failure, the packet is subject to additional retransmission delay(s), d_{HARQ}^{RTT}, until either it is decoded successfully or the maximum number of retransmissions is reached. In line with [10], the minimum retransmission delay of $d_{HARQ}^{RTT} = 4$ TTIs is assumed.

2.3 Problem Formulation

The objective is to maximize the network capacity of serving both URLLC and eMBB services. The URLLC capacity is defined as the maximum served average URLLC traffic L^{llc}, while still ensuring the packets are successfully delivered with the reliability of P_{target} within the given latency budget of T_{target} , expressed as $P(Y \leq T_{target}) \geq P_{target}$. For eMBB, maximizing the well-known Proportional-Fair (PF) utility function is assumed [21]. Dropping notations t and c for the ease of presentation, for a cell with D_{tot} PRBs, the resource allocation problem is formulated as:

$$\begin{aligned}
&\max_{b_{u/m}^j} \sum_{u\in\mathbf{U}} a_u R_u^{llc} + \sum_{m\in\mathbf{M}} \log \bar{R}_m^{mbb}, \\
&\text{Sub. to:} \sum_{u\in\mathbf{U}} b_u^j + \sum_{m\in\mathbf{M}} b_m^j \leqslant 1,\ \forall j \in \{1,\cdots,D_{tot}\}, \\
&\quad \sum_{j=1}^{D_{tot}} b_{u/m}^j \geq \min(R_{u/m}^{llc/mbb}, 1)\cdot b_{u/m}^{\min},\ \forall u,m, \\
&\quad R_u^{llc} \leqslant Q_u^{llc}\ \ \forall u, \\
&\quad b_{u/m}^j \in \{0,1\}\ \ \forall u,m,j,
\end{aligned} \tag{B.2}$$

where the binary variable b_i^j ($i \in \{u,m\}$, $j \in \{1,\cdots,D_{tot}\}$) indicates if the j-th PRB is allocated to i-th UE. The achievable rate of the u-th URLLC and the average throughput of m-th eMBB UEs are denoted by R_u^{llc} and $\bar{R}_m^{mbb}$, respectively. The minimum control channel overhead of the i-th UE is denoted by $b_i^{\min}$. The variable a_u is the u-th URLLC user QoS indicator chosen to satisfy the low-latency constraint. A larger a_u value indicates it is higher priority UE. Buffered data of the u-th URLLC user is represented by Q_u^{llc}. The first constraint in (B.2) ensures that each PRB is assigned to maximum one UE (single-user transmission). The second constraint guarantees that each scheduled UE has been assigned the minimum required number of PRBs to include the scheduling grant. Finally, the third constraints takes into account that the URLLC users have rather small amounts of buffered data to be served per scheduling interval. Problem (B.2) is a non-linear integer optimization can be solved using brute-force algorithm with complexity $\mathcal{O}\left(D_{tot}^{|\mathbf{U}|+|\mathbf{M}|}\right)$. This is too high complexity for practical network implementations as the URLLC scheduling decision needs to be taken every TTI on a fast basis.

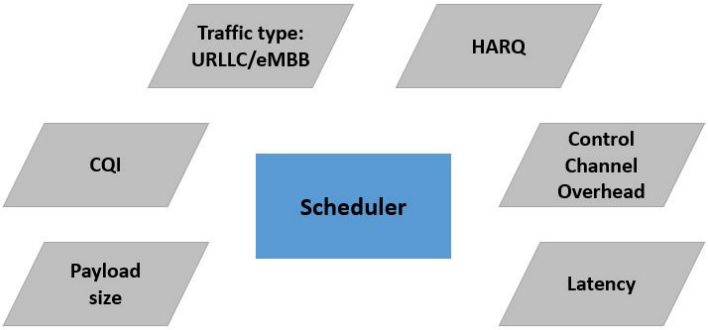

Fig. B.1: Parameters affecting scheduling decision.

3 Proposed Packet Scheduling Solution

A low-complexity packet scheduling algorithm that is aware of traffic, latency, control channel, HARQ, and radio channel is proposed as schematically presented in Fig.B.1. In line with [10–13], to reduce the queuing delay and enhance the reliability, URLLC payloads are scheduled first. After scheduling URLLC, eMBB traffic is served on the remaining PRBs.

3.1 URLLC Scheduling

URLLC payloads are scheduled in the following order.

Pending HARQ Retransmission: First, we assign the highest priority to HARQ retransmissions by scheduling them immediately over the set of PRBs with the highest CQI values. Thus, additional queuing delay is avoided as the payloads are already subjected to retransmission delay(s) of d_{HARQ}^{RTT}. By scheduling HARQ retransmissions over the best set of PRBs, we aim at increasing the reliability and minimizing the probability of further retransmissions.

Buffered URLLC Packets: Buffered URLLC payloads are scheduled thereafter. A low complexity time/frequency domain scheduler is applied as follows. First, the time-domain (TD) scheduler selects a group of UEs that can be fully scheduled over the available PRBs. Buffered payloads that are closer to the latency target (i.e. have lower latency budget) are prioritized by the TD scheduler. The number of required PRBs for each payload is estimated from the reported wide-band CQI. The selected UEs are thereafter scheduled by the FD scheduler.

The FD scheduler utilizes multi-user radio channel-aware diversity mechanisms to achieve good performance. We utilize throughput to average (TTA)

metric for scheduling URLLC payloads. Lets assume that r_u^p denotes the achievable throughput (TP) of PRB p for the u-th UE. The scheduler selects user $\hat{u}$ for being scheduled on PRB p which maximizes

$$\hat{u} = \max_u \frac{r_u^p}{\bar{\mathbf{r}}_u}, \tag{B.3}$$

where $\bar{\mathbf{r}}_u$ is the instantaneous full-bandwidth TP. Normalizing the achievable rate by the full-bandwidth TP, enhances fairness among the UEs and the probability to access to relatively good channels for all UEs [21]. As the rates of increase in TP is higher in low-SINR regimes [22], moderate and low-SINR UEs receive higher opportunity to occupy relatively better frequency-selective channel variations. Thus, scheduling based on (B.3) not only enhances the reliability of low-SINR UEs, but also fewer number of resources are needed to schedule the total payloads.

After UEs are scheduled in FD, the scheduler checks if it is possible to schedule more UEs on the remaining PRBs. The procedure is repeated until all buffered UEs are scheduled or there are not enough PRBs to schedule a full URLLC payload. For cases with insufficient PRBs for a full payload, at most one URLLC payload is segmented and scheduled over the remaining PRBs. To further reduce the cost of segmentation, UEs in good channel conditions (i.e. lower control channel overhead) are prioritized for segmentation. Details of the proposed scheduling is summarized in Algorithm 1.

Algorithm 1 Proposed algorithm for URLLC packet scheduling

1: Schedule the HARQ retransmission over PRBs with the highest CQI values.
2: **while** Unscheduled UEs and enough PRBs **do**
3: Select a group of UEs with the lowest latency budget that can be fully scheduled.
4: For each selected UE and the available PRB, create pairs of UE/PRB and calculate the corresponding scheduling metric based on (B.3).
5: Sort pairs in the descending order of metric.
6: Allocate PRBs to UEs with the highest metric values, up to the required PRBs for each payload yields.
7: Remove if there is a segmented payload.
8: Update available PRBs.
9: **end while**
10: **if** Still unscheduled URLLC payload(s) and enough PRBs to partially schedule one payload **then**
11: Select the UE with the highest TP and schedule it over remaining PRBs.
12: **end if**

3.2 eMBB Scheduling

After scheduling URLLC, eMBB UEs are scheduled on the remaining PRBs according to the PF metric. PRB p is assigned to UE $\hat{m}$ with the highest metric [21]

$$\hat{m} = \max_{m} \frac{r_m^p}{\bar{\mathbf{R}}_m}, \tag{B.4}$$

where $\bar{\mathbf{R}}_m$ is the m-th user average delivered throughput in the past, calculated by a moving average filter.

4 Simulation Results

4.1 Simulation Methodology

The performance of the proposed solution is evaluated by simulations using a highly detailed system level simulator that includes the 5G NR radio resource management functionalities as described in Section 2. The simulation methodology is based on 3GPP 5G NR mathematical models and assumptions [1, 14, 23]. The assumed network configuration and default simulation parameters are summarized in Table B.1. At least five million URLLC packet transmissions are simulated to obtain statistical reliable results. This results in statistically reliable results with the confidence level of 95% for the 99.999% percentile of the latency [10]. For URLLC, the key performance indicator (KPI) is the one-way achievable latency with 99.999% reliability. For eMBB, the average cell TP is considered.

The results are compared against recent URLLC studies with PF scheduling [10, 11]. A comparison versus the well-known modified largest weighted delay first (M-LWDF) algorithm is also included. The M-LWDF scheduler is expressed as [21]

$$\hat{u} = \max_{u} \frac{-\log P_{target}}{T_{target}^u} d_{HOL}^u \frac{r_u^p}{\bar{\mathbf{R}}_m}, \tag{B.5}$$

where d_{HOL}^u is the head of line delay of user u. For both the PF and M-LWDF algorithms, URLLC UEs are scheduled first. eMBB traffic is served over the remaining PRBs. The network does not discard delayed packets.

4.2 Performance Results

Fig. B.2 depicts the complementary cumulative distribution function (CCDF) of URLLC latency for different offered URLLC loads from 4 to 14 Mbps/cell. At low offered loads, the latency performance is mainly affected by the transmission delay, processing times, and HARQ-RTT. URLLC payloads usually occupy only part of the available bandwidth and a few UEs compete for the

Table B.1: Default simulation assumptions.

Description	**Assumption**
Environment	3GPP Urban Macro (UMa); 3-sector BSs with 500 meters inter-site distance. 21 cells.
Propagation	Urban Macro-3D.
Carrier	2 GHz (FDD), 20 MHz carrier bandwidth.
PHY numerology	30 kHz sub-carrier spacing configuration. PRB size of 12 sub-carrier (360 kHz).
TTI sizes	0.143 msec (4-symbols mini-slot).
MIMO	Single-user 2x2 closed loop single-stream (Rank-1) configuration. LMMSE-IRC receiver.
CSI	Periodic CSI every 5 msec, with 2 msec latency.
MCS	QPSK to 64 QAM, with same encoding rates as specified for LTE. Turbo codes.
Link adaptation	Dynamic MCS with 1% and 10% BLER for URLLC and eMBB, respectively.
HARQ	Asynchronous HARQ, Chase combining. HARQ-RTT=4 TTIs, max. 6 retransmissions.
User distribution	2100 URLLC and 210 eMBB UEs (Average 100 URLLC and 10 eMBB UEs per cell).
Traffic model	FTP3 downlink traffic with $B = 50$ bytes data for URLLC. Full buffer for eMBB.
Link-to-system (L2S) mapping	Based on MMIB mapping [24].

resources. Thus, access to relatively good channels is possible for most UEs. Therefore, all scheduling methods have the same performance at low loads.

As the offered load increases, the queuing delay becomes more dominant. It is observed that the proposed solution provides significant latency improvement as the load increase. As an example, at 12 Mbps/cell load, the latency at 10^{-5} outage probability with PF, M-LWDF and the proposed algorithm is 4.5, 2.92 and 1.38 msec, respectively. This is equivalent to 70% and 53% latency gain in comparison with PF and M-LWDF scheduling, respectively. The proposed algorithm also shows a robust behaviour over the offered load variations, where the latency increases from 1.20 to 1.56 msec when the load is increased from 4 to 14 Mbps. In comparison, the latency increase corresponding to the same load increase for the PF and M-LWDF algorithm is 1.21 to 69 msec and 1.20 to 10.45 msec, respectively.

Fig. B.3 presents the CCDF of the combined queuing and frame alignment delay for different offered loads. As expected, the queuing delay increases with the offered load. The Figure shows the superior performance of the pro-

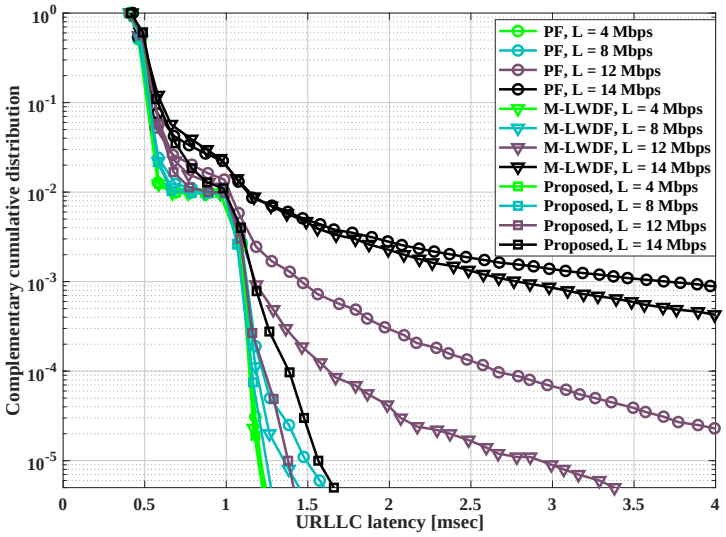

Fig. B.2: URLLC latency distribution for different URLLC offered loads and scheduling methods.

posed algorithm in reducing the tail of the queuing delay which is important for URLLC traffic. For example, at 12 Mbps offered load only 0.01% of the payloads experience more than 0.5 msec queuing and frame alignment delay. While for M-LWDF and PF, it increases to 0.23% and 0.53%, respectively.

Table B.2 presents the URLLC latency and the average eMBB cell TP for different scheduling and offered URLLC traffic settings. As the URLLC traffic is always prioritized over the eMBB, the average eMBB TP decreases when increasing the URLLC load. It can be seen from the table that the proposed solution improves both the URLLC latency and eMBB TP. At 14 Mbps URLLC load, it provides 98% URLLC latency reduction as well as 12% increase in eMBB TP in comparison to PF. Gains of 84% URLLC latency reduction and 11% eMBB TP enhancement are achieved over the M-LWDF. The performance benefits come as the results of: (i) considering the latency budget as the main scheduling parameters for URLLC (prioritizing UEs with the lowest latency budget). (ii) reducing the control channel overhead by single-TTI transmission of URLLC payloads, (iii) efficient FD multiplexing of URLLC UEs that results in fewer number of allocated resources to schedule the URLLC payloads.

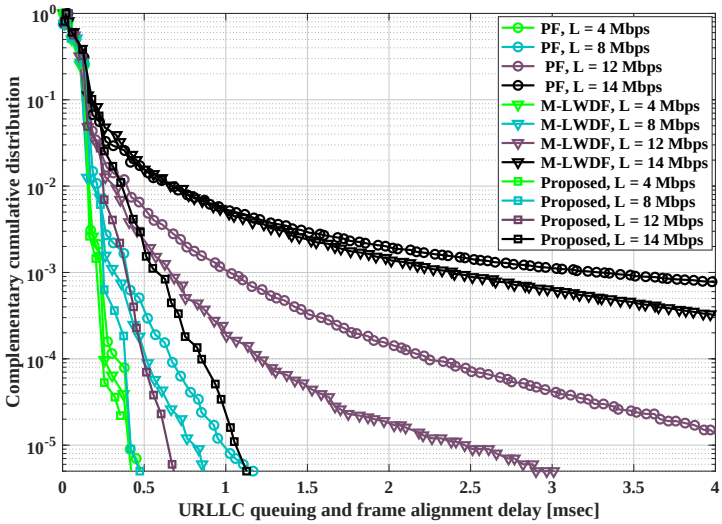

Fig. B.3: Queuing and frame alignment delay for different offered loads and scheduling methods.

5 Conclusion

We studied the problem of resource allocation for mixed URLLC and eMBB traffic in 5G NR multi-service networks. A latency-QoS, control channel, HARQ, and radio channel aware scheduling algorithm is proposed to enhance the performance of both URLLC and eMBB traffic. The proposed algorithm exploits the gains of frequency-selective multi-user scheduling while avoiding unnecessary and costly segmentation of URLLC payloads over multiple transmissions. The solution benefits from low computational complexity and is attractive for practical network implementation. Results show significant latency improvement of URLLC traffic as well as higher average eMBB throughput. As an example, at 14 Mbps URLLC offered load, the latency of URLLC at the 10^{-5} outage level is improved by 98% compared state of the art proportional fair scheduling and also the average eMBB throughput is increased by 12%.

Acknowledgement

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.

Table B.2: Network performance for different URLLC offered loads

Scenario			Offered URLLC load [Mbps]					
			4	**8**	**10**	**12**	**14**	**15**
URLLC latency at the outage probability of 10^{-5} [msec]	PF		1.21	1.5	2.3	4.5	69	358
	M-LWDF		1.2	1.36	1.63	2.92	10.45	22.5
	Proposed		1.2	1.24	1.31	1.38	1.56	2.27
	Relative improvement to	PF	0 %	18 %	57 %	70 %	98 %	99.4 %
		M-LWDF	0 %	9 %	20 %	53 %	84 %	90 %
Average eMBB cell throughput [Mbps/cell]	PF		34.6	25.3	20.8	16.2	11.5	9.2
	M-LWDF		34.6	25.3	20.8	16.3	11.64	9.3
	Proposed		34.7	25.6	21.3	17.07	12.9	10.8
	Relative improvement to	PF	0 %	1.3 %	2.5 %	5.3 %	12 %	17.3 %
		M-LWDF	0 %	1.3 %	2.5 %	4.7 %	11 %	16.1 %

References

[1] 3GPP Technical Specification 38.300, "NR and NG-RAN overall description; stage-2," Version 15.5.0, March 2019.

[2] 3GPP Technical Specification 23.501, "Technical specification group services and system aspects, system architecture for the 5G system," Release 15, December 2017.

[3] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail, risk and scale," *CoRR*, 2018. [Online]. Available: http://arxiv.org/abs/1801.01270

[4] P. Popovski, J. J. Nielsen, C. Stefanovic, E. d. Carvalho, E. Strom, K. F. Trillingsgaard, A. Bana, D. M. Kim, R. Kotaba, J. Park, and R. B. Sorensen, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Network*, vol. 32, no. 2, pp. 16–23, March 2018.

[5] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, 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.

[6] A. Destounis, G. S. Paschos, J. Arnau, and M. Kountouris, "Scheduling URLLC users with reliable latency guarantees," in *2018 16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2018, pp. 1–8.

[7] E. Khorov, A. Krasilov, and A. Malyshev, "Radio resource and traffic management for ultra-reliable low latency communications," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.

[8] ——, "Radio resource scheduling for low-latency communications in lte and beyond," in *2017 IEEE/ACM 25th International Symposium on Quality of Service (IWQoS)*, June 2017, pp. 1–6.

[9] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–5.

[10] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28 912–28 922, 2018.

[11] ——, "Multiplexing of latency-critical communication and mobile broadband on a shared channel," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.

[12] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38 451–38 463, 2018.

[13] A. Anand, G. D. Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, April 2018, pp. 1970–1978.

[14] 3GPP Technical Report 38.913, "Study on scenarios and requirements for next generation access technologies," Version 14.1.0, March 2017.

[15] 3GPP Technical Report 38.802, "Study on new radio access technology physical layer aspects," Version 14.0.0, March 2017.

[16] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 210–217, March 2018.

[17] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen, "Centralized joint cell selection and scheduling for improved URLLC performance," *In Proc. 29th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, September, 2018.

[18] K. I. Pedersen, S. R. Khosravirad, G. Berardinelli, and F. Frederiksen, "Rethink hybrid automatic repeat request design for 5G: Five configurable enhancements," *IEEE Wireless Communications Magazine*, vol. 24, no. 6, pp. 154–160, December 2017.

[19] 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.

[20] 3GPP Technical Documents R1-1808449, "IMT-2020 self-evaluation: UP latency analysis for FDD and dynamic TDD with UE processing capability 2 (URLLC)," August 2018.

[21] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 2, pp. 678–700, Second 2013.

[22] G. Ku and J. M. Walsh, "Resource allocation and link adaptation in LTE and LTE-Advanced: A tutorial," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1605–1633, third-quarter 2015.

[23] 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.

[24] T. L. Jensen, S. Kant, J. Wehinger, and B. H. Fleury, "Fast link adaptation for MIMO OFDM," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 8, pp. 3766–3778, October 2010.

Paper C

5G URLLC Performance Analysis of Dynamic-Point Selection Multi-User Resource Allocation

Ali Karimi, Klaus I. Pedersen, and Preben Mogensen

The paper has been published in the
16th International Symposium on Wireless Communication Systems (ISWCS), 2019.

The layout has been revised.

Abstract

This paper studies dynamic point selection (DPS) and frequency-selective multi-user scheduling to improve ultra-reliable low-latency communication (URLLC) for the fifth generation New Radio (5G NR) systems. DPS is a special type of multi-channel access scheme enhances the network performance by enabling dynamic transmission point selection on a fast time-scale. The achieved gain from frequency-selective URLLC scheduling is further studied by investigating a low-complexity resource allocation algorithm. Extensive 5G NR system-level simulation results show that DPS achieves 30% *improvement of URLLC latency. Our analyses also indicate that for DPS, user-specific clustering with 3-dB power range achieves the major improvement of URLLC latency.*

1 Introduction

Third generation partnership project (3GPP) has introduced ultra-reliable low-latency communications (URLLC) as a new service class in the fifth generation New Radio (5G NR) [1, 2]. URLLC is envisioned to support a wide range of mission critical applications such as industrial automation, E-health, and vehicular communications, with strict quality of service (QoS) requirements in terms of both reliability (99.999%) and latency (one millisecond) [1, 3]. Lots of studies have addressed challenges that arise from such stringent requirements. As an essential baseline for enabling low-latency communications, the use of short time transmission intervals (TTIs) and flexible frame structure has been investigated in [4]. Dynamic link adaptation and QoS-aware resource allocation of URLLC and enhanced mobile broadband (eMBB) traffic are studied in [5, 6]. The authors of [7, 8] present a survey of the theoretic principles of URLLC and discuss several important enablers for reliable communications. Among the promising solutions enabling URLLC, utilization of massive multiple-input multiple-output (massive MIMO) antennas is investigated in [9]. The use of centralized radio access network (C-RAN) architecture is discussed [10, 11]. The work in [12, 13] present a survey of reliability enhancement of URLLC services through multi-channel access (MCA) solutions.

This paper studies the performance of dynamic point selection (DPS) multi-user resource allocation for URLLC services. DPS is a special case of the MCA family, which provides dynamic transmission point selection on a TTI basis based on channel and load conditions [14]. It is a key feature to mitigate stochastic variations of fading channels for cell-edge users and to enhance the spectral efficiency by enabling fast switching between serving cells.

The concept of DPS has earlier been studied for eMBB traffic in LTE sys-

tems to improve the average network capacity [15, 16]. However, given the many differences between LTE and NR, the concept of DPS needs to be revisited to assess its potential performance for URLLC cases. Our starting point is the so-called spectrum efficient DPS, where the users are scheduled by the cell with the highest instantaneous throughput (TP), offering a simple, yet efficient, diversity mechanism. This solution is further extended by pairing it with a latency-aware multi-user diversity resource allocation policy. The proposed solution takes the overhead of control channel transmissions carrying scheduling grants explicitly into account, as well as potential effects of segmentation of the URLLC payloads. The study is conducted for a highly detailed system model in line with 3GPP NR specifications. The model comprises the NR radio access network protocol stack, time-variant URLLC traffic models, a realistic three-dimensional (3D) radio propagation channel, MIMO transmission, dynamic link adaptation, hybrid automatic repeat request (HARQ) retransmission, etc.

Performance results from 3GPP 5G NR compliant system-level simulations are presented to evaluate the performance of the proposed schemes. The results reveal that both DPS and frequency-selective scheduling offer significant latency reduction.

The rest of the paper is organized as follows: Section 2 presents an overview of the system model and network deployment. The proposed packet scheduling algorithm is discussed in Section 3. Simulation methodology and performance results are presented in Section 4. Finally, Section 5 concludes the paper.

2 Setting the Scene

2.1 System Model

We study the downlink (DL) performance for the frequency division duplexing (FDD) mode in line with the 5G NR specifications as outlined in [1, 17]. As in [5], a wide-area urban macro (UMa) scenario of $C = 21$ cells deployed in a three-sectorized manner is assumed. A set of U URLLC user equipments (UEs) are uniformly distributed in the network area. Sporadic traffic is assumed for each UE where bursts of small payloads of 50 bytes arrive at the network following a Poisson point process with an average arrival rate of λ [payload/sec]. The average offered load per cell equals $L = C^{-1} \times U \times \lambda \times 50 \times 8$ [bps/cell].

The UEs are dynamically multiplexed on a shared channel with 20 MHz bandwidth using orthogonal frequency division multiple access (OFDMA) with 30 kHz sub-carrier spacing. A short mini-slot time transmission interval (TTI) of 4 OFDM symbols (≈ 0.143 msec) and physical resource block (PRB)

of 12 sub-carriers are assumed.

Both cells and UEs have two transmit/receive antennas. Linear minimum mean square error interference rejection combining (LMMSE-IRC) receiver is assumed at the UEs to suppress the received noise plus interference.

2.2 Cell Connectivity and DPS Procedure

For the baseline (no DPS) scenario, each UE measures cells it can hear and connects to the cell corresponding to the highest received average reference signal received power (RSRP).

Dynamic user-centric clustering is assumed for DPS case. The UE connects to a cluster of maximum Q cells that are within a RSRP power window of W dB as compared to the strongest cell. We denote $\Gamma(u)$ as the set of cells in the cluster for UE u. Channel state information (CSI) measurements are performed periodically for the connected cells and the UE reports channel quality indicator (CQI) of the best cell to the network. Targeting to maximize the instantaneous user TP, UE u reports cell $\hat{c}$ with the highest spectrum efficient metric

$$\hat{c} = \arg \max_{c \in \Gamma(u)} \bar{\mathbf{r}}_c^u, \tag{C.1}$$

where $\bar{\mathbf{r}}_c^u$ is the estimated full-band TP of the u-th UE served by cell c.

Two types of CQI measurement are performed for the selected serving cell. i) The UE reports one wide-band CQI. ii) One CQI value per a sub-channel of eight PRBs. The CQI is subject to reporting and network processing delay before being applied for the DL transmission decisions.

For each scheduling interval, both the user-specific physical downlink control channel (PDCCH) and the actual data are transmitted on the assigned PRBs. In line with [18, 19], the aggregation level of the PDCCH is dynamically adjusted based on the reported CQI to guarantee low-probability of failure. Dynamic link adaption is adopted for data transmission. The well-known outer-loop link adaptation offset is applied to achieve 1% block error rate of the first data transmission [6, 11]. The UE feeds back a negative acknowledgement (NACK) in case of packet failure and the corresponding HARQ retransmission is scheduled by the network. HARQ Chase-combing is assumed at the UE to increase the quality of received signal by maximum ratio combining (MRC) of the multiple received packets [20].

2.3 URLLC Latency Components

The DL one-way latency (Y) of a URLLC payload is defined from the time that the payload arrives at the network until it is successfully received at the UE. If the payload is decoded correctly within the first transmission, the

latency equals:

$$\Upsilon = d_{fa,q} + d_{bsp} + d_{tx} + d_{uep}, \tag{C.2}$$

where $d_{fa,q}$ is the frame alignment and queuing delay. The transmission time is denoted by d_{tx}. The processing time at the base station and user-end are denoted by d_{bsp} and d_{uep}, respectively. The frame alignment delay is a random variable with uniform distribution between zero and one TTI. The queuing delay is the time the packet is buffered before getting scheduled at the physical layer. Depending on the payload size, CQI, and the number of available resources, the transmission time varies between one and multiple TTIs. In line with [21], the processing times are assumed to be constant equal to $d_{bsp} = 2.75$ and $d_{uep} = 4.5$ OFDM symbols, respectively. In case of failure in data transmission, the packet is subject to additional HARQ round-trip-time (HARQ-RTT) retransmission delay(s) (d_{HARQ}). A minimum retransmission delay of $d_{HARQ} = 4$ TTIs is assumed [6].

3 Proposed URLLC Resource Allocation Algorithm

Our target is to maximize the URLLC capacity subject to satisfying both the reliability and latency constraints. The applied radio resource management procedure is as follows. As discussed in Section 2.2, each UE dynamically determines the serving cell and periodically reports the corresponding CQI (wide-band/sub-band) to the network. The active UEs with data are allocated resources in each scheduling interval. Building on [5], a low-complexity resource allocation algorithm is applied to schedule the buffered UEs.

To minimize additional queuing delay, first the HARQ retransmissions are scheduled. For cases with sub-band CQI, the HARQ payloads are scheduled over the set of PRBs with the highest CQI values to enhance the reliability of retransmissions.

Afterwards, pending URLLC payloads are allocated. The time-domain (TD) scheduler selects a subset of UEs closer to the latency deadline which can be fully scheduled on the available resources. The selection metric is expressed as follows

$$\dot{u} = \arg \min_{u \in \Xi(c)} \{\Upsilon^u_{tar} - \Upsilon^u_{cur} \mid R^{uc} \leqslant D^c_{tot}\}, \tag{C.3}$$

where $\Xi(c)$ is the set of active UEs of c-th cell. The target and current latencies of the u-th UE are denoted by Υ^u_{tar} and Υ^u_{cur}, respectively. The number of available PRBs at cell c and that of required to schedule (both the data and PDCCH) UE u are presented by D^c_{tot} and R^{uc}, respectively. The value of R^{uc} is estimated from the reported wide-band CQI. After selecting UE $\dot{u}$, the scheduler updates the number of available PRBs as $D^c_{tot} = D^c_{tot} - R^{\dot{u}c}$ and search for other schedulable candidate UEs.

For cases with wide-band CQI, the TD selected UEs are randomly allocated over the entire bandwidth. For scenarios with available sub-band CQI, the selected UEs are frequency-domain (FD) multiplexed by allocating resources based on the throughput to average (TTA) metric [5]. PRB p is assigned to UE $\hat{u}$ with the highest TTA metric:

$$\hat{u} = \arg\max_{u \in \Pi} \frac{r_u^p}{\bar{\mathbf{r}}_u}, \tag{C.4}$$

where Π denotes the set of UEs selected by TD scheduler. Variables r_u^p and $\bar{\mathbf{r}}_u$ represent the u-th UE's achievable TP of PRB p and the instantaneous full-bandwidth TP in current TTI.

Finally, the scheduler checks if there are still available resources to schedule additional UE(s). In case of not having enough PRBs to allocate a full payload, only one URLLC payload is segmented and transmitted on the remaining PRBs. To minimize the cost of PDCCH transmission, an UE with the lowest PDCCH overhead (i.e. higher CQI value) is prioritized.

Table C.1: Default simulation assumptions.

Description	**Assumption**
Environment	3GPP Urban Macro (UMa); 3-sector BSs with 500 meters inter-site distance. 21 cells.
Propagation	Urban Macro-3D
Carrier	2 GHz (FDD), 20 MHz carrier bandwidth
PHY numerology	30 kHz sub-carrier spacing configuration. PRB size of 12 sub-carrier (360 kHz).
TTI sizes	0.143 msec (4-symbols mini-slot).
MIMO	Single-user 2x2 closed loop single-stream (Rank-1) configuration. LMMSE-IRC receiver.
CSI	Periodic CSI every 5 msec, with 2 msec latency.
MCS	QPSK to 64 QAM, with same encoding rates as specified for LTE.
Link adaptation	Dynamic MCS with 1% BLER of initial transmission.
HARQ	Asynchronous HARQ with Chase-combining. HARQ-RTT=4 TTIs.
User distribution	2100 URLLC UEs (Average 100 UEs per cell).
Traffic model	FTP3 downlink traffic with payload size of 50 bytes.
Link-to-system (L2S) mapping	Based on MMIB mapping [22].

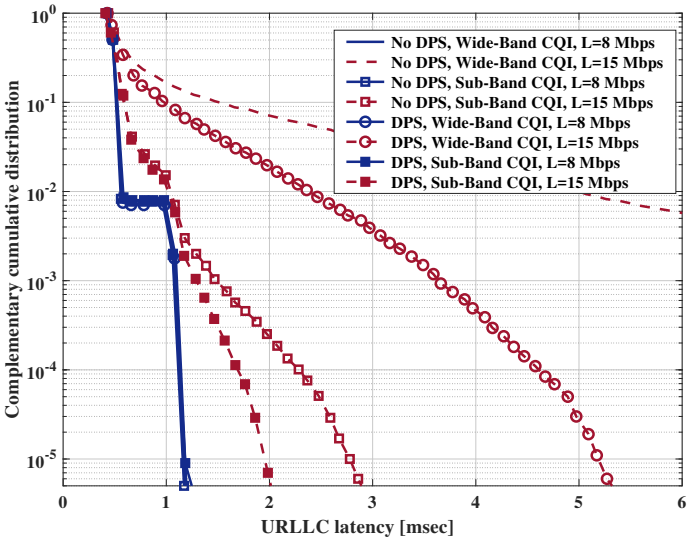

Fig. C.1: URLLC latency distribution for different URLLC offered loads and scheduling methods with $Q = 2$ cells $W = 10$ dB.

4 Simulation Results

4.1 Simulation Methodology and Assumptions

The results are generated by running dynamic system-level simulations following 5G NR methodology in 3GPP [1, 17]. Table C.1 summarizes the network settings and simulation parameters. The key performance indicator (KPI) is the one-way latency with 99.999% reliability. The simulation time is set so at least five million packet transmissions are performed, providing reliable results for the 99.999% percentile of the latency [6].

4.2 Performance Results

Fig. C.1 plots the complementary cumulative distribution function (CCDF) of the URLLC latency for different offered loads and scheduling strategies. At eight Mbps offered load, all schemes have similar performance at $1 - 10^{-5}$ reliability, with latencies between 1.15 to 1.2 msec. At such low offered load, there are only a few active UEs in each scheduling interval. As a consequence, lower levels of inter-cell interference and queuing delay are experienced. Therefore, processing/transmission times, and HARQ-RTT are the dominant factors for the URLLC latency.

Notable latency degradation occurs when increasing the load to 15 Mbps

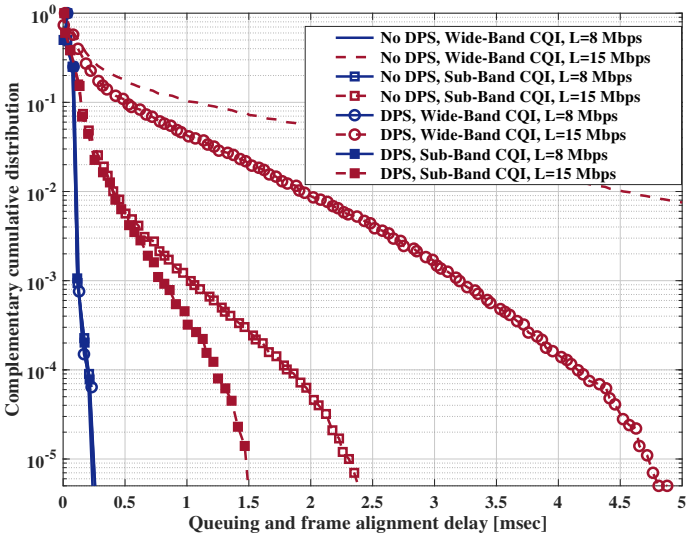

Fig. C.2: Queuing and frame alignment delay for different offered loads and scheduling methods with $Q = 2$ cells $W = 10$ dB.

as a consequence of higher queuing delay. Here, the latency performance varies depending on the used scheduling policy. Considering the baseline (no DPS) with wide-band CQI, the outage reliability of 10^{-5} is achieved at 12.2 msec for 15 Mbps offered load. The latency decreases to 2.77 msec by exploiting frequency-selective scheduling. Around 30% improvement is achieved with DPS so the latency is reduced to 1.95 msec. It can be seen that the combination of DPS and frequency selective scheduling results in 85% latency reduction as compared to the baseline with wide-band CQI. The superior resource allocation by DPS and frequency-selective multiplexing leads to lower number of required PRBs to transmit both data and the PDCCH. As a consequence, the generated inter-cell interference and the queuing delay is decreased.

Fig. C.2 depicts the CCDF of the queuing plus frame alignment delay. Although some temporary queuing is observed at low load regimes, the queuing delay has a major impact on the latency degradation when the load increases. Higher packet arrival rate along with the excessive resources required to mitigate inter-cell interference lead to significant negative impact on queuing delay. Fig. C.2 shows clear advantages of DPS and frequency-selective scheduling reducing the tail of queuing delay. At 15 Mbps load, 10% of the payloads for the baseline wide-band CQI scenario experience more than one msec queuing delay. With DPS, the number of queued packets decreases by a factor of 2.5. This is further reduced by applying sub-band

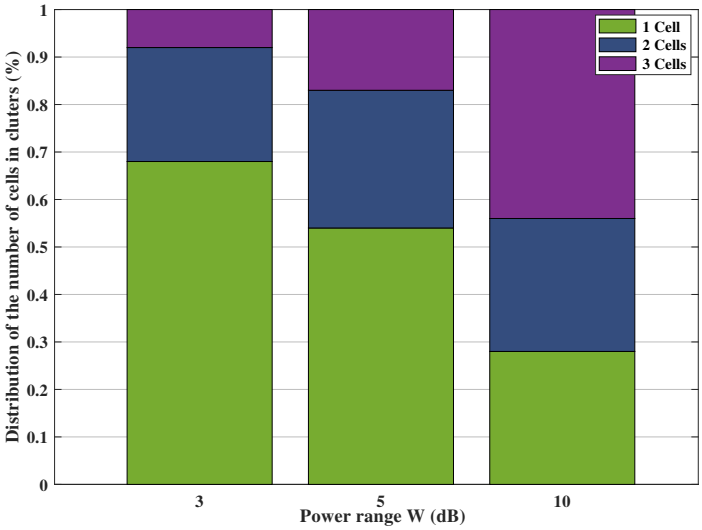

Fig. C.3: Distribution of the number of cells in each cluster with respect to different power ranges W.

scheduling to 0.1% and 0.03% for no DPS and DPS cases, respectively.

4.3 Cluster Variables Analysis

Fig. C.3 shows the dynamic cluster size distribution for different values of W, for $Q = 3$. We observe that with $W = 3$ dB, only 32% of the UEs have more than one cell in their cluster, while only 8% of UEs have three cells. As expected, the number of cells in the cluster increases with W. Assuming $W = 10$ dB, in 72% of cases there are at least two cells in clusters while, 44% of UEs have three cells in their cluster sets.

The impact of different cluster parameters is pictured in Fig. C.4. As can be seen, the major improvement of DPS is achieved for cases with the power range of $W = 3$ dB, where only 32% of clusters have more than one cells. The results indicate that higher value of the power range does not provide additional latency gain. It is less probable for the cells with relatively lower signal strength to provide sufficient spectral efficiency. Our observations confirm that DPS affects mainly cell-edge UEs that receive similar signal power from neighbouring cells.

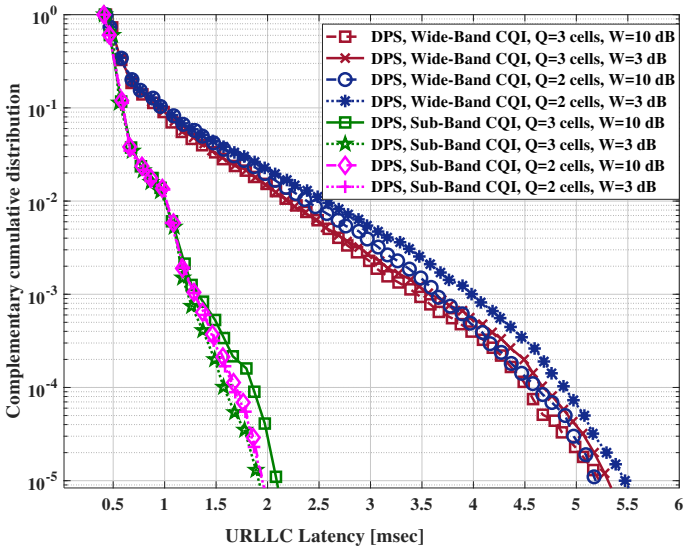

Fig. C.4: URLLC latency distribution for different cluster parameters with $L = 15$ Mbps offered load.

5 Conclusion

We have studied frequency-selective multi-user scheduling and DPS performance of URLLC in 5G NR. Extensive system-level simulations show significant reduced latency of URLLC services at high load scenarios. As an example at 15 Mbps offered load, DPS achieves 30% latency improvement at $1 - 10^{-5}$ reliability. Exploiting the benefits of both DPS and frequency-selective scheduling offers 85% latency improvement as compared to wide-band CQI scheduling. The results show that DPS is mainly beneficial for cell-edge UEs where major improvement is achieved for dynamic user-specific clustering with the power range of $W = 3$ dB. Future studies should examine load-aware DPS algorithms, impact of non-ideal backhaul, and channel quantization error on the URLLC performance.

Acknowledgement

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 Specification 38.300, "NR and NG-RAN overall description; stage-2," Version 2.0.0, December 2017.

[2] 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.

[3] 3GPP Technical Specification 23.501, "Technical specification group services and system aspects, system architecture for the 5G system," Release 15, December 2017.

[4] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, 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.

[5] A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Pocovi, and P. Mogensen, "Efficient low-complexity packet scheduling algorithm for mixed URLLC and eMBB traffic in 5G," *In Proc. 2019 IEEE 89th Vehicular Technology Conference - VTC2019-Spring*, May 2019.

[6] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28 912–28 922, 2018.

[7] P. Popovski, J. J. Nielsen, C. Stefanovic, E. d. Carvalho, E. Strom, K. F. Trillingsgaard, A. S. Bana, D. M. Kim, R. Kotaba, J. Park, and R. B. Sorensen, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Network*, vol. 32, no. 2, pp. 16–23, March 2018.

[8] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, October 2018.

[9] T. K. Vu, C. Liu, M. Bennis, M. Debbah, M. Latva-aho, and C. S. Hong, "Ultra-reliable and low-latency communication in mmWave-enabled massive MIMO networks," *IEEE Communications Letters*, vol. 21, no. 9, pp. 2041–2044, Sep. 2017.

[10] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen, "5G centralized multi-cell scheduling for URLLC: Algorithms and system-level performance," *IEEE Access*, vol. 6, pp. 72 253–72 262, 2018.

[11] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen, "Centralized joint cell selection and scheduling for improved URLLC performance," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, September 2018, pp. 1–6.

[12] N. H. Mahmood, D. Laselva, M. C. F. D. Palacios, M. Emara, D. M. Kim, and I. de-la Bandera, "Multi-channel access solutions for 5G new radio," *IEEE Wireless Comm. Magazine., Submitted.*, 2018.

[13] N. H. Mahmood, A. Karimi, G. Berardinelli, K. I. Pedersen, and D. Laselva, "On the resource utilization of multi-connectivity transmission for URLLC services in 5G New Radio," *CoRR*, vol. abs/1904.07963, 2019.

[14] S. Bassoy, H. Farooq, M. A. Imran, and A. Imran, "Coordinated multi-point clustering schemes: A survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 743–764, Secondquarter 2017.

[15] R. Gupta, S. Kalyanasundaram, and B. Natarajan, "Dynamic point selection schemes for LTE-A networks with load imbalance," in *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, September 2015, pp. 1–5.

[16] R. Gupta, S. Kalyanasundaram, B. Natarajan, and M. Sen, "Performance analysis of enhanced dynamic point selection CoMP scheme for heterogeneous networks," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.

[17] 3GPP Technical Report 38.913, "Study on scenarios and requirements for next generation access technologies," Version 14.1.0, March 2016.

[18] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 210–217, March 2018.

[19] K. I. Pedersen, M. Niparko, J. Steiner, J. Oszmianski, L. Mudolo, and S. R. Khosravirad, "System level analysis of dynamic user-centric scheduling for a flexible 5G design," in *2016 IEEE Global Communications Conference (GLOBECOM)*, December 2016, pp. 1–6.

[20] 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.

[21] 3GPP Technical Documents R1-1808449, "IMT-2020 self-evaluation: UP latency analysis for FDD and dynamic TDD with UE processing capability 2 (URLLC)," August 2018.

[22] T. L. Jensen, S. Kant, J. Wehinger, and B. H. Fleury, "Fast link adaptation for MIMO OFDM," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 8, pp. 3766–3778, October 2010.

Paper D

Centralized Joint Cell Selection and Scheduling for Improved URLLC Performance

Ali Karimi, Klaus I. Pedersen, Nurul Huda Mahmood, Jens Steiner, and Preben Mogensen

The paper has been published in the
IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2018.

Abstract

A centralized joint cell selection and scheduling policy for ultra-reliable low latency communication (URLLC) is studied in this paper for the 5G new radio (NR). A low complexity cell association and scheduling algorithm is proposed, while maintaining attractive performance benefits. By being able to centrally control from which cells the different users are instantaneously scheduled, we show that the undesirable queuing delays for URLLC traffic can be significantly reduced. The proposed solution is evaluated in a realistic multi-cell, multi-user, dynamic network setting in line with the 5G NR system design specifications, and calibrated against 3GPP NR assumptions. The presented performance results show promising gains, where the proposed centralized solution can accommodate 38% *higher traffic offered load than a traditional distributed network implementation, while still fulfilling the challenging reliability and latency targets for URLLC.*

1 Introduction

Research on 5G New Radio (NR) is rapidly progressing with 3GPP having released the first 5G specifications [1]. The ambitions for 5G NR are high, aiming for enhanced support for multiplexing of diverse services such as enhanced mobile broadband (eMBB) and ultra-reliable low latency communication (URLLC) [2], [3]. Meeting the most stringent requirement for URLLC of 1 msec one-way latency in the radio access network with 99.999% reliability to fulfil the IMT2020 target is very challenging. In addition to this stringent requirement, 3GPP have also defined other classes of URLLC requirements as part of the 5G quality of service (QoS) indicator (5QI) with latency budgets of, for instance 5, 10, and 20 msec, as well as reliability targets from 99% to 99.999% [4].

The number of studies towards fulfilling the URLLC targets has rapidly increased during recent years, and hence it would be too exhaustive to quote all. As a few examples, there have been studies of dynamic link adaptation for URLLC in [5] and [6], hybrid automatic repeat request (HARQ) enhancements in [7] and [8], use of different diversity mechanisms in [9], and the benefits of short transmission time intervals (TTIs) and different scheduling policies in [5], [10]. From the system-level URLLC studies, and also more fundamental traffic theory studies, it has been found that the stringent URLLC targets can at most be met up to the point where payloads arriving at the transmitter-end start to be subjected to queuing delays (i.e. not being scheduled immediately upon arrival) [11]. This is of relevance even if the queuing only happens with moderate probabilities in the different cells, when subject to ultra-reliability constraints of e.g. 99.999%. Even for regular macro cellular deployments with spatial uniform traffic, where each user has Poisson

arrival traffic, the instantaneously experienced traffic load per cell is found to be highly time-variant and non-uniform. Meaning that some cells may experience temporary high loads (and consequently queuing delays for some packets), while other cells have low offered traffic and excess capacity available. This is the problem that we address in this study.

Our hypothesis is that a centralized radio access network (C-RAN) with fast joint cell selection and scheduling can improve the URLLC performance. By being able to schedule users from different cells on a per TTI granularity, we can reduce the probability of queuing delays, and thereby improve the performance. The concept of C-RAN multi-cell scheduling has been earlier studied for LTE systems with eMBB traffic, aiming at improving the average experienced user file-down load performance; see e.g. [10]. In this study, we build on those previous studies, and extend the concept to NR URLLC cases, which involves new problems that to the best of our knowledge have not yet been addressed. Our objective is to present a solution that is practically feasible, and therefore must be aligned with the 5G NR system design constraints, and naturally taking advantage of the many new degrees of freedom that comes with the NR. A highly realistic system model is adopted, including time-variant traffic, detailed physical (PHY) layer and medium access control (MAC) functionalities, etc. Based on that, a joint cell selection and packet scheduling method is developed with reasonable complexity. A hierarchical solution is suggested, where the first step involves deciding which users are scheduled at which cells, followed by parallel resource allocation and link adaptation for the individual cells. Although numerous studies have investigated different aspects of cell association and packet scheduling in wireless networks, most of the contributions are based on theoretical results, unrealistic assumptions, and simplified simulations. The performance evaluation and application of proposed contributions on practical systems which consider the network limitations and imperfections is still an open research area [12]. In this work, the proposed method is evaluated in a dynamic multi-user, multi-cell setting with high degree of realism. Due to the complexity of the 5G NR system and the addressed problems, we rely on advanced system-level simulations for results generation. Those simulations are based on commonly accepted mathematical models, calibrated against the 3GPP 5G NR assumptions [1], [2], making sure that reliable results are generated.

The remainder of this paper is organized as follows: the system model and problem formulation are elaborated in Section 2. Section 3 discusses the proposed cell association and scheduling algorithm. Simulation methodology and performance results evaluating the proposed algorithms are presented in Section 4 and Section 5, respectively. Section 6 concludes the paper.

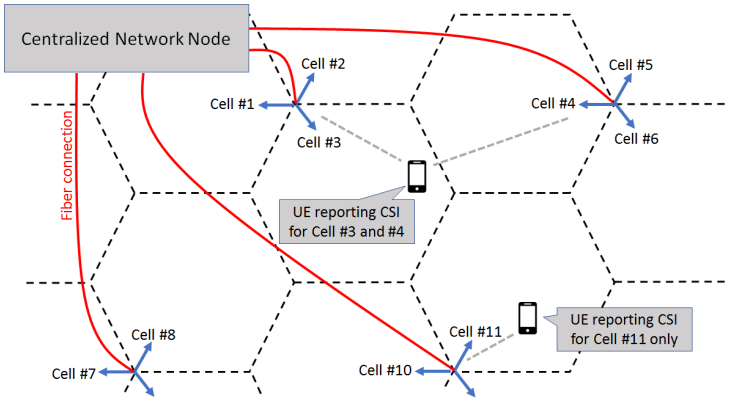

Fig. D.1: Network deployment with centralized network element.

2 Setting the Scene

2.1 Basic System Model

We adopt the 5G NR assumptions as outlined in [1, 2], focusing primarily on the downlink performance. A C-RAN as illustrated in Fig. D.1 is assumed, where the centralized unit is connected with zero-latency (ideal) fiber connections to the remote radio heads (RRHs). The RRHs host the physical (PHY) layer functionalities, while the centralized network node contains the remaining layers of the NR radio access network protocol stack, including the MAC that holds the scheduling functionality. The RRHs are placed to form a traditional three sector macro-cellular deployment with 500 meters inter-site distance. There are C cells in the network. User equipments (UEs) are randomly placed in the network area, following a spatially uniform distribution. An open-loop traffic model is assumed, where bursts of payload sizes of B bytes that arrive for each UE at the centralized network node according to a uniform Poisson arrival point process with arrival rate of λ. In line with the 3GPP NR assumptions for URLLC, we assume $B = 50$ bytes [2]. Thus, the average offered traffic load per cell equals $L = N \cdot B \cdot \lambda$, where N is the average number of UEs per cell.

UEs measure the average received power from the cells, and perform channel state information (CSI) measurements on the Q strongest cells that are within a power-window of W dB (as compared to the strongest received cell) – denoted the CSI measurement set. The value of Q helps limiting the CSI feedback overhead, while W helps ensuring that only relevant cells with sufficiently good quality are in the CSI measurement set [13]. Hence, as illustrated in Fig. D.1, some UEs may report CSI information corresponding

to a single cell only, while other UEs report multi-cell CSI information. The CSI reports are periodically reported to the network, and made available at the centralized network node.

UEs are dynamically multiplexed on a shared channel, using orthogonal frequency division multiple access (OFDMA). We assume the setting with 15 kHz sub-carrier spacing. UEs are scheduled within a short TTI of only 2 OFDM symbols, corresponding to a mini-slot of 0.143 msec. In the frequency domain, users can be multiplexed on a physical resource block (PRB) resolution of 12 sub-carriers. The MAC schedules UEs only from the cells that they report CSIs for, and not from multiple cells per time-instant. UEs are dynamically scheduled, using a user-centric downlink control channel for transmitting the scheduling grant [14]. This includes informing the users on which resources they are scheduled, which modulation and coding scheme (MCS) is used, etc. Hence, dynamic link adaptation (based on the CSI feedback) is applied. Due to the small data packet sizes in URLLC, the required resources to transmit control channel information are comparable to those required for payload. In case a transmission fails, asynchronous HARQ with Chase combining is assumed [15].

2.2 Problem Formulation

The overall objective is to maximize the served average traffic load (L), while still ensuring that all payloads are delivered within a given latency budget, T_{target}, with a reliability of P_{target}, expressed as $P(\tau \leq T_{target}) \geq P_{target}$, where τ denotes the one-way latency defined as the time when a payload is received at the network until it is correctly decoded by the UE. We map this into a joint cell selection and scheduling problem by defining a scheduling matrix $\mathbf{M}$ with dimension $U \times C$, where U is the number of active users in the network with pending data to transmit. Note that U is time-variant due to the dynamic traffic model. The scheduling matrix includes scheduling metrics (m_{uc}) for the individual users that are used to decide the cell associations and scheduling decisions. It is assumed that $m_{uc} = 0$ for cells where UE u does not report CSI values (i.e. cells where the UE is not schedulable). Given $\mathbf{M}$, the objective is to

$$\max_{x_{uc}} \sum_{u=1}^{U} \sum_{c=1}^{C} m_{uc} x_{uc}, \tag{D.1}$$

where $x_{uc} \in \{0, 1\}$ is a binary variable that equals one if the u-th UE is scheduled from cell c, and otherwise zero. As each UE can only be scheduled from at most one cell per TTI, $\sum_{c=1}^{C} x_{uc} \leqslant 1, \forall u$. The summation over the number of required PRBs (R_{uc}) by the UEs associated to the same cell should be less

than those of available at the cell (D_{total}), $\sum_{u=1}^{U} x_{uc} R_{uc} \leqslant D_{total}, \forall c$. Reporting maximum Q-best CSIs, the solution can be found by iteratively solving (D.1) over all possible values of x_{uc} with the complexity of $\mathcal{O}((Q+1)^U)$ [16], which is still considered to be too high for practical C-RAN implementations.

3 Proposed Algorithm

In this study, we propose a hierarchical joint cell selection and scheduling (i.e. resource allocation) algorithm. In the first step, it is decided which users are scheduled from which cells. Secondly, the users assigned to the different cells are allocated PRBs and MCS (according to CSI information) without exploiting any centralized multi-cell/multi-user knowledge. At each TTI, the algorithm in the centralized network node has the following information available:

1. Which users have pending HARQ retransmissions that can be scheduled.
2. Which users have new data pending for transmission.
3. Buffering delay.
4. From which cells the users are schedulable (i.e. corresponding to the UEs CSI measurement set).
5. Required number of PRBs for scheduling the users with pending data in the different candidate cells for each UE.

Notice that 5) is calculated from the knowledge of the CSI and pending data in the network for the UE. Due to the randomness of packet arrival, each RRH only transmits when the assigned users have data to send which is different from the assumption in [17, 18]. Recall here that for the considered URLLC case, the incoming payloads have size $B = 50$ bytes only. So in most cases, one full packet (size B) requires only part of the available bandwidth and can be fitted into one transmission. And even in short TTIs, multiple users can be scheduled.

As we are dealing with URLLC cases, we prioritize scheduling of pending HARQ retransmissions. Those are transmitted to the UEs from the cell with highest reported CSI. HARQ retransmissions are always sent with same MCS and on the same number of PRBs as the original transmission; i.e. Chase combining assumed [15]. Giving higher priority to HARQ transmission has the advantages of first: it avoids excessive queueing delay to meet low-latency requirement of URLLC transmission, and second: serving the UE by the cell with the highest CSI significantly enhances the reliability by increasing the

probability of decoding the message correctly without further retransmissions. After the initial step of scheduling pending HARQ retransmissions, cell c has $D_c \in [0, 1, 2, ..., D_{total}]$ unused PRBs. If no HARQ retransmissions in cell c, $D_c = D_{total}$.

To schedule UEs with new data, an iterative modified matrix elimination method similar to [13] is adopted. First, the scheduling matrix $\mathbf{M}$ is calculated, where m_{uc} is set to zero if no CSI information is available from user u for cell c (i.e. cells that are not schedulable for the UE). During each iteration, we find the highest scheduling metric m_{uc}. If there are enough PRBs to serve the UE, the user u is connected to cell c. To avoid user u from being co-scheduled by other cells, other non-zero elements in u-th row of $\mathbf{M}$ are set to zero. The available number of PRBs at cell c is updated as $D_c = D_c - R_{uc}$. If there are not enough resources at cell c, m_{uc} is set to 0 meaning that the UE u can not be scheduled from cell c. This procedure is repeated until matrix $\mathbf{M}$ has all zero entries. The complexity of this method is $\mathcal{O}(U^3)$ [13].

A more computational efficient method can be achieved by the proposed sequential cell association algorithm where the UE-cell assignment is performed according to scheduling metric. The UE-cell pairs having the higher scheduling metric are chosen first. Details of the proposed algorithm are presented in Algorithm 1. For the case that all UEs have new data and report Q CSIs, the approximated computational complexity of Algorithm 1 is $\mathcal{O}(Q \cdot U \log(Q \cdot U))$, while presenting the same performance as that of matrix elimination method.

In this paper, we investigate two different channel and channel-delay aware scheduling metrics namely maximum throughput (Max-TP) and throughput-delay (TP-Delay), respectively. The Max-TP aims at maximizing the cell TP by prioritizing the user that reports higher TP.

To guarantee service for users with low CSI, inspired from the well known *Modified Largest Weighted Delay First (MLWDF)* algorithm [19], we define TP-Delay metric as:

$$m_{uc} = \begin{cases} TP_{uc} & \text{if } \tau_u \leqslant 0.5 \text{ msec}, \\ \dfrac{\tau_u \cdot TP_{uc}}{\psi} & \text{if } \tau_u > 0.5 \text{ msec}, \end{cases}$$

where τ_u and TP_{uc} represent the u-th UE head of line queuing delay and supported TP from c-th cell, respectively. ψ is equal to 1 OFDMA symbol. Increasing m_{uc} with delay, enhances the probability of scheduling users with queued data.

Algorithm 1 Proposed Algorithm for Cell Association

1: Create a vector of available PRBs at cells.
2: Schedule the HARQ transmission with the highest reported CSI and update the available number of PRBs at the serving cells.
3: For each UE that has new data, define pairs consisting of the UE and its corresponding cell candidates which the UE is schedulable.
4: Create list **S** of candidate pairs.
5: Sort candidate pairs of **S** according to the defined scheduling metric.
6: **while** Unscheduled UEs at **S** and enough PRBs at cells **do**
7: Select the first pair (u, c) of list **S**.
8: **if** $R_{uc} \leq D_c$ **then**
9: Assign UE u to cell c.
10: Update the number of available PRBs at cell c as $D_c = D_c - R_{uc}$.
11: Remove pairs corresponding to u from **S**.
12: **else**
13: Remove pair (u, c) from **S**.
14: **end if**
15: **end while**

4 Simulation Methodology

Extensive dynamic system-level simulations are conducted, following the 5G NR methodology in 3GPP [1, 3], assuming a macro-cellular multi-cell scenario in line with outlined system model in Section II-A. The default simulation assumptions are summarized in Table D.1. For each transmission, the SINR is calculated per sub-carrier symbol. Single-user 2×2 MIMO with Rank 1 is assumed for all transmissions, with UE interference rejection combining (IRC) receiver. 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 [20, 21]. Given the MMIB and the used modulation and coding rate of the transmission, the error probability of the transmission is determined from look-up tables that are obtained from extensive link level simulations. For failed transmission, the UE feeds-back a negative acknowledgement, triggering a HARQ retransmission. Dynamic link adaptation (LA) is assumed by setting the MCS for each transmission based on the users reported CSI. The MCS is adjusted to reach an average block error rate (BLER) target of 1%. This is achieved by using the well-known outer loop link adaptation (OLLA) algorithm, where the received CSI values are offset by certain factor (a.k.a. the OLLA offset) calculated in accordance to the received HARQ Ack/Nacks from past transmissions [5, 14]. Given the burstiness of the URLLC traffic, also the experienced interference at the UEs will be highly time-variant. Each transmission includes both the actual data transmission,

Table D.1: Default Simulation Assumptions.

Description	Assumption
Environment	3GPP Urban Macro (UMa); 3-sector base stations with 500 meters inter-site distance. 21 cells. 3D radio propagation model.
Carrier	10 MHz carrier bandwidth at 2 GHz (FDD)
PHY numerology	15 kHz subcarrier spacing configuration [1]. 10 MHz carrier bandwidth with 50 PRBs.
TTI sizes	0.143 msec (2-symbol mini-slot).
MIMO	Single-user 2x2 closed loop single-stream (Rank-1) MIMO and UE MMSE-IRC receiver. Cross polarized antennas.
CSI	Periodic CSI every 5 msec, with 2 msec latency. UEs report CSI for up to Q strongest received cells that are within a power receive window of W dB. Default is $Q = 1$ and $W = 10$ dB.
Data channel modulation and coding	QPSK to 64QAM, with same encoding rates as specified for LTE. Turbo codes.
Link adaptation	Dynamic MCS selection with 1% initial BLER target.
HARQ	Asynchronous HARQ with Chase Combining. The HARQ RTT equals minimum 4 TTIs.
Traffic model	Poisson arrival of $B = 50$ bytes data bursts from each UE.
Scheduling	Max-TP, TP-Delay.
Link-to-system (L2S) mapping	Based on the mean mutual information per coded bit (MMIB) mapping methodology.

but also a downlink control channel with the actual scheduling grant. The overhead from the scheduling grant is accounted in the simulations in line with the studies [14]. Similarly, the control channel overhead is accounted for in the proposed algorithm when computing how many PRBs are needed for certain UEs to be served from a given cell.

5 Performance Results

Fig. D.2 depicts the complementary cumulative distribution function (CCDF) of the URLLC latency for different proposed methods in a network with the offered load of 3.5 Mbps/cell.

Performance of the centralized Algorithm 1 is compared against the dis-

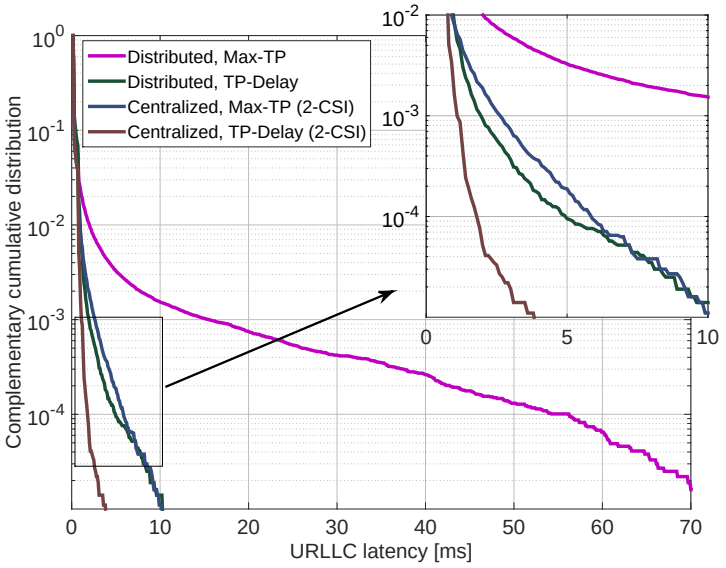

Fig. D.2: URLLC latency distribution with 3.5 Mbps/cell offered load.

tributed case. For distributed implementation, UEs are scheduled from a unique cell, reporting only CSI for that cell. While in the centralized scenario, each UE reports maximum $Q = 2$ CSIs within the power window of $W = 10$ dB and the cell association is performed based on the proposed Algorithm 1. It can be seen from the figure that the centralized cell association methods significantly outperform the distributed ones. The improved performance of the centralized solution is mainly due to the decrease in queuing delay by serving buffered UEs with secondary cells. However by switching the UE to a secondary cell, the required number of resources to send the data and consequently the generated interference increases. Our simulation results show that still significant latency improvement are achieved. For example, considering Max-TP case, the outage probability of 10^{-4} is achieved at 5.8 and 54 msec for the centralized and distributed scenarios, respectively. For the TP-Delay case, it is 1.8 msec for the centralized and 5 msec for the distributed scheduling. Moreover, the performance improvement with the TP-Delay scenarios highlights the importance of channel-delay aware scheduling for URLLC.

Fig. D.3 shows the maximum supported traffic fulfilling the outage probability of 10^{-4} within different latency budgets of 2 and 5 msec. At 2 msec, the centralized TP-Delay provides significant gain of 63% and 38% over distributed scheduling with Max-TP and TP-Delay scenarios, respectively. The

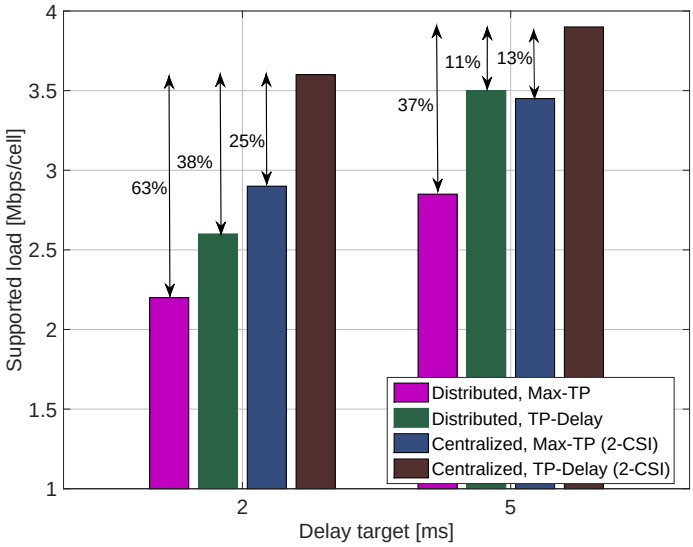

Fig. D.3: Network supported load for different latency budgets and outage probability of 10^{-4}.

achieved gain is due to more efficient use of resources at other low loaded cells (load balancing) and give higher scheduling priority to the queued users (specially those who report low CSIs). The maximum supported load increases as the latency budget is relaxed from 2 to 5 msec. As the network load increases, we observe that the gain from cell association decreases since most of the cell capacity is occupied with primary UEs, consequently there are less resources for secondary UEs.

6 Conclusion

In this paper, we have proposed an efficient centralized cell association and scheduling algorithm for URLLC. The proposed solution is of low complexity and is therefore suitable for practical implementations. The method incorporates an element of fast per TTI inter-cell load balancing as users can be scheduled from different cells on a fast basis. This reduces the probability of experiencing queuing delays that might otherwise happen for a traditional distributed network implementation.

Advanced 5G NR compliant system-level performance results with high degree of realism confirm that our proposed centralized solutions offer attractive performance gains over traditional distributed state of the art solutions. As an example, it is shown that maximum tolerable offered load can be

increased by 38% for the centralized case, while still fulfilling the URLLC requirement. The results also highlight the significant impact of channel-delay aware scheduling.

In our future work, we will therefore further study various related enhancements. Among others, we will generalize the centralized algorithm to allow segmentation of URLLC payload transmissions to happen over multiple TTIs, consider cases where UEs may report CSI reports from more than 2 cells, as well as taking advantage of the centralized architecture also for inter-cell interference control.

Acknowledgement

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 Specification 38.300, "NR and NG-RAN overall description; stage-2," Version 15.5.0, March 2019.

[2] 3GPP Technical Report 38.913, "Study on scenarios and requirements for next generation access technologies," Version 14.1.0, March 2017.

[3] 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.

[4] 3GPP Technical Specification 23.501, "Technical specification group services and system aspects, system architecture for the 5G system," Release 15, December 2017.

[5] G. Pocovi, B. Soret, K. I. Pedersen, and P. Mogensen, "MAC layer enhancements for ultra-reliable low-latency communications in cellular networks," in *2017 IEEE ICC Workshops*, May 2017, pp. 1005–1010.

[6] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji, and R. Jantti, "Link adaptation design for ultra-reliable communications," in *2016 IEEE International Conference on Communications*, May 2016, pp. 1–5.

[7] K. I. Pedersen, S. R. Khosravirad, G. Berardinelli, and F. Frederiksen, "Rethink hybrid automatic repeat request design for 5G: Five configurable enhancements," *IEEE Wireless Communications Magazine*, vol. 24, no. 6, pp. 154–160, December 2017.

[8] G. Berardinelli, S. R. Khosravirad, K. I. Pedersen, F. Frederiksen, and P. Mogensen, "On the benefits of early HARQ feedback with non-ideal prediction in 5G networks," in *2016 International Symposium on Wireless Communication Systems (ISWCS)*, September 2016, pp. 11–15.

[9] D. Ohmann, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Diversity trade-offs and joint coding schemes for highly reliable wireless transmissions," in *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, September 2016, pp. 1–6.

[10] Q. Liao, P. Baracca, D. Lopez-Perez, and L. G. Giordano, "Resource scheduling for mixed traffic types with scalable TTI in dynamic TDD systems," in *2016 IEEE Globecom Workshops (GC Wkshps)*, December 2016, pp. 1–7.

[11] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–5.

[12] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schober, and L. Hanzo, "User association in 5G networks: A survey and an outlook," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1018–1044, Second-quarter 2016.

[13] V. Fernández-López, K. I. Pedersen, B. Soret, J. Steiner, and P. Mogensen, "Improving dense network performance through centralized scheduling and interference coordination," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4371–4382, May 2017.

[14] K. I. Pedersen, M. Niparko, J. Steiner, J. Oszmianski, L. Mudolo, and S. R. Khosravirad, "System level analysis of dynamic user-centric scheduling for a flexible 5G design," in *2016 IEEE Global Communications Conference (GLOBECOM)*, December 2016, pp. 1–6.

[15] 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.

[16] D. Liu, Y. Chen, K. K. Chai, T. Zhang, and M. Elkashlan, "Opportunistic user association for multi-service HetNets using Nash bargaining solution," *IEEE Communications Letters*, vol. 18, no. 3, pp. 463–466, March 2014.

[17] H. Zhou, S. Mao, and P. Agrawal, "Approximation algorithms for cell association and scheduling in femtocell networks," *IEEE Transactions on Emerging Topics in Computing*, vol. 3, no. 3, pp. 432–443, September 2015.

[18] J. Ghimire and C. Rosenberg, "Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1340–1351, March 2013.

[19] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 39, no. 2, pp. 150–154, February 2001.

[20] K. Brueninghaus, D. Astely, T. Salzer, S. Visuri, A. Alexiou, S. Karger, and G. Seraji, "Link performance models for system level simulations of broadband radio access systems," in *2005 IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 4, September 2005, pp. 2306–2311.

[21] R. Srinivasan, J. Zhuang, L. Jalloul, R. Novak, and J. Park, "IEEE 802.16m evaluation methodology document (EMD)," Tech. Rep., July 2008.

Paper E

5G Centralized Multi-Cell Scheduling for URLLC: Algorithms and System-Level Performance

Ali Karimi, Klaus I. Pedersen, Nurul Huda Mahmood, Jens Steiner, and Preben Mogensen

The paper has been published in the
IEEE Access, 2018

Abstract

We study centralized radio access network (C-RAN) with multi-cell scheduling algorithms to overcome the challenges for supporting ultra-reliable low-latency communications (URLLC) in the fifth generation New Radio (5G NR) networks. Low complexity multi-cell scheduling algorithms are proposed for enhancing the URLLC performance. In comparison to the conventional distributed scheduling, we show that the C-RAN architecture can significantly reduce undesirable queuing delay of URLLC traffic. The gain of user scheduling with different metrics and the benefit of packet segmentation are analysed. The performance of the proposed solutions is evaluated with an advanced 5G NR compliant system-level simulator with high degree of realism. The results show that the centralized multi-cell scheduling achieves up to 60% *latency improvement over the traditional distributed scheduling while fulfilling the challenging reliability of URLLC. It is shown that segmentation brings additional performance gain for both centralized and distributed scheduling. The results also highlight the significant impact of channel-delay aware scheduling of URLLC payloads.*

1 Introduction

1.1 Setting the Scene

The third generation partnership program (3GPP) has recently released the first specifications for the fifth generation (5G) radio system, also known as the 5G New Radio (NR) [1]. The 5G NR is designed to fulfil the IMT2020 requirements [2–4], being able to support a diverse set of services with different characteristics and quality-of-service (QoS) targets. One of the challenging service categories is ultra-reliable low-latency communication (URLLC), where the most stringent requirement is 1 msec one-way latency in the radio access network with 99.999% reliability. However, the 5G NR is also designed to support other classes of URLLC requirements as defined in the 5G QoS class indices (5QI) with latency budgets of, for instance 5, 10, and 20 msec, as well as reliability targets from 99% to 99.999% [5].

Meeting the URLLC requirements is obviously a challenging task, especially when considering a highly dynamic multi-cell and multi-user system. Our hypothesis is that a centralized radio access network (C-RAN) architecture with fast multi-cell scheduling is an attractive solution for improving the downlink latency of URLLC, while still fulfilling the reliability requirements. We validate this hypothesis in this paper, starting with a compact overview of previous URLLC studies, followed by further crystallization of our contributions.

1.2 Related Studies

A large number of URLLC related studies have been published during recent years, so it would be too exhaustive to quote all here. Hence, only some relevant examples of which are summarized in the following. The authors in [6] discuss the principles and enablers of URLLC by considering different design aspects. A recent overview paper has been published in [7], focusing on the medium access (MAC) and physical (PHY) layer enablers considered for NR standardization to make URLLC come true. There have been numerous studies on dynamic link adaptation for URLLC in [8], [9], diversity and coding techniques [10], hybrid automatic repeat request (HARQ) enhancements in [11, 12], and variable transmission time intervals (TTIs) [13], [14]. An overview of the scheduler options in 5G NR is provided in [15], including descriptions of new scheduling formats and degrees of freedom added to facilitate URLLC and other services. In [16], the authors study the effect of power allocation for URLLC vehicle-to-vehicle transmission. Several studies also find that queuing delay is a major threat for fulfilling URLLC requirements [17, 18]. As an example, even for homogeneous macro cellular deployments with spatial uniform traffic and Poisson arrival data bursts, some cells may likely experience temporary high loads, and consequently cause queuing delays that can exceed the maximum tolerable latency.

Centralized multi-cell scheduling has been studied earlier for LTE systems with mobile broadband (MBB) traffic for improving the average user experienced data rates [19]. However, to the best of our knowledge, there are very few 5G NR studies of centralized multi-cell scheduling for URLLC use cases. The study in [20] is one such example. Numerous studies have also investigated different cell association and packet scheduling methods in wireless networks. Most of the contributions are proposed for MBB traffic, based on theoretical results and mostly with high computational complexity [21, 22]. The performance evaluation of proposed contributions on practical systems without simplified assumptions and by considering the network limitations and imperfections is still an open research area [23, 24].

1.3 Our Contribution

In the 5G era, C-RAN architectures are expected to gain further popularity, especially in areas where fiber availability is present to realize fronthaul connections with practically zero latency becomes a viable option [25]. Thereby, allowing centralization of resource management procedures to overcome some of the challenges for supporting URLLC. Centralized multi-cell scheduling offers numerous benefits such as increased diversity (e.g. if using dynamic point selection [26]) and the ability to reduce queuing delays as individual users data can be flexibly scheduled from different cells, as com-

pared to more traditional distributed network architectures where users are scheduled from their single serving cell all the time.

We build on the quoted studies and propose improved centralized multi-cell scheduling algorithms for the 5G NR to enhance the URLLC performance. The starting point for the study is a realistic system model in line with the 3GPP NR specifications, adopting the advanced performance assessment models used in 3GPP. The system model comprises a multi-cell deployment with dynamic user traffic models, three-dimensional (3D) channel propagation, the 5G NR protocol stack, flexible frame structure, scheduling, link adaptation, HARQ, MIMO transmission and reception, etc. The dynamic varying overhead from sending scheduling grants to the users is taken explicitly into account. As compared to the our earlier study in [20], enhanced multi-cell scheduling algorithms are proposed and a more detailed system-level performance assessment is presented. In our search for such algorithms, we prioritize solutions of the modest complexity that are feasible for C-RAN architecture implementations, offering additional insight on the trade-offs between achievable performance and the use of sub-optimal algorithms with acceptable complexity.

Attractive multi-cell scheduling algorithms are presented, including cases with/without segmentation of the URLLC payloads over multiple transmission opportunities. That is, without segmentation, only the full URLLC payloads of modest size 50 bytes are scheduled, while for cases with segmentation, we allow that a URLLC payload is segmented so it is transmitted over multiple TTIs. Cases without segmentation have the advantage of aiming for single-shoot transmission of URLLC payloads, at the cost of not always being able to utilize all transmission resources as there may be insufficient resources to transmit full URLLC payloads. On the contrary, use of segmentation allows better utilization of radio resources, but at the expense of (i) higher control channel overhead as each transmission is accompanied with scheduling grant, as well as (ii) possibility of errors at each transmission. The trade-offs between allowing segmentation vs no segmentation therefore signify an interesting problem, which to the best of our knowledge has not yet been fully addressed. In summary, our main contributions in this article are:

- Adopting a highly detailed 5G NR compliant system-model with detailed representation of a macro cellular environment and the many performance determining C-RAN mechanisms for studying URLLC.
- Attractive sub-optimal centralized multi-cell scheduling algorithms for enhancing the URLLC system-level performance of acceptable computational complexity, including cases with/without segmentation of URLLC payloads.
- State-of-the-art system-level performance analysis of centralized multi-

cell scheduling performance for URLLC cases by means of advanced system-level simulations.

Given the complexity of the considered system-model and related scheduling problems, mainly heuristic methods are applied in deriving the proposed algorithms. The corresponding performance analysis is conducted in a dynamic multi-user, multi-cell setting with high degree of realism. Due to the complexity of the system model, we rely on advanced system-level simulations for results generation. Those simulations are based on commonly accepted mathematical models, calibrated against the 3GPP 5G NR assumptions [2], making sure that statistical reliable results are generated.

The rest of the paper is organized as follows: In Section 2, we outline the system model and a more detailed problem formulation of the multi-cell scheduling challenge for URLLC. In Section 3 the proposed multi-cell scheduling algorithms are presented. The system-level simulation methodology appears in Section 4, followed by performance results in Section 5. Finally, the study is concluded in Section 6.

2 System Model and Problem Formulation

In line with [20] and [17], and the 3GPP NR specifications [27], we outline the assumed system model in the following, as well as present the problem formulation in greater details.

2.1 Network topology and Traffic Model

We consider C-RAN architecture as depicted in Fig. E.1 comprises of one centralized unit (CU) controlling several remote radio heads (RRHs) in a large geographical area. Ideal loss-less and zero-latency communication via fiber optic cables is assumed between the CU and RRHs. The interface between the CU and the RRHs corresponds to split option-7 [28], also known as the F2 interface that can be realized with the common public radio interface (CPRI), or the enhanced CPRI (eCPRI). In line with the 3GPP defined NR architecture (see [1] and [29]), the CU hosts all the radio access network protocols from the higher PHY and upwards. Hence, including the service data adaptation protocol (SDAP), packet data convergence protocol (PDCP), radio link control (RLC), and MAC that holds the scheduling responsibility, as well as the control plane protocol and radio resource control (RRC) functionality. Thus, the RRH only includes the lower PHY functions.

The 3GPP urban macro (UMa) deployment is assumed where the RRHs are deployed in a sectorized macro cellular deployment with 500 meters inter-site distance, each hosts three sectors (cells) [2, 17]. A set of $\overline{U}$ URLLC users (UE) are randomly placed in the network area with uniform distribution. A

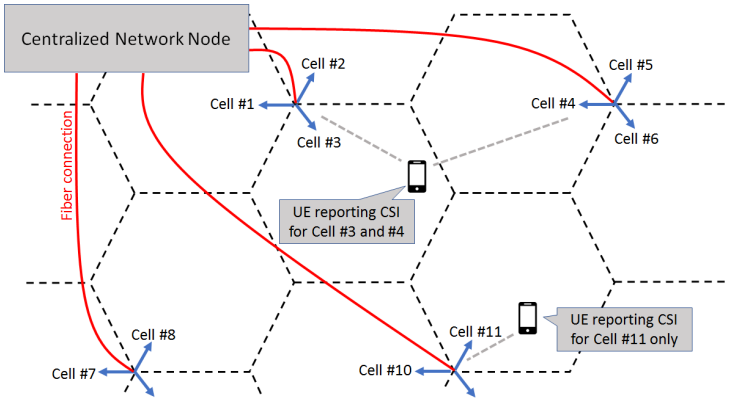

Fig. E.1: Network deployment with network element.

birth-dead traffic model is assumed for each URLLC UE in which a burst of small payloads of B bytes arrive at the CU according to the Poisson distribution with an average arrival rate of λ packet per second. This traffic model is known as FTP3 in 3GPP [27]. The average offered load per cell equals to $L = 8 \cdot \overline{U} \cdot B \cdot \lambda / C$ bps/cell, where C denotes the number of cells in the network area.

2.2 Basic Radio Assumptions

In line with [19] and [20], each UE measures the average reference symbol received power (RSRP) from the cells that it can hear and creates its channel state information (CSI) measurement set of maximum Q ($Q \geq 1$) cells it can connect to. The measurement set contains the cell with the highest received power denoted as the primary cell. It also includes up to the $Q-1$ other strongest secondary cells within the power range of W dB as compared to the primary one.

The UE measures the channel and interference for each of the cells in the CSI measurement set and reports the CSI to the network. The value of Q limits the computational complexity of CSI measurement as well as the CSI feedback overhead. Parameter W helps to control that the measurement set contain cells with sufficiently good channel quality.

Users are dynamically time-frequency multiplexed on a shared channel, using orthogonal frequency division multiple access (OFDMA). A 15 kHz sub-carrier spacing is assumed, where one physical resource block (PRB) equals 12 sub-carriers. A short TTI size of 0.143 msec, corresponding to a mini-slot of 2 OFDM symbols is assumed. The minimum scheduling resolution is one TTI (time-domain) and one PRB (frequency domain). Considering

10 and 20 MHz bandwidth (BW) configurations, the total number of available PRBs equals to $D_{total} = 50$ and $D_{total} = 100$ PRBs, respectively.

The network is only allowed to schedule a user from a cell that belongs to the user's CSI measurement set, and only from one cell per TTI. Whenever the MAC schedules a user on a certain set of resources, both a user-specific scheduling grant on the physical downlink control channel (PDCCH) and the actual transport block (data) on the physical downlink shared channel (PDSCH) are transmitted. In line with [15] and [17], the scheduling grant on the PDCCH is transmitted with aggregation levels of one to eight (or even 16) to ensure good reception quality at the UE. The data transmission on the PDSCH relies on fast link adaptation where the effective coding rate and modulation scheme is set per transmission (and communicated to the UE as part of the scheduling grant).

The link adaptation for PDCCH (i.e. setting of the aggregation level) and PDSCH is based on the received CSI from the user. As the CSI is subject to reporting delays (and other imperfections), we rely on the well-known outer loop link adaptation (OLLA) to control the block error rate (BLER). As in [8] and [17], the OLLA is set to 1% BLER for the first PDSCH transmission. If the UE fails to correctly decode a downlink scheduled data transmission, it will feed back a negative acknowledgement (NACK), and the network will later schedule a corresponding HARQ retransmission. Asynchronous HARQ is assumed for the 5G NR [11]. Conventional Chase combining [30] is assumed to combine the signals received over multiple transmissions.

2.3 Latency Procedure

The downlink one-way user latency (Υ^{tot}) is defined from the time a packet arrives at the CU, until it is successfully received at the UE. If the UE decodes the packet correctly in the first transmission, the latency equals the first transmission delay (Υ^0) expressed as:

$$\Upsilon^0 = d^0_{q,fa} + d_{cup} + d_{tx} + d_{uep}, \tag{E.1}$$

where $d^0_{q,fa}$ denotes the queuing and frame alignment delay of initial transmission, d_{tx} is the payload transmission time. Processing time at the CU and UE are denoted by d_{cup} and d_{uep}, respectively. If the message is erroneously decoded, the packet is subject to HARQ retransmission(s) until either it is decoded successfully or the maximum retransmissions (ϱ) is reached. In this case, Υ^{tot} can be formulated as:

$$\begin{aligned} \Upsilon^{tot} &= \Upsilon^0 + \sum_{i=1}^{r} \Upsilon^i, \\ \Upsilon^i &\triangleq d^i_{q,fa} + d^{RTT}_{HARQ}, \end{aligned} \tag{E.2}$$

where $r \in [1, \cdots, \varrho]$ and Υ^i denote the number of retransmissions and the i-th retransmission delay ($i \geq 1$). The HARQ round trip time is denoted by d_{HARQ}^{RTT}. In line with [17], we assume that the minimum retransmission delay is equal to $d_{HARQ}^{RTT} = 4$ TTIs.

The processing times (d_{cup} and d_{uep}) are considered to be constant with the length of 3 OFDM symbols at both the network and the receiver end [31]. The transmission time is a discrete random variable. Depending on the packet size, channel quality, and the number of assigned PRBs, d_{tx} varies from one to multiple TTIs. The frame alignment delay is a random variable with uniform distribution between 0 and 1 TTI. The queuing delay is defined as the waiting time for getting scheduled at physical layer. It is a random variable and depends on various network parameters such as the payload size, channel quality and required QoS, number of available resources, network load, and the scheduling algorithm.

It has earlier been attempted to study the effect of queuing delays by adopting multi-class queuing network models as considered [32], [33]. For such models users connected to the same cell are categorized in Γ different classes $\mathbf{k} = \{k_1, k_2, ..., k_\Gamma\}$ where members of each class share the same signal to interference plus noise ratio (SINR). On a TTI basis, the packet arrival of k-th class is modelled as a Poisson distribution with the average of $\lambda_k = \overline{u}_k \times \lambda_{TTI}$. $\overline{u}_k$ and λ_{TTI} are the number of UEs in k-th class and the user average packet arrival rate in each TTI, respectively. Note that $\overline{u}_k$ changes with channel variation. Although such models do offer some valuable insight, they fail to fully capture all performance-determining factors of the system model, and in particularly interference coupling between cells, causing random SINR fluctuations.

In a time instance, assume there are u_k UEs with pending data in k-th class, each requires r_k PRBs to transmit the packet. One or some of the UEs are subjected to queuing/multiple TTI transmission delay if

$$\sum_{k=1}^{\Gamma} u_k r_k > D_{total}.$$

2.4 Problem Formulation

The CU has the following information available at each TTI:

1. Which users have pending HARQ retransmissions.
2. Which users have new data and the corresponding buffering delay.
3. From which cells the users are schedulable (i.e. corresponding to the UEs CSI measurement set).

4. An estimate of the number of PRBs for transmission of both the data and PDCCH for the cells in the CSI measurement set.

The overall objective is to maximize the tolerable average served traffic load L, while still ensuring that all payloads are delivered within a given latency budget, T_{target}, with a reliability of P_{target}, expressed as $P(Y^{tot} \leq T_{target}) \geq P_{target}$. In order to minimize the undesirable control channel overhead that unavoidable comes from segmentation of a payload over multiple TTIs, we first aim for single TTI transmission of the full URLLC payloads. For a multi-cell multi-user network of U UEs with pending data and C cells, we formulate a joint scheduling problem by defining the scheduling matrix $\mathbf{M} \in \mathbb{R}_{+}^{U \times C}$. Element m_{uc} of $\mathbf{M}$ is the scheduling metric for user u on cell c used for multi-cell scheduling decisions. It is assumed that $m_{uc} = 0$ for cells that are not included in the CSI measurement set of UE u. Given $\mathbf{M}$, our objective is expressed as:

$$\begin{aligned} &\max_{x_{uc}} \sum_{u=1}^{U} \sum_{c=1}^{C} x_{uc} m_{uc}, \\ &\text{Subject to:} \quad \sum_{u=1}^{U} x_{uc} R_{uc} \leqslant D_{total}, \ \ \forall c, \\ &\qquad\qquad \sum_{c=1}^{C} x_{uc} \leqslant 1, \ \ \forall u, \\ &\qquad\qquad x_{uc} \in \{0,1\} \ \ \forall u, c, \end{aligned} \tag{E.3}$$

where R_{uc} denotes the estimated number PRBs to schedule UE u from cell c. Binary variable x_{uc} equals one if the u-th UE is scheduled from cell c, and otherwise zero. The first constraint is to guarantee that the summation over the number of required PRBs by the UEs associated to the same cell does not exceed total number available PRBs (D_{total}). The second constraint ensures that each UE is scheduled from at most one cell per TTI.

Note that (E.3) is a mixed linear integer problem which can be solved using brute-force algorithm with complexity $\mathcal{O}((Q+1)^U)$ [34]. As an example, for $U = 30$ active user in a TTI and $Q = 2$ CSIs, the complexity of optimal solution equals $3^{30} \sim 2 \times 10^{14}$. However, this is too high for practical C-RAN implementations as the scheduling decision needs to be taken every TTI and in a fast basis.

3 Proposed Multi-Cell Scheduling

A low-complexity hierarchical joint multi-cell scheduling is proposed according to the following steps. First, pending HARQ packets and full URLLC payloads are scheduled. Finally, segmentation is applied.

3.1 Pending HARQ and Full Payload Packet Scheduling

Pending HARQ Retransmissions: We assign the highest priority to pending HARQ retransmissions. HARQ retransmissions are scheduled immediately and from the cell which provides the best CSI. Giving the highest priority to HARQ avoids additional queuing delay of HARQ retransmissions as they are already subject to additional retransmission delay(s) of d^{RTT}_{HARQ}. Also, the probability of successful decoding increases by scheduling the UE from the cell with highest channel quality. Thus, we reduce the probability of further retransmission delays.

Buffered URLLC Packets: After scheduling of HARQ retransmissions, buffered packets are scheduled on the remaining PRBs. A modified matrix elimination method inspired by [19] for URLLC is adapted as follows. Based on the reported CSIs, the elements of the scheduling matrix $\mathbf{M}$ and the corresponding required number of PRBs are calculated (recall that $m_{uc} = 0$ if the c-th cell is not included in the CSI measurement set of u-th UE). If there are not enough PRBs at cell c to transmit the full payload of UE u, the corresponding scheduling metric is set to 0 meaning that UE u can not be scheduled from cell c.

At each step, the highest scheduling metric m_{uc} is selected. If there are enough PRBs at the candidate cell c to transmit the payload of UE u, the UE u is scheduled with cell c and the CU updates the number of its available PRBs as $D_c = D_c - R_{uc}$, otherwise sets $m_{uc} = 0$. To avoid user u from being co-scheduled by the other cell, the u-th row of $\mathbf{M}$ is removed. The procedure is repeated until the matrix $\mathbf{M}$ has all zero entries. The complexity of this method is $\mathcal{O}(U^3)$ [19].

A computationally efficient implementation of this method can be achieved by a sequential method as described in Algorithm 1. The approximated computational complexity of Algorithm 1 is $\mathcal{O}(Q \cdot U \log(Q \cdot U))$, while presenting the same performance as that of the matrix elimination method. It can be seen that the complexity of Algorithm 1 is significantly lower than that of the brute-force solution, making it attractive for practical C-RAN implementation.

Three different scheduling metrics are considered. Maximum throughput (Max-TP), proportional fair (PF), and throughput-delay (TP-Delay). The Max-TP aims at maximizing the achievable cell TP by prioritizing UEs reporting higher TP. In this case, the scheduling metric is defined as $m_{uc} = TP_{uc}$, where TP_{uc} is the predicted TP of the u-th UE if served by c-th cell. In line with [8, 17, 35], we also consider the well-known proportional fair (PF) metric:

$$m_{uc} = \frac{TP_{uc}}{\overline{TP}_u}$$

where $\overline{TP}_u$ is the average delivered throughput in the past.

Algorithm 1 Proposed Algorithm for Cell Association

1: Create a vector of available PRBs at cells.
2: Schedule the HARQ transmission through the cell with the highest reported CSI and update the available number of PRBs at the serving cells.
3: For each UE that has new data, define pairs consisting of the UE and its corresponding cell candidates which the UE is schedulable.
4: Create list **s** of candidate pairs.
5: Sort candidate pairs of **s** according to the defined scheduling metric.
6: **while** Unscheduled UEs at **s** and enough PRBs at cells **do**
7: Select the first pair (u, c) of list **s**.
8: **if** $R_{uc} \leq D_c$ **then**
9: Assign UE u to cell c.
10: Update the number of available PRBs at cell c as $D_c = D_c - R_{uc}$.
11: Remove pairs corresponding to u from **s**.
12: **else**
13: Remove pair (u, c) from **s**.
14: **end if**
15: **end while**

Inspired from the well-known *Modified Largest Weighted Delay First (ML-WDF)* algorithm [36], we finally define the TP-Delay metric as:

$$m_{uc} = \begin{cases} TP_{uc} & \text{if } \tau_u \leqslant 0.5 \text{ msec}, \\ \dfrac{\tau_u \cdot TP_{uc}}{\psi} & \text{if } \tau_u > 0.5 \text{ msec}, \end{cases}$$

where τ_u represents the u-th UE head of line queuing delay and ψ is equal to the time of 1 OFDMA symbol in msec. The metric increases with queuing delay and thus increases the probability of scheduling UEs with queued data.

After completion of Algorithm 1, users that can be scheduled with their full URLLC payload (one packet) have been assigned. However, there may still be some unused PRBs at some cells that could be utilized, although being insufficient to accommodate transmission of full URLLC payloads. The advantage of allowing segmentation is that higher PRB utilization is achieved, but at the cost of more generated interference because of the higher PRB utilization. Moreover, recall that to allow transmission from a cell to a UE, the available PRBs at the cell should be enough for transmission of both the PDCCH and the segmented URLLC payload at the PDSCH. The minimum required allocation size (R_{uc}^{min}) for the link between u-th UE and c-th cell is a function of the experienced SINR at the UE (obtained through the CSI). Table E.1 depicts mapping of the SINR to the required number resource elements (REs) for the transmission of PDCCH and related reference signals. As the

segmentation involves additional cost in terms of higher control overhead, at most one UE is segmented per cell and scheduled over remaining PRBs. Users in good channel conditions (i.e. lower control channel overhead) are also prioritized for segmentation. Algorithm 2 is a method to allow segmentation over the cells with sufficient number of remaining PRBs (after having executed Algorithm 1), transmitting a segmented URLLC payload.

Table E.1: Mapping SINR to CCH overhead and minimum allocation size

SINR [dB]	CCH overhead (REs)	Min. Alloc. Size (PRBs)
$[4.2, \infty)$	$1 \times 36 = 36$	4
$[0.2, 4.2)$	$2 \times 36 = 72$	6
$[-2.2, 0.2)$	$4 \times 36 = 144$	10
$(-\infty, -2.2)$	$8 \times 36 = 288$	20

Algorithm 2 Proposed Algorithm for Segmentation

1: Create a vector of available PRBs at cells.
2: For each of the unscheduled UE, define pairs consisting of the UE and its corresponding cell candidates which have available RBs more than that of minimum required by the UE.
3: Create list **s** of candidate pairs.
4: Sort candidate pairs of **s** according to throughput.
5: **while** Unscheduled UEs at **s** and enough PRBs at cells **do**
6: Select the first pair (u, c) of list **s**.
7: **if** $R_{uc}^{min} \leq D_c$ **then**
8: Assign UE u to cell c.
9: Remove pairs corresponding to u-th UE from **s**.
10: Remove pairs corresponding to c-th cell from **s**.
11: **else**
12: Remove pair (u, c) from **s**.
13: **end if**
14: **end while**

4 Simulation Methodology

The performance of the proposed algorithms is evaluated by extensive system-level simulations following the 5G NR methodology in [1, 3]. The simulations methodology is based on commonly accepted mathematical models and is calibrated against 3GPP 5G NR assumptions [1, 2]. Table E.2 summarizes the network configuration and default simulation parameters. The network operates at a carrier frequency of 2 GHz with 10 and 20 MHz bandwidth.

Table E.2: Default simulation assumptions.

Description	Assumption
Environment	3GPP Urban Macro (UMa); 3-sector RRHs with 500 meters inter-site distance. 21 cells.
Propagation	Urban Macro-3D.
Carrier	2 GHz (FDD).
PHY numerology	15 kHz sub-carrier spacing configuration. PRB size of 12 sub-carriers (180 kHz). 24 REs in each PRB (4 REs are reserved transmission of the reference symbols. 20 REs for data) 10 and 20 MHz carrier bandwidth with 50 and 100 PRBs, respectively.
TTI sizes	0.143 msec (2-symbols mini-slot).
MIMO	Single-user 2x2 closed loop single-stream (Rank-1) configuration with $+45/-45$ cross polarization antennas at the cell, $0/90$ isotropic antenna at the UE. MMSE-IRC receiver.
CSI	Periodic CSI every 5 msec, with 2 msec latency. UEs report CSI for up to Q strongest received cells that are within a power receive window of W dB. In distributed scenario, $Q = 1$ and in centralized scheduling, default is $Q = 2$ and $W = 10$ dB.
Data channel modulation and coding	QPSK to 64 QAM, with same encoding rates as specified for LTE. Turbo codes.
Link adaptation	Dynamic MCS selection with 1% initial BLER target.
HARQ	Asynchronous HARQ with Chase combining. The HARQ RTT equals minimum 4 TTIs.
User distribution	2100 (4200) URLLC UEs uniformly distributed over the network area (Average 100 (200) UEs per cell).
Traffic model	FTP3 downlink traffic with Poisson arrival of $B = 50$ bytes data bursts from each UE.
Scheduling	Max-TP, TP-Delay, PF.
Link-to-system (L2S) mapping	Based on effective exponential SINR mapping (EESM).

The simulator resolution is one OFDM symbol and includes all 5G NR radio resource management functionalities outlined in Section. 2.

The network consists of $C = 21$ macro cells in a three sector cellular deployment with 500 meters inter site distance. Closed-loop 2×2 single-user MIMO with rank one is assumed for all the transmissions. Each cell is configured with one panel set with $-45/+45$ degree polarization. At the UE-side, antenna polarization is $0/90$. 3GPP urban macro-3D channel model is considered [37].

A dynamic birth-death traffic model is assumed where for each UE finite-length payloads of $B = 50$ bytes are generated following a homogeneous Poisson distribution with the average of λ packet per second. Each UE performs the channel and interference estimation of the cells in the CSI measurement set periodically every 5 msec. The CSI reports are subject to 2 msec delay before being applied at the CU. In distributed scenario, each UE reports one CSI corresponding to the cell with highest RSRP value. For the centralized case, the default values of measurement set size and the window size are $Q = 2$ and $W = 10$ dB, respectively.

To suppress the noise and received interference, the UE exploits linear minimum-mean square error interference rejection combining (MMSE-IRC) receiver. After each transmission the effective SINR for each of the assigned REs is calculated and the effective exponential SINR mapping (EESM) is computed over all the scheduled RBs [38]. The calculated EESM value along with the knowledge of transmitted MCS are used to determine the probability of packet failure from detailed look-up tables that are obtained from extensive link level simulations.

The key performance indicator (KPI) for URLLC is defined as the one way achievable latency with different reliability target (i.e. 99.99%). The network URLLC capacity is defined as the maximum supported load at which the defined reliability and latency is satisfied. The simulations runs over more than 5 million packet transmissions generating results with the confidence level of 95% for the 99.999% percentile of the latency [17].

5 Simulation Results

5.1 Performance of Algorithm 1

Fig. E.2 depicts the complementary cumulative distribution function (CCDF) of the URLLC latency for a network with $BW = 10$ MHz bandwidth and the offered load of $L = 3.5$ Mbps/cell. The performance of the centralized Algorithm 1 is compared against that of the distributed one under different scheduling metrics. As can be seen, the centralized multi-cell scheduling significantly outperforms the distributed one. The improved latency perfor-

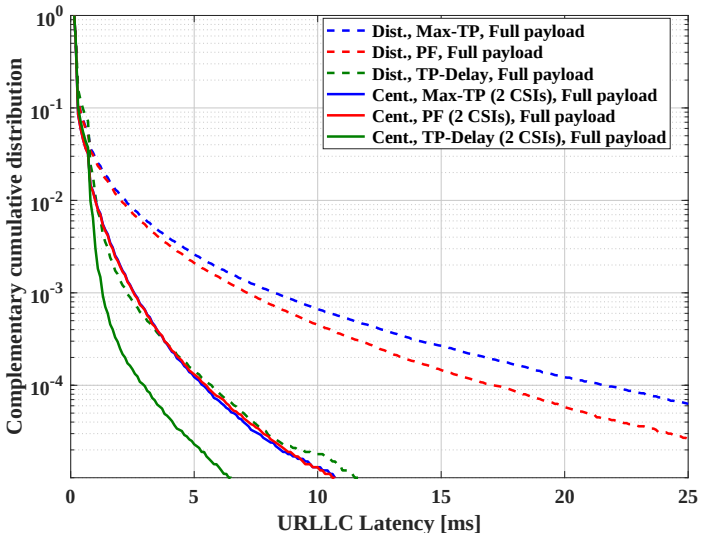

Fig. E.2: URLLC latency distribution with L = 3.5 Mbps/cell, W = 10 dB, BW=10 MHz.

mance is mainly due to the decrease in queuing delay by exploiting available resources at secondary cells to serve more UEs. With Max-TP, the outage probability at 10^{-4} is achieved at 5.1 and 22 msec for the centralized and distributed solutions, respectively. Considering the PF metric, the latency of 17 msec for distributed solution decreases to 5.1 msec with centralized scheduling. Finally, for TP-Delay the latency improves from 5.7 msec to 3 msec. In comparison to previous studies with PF scheduling [17, 35], the TP-Delay scheduling metric provides better latency performance. At an outage probability of 10^{-4}, it achieves more than 66% and 41% latency gain under the distributed and centralized scheduling, respectively. The superior performance of the TP-Delay metric highlights the importance of channel-delay aware scheduling for URLLC. Putting the results into further perspective, it is worth noticing that end-user throughput gains of 40% from using centralized multi-cell scheduling for LTE are reported in [19, 39] for mobile broadband file download.

5.2 Performance of Algorithm 2

Now, we compare the performance of Algorithm 1 with the case where Algorithm 2 (segmentation) is also applied over the remaining PRBs after executing Algorithm 1. Figs. E.3 and E.4 present the CCDF of the URLLC latency for a network with 10 and 20 MHz bandwidth and different average loads of $L = 3.5$ Mbps/cell and $L = 8.5$ Mbps/cell, respectively. The re-

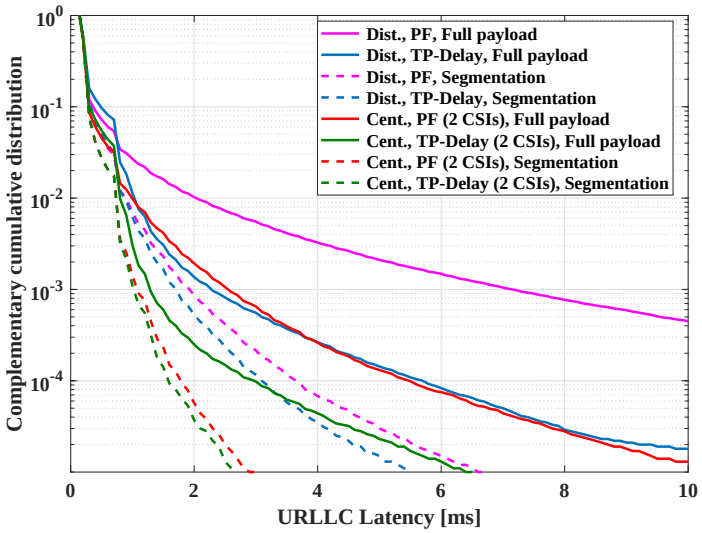

Fig. E.3: URLLC latency distribution with L = 3.5 Mbps/cell, W = 10 dB, BW = 10 MHz.

sults confirm that segmentation brings additional benefit for both centralized and distributed scheduling. For $BW = 10$ MHz system, it achieves significant improvements of 83% and 67% under PF for distributed and centralized scheduling. The results with TP-Delay show an improvement of 45%. The improved performance is due to the efficient utilization of all the available PRBs, thus reducing the queued data size. It is especially beneficial for low SINR UEs as they usually require large number of PRBs, which may be challenging to fit into one TTI. The main benefit of segmentation comes from applying it over the primary cells. It is usually less efficient to transmit a small part of the payload over a secondary cell as the performance degradation due to the generated interference by transmission of PDCCH becomes comparable to the achieved gain of transmitting part of the message.

Comparing 10 and 20 MHz bandwidth configuration reveals that by doubling the bandwidth, the maximum supported load that can achieve the same latency budget is more than doubled. For example, considering centralized TP-Delay scenario, 5 msec latency at the outage probability of 10^{-4} is achieved supporting $L = 4$ Mbps/cell and $L = 9.3$ Mbps/cell for 10 and 20 MHz bandwidth, respectively. Similar findings are reported in [17, 32].

Table E.3 compares the latency performances of distributed and centralized scheduling at different loads and latency budgets at an outage probability of 10^{-4}. Centralized scheduling achieves $30\% - 60\%$ improvement with respect to that of distributed one. At low latency regimes (equivalent to low

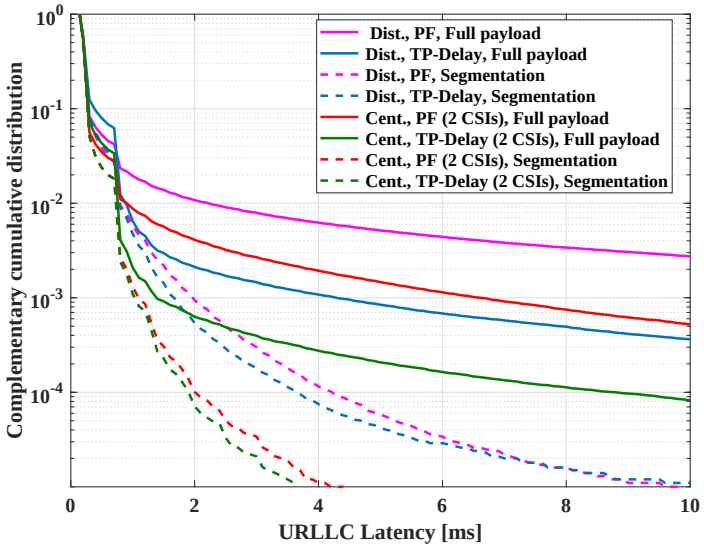

Fig. E.4: URLLC latency distribution with L = 8.5 Mbps/cell, W = 10 dB, BW = 20 MHz.

Scenario	Delay [msec]					
	BW = 10 MHz			BW = 20 MHz		
Dist., Seg.	2	5	10	2	5	10
Cent. (2 CSIs), Seg.	1.3	2.16	6	1.38	2.3	3.95
Improvement (%)	35%	57%	40%	31%	54%	60%

Table E.3: Network performance of TP-Delay scheduling for different latency budgets and at the outage probability of 10^{-4}.

network loads), the effect of transmission delay, processing time, and HARQ RTT are dominant. As the average offered load increases, queuing delay becomes more dominant and thus the gain of centralized scheduling increases.

5.3 CSI Measurement Set Sensitivity

We next investigate the performance sensitivity versus the settings for the UEs CSI measurements (namely Q and W parameters), particularly assessing how many cells shall be considered by the centralized multi-cell scheduling algorithm for each UE. Fig. E.5 illustrates the percentage of UEs having either one, two or three cells in its CSI measurement set depending on the value of W, for $Q = 3$. As expected, by increasing the value of the window size (W) the percentage of UEs with a CSI measurement size of two or three increases. For example, increasing the window size from $W = 2$ dB to $W = 15$ dB, the

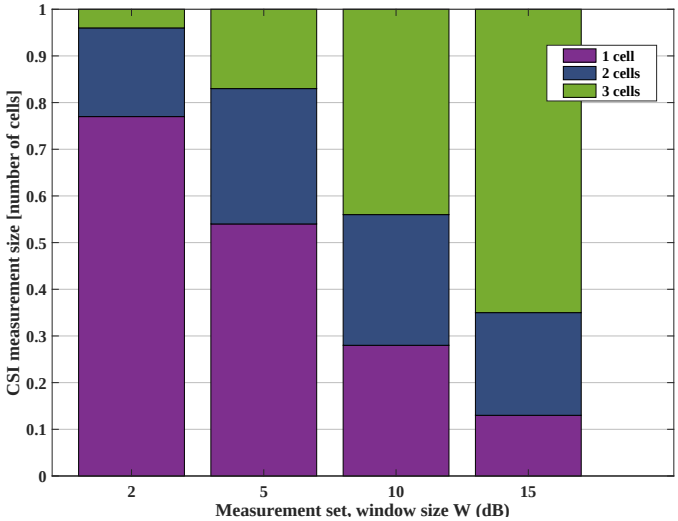

Fig. E.5: Distribution of the number of cells each user connects to, with different window size W, Q = 3, BW = 10 MHz.

percentage of UEs with a CSI measurement size greater than one increases from 23% to 87%, respectively, i.e. those UEs that are subjected to multi-cell scheduling. The effect of Q and W on the URLLC performance is presented in Fig. E.6.

It is interesting to note that the major improvements of the URLLC latency performance are achieved with $Q = 2$ cells and $W = 2$ dB, despite that only 23% of the UEs have a CSI measurement size of two, and thus 77% of the UEs are scheduled always from their primary cell. Increasing W to 5 dB or 10 dB results in additional performance benefits. Increasing W beyond 10 dB results in no additional gains, but rather a risk of experiencing some performance losses as cells with too weak signal strength are included in the UEs CSI measurement set. Increasing Q from 2 to 3, at the best results in minor additional benefits. The former observation partly relates to our assumption of having UEs with two receive antennas and MMSE-IRC receiver type, and thus being able to maximum suppress the interference from one dominant interfering cell. Hence, for $Q = 2$, the UE may be able to suppress the interference from its primary cell if being scheduled from its secondary cell. While if $Q = 3$, it cannot suppress the interference from both its primary cell and the strongest secondary cell, if being scheduled from the weakest secondary cell.

Fig. E.7 shows the empirical CDFs of the predicted TP for the cells in the CSI measurement set for $Q = 3$, $W = 10$ dB, and different offered loads.

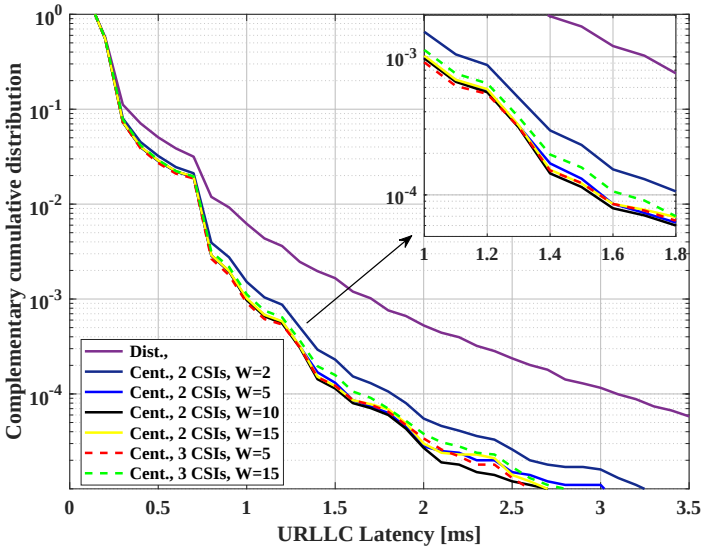

Fig. E.6: URLLC latency distribution for TP-Delay, segmentation scheduling with L = 3.5 Mbps/cell, BW = 10 MHz.

As expected, the highest TP is observed the 1st cell (primary) where the UE receives the strongest RSRP. The supported throughput for the second and third strongest cells is clearly much lower, and hence further illustrates why the benefits of setting $Q = 3$, as compared to $Q = 2$, are marginal, and in most cases not worth considering. Hence, based on the reported findings in Figs. E.5-E.7, we recommend using $W \in [5\ 10]$ dB and $Q = 2$. Referring to the complexity expressions for the centralized multi-cell scheduling algorithms in Section 3, using $Q = 2$ (instead of $Q = 3$) also helps significantly reduce the complexity of centralized multi-cell scheduling algorithms. Similarly, the UE complexity, and uplink CSI reporting overhead is obviously more attractive for $Q = 2$, as compared to $Q = 3$.

6 Conclusion

In this paper, we have investigated centralized multi-cell scheduling of URLLC for 5G NR. Dynamic algorithms including the case with/without segmentation of URLLC payloads are proposed to improve the latency and reliability of URLLC. The solutions have low computational complexity and are attractive for practical C-RAN implementations.

The performance of the proposed solutions is evaluated by performing a variety of simulations using a highly detailed advanced 5G NR compliant

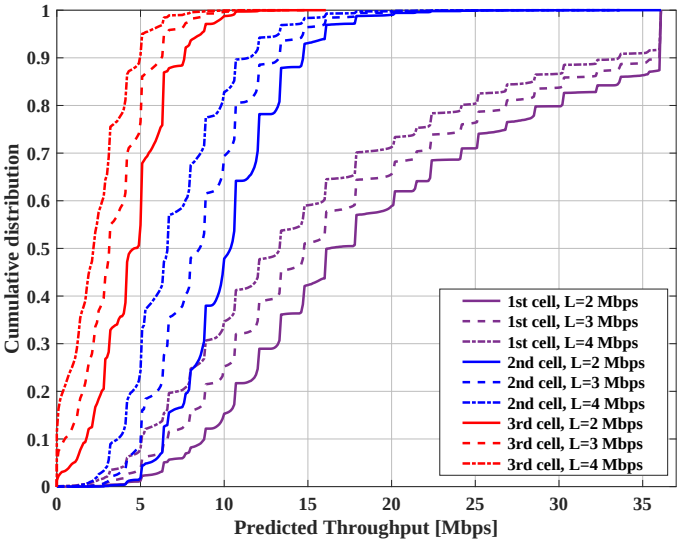

Fig. E.7: User predicted throughput of the cells in the measurement set for different load, W=10 dB, BW=10 MHz.

system-level simulator. Results show that the proposed centralized multi-cell scheduling solutions provide significant latency performance gains of up to 60% over traditional distributed solutions. We showed that the major improvement of URLLC latency is achieved for the case with the UE CSI measurement size of $Q = 2$ cells within a power window of $W \in [5\ 10]$ dB. The results also illustrates that segmentation can reduce the queued data and bring significant URLLC latency improvement for both centralized and distributed scheduling. Finally, the importance of channel-delay aware scheduling for URLLC is shown. Future studies could examine the performance of the optimal solution, investigate more advanced interference coordination and multi-cell scheduling techniques for URLLC.

References

[1] 3GPP Technical Specification 38.300, "NR and NG-RAN overall description; stage-2," Version 15.5.0, March 2019.

[2] 3GPP Technical Report 38.913, "Study on scenarios and requirements for next generation access technologies," Version 14.1.0, March 2017.

[3] 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.

[4] 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, December 2014.

[5] 3GPP Technical Specification 23.501, "Technical specification group services and system aspects, system architecture for the 5G system," Release 15, December 2017.

[6] P. Popovski, J. J. Nielsen, C. Stefanovic, E. d. Carvalho, E. Strom, K. F. Trillingsgaard, A. Bana, D. M. Kim, R. Kotaba, J. Park, and R. B. Sorensen, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Network*, vol. 32, no. 2, pp. 16–23, March 2018.

[7] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner, and Z. Li, "Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements," *IEEE Network*, vol. 32, no. 2, pp. 8–15, March 2018.

[8] G. Pocovi, B. Soret, K. I. Pedersen, and P. Mogensen, "MAC layer enhancements for ultra-reliable low-latency communications in cellular networks," in *2017 IEEE ICC Workshops*, May 2017, pp. 1005–1010.

[9] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji, and R. Jantti, "Link adaptation design for ultra-reliable communications," in *2016 IEEE International Conference on Communications*, May 2016, pp. 1–5.

[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," in *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, September 2016, pp. 1–6.

[11] K. I. Pedersen, S. R. Khosravirad, G. Berardinelli, and F. Frederiksen, "Rethink hybrid automatic repeat request design for 5G: Five configurable enhancements," *IEEE Wireless Communications Magazine*, vol. 24, no. 6, pp. 154–160, December 2017.

[12] G. Berardinelli, S. R. Khosravirad, K. I. Pedersen, F. Frederiksen, and P. Mogensen, "On the benefits of early HARQ feedback with non-ideal prediction in 5G networks," in *2016 International Symposium on Wireless Communication Systems (ISWCS)*, September 2016, pp. 11–15.

[13] Q. Liao, P. Baracca, D. Lopez-Perez, and L. G. Giordano, "Resource scheduling for mixed traffic types with scalable TTI in dynamic TDD

systems," in *2016 IEEE Globecom Workshops (GC Wkshps)*, December 2016, pp. 1–7.

[14] K. I. Pedersen, M. Niparko, J. Steiner, J. Oszmianski, L. Mudolo, and S. R. Khosravirad, "System level analysis of dynamic user-centric scheduling for a flexible 5G design," in *2016 IEEE Global Communications Conference (GLOBECOM)*, December 2016, pp. 1–6.

[15] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 210–217, March 2018.

[16] C. F. Liu and M. Bennis, "Ultra-reliable and low-latency vehicular transmission: An extreme value theory approach," *IEEE Communications Letters*, vol. 22, no. 6, pp. 1292–1295, June 2018.

[17] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28 912–28 922, 2018.

[18] E. Khorov, A. Krasilov, and A. Malyshev, "Radio resource and traffic management for ultra-reliable low latency communications," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.

[19] V. Fernández-López, K. I. Pedersen, B. Soret, J. Steiner, and P. Mogensen, "Improving dense network performance through centralized scheduling and interference coordination," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4371–4382, May 2017.

[20] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen, "Centralized joint cell selection and scheduling for improved URLLC performance," *In Proc. 29th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, September, 2018.

[21] J. Ghimire and C. Rosenberg, "Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1340–1351, March 2013.

[22] H. Zhou, S. Mao, and P. Agrawal, "Approximation algorithms for cell association and scheduling in femtocell networks," *IEEE Transactions on Emerging Topics in Computing*, vol. 3, no. 3, pp. 432–443, September 2015.

[23] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schober, and L. Hanzo, "User association in 5G networks: A survey and an outlook," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1018–1044, Second-quarter 2016.

[24] Z. Qi, T. Peng, and W. Wang, "Distributed resource scheduling based on potential game in dense cellular network," *IEEE Access*, vol. 6, pp. 9875–9886, 2018.

[25] T. Pfeiffer, "Next generation mobile fronthaul and midhaul architectures [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 7, no. 11, pp. B38–B45, November 2015.

[26] R. Agrawal, A. Bedekar, R. Gupta, S. Kalyanasundaram, H. Kroener, and B. Natarajan, "Dynamic point selection for lte-advanced: Algorithms and performance," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2014, pp. 1392–1397.

[27] 3GPP Technical Report 38.802, "Study on new radio access technology physical layer aspects," Version 14.0.0, March 2017.

[28] 3GPP Technical Report 38.804, "Study on new radio access technology; radio interface protocol aspects," Version 14.0.0, March 2017.

[29] 3GPP Technical Specification 38.401, "Technical specification group radio access network; NG-RAN; architecture description," Version 15.1.0, March 2018.

[30] 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.

[31] 3GPP Technical Specification 36.213, "Physical layer procedures; Evolved universal terrestrial radio access," Version 15.1.0, March 2018.

[32] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–5.

[33] A. Anand, G. D. Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, April 2018, pp. 1970–1978.

[34] D. Liu, Y. Chen, K. K. Chai, T. Zhang, and M. Elkashlan, "Opportunistic user association for multi-service HetNets using Nash bargaining solution," *IEEE Communications Letters*, vol. 18, no. 3, pp. 463–466, March 2014.

[35] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38 451–38 463, 2018.

[36] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 39, no. 2, pp. 150–154, February 2001.

[37] 3GPP Technical Report 36.814, "Further advancements for E-UTRA physical layer aspects," Version 9.2.0, March 2017.

[38] S. N. Donthi and N. B. Mehta, "An accurate model for EESM and its application to analysis of CQI feedback schemes and scheduling in LTE," *IEEE Transactions on Wireless Communications*, vol. 10, no. 10, pp. 3436–3448, October 2011.

[39] V. Fernandez-Lopez, B. Soret, and K. I. Pedersen, "Joint cell assignment and scheduling for centralized baseband architectures," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–5.

Paper F

Low-Complexity Centralized Multi-Cell Radio Resource Allocation for 5G URLLC

Ali Karimi, Klaus I. Pedersen, and Preben Mogensen

The paper has been submitted to the
IEEE Wireless Communications and Networking Conference (WCNC), 2020.

Abstract

This paper addresses the problem of downlink centralized multi-cell scheduling for ultra-reliable low-latency communications in a fifth generation New Radio (5G NR) system. We propose a low-complexity centralized packet scheduling algorithm to support quality of service requirements of URLLC services. Results from advanced 5G NR system-level simulations are presented to assess the performance of the proposed solution. It is shown that the centralized architecture significantly improves the URLLC latency. The proposed algorithm achieves gains of 99% and 90% URLLC latency reduction in comparison to distributed scheduling and spectral efficient dynamic point selection.

1 Introduction

Supporting ultra-reliable low-latency communications (URLLC) is one of the major goals of the fifth generation New Radio (5G NR) systems [1, 2]. The third generation partnership project (3GPP) has defined several quality of service (QoS) requirement levels for URLLC. The most extreme one has a short latency budget of one millisecond (msec) for transmission of small payloads with 99.999% reliability [3].

Recently, lots of attention has been focused on developing the theoretical bases and engineering-related protocols for URLLC. In [4] and [5], the authors investigate the principles of finite-block length (FBL) wireless data transmission. Motivated by the results from FBL communications, multi-user resource allocation and optimization of data and metadata transmission for URLLC are studied in [6] and [7], respectively. The research in [8] outlines several challenges that 5G NR faces to enable URLLC. Especially, extreme latency requirements of URLLC services need network design to support low processing/transmission times, short hybrid automatic repeat-request (HARQ) delay, and also to accomplish with unprecedented queuing delay [9, 10]. The use of flexible numerology with dynamic transmission times and user-centric adjustment of control information are investigated in [11]. The studies in [12] and [13] present performance analysis of HARQ retransmission for URLLC.

Several studies have found that URLLC performance is overshadowed even with temporary queuing delays [14, 15]. This is further exacerbated with growing traffic. The authors in [16] present a method to reduce the network traffic by dropping delayed payloads. Dynamic link adaptation is exploited in [14] to effectively allocate the resources. An efficient packet scheduling algorithm is proposed in [17] to reduce the tail of queuing delay for URLLC.

In this paper, we investigate the potentials of centralized radio access network (C-RAN) to reduce the queuing delay and enhance URLLC perfor-

mance in the downlink (DL). Motivated by previous study in [15] we discuss C-RAN architecture with multi-cell multi-user scheduling, where each user can be scheduled from a cluster of connected cells. An optimization problem is formulated to maximize the network capacity of supporting URLLC payloads. As the optimal solution requires high computational complexity, a heuristic, low-complexity algorithm is proposed. In comparison to [15], this includes an improved scheduling metric and an enhanced frequency-selective multi-user resource allocation.

We present numerical results from dynamic advanced system-level simulations with high degree of realism. The results show that proposed centralized scheduling provides outstanding URLLC performance in comparison to those of distributed scheduling [17] and the spectral-efficient dynamic point selection (DPS) [18].

The paper is organized as follows: We present the system model in Section 2. The optimization problem and the proposed solution are discussed in Section 3 and 4, respectively. Simulation methodology and performance results are presented in Section 5. Finally, Section 6 concludes the paper.

2 Setting the Scene

2.1 System Model

As outlined in [15, 18], we focus on the C-RAN architecture for the DL transmission operating in frequency division duplexing (FDD) mode. A C-RAN is connected to seven radio remote heads (RRHs). Each RRH has three cells with sectorized deployment and is responsible for physical layer functionalities. The upper layer network protocols are hosted in the C-RAN.

A set of U URLLC user equipments (UEs) are randomly placed over the entire geographical area covered by the network with uniform distribution. Each UE is subject to the DL transmission of B bytes data packets that need to be successfully received within a short latency target of Υ^{tar}. Packet arrival is modelled as a Poisson point process with average of λ [payload/sec/UE]. The average offered load is thus equivalents to $L = C^{-1} \times U \times \lambda \times B \times 8$ [bps/cell], where $C = 21$ is the total number of cells.

The UEs are multiplexed on a shared channel of 20 MHz bandwidth using orthogonal frequency division multiple access (OFDMA) with 30 kHz sub-carrier spacing. The transmission time interval (TTI) equals two or four OFDM symbols ($\approx$ 0.072 and 0.143 msec). A physical resource block (PRB) of 12 sub-carriers is assumed as the minimum physical allocation unit.

2.2 Channel Measurement and Cell Connectivity

Each UE periodically performs measurements by estimating the received signal reference power (RSRP) of the cells. The UE connects to one/multiple serving cell(s) as follows:

Baseline (Distributed scheduling)

The UE connects to only the cell with the highest average RSRP value. Channel state information (CSI) measurement is performed for the connected cell and the corresponding post-receiver channel quality indicator (CQI) is reported to the network.

DPS

As outlined in [18], the UE connects to a cluster of maximum Q cells that are within a RSRP power window of W dB from the cell with the highest RSRP value. The UE periodically measures the CSIs of the connected cells and reports the CQI for the cell with the highest instantaneous estimated throughput (TP).

Centralized Scheduling

The UE performs CSI measurements for the cells in its cluster. The CQIs are reported to the network.

2.3 URLLC Latency Components

Focussing on the one-way latency (Y) in the DL transmission, the latency components are given by:

$$Y = d_{fa,q} + d_{bsp} + d_{tx} + d_{uep} + d_{HARQ}, \tag{F.1}$$

where $d_{fa,q}$ denotes the frame alignment and queuing delay of the initial transmission. The processing times required for scheduling the payload at the network and that for decoding the data at the UE are presented by d_{bsp} and d_{uep}, respectively. The transmission time is denoted by d_{tx}. The imposed delay by HARQ retransmission(s) is denoted by d_{HARQ}. We set $d_{HARQ} = 0$ if the packet is received correctly within the first transmission. In case of failure, a minimum HARQ delay equal to 12 OFDM symbols is assumed [14]. The processing times d_{bsp} and d_{uep} are constant, equal to $d_{bsp} = 2.75$ and $d_{uep} = 3.25$ OFDM symbols, respectively [19].

3 Problem Formulation

Let us assume D^t and $\mathbf{\Gamma}_u$ represent the total number of PRBs and the set of cells in the measurement cluster of UE u (those the UE can be scheduled from). The set of assigned PRBs to UE u from cell c ($c \in \mathbf{\Gamma}_u$) is denoted by $\mathbf{P}_{uc}$ and the corresponding achievable data rate is shown by $R_{\mathbf{P}_{uc}}$. For each TTI, we formulate an optimization problem for cell and PRB allocation as follows:

$$
\begin{aligned}
&\max_{x_{uc}, p_u^j} \sum_u \sum_c a_u x_{uc} R_{\mathbf{P}_{uc}} \\
&\text{Sub. to C1: } \sum_c x_{uc} \leq 1, \qquad \forall u, \\
&\qquad \text{C2: } \sum_j p_u^j \geq x_{uc} p_{uc}^{\min}, \quad \forall u, c, \\
&\qquad \text{C3: } R_{\mathbf{P}_{uc}} \leqslant Q_u, \qquad \forall u, \\
&\qquad \text{C4: } \sum_u x_{uc} p_u^j \leq 1, \quad \forall c, j, \\
&\qquad \text{C5: } x_{uc}, p_u^j \in \{0, 1\}, \ \ \forall u, c, j,
\end{aligned} \tag{F.2}
$$

where x_{uc} is a binary variable accounting for the association of UE u to cell c. Binary variable p_u^j is the PRB allocation indicator equals one if j-th PRB ($1 \leq j \leq D^t$) is assigned to UE u, otherwise it is zero. Parameters a_u is a weighting factor responsible for guaranteeing latency requirements of UE u. $p_{uc}^{\min}$ is the number of required PRBs to transmit the control information if UE u is scheduled by cell c. The buffered data of u-th UE is denoted by Q_u.

The first constraint in (F.2) indicates that each UE can be scheduled from maximum one cell per TTI. The second constraint ensures the minimum allocated resources for sending the scheduling grants. Constraint C3 imposes that the achievable rate of each UE is bounded by that of the buffered data. Finally, the orthogonality of resource allocation is ensured by C4. The problem (F.2) is a non-convex combinatorial optimization that is difficult to solve in polynomial time and within the stringent processing time requirement of URLLC services [15, 17]. Assuming each UE is connected to Q cells, the cell selection has the complexity of $\mathcal{O}((Q+1)^U)$. The complexity further increases by considering PRB allocation in frequency domain.

4 Proposed Solution

We present a heuristic low-complexity solution for (F.2). The following describes the steps of the proposed solution.

In line with [17], we first prioritize scheduling HARQ retransmissions. The C-RAN schedules HARQ retransmissions immediately from the cells with the best link quality (highest TP) and over a set of PRBs with the highest CQI values, aiming to avoid additional queuing delay and minimize the need for more retransmission(s).

After allocating HARQ retransmissions, UEs with pending data are scheduled. We split the problem (F.2) into two sub-problems of cell and PRB allocation. First, the C-RAN determine the UEs and cells allocations. Afterwards, frequency domain (FD) PRB allocation is performed for the UEs allocated to the same cell.

For a given UE, the achievable data rate depends on CSI and the assigned PRBs. Dealing with small URLLC payloads, the overhead of transmitting control information is not negligible and can not be ignored [14, 20]. To manage the cost of sending multiple physical downlink control channels (PDCCHs), we aim for scheduling full data payloads in one transport block and not to split them over multiple TTIs (aka as packet segmentation). As detailed in [14], the PDCCH size is adjusted based on the wide-band CQI to ensure low-failure probability. We define pair(s) of each UE with pending data and the cell(s) in its measurement set. For a given pair (u, c) of UE u and cell c, we denote P_{uc} as the estimated number of required PRBs for sending both the PDCCH and data. The length of the data block can be estimated from the wide-band CQI.

The UE/cell allocation is performed sequentially by selecting a UE/cell pair with the highest scheduling metric. For a pair (u, c), the scheduling metric m_{uc} is defined as:

$$m_{uc} = \Upsilon_u + \frac{TP_{uc}}{TP_{\mathbf{\Gamma}u}}, \tag{F.3}$$

where Υ_u [TTI] is the head-of-line delay for UE u. We denote TP_{uc} as the estimated full-bandwidth TP of UE u if served by cell c. Variable $TP_{\mathbf{\Gamma}_u}$ is the sum TPs of all serving cells in the measurement set of UE u that equals $TP_{\mathbf{\Gamma}_u} = \sum_{c \in \mathbf{\Gamma}_u} TP_{uc}$. The role of Υ_u in (F.3) is to minimize the outage latency by prioritizing UEs which are closer to the latency target. The term TP_{uc} exploits channel aware scheduling. For a given UE, a cell with the best CSI (primary cell) receives higher priority for being selected. Finally, the term $TP_{\mathbf{\Gamma}_u}$ acts as normalizing factor. Cell-centred UEs (which usually connect to one cell) will get a higher chance for being scheduled from the primary cells. On the other hand, cell-edge UEs (that usually have multiple connections with similar signal strengths) are assigned relatively lower values as they have a chance for being served by the secondary cells (if the primary cells are overloaded).

The C-RAN sequentially selects a pair (u, c) with the highest scheduling metric m_{uc} and assign the UE to the cell if the number of available PRBs at

cell c (D_c) is enough for scheduling a full payload (i.e. $D_c \geq P_{uc}$). Afterwards the C-RAN removes the other pairs associated with the selected UE and update the number of available PRBs at the serving cell. The procedure continues until all UEs are scheduled or there are not enough resources at the serving cells. The approximated complexity of proposed cell allocation is $\mathcal{O}(Q \cdot U \log(Q \cdot U))$.

For each cell, the selected UEs are multiplexed in FD by comparing per-PRB metric. Throughput to average (TTA) [17] is adopted. j-th PRB is assigned to UE $\hat{u}$ (i.e. $p^j_{\hat{u}} = 1$) that has the highest metric of

$$\hat{u} = \arg\max_{u \in \Pi_c} \frac{r^j_u}{\bar{\mathbf{r}}_u}, \tag{F.4}$$

where Π_c denotes the set of UEs allocated to cell c. Variables r^j_u and $\bar{\mathbf{r}}_u$ are the u-th UE's achievable TP of PRB j and the instantaneous full-bandwidth TP, respectively.

Finally, the C-RAN checks if it is possible to schedule more UEs on the remaining PRBs after FD allocation. If there are not sufficient PRBs to schedule a full payload, at most one payload is segmented and transmitted over the available PRBs. A UE with the lowest PDCCH overhead is prioritized. Algorithm 1 summarizes details of the proposed solution.

5 Numerical Evaluations

Performance results are obtained by running advanced 5G NR system-level simulations in line with the 3GPP NR guidelines as outlined in Section 2 [1, 23]. The summarized assumptions are presented in Table F.1. The simulation time is set so at least five million packets to obtain reliable results. We compare the proposed C-RAN scheduling against the distributed allocation [17] and DPS [18]. In line with [18], for centralized scheduling and DPS we set $Q = 2$ cells and $W = 3$ dB. Applying these settings, only 32% of UEs have two cells in their measurement set while 68% of them connect to only one cell [18]. This further reduces the complexity of cell allocation at the C-RAN as well as the overhead and the complexity of channel measurement at the user-side.

5.1 One millisecond latency performance

First we focus on achieving one msec latency at the outage probability of 10^{-5}, assuming $B = 32$ bytes and a short TTI size of two OFDM symbols. The complementary cumulative distribution function (CCDF) of the latency is depicted in Fig. F.1. At low-offered loads, one msec latency at 10^{-5} outage

Algorithm 1 Proposed centralized scheduling.

1: Schedule the HARQ retransmissions. Update the available PRBs at the serving cells.
2: Create pairs of the buffered UEs and the cells in their measurement sets.
3: For each pair, calculate P_{uc} and m_{uc}.
4: Create list **S** of the pairs, sort **S** in decreasing order of the scheduling metric (F.3).
5: **while** enough PRBs at cells and **S** is not empty **do**
6: Select pair (u, c) as the first element of **S**.
7: **if** $P_{uc} \leq D_c$ **then**
8: Assign UE u to cell c.
9: Update UE $D_c = D_c - P_{uc}$.
10: Remove pairs associated with UE u from **S**.
11: **else**
12: Remove pair (u, c) from **S**, put it in the unscheduled list $\overline{\mathbf{S}}$
13: **end if**
14: **end while**
15: **for** $c = 1 : C$ **do**
16: Calculate the FD scheduling metric (F.4) for the available PRBs and the users assigned to cell c.
17: Sort PRBs in decreasing order of (F.4).
18: Allocate PRBs to UEs. Remove if there is segmented payload.
19: **if** there are unallocated PRBs **then**
20: Schedule a full payload from $\overline{\mathbf{S}}$ or segment a UE with low PDCCH usage if PRBs are not enough.
21: **end if**
22: **end for**

is fulfilled for all the scheduling methods. The reason is that the probability of experiencing queuing delay is very low. Therefore, the performance is mainly impacted by the transmission time, processing delays, and HARQ delay.

The probability of the queuing increases with the offered load. The centralized scheduling shows clear latency benefits by instantly offloading the congested cells. The results in Fig. F.1 show that one msec latency can not be achieved for the distributed and DPS if the traffic exceeds 10 Mbps, whereas the C-RAN can tolerate up to 12 Mbps load while fulfilling URLLC requirements. This is equivalent to 20% improvement in the network capacity.

Fig. F.2 illustrates the CCDF of the queuing and frame alignment. As expected, the queuing delay is negligible at low-load so that most of the packets are immediately allocated without waiting for receiving resources. We observe a longer tail of queuing when increasing the offered traffic, in-

Table F.1: Default simulation assumptions.

Description	Assumption
Environment	3GPP Urban Macro (UMa); 3-sector RRHs with 500 meters inter-site distance. 21 cells.
Propagation	Urban Macro-3D
Carrier	2 GHz (FDD), 20 MHz carrier bandwidth
PHY numerology	30 kHz sub-carrier spacing. PRB size of 12 sub-carriers (360 kHz).
TTI sizes	0.072 and 0.143 msec corresponding to 2 and 4 OFDM symbols mini-slot, respectively.
MIMO	Single-user 2x2 closed loop single-stream (Rank-1) configuration. LMMSE-IRC receiver.
CSI	Periodic CSI measurement every 5 msec, with 2 msec latency for being applied in DL transmission. One CQI per a sub-channel of eight PRBs.
MCS	QPSK to 64QAM, with same encoding rates as specified for LTE.
Link adaptation	Dynamic MCS with outer-loop link adaptation to achieve 1% BLER of initial transmission.
HARQ	Asynchronous HARQ with Chase-combining [21]. HARQ delay equals 12 OFDM symbols.
User distribution	2100 URLLC users (Average 100 users per cell).
Traffic model	FTP3 downlink traffic with payload sizes of $B = 32$ and $B = 50$ bytes.
Link-to-system (L2S) mapping	Based on MMIB mapping [22].

dicating that some of cells become temporarily overloaded. For 12 Mbps, we observe that in 99.999% of the cases the queuing delay is below 0.3 msec for the C-RAN. Whereas, this increases to 1.4 and 2.5 msec for DPS and the distributed scheduling, respectively.

5.2 Beyond one millisecond latency

For cases with larger payload sizes, higher offered loads, and more relaxed latency requirements (e.g. two or five msec), we configure the TTI size to four OFDM symbols. So that a full payload of UEs with low CSIs can fit into one TTI without segmentation. Fig. F.3 and Fig. F.4 plot the CCDFs of the latency and the queuing plus frame alignment delay for $B = 50$ bytes. In Fig. F.3, the outage reliability of 10^{-5} is achieved at 1.25 msec for distributed scheduling, whereas it is 2.8 msec in Fig. F.1. That is because for the same volume of traffic, the packet arrival rate for $B = 50$ bytes case is 36% less than that of $B = 32$ bytes. This leads to lower overhead of sending PDCCHs as well as

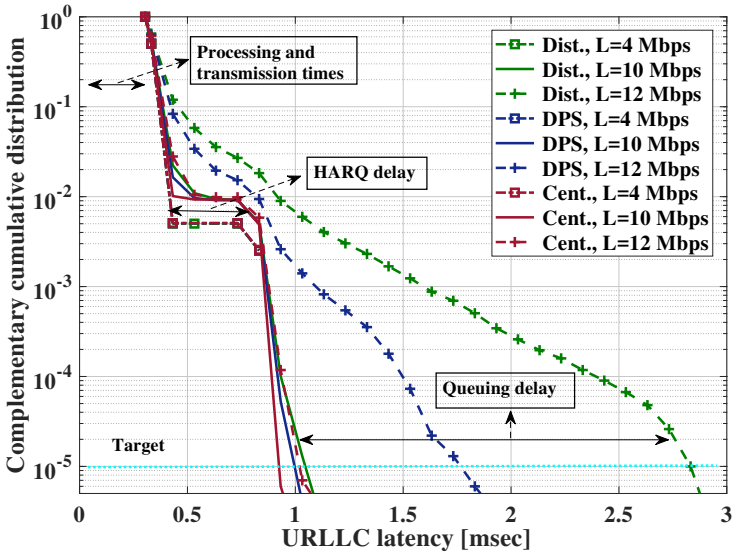

Fig. F.1: URLLC latency for different scheduling methods assuming $B = 32$ and two OFDM symbols TTI size.

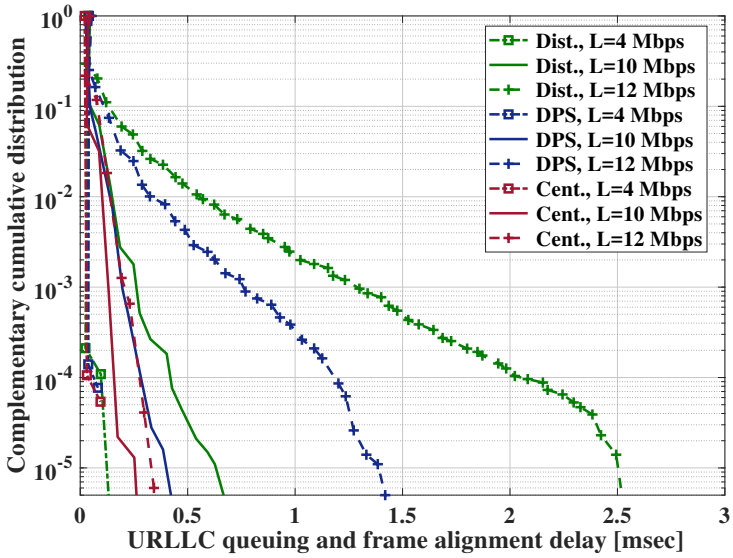

Fig. F.2: Queuing and frame alignment delay for different scheduling methods assuming $B = 32$ and two OFDM symbols TTI size.

the queuing delay and overall latency.

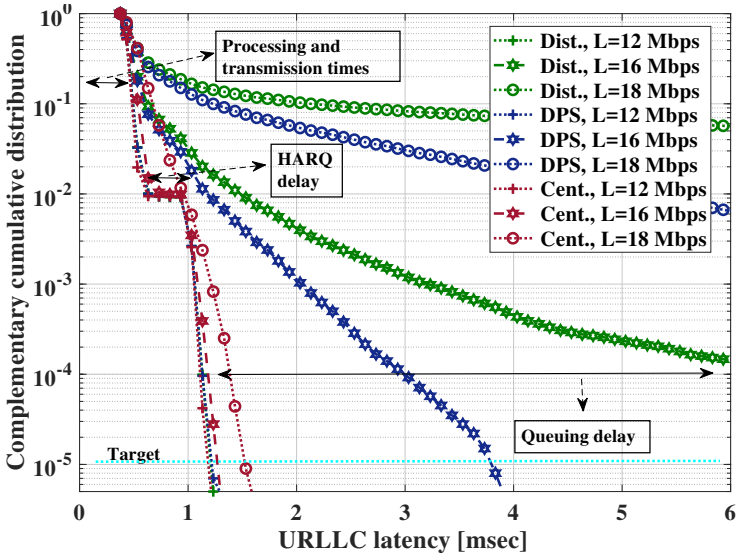

Fig. F.3: URLLC latency for different scheduling methods assuming $B = 50$ and four OFDM symbols TTI size.

In comparison to the distributed case, the DPS provides better channel quality specially for the cell-edge UEs. This results in lower latency. For 18 Mbps load, the latency at the outage level of 10^{-5} is 45 msec for the distributed case. It is reduced to 16 msec by applying DPS. We observe that the C-RAN further boosts the performance by reducing the latency to only 1.5 msec. The proposed centralized solution not only benefits from fast-fading channel variations but also exploits the gain of fast load balancing. The superior performance of the C-RAN in improving the queuing delay is highlighted in Fig. F.4. At 18 Mbps offered traffic, 20% and 15% of the payloads for the distributed and DPS scheduling experience more than 0.5 msec queuing delay. As compared to then, only 0.5% of the packets experience queuing delay when applying the proposed centralized algorithm.

Table F.2 summarizes the URLLC latency at the outage level of 10^{-5} for different resource allocation methods. The results show the advantage of the C-RAN at high offered traffic. As an example, for 20 Mbps load gains of 99% and 88% latency reduction is achieved in comparison to the distributed scheduling and DPS, respectively.

Table F.2: URLLC performance at the outage of 10^{-5} for different numerologies and offered loads.

Scenario	TTI = 2 symbols, B = 32 bytes Load [Mbps/cell]				TTI = 4 symbols, B = 50 bytes Load [Mbps/cell]			
	4	8	10	12	12	16	18	20
Distributed [msec]	0.83	0.92	1.05	2.8	1.23	7.8	45	500
DPS [msec]	0.83	0.92	1	1.75	1.23	3.83	16.4	43
Centralized [msec]	0.83	0.89	0.91	1	1.13	1.23	1.53	4.95
Relative gain to Distributed	0%	3%	13%	64%	8%	84%	97%	99%
Relative gain to DPS	0%	3%	9%	42%	8%	68%	90%	88%

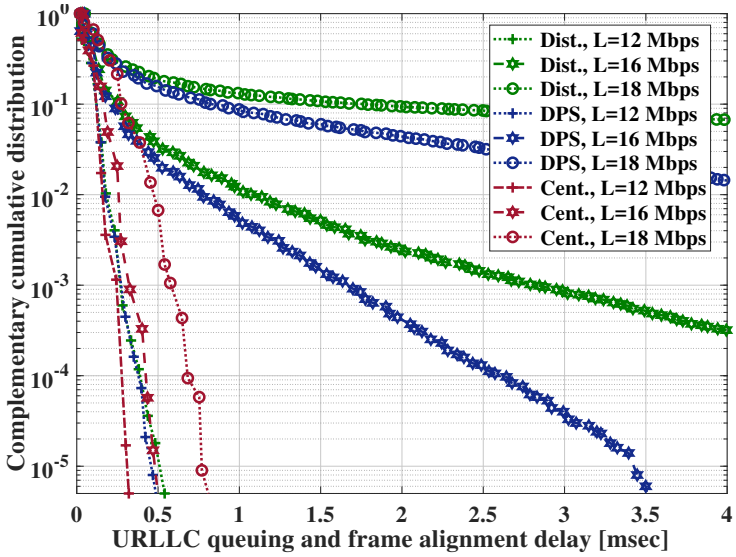

Fig. F.4: Queuing and frame alignment delay for different scheduling methods assuming $B = 50$ and four OFDM symbols TTI size.

6 Conclusion and Future Studies

This paper studied DL URLLC performance optimization through the use of C-RAN. We proposed an attractive low-complexity multi-cell scheduling algorithm to improve the network performance of supporting URLLC in 5G NR. Extensive numerical results from 3GPP compliant advanced system-level simulator were presented. The results confirm the high potential of the C-RAN to tackle the undesired queuing delay of URLLC payloads. Performance results show that the proposed solution can achieve remarkable gains of up to 99% and 90% URLLC latency improvements in comparison to the distributed scheduling and DPS, respectively.

As future work, it is of high interest to incorporate dynamic inter-cell interference coordination techniques to further boost URLLC capacity.

References

[1] 3GPP Technical Specification 38.300, "NR and NG-RAN overall description; stage-2," Version 15.5.0, March 2019.

[2] 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.

[3] 3GPP Technical Specification 23.501, "Technical specification group services and system aspects, system architecture for the 5G system," Release 15, December 2017.

[4] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, July 2014.

[5] P. Popovski, . Stefanović, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Transactions on Communications*, pp. 1–1, 2019.

[6] W. R. Ghanem, V. Jamali, Y. Sun, and R. Schober, "Resource allocation for multi-user downlink URLLC-OFDMA systems," *CoRR*, vol. abs/1901.05825, 2019. [Online]. Available: http://arxiv.org/abs/1901.05825

[7] A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Berardinelli, and P. Mogensen, "On the multiplexing of data and metadata for ultra reliable low latency communications in 5G," *Submitted to IEEE Transactions on Vehicular Technology*, 2019.

[8] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail, risk and scale," *CoRR*, 2018. [Online]. Available: http://arxiv.org/abs/1801.01270

[9] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 127–141, January 2018.

[10] A. Anand and G. de Veciana, "Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2411–2421, November 2018.

[11] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, 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.

[12] J. P. B. Nadas, O. Onireti, R. D. Souza, H. Alves, G. Brante, and M. A. Imran, "Performance analysis of hybrid ARQ for ultra-reliable low latency communications," *IEEE Sensors Journal*, pp. 1–1, 2019.

[13] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2475–2485, November 2018.

[14] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28 912–28 922, 2018.

[15] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen, "5G centralized multi-cell scheduling for URLLC: Algorithms and system-level performance," *IEEE Access*, vol. 6, pp. 72 253–72 262, 2018.

[16] E. Khorov, A. Krasilov, and A. Malyshev, "Radio resource and traffic management for ultra-reliable low latency communications," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.

[17] A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Pocovi, and P. Mogensen, "Efficient low-complexity packet scheduling algorithm for mixed URLLC and eMBB traffic in 5G," *In Proc. 2019 IEEE 89th Vehicular Technology Conference - VTC2019-Spring*, May, 2019.

[18] A. Karimi, K. I. Pedersen, and P. Mogensen, "5G URLLC performance analysis of dynamic-point selection multi-user resource allocation," *in Proc. 2019 International Symposium on Wireless Communication Systems (ISWCS)*, August, 2019.

[19] 3GPP Technical Documents R1-1813120, "Discussion on the RAN2 LS on TSN requirements evaluation," November 2018.

[20] N. H. Mahmood, A. Karimi, G. Berardinelli, K. I. Pedersen, and D. Laselva, "On the resource utilization of multi-connectivity transmission for URLLC services in 5G new radio," *in Proc. 2019 IEEE Wireless Communications and Networking Conference (WCNC) Workshops*, April 2019.

[21] 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.

[22] T. L. Jensen, S. Kant, J. Wehinger, and B. H. Fleury, "Fast link adaptation for MIMO OFDM," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 8, pp. 3766–3778, October 2010.

[23] 3GPP Technical Report 38.913, "Study on scenarios and requirements for next generation access technologies," Version 14.1.0, March 2017.

Part IV

Data and Control Channel Scheduling, PDCP Packet Duplication

Data and Control Channel Scheduling, PDCP Packet Duplication

This chapter presents a theoretical framework to design and analyse downlink (DL) radio resource allocation for Ultra-Reliable Low-Latency Communications (URLLC).

1 Problem Description

In the previous part, we studied practical system design solutions for URLLC through the 3GPP proposed framework for the fifth generation New Radio (5G NR). In this part, we take one step forward and focus on the theoretical bases for URLLC. The objective is to achieve an in-depth understanding of different components and transmission aspects associated with URLLC. Therefore, we evaluate current system performance and obtain new insights for alternative transmission design and network optimization.

Successful scheduled transmission in the DL direction depends on correctly decoding of both data and control channel (also called metadata). Metadata has information about user ID, MCS, precoding matrix, location of the resources, etc. Since URLLC is characterized by small data payloads, the packet size is comparable to metadata. Also, both data and metadata need to be encoded with low-failure probability to fulfil reliability targets [1]. While most of the URLLC studies mainly focus on data transmission, metadata has a considerable impact on resource utilization and success probability [2, 3].

The well-known Shannon formula can not be used for system design and performance evaluation for short packet transmissions [4]. Recent studies indicate that for finite codewords, there is a loss in the achievable rate that is proportional to the square root of the blocklength [5, 6]. For URLLC resource optimization, it is essential to take into account the impact of finite blocklength codewords and consider both data and metadata.

In this part, we first address such issues and study the problem of DL resource allocation for URLLC. Two multiplexing schemes based separate coding and joint encoding of data and metadata are investigated. The first option offers lower complexity for decoding the data block, while the second one is more resource-efficient. The success probability and resource usage of each scheme are analysed. We study a resource allocation solution to minimize the combined resources and satisfy both reliability and latency constraints. Therefore, dynamic link adaptation is investigated to optimize the consumed resources while ensuring the URLLC targets. The performance, costs, and benefits of each scheme are evaluated and discussed for URLLC.

Secondly, packet data convergence protocol (PDCP) packet duplication is addressed as one of the enablers for URLLC. Here, packet duplication is performed at the PDCP layer. The data independently goes through lower layers. So different modulation and coding scheme (MCS)s can be adopted at transmitters. The received packets are decoded individually at the mobile terminal and sent to the upper layers. The PDCP layer forwards the first correctly received packet to the upper layer and subsequently discards any later copies. Data duplication is one of the potential candidates to enhance reliability [7, 8]. However, the benefits and costs of this scheme need to be further understood for practical applications. Resource usage and reliability enhancement of this feature need to be calculated and compared with single-node connectivity.

2 Objectives

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

- Study the performance of different data and metadata multiplexing schemes.
- Design and optimize link adaptation for joint data and metadata transmission.
- Study the fundamentals of PDCP packet duplication for URLLC.

3 Included Articles

This part includes the following articles

Paper G. On the Multiplexing of Data and Metadata for Ultra-Reliable Low-Latency Communications in 5G

This paper presents an in-depth theoretical study of URLLC link adaptation for downlink transmission. Two recent multiplexing approaches for transmitting data and metadata are investigated. Those include *in-resource control signalling* and *joint encoding of data and metadata*. The paper evaluates the outage reliability and resource usage for each scheme. Throughout the analyses, a comprehensive transmission scenario is assumed that includes modelling the required resources and failure probabilities for transmitting data and metadata. Moreover, probability of error in uplink feedback channel (e.g., NACK to ACK, DTX to ACK, etc.) are taken into account to enhance the accuracy of results.

Optimization problems are formulated for joint link adaptation and resource allocation to minimize the consumed resources while ensuring fulfilments of the URLLC QoS requirements. It is shown that the problems are combinatorial non-convex optimizations difficult to solve in polynomial time. Low-complexity algorithms based on successive convex optimization are proposed. Performance of the proposed solutions is analyzed via extensive numerical simulations. Sensitivity analysis is conducted to determine the performance for various transmission parameters, including different data and metadata size, SNR levels, and probability of uplink errors.

Paper H. On the Resource Utilization of Multi-Connectivity Transmission for URLLC Services in 5G New Radio

This paper studies PDCP packet duplication for URLLC. An analytical survey of reliability enhancement of data duplication is presented. The success probability and resource usage of single-node transmission and data duplication are computed. To meet URLLC targets, 5G NR mini-slot configuration is adopted. This provides a sufficient time budget for initial transmission and one retransmission if the first one fails. In case of failure to decode the data, Chase combing is assumed to boost the received signal power after HARQ. The consumed resources and corresponding errors of sending both data and metadata are taken into account. The analytical performance results of PDCP duplication are compared and discussed against single-node transmission.

4 Main Findings and Recommendations

Joint link adaptation for data and control information

As shown in Paper G, metadata has a significant impact on the successful transmission probability and resource efficiency. It is essential to have reliable feedback channel and include imperfections for optimum system design. Due to the tradeoff between reliability and spectrum efficiency, dynamic link adaptation is desirable to fulfil URLLC and improve the resource usage. To achieve the maximum benefits, resource allocation and link adaptation need to be conducted for jointly minimizing data and metadata transmission, which is non-convex combinatorial optimization. While benefiting from low computational complexity, the paper proposes a near-optimal solution. The provided solution is a function of data size, metadata size, SNR, HARQ combining possibility, and error in uplink channel. It achieves up to 27% of resource improvement in comparison to one shot transmission [9]. Performing joint link adaptation provides higher benefits in low-SNR regimes and for cases with relatively smaller payload sizes. Due to the possibility of sending information over longer blocklength, joint encoding has better spectrum efficiency in comparison to in-resource signalling. The gain decreases when two transmissions are allowed. The aforementioned gain comes at the cost of higher computational complexity (processing delay) at the user-end for blindly decoding the received messages, even those which are not sent to the intended user. Therefore, in-resource control signalling is recommended for URLLC.

PDCP packet duplication

Provided results in paper H indicate the data duplication offers significant reliability enhancements in order of magnitudes with the price of additional resource usage. It is shown that even for the case that the mobile terminal receives similar signal strength from transmitting cells (cell-edge users), achieving the success level of 99.999% with single connectivity is between 45% to 50% more resource-efficient [10]. Therefore, for cases that the channel and interference is well-known at both transceivers, single-connectivity shows to be a more spectrum efficient solution. In such cases, it is recommended to transmit the payloads from a single base station with more conservative MCSs (i.e., lower BLER targets).

References

[1] P. Popovski, . Stefanović, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A. Bana, “Wireless access in ultra-reliable low-latency com-

munication (URLLC)," *IEEE Transactions on Communications*, pp. 1–1, 2019.

[2] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28 912–28 922, 2018.

[3] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner, and Z. Li, "Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements," *IEEE Network*, vol. 32, no. 2, pp. 8–15, March 2018.

[4] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 72–78, June 2017.

[5] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[6] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, July 2014.

[7] N. H. Mahmood, D. Laselva, D. Palacios, M. Emara, M. C. Filippou, D. M. Kim, and I. de la Bandera, "Multi-channel access solutions for 5G New Radio," *CoRR*, vol. abs/1902.01948, 2019. [Online]. Available: http://arxiv.org/abs/1902.01948

[8] D. S. Michalopoulos and V. Pauli, "Data duplication for high reliability: A protocol-level simulation assessment," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, May 2019, pp. 1–7.

[9] A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Berardinelli, and P. Mogensen, "On the multiplexing of data and metadata for ultra reliable low latency communications in 5G," *Submitted to IEEE Transactions on Vehicular Technology*, 2019.

[10] N. H. Mahmood, A. Karimi, G. Berardinelli, K. I. Pedersen, and D. Laselva, "On the resource utilization of multi-connectivity transmission for URLLC services in 5G new radio," *in Proc. 2019 IEEE Wireless Communications and Networking Conference (WCNC) Workshops*, April 2019.

Paper G

On the Multiplexing of Data and Metadata for Ultra-Reliable Low-Latency Communications in 5G

Ali Karimi, Klaus I. Pedersen, Nurul Huda Mahmood, Gilberto Berardinelli, and Preben Mogensen

The paper has been submitted to the
IEEE Transaction on Vehicular Technology, 2019.

Abstract

This paper addresses the problem of downlink radio resource management for ultra-reliable low-latency communications (URLLC) in the fifth generation (5G) systems. To support low-latency communications, we study the performance of two multiplexing schemes namely in-resource control signalling and joint encoding of data and metadata. In the former, the metadata and data are separately encoded and the metadata is sent at the beginning of transmission time prior to the data. Thus, it benefits from a low-complexity receiver structure to decode the data block. Whereas the latter takes the advantages of transmitting a larger blocklength to enhance the reliability and improve the spectrum efficiency by jointly encoding data and metadata. Dealing with small URLLC payloads, the size and the error of sending metadata are not negligible and have a significant impact on the system performance and reliability of transmission. For each scheme, we derive expressions for the outage reliability and resource usage by taking into account the impacts of the finite blocklength payloads, the overhead and the error of sending metadata, and the probability of error in the feedback channel. We propose a novel framework for joint data and metadata link adaptation and resource allocation to minimize the number of allocated resources while ensuring the stringent URLLC quality of service requirements. An optimization problem is formulated for each scheme that is non-convex, combinatorial problem, difficult to solve in polynomial time. Solutions based on successive convex optimization are proposed. Numerical evaluations show that the proposed algorithms perform close to the optimal solution and demonstrate remarkable gains of up to **27**% *improvement in resource usage. Finally, we present sensitivity analysis of the results for various network parameters.*

1 Introduction

Recently, the third generation partnership project (3GPP) has introduced the first release of the fifth generation New Radio (5G NR) [1]. Unlike long term evolution (LTE) network, which was primarily serving mobile broadband (MBB) and machine type communication (MTC) services, 5G NR is designed to additionally support ultra-reliable low-latency communication (URLLC) [2]. As the name suggests, URLLC targets applications requiring high reliability with low-latency for emerging use cases like industrial automation, intelligent transport systems, and haptic communication [3–5]. A typical URLLC target is one-way reliability of 99.999% for a data unit of 32 bytes to be delivered within a tight budget of one millisecond [2, 6].

1.1 Brief Overview of the State of the Art on URLLC

In recent years, extensive research efforts have been made to enable URLLC in 5G NR. As highlighted in [7], the current LTE network has not been designed to support the stringent URLLC targets. As one of the main building blocks to reduce the latency, flexible frame structure and user scheduling over short transmission time intervals (TTIs) are discussed in [8]. Performance analysis of URLLC through advanced system-level simulations are investigated for the downlink (DL) and uplink (UL) transmission direction in [9] and [10], respectively. A low-complexity multiplexing solution for the coexistence of URLLC and enhanced MBB (eMBB) traffic is presented in [11]. Multi-user pre-emptive scheduling algorithms are analysed [12, 13]. The works in [14] and [15] study centralized radio access network (C-RAN) architecture and dynamic point selection to reduce the obstructive queuing delay of URLLC payloads. Reliability enhancement by means of data packet duplication is presented in [16]. To overcome the timely handshaking procedures of grant-based scheduling in UL, the studies in [17] and [18] investigate grant free and semi-grant free access protocols, respectively. UL Multi-cell reception design has been extensively addressed in [19] by comparing the achievable capacity of various receiver-combining techniques. Finally, performance evaluation of URLLC in unlicensed spectrum is presented in [20].

As URLLC mainly entails transmission of small payloads, applying the well-known Shannon's Capacity under asymmetric assumptions (which is valid when the codeword size tends to infinity) is not an appropriate approach for radio resource allocation and performance evaluation [21, 22]. Information-theoretic principles of finite blocklength (FBL) communications are studied in [23, 24]. It is shown that applying the law of large numbers for averaging channel distortions and noise is not applicable for FBL. The achievable rate is then subject to a rate penalty from Shannon's Capacity which is proportional to the square root of encoded blocklength [24, 25].

Taking into account the results from FBL communication theory, many researches have studied several radio resource managements techniques and investigated different URLLC enablers to further boost the 5G performance. Particularly, optimal power allocation and joint power and subcarrier assignment for DL multi-user networks are proposed in [26] and [27], respectively. The authors in [28] use results from multi-class queuing theory to design and analyse network performance of URLLC. The authors in [29] and [30] study the effective capacity for FBL regime and propose a bandwidth assignment policy for joint UL and DL transmission to guarantee the end-to-end (E2E) latency.

To improve the spectral efficiency and enable massive connectivities, network design through non-orthogonal multiple access (NOMA) is discussed for URLLC in [31, 32]. Performance comparison of orthogonal multiple access

(OMA) and NOMA is provided in [33]. The study in [34] presents a hybrid multiple access solution based on the machine learning techniques. The authors in [35] focus on mobile edge computing and user-server association to ensure low-latency communication. Reliability enhancements of cooperative communications are vastly investigated in [36–40]. Finally, several resource allocation techniques are surveyed in [41–43] to handle URLLC for vehicular communications.

As a well-known technique to enhance the reliability and spectral efficiency, throughput analysis of FBL hybrid automatic repeat request (HARQ) is investigated in [44]. Triggered by [44], several works have looked at various HARQ retransmission protocols for FBL communications (see e.g. [45–49]). Interesting power allocation strategies are proposed to minimize the energy consumption of Chase-Combining HARQ (CC-HARQ) for Rayleigh and Nakagami-m block-fading channels in [45] and [46], respectively. On the other hand, resource allocation for incremental redundancy HARQ (IR-HARQ) is discussed in [47, 48]. However, due to the tight latency targets of URLLC applications and the time requirement of HARQ procedures, HARQ performance is limited to usually one retransmission for URLLC [50, 51].

Successful scheduled data communication in DL is conditioned on the reliable transmission of metadata (also referred as control channel information or scheduling grants). Aiming to maximize the throughput for eMBB services, large data packets are usually scheduled with higher modulation and coding schemes (MCSs). This results in a medium block error rate (BLER) (around 10%) of the initial transmission [52]. The possible errors are then recovered by HARQ/ARQ procedures. Hence, network performance and transmission reliability mainly depend on how the data packets are scheduled. However, the situation is different for URLLC cases as the metadata become more dominant, and transmissions are subject to stricter outage constraints [51, 53]. While most of the existing research has focused on data scheduling (see e.g., [26- 48] and [54, 55]), the impacts of metadata overhead, the corresponding error probability, and probability of erroneous decoding feedback signals are not negligible and may not be ignored for URLLC. The study in [51] discusses the bounds and the reliability tradeoffs between the transmission of data and metadata. It is shown in [56] that low-error transmission of metadata is essential to support URLLC. As enhancing the reliability usually comes at the cost of additional resource usage, a new transmission design is required to efficiently manage the resource allocation for both data and metadata while ensuring the fulfilments of URLLC requirements.

1.2 Main Contributions

Motivated by the state-of-the-art URLLC studies, this paper provides a thorough analysis of URLLC with non-ideal control channel transmission. To

support low-latency communication, two recent multiplexing schemes namely *in-resource control signalling* [8] and *joint encoding of data and metadata* [53] are studied. The main idea of the former is to separately allocate the control information at the beginning of the DL sub-frame. Thereby, benefiting from low-computational complexity for decoding data as well as reduced processing time. The latter achieves the enhanced reliability and spectral efficiency gains of transmitting larger blocklength by jointly encoding data and control channel in a single codeword.

For each scheme, we derive expressions for the outage probability and resource usage. The impacts of overhead and errors of sending metadata are explicitly considered. We concentrate on reducing the resource usage by proposing a new radio resource allocation technique based on joint data and metadata link adaptation for URLLC services. An optimization problem is formulated for each scheme to minimize the number of allocated resources while satisfying URLLC requirements. To the best of our knowledge, the problems have not been investigated in the existing literature. The formulated problems are mixed non-convex combinatorial optimizations, which are difficult to solve in polynomial time. Solutions based on successive convex optimization are therefore proposed. Numerical results indicate that the proposed algorithms significantly improve resource efficiency and achieve near-optimal performance. Finally, we provide extensive results and discussions on the impact of various network parameters on the solution's performance.

The rest of this paper is organized as follows. The system model and the basic assumptions are presented in Section 2. In Section 3 and 4, we discuss the problem formulation and present the solution, respectively. Numerical results are provided in Section 5. Finally, Section 6 concludes the paper. Table G.1 includes a list of the main symbols used in this paper.

2 System Model and Basic Transmission Assumptions

We focus on DL performance assuming orthogonal frequency division multiple access (OFDMA) transmission in which a base station (BS) serves user-equipments (UEs) with packets of D bytes. URLLC requires reliable transmission within a time budget in the order of millisecond(s) and very low outage probability target of P_{out}^{tar}. To reduce the transmission time and achieve the extreme latency requirement of one millisecond, we adopt 5G NR flexible numerology with the capability of mini-slot scheduling. Depending on the payload size and service requirements, the TTI varies between $1-14$ OFDM symbols and the sub-carrier spacing can be configured from 15 kHz up to 240 kHz [57]. Assuming a mini-slot length of two to four OFDM symbols with $15-30$ kHz sub-carrier spacing and by taking into account the packet trans-

Table G.1: List of Symbols

Symbol	Definition
General Symbols	
M	Metadata size [byte]
D	Data size [byte]
γ	SNR
P_{out}^{tar}	Outage reliability target
P_e^{am}	Probability of decoding ACK as DTX
P_e^{ma}	Probability of decoding DTX as ACK
P_e^{na}	Probability of decoding NACK as ACK
P_e^{an}	Probability of decoding ACK as NACK
$\mathbb{R}$	Real numbers
$\mathbb{R}^+$	Positive real numbers
$\mathbb{N}$	Positive integer numbers
In-resource control signalling	
m_i	Metadata blocklength in i-th transmission
$P_e^{m_i}$	Metadata BLER in i-th transmission
d_i	Data blocklength in i-th transmission
$P_e^{d_i}$	Data BLER in i-th transmission
$P_e^{d_{12}}$	Data BLER upon HARQ retransmission
N^{It}	Average number of resource usage
P_{out}^{It}	The over all outage probability
Joint encoding of data and metadata	
n_i	Codeword length in i-th transmission
$P_e^{J_i}$	BLER in i-th transmission
N^{Jt}	Average number of resource usage
P_{out}^{Jt}	The over all outage probability

mission/processing times and retransmission delay, this leaves enough time budget for a single retransmission (if the initial transmission fails) [50, 51, 58].

Transmitting the D bytes of data requires preceding transmission of M bytes of metadata carrying transceiver/transmission specific information such as device-ID, adopted MCS, precoding matrix information, allocated physical resource block for DL and UL direction, etc. Two proposals are investigated for multiplexing of data and metadata. *In-resource control signalling* along with front-loading of demodulation reference signals is proposed in [7, 8]. Following 5G NR user-centric design for dynamic scheduling of URLLC UEs, whenever the network schedules a UE, the corresponding control information is separately encoded and sent at the beginning of the transmitted packet. It allows the receiver to start processing the metadata while data is still on the air, estimate the channel, and may enable fast HARQ feedback prior to actual

data decoding [59, 60].

The second approach, *joint encoding/decoding of data and metadata* is proposed in [53, 61] aiming to reduce error and spectrum inefficiency initiated by FBL codewords. The main idea is to combine metadata and data in a single packet of size $M + D$ bytes. It is shown that packet transmission with relatively larger blocklength is more reliable and spectrally efficient [24]. However, this concept suffers from high computational-complexity as the UE is required to decode all the messages, even when it is not the intended destination. Thus, it is a trade-off between spectral efficiency and complexity (additional UE processing time and energy).

In line with [30, 62, 63], we analyse URLLC performance using the results from the FBL theory in quasi-static channels [24]. That is, for a payload of b bits information mapped to a codeword of length n channel uses, the error probability ϵ is well approximated as

$$\epsilon \approx E\left(n, \gamma, b\right) \triangleq Q\left(\frac{nC(\gamma) - b}{\sqrt{nV(\gamma)}}\right), \tag{G.1}$$

where $C(\gamma) = \log_2(1+\gamma)$ is the Shannon capacity of complex AWGN channels for a given signal to noise ratio (SNR) γ. $Q(\cdot)$ is the Gaussian Q-function $\left(Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-\frac{u^2}{2}) \mathrm{d}u\right)$ and $V(\gamma) = \frac{1}{\ln^2 2}\left(1 - \frac{1}{(1+\gamma)^2}\right)$ is the channel dispersion factor [24]. Performing periodic channel state information (CSI) measurements, we assume CSI knowledge is available at both transceivers [26, 27, 33]. As the URLLC latency target is smaller than the channel coherence time in most of the applications, it is assumed that the channel remains constant during the initial transmission and the likely one additional retransmission [44, 47, 60].

Using (G.1), the minimum blocklength $n^{\min}$ satisfying the outage probability ϵ_0 is related to the SNR and the payload size, which can be expressed as

$$f(n^{\min}, \gamma, b, \epsilon_0) \triangleq n^{\min} C(\gamma) - b - \sqrt{n^{\min} V(\gamma)} Q^{-1}(\epsilon_0) = 0, \tag{G.2}$$

where $Q^{-1}(\cdot)$ is the inverse of Gaussian Q-function. The function $f(\cdot)$ is convex and has a quadratic form with respect to $\sqrt{n^{\min}}$. Solving (G.2) (noting that only the positive solution is valid) and after some algebraic manipulations we have

$$n^{\min} = \frac{b}{C(\gamma)} + \frac{(Q^{-1}(\epsilon_0))^2 V(\gamma)}{C^2(\gamma)} \times \left[1 + \left(1 + \frac{4C(\gamma)b}{(Q^{-1}(\epsilon_0))^2 V(\gamma)}\right)^{1/2}\right]. \tag{G.3}$$

3 Problem Formulation

3.1 In-Resource Control Signalling

Assuming separate encoding of metadata and data, the probability of success in the initial transmission is conditioned upon correct decoding of both the metadata and data. Let m_i and d_i respectively denote the number of allocated resources (i.e., subcarrier symbols) to metadata and data in the i-th transmission round ($i \in \{1, 2\}$). The number of resources in the initial transmission equals

$$N^{I_1} = m_1 + d_1. \tag{G.4}$$

The probability of success $P_{succ}^{I_1}$ and the corresponding outage probability $P_{out}^{I_1}$ of the first transmission are given by

$$\begin{aligned} P_{succ}^{I_1} &= \left(1 - P_e^{m_1}\right)\left(1 - P_e^{d_1}\right), \\ P_{out}^{I_1} &= 1 - P_{succ}^{I_1} = P_e^{d_1} + P_e^{m_1} - P_e^{d_1} P_e^{m_1}, \end{aligned} \tag{G.5}$$

where $P_e^{d_1}$ and $P_e^{m_1}$ denote the error probability of decoding data and metadata that are scheduled over codeword sizes of d_1 and m_1, respectively. After each successful transmission, the UE feeds back an acknowledgement (ACK) signal to the network. Three possible outcomes of the initial transmission may trigger retransmission.

Decoding ACK as negative acknowledgement (NACK)

It happens when the UE successfully decodes both the metadata and data and sends ACK. But, the ACK message is decoded as NACK at the network. Thereby, an unnecessary HARQ retransmission is scheduled with the same data blocklength d_1. The received copy is discarded after being decoded by the UE and has no impact on the outage reliability. But, it increases the resource utilization by

$$N_{an}^{I_1} = P_{succ}^{I_1} P_e^{an} (m_2 + d_1), \tag{G.6}$$

where P_e^{an} is the error probability of decoding ACK as NACK.

Failure to decode the data

The second case occurs when the UE receives the metadata but fails to decode the data. It will then feed back a NACK. Correct decoding of the NACK by BS will trigger scheduling of the corresponding HARQ retransmission. Otherwise, if the BS decodes the NACK as ACK, it and assume successful

transmission and terminates of procedure. This results in outage for URLLC applications.

Two retransmission protocols of IR-HARQ and CC-HARQ can be considered. Using CC-HARQ, the same codeword as the initial transmission is sent over retransmission round [46]. The UE combines multiples of received data packets upon maximum ratio combining (MRC) to enhance the desired signal power and increase successful decoding probability. With IR-HARQ, data bits are encoded to a parent codeword of length dL channel uses, where L is the maximum number of transmissions. [44]. The parent codeword is split into sub-codewords of d symbols. A new sub-codeword is consecutively transmitted if the UE fails to correctly decode previous received concatenated samples. The success probability following HARQ retransmission is obtained as [64]

$$P_{succ}^{I_{2a}} = \left(1 - P_e^{m_1}\right)\left(1 - P_e^{na}\right)\left(1 - P_e^{m_2}\right)\left(P_e^{d_1} - P_e^{d_{12}}\right), \tag{G.7}$$

where P_e^{na} is the error probability of decoding of NACK as ACK. Variable $P_e^{d_{12}}$ is the data error probability after HARQ retransmission combining equals $E(d_1, 2\gamma, D)$ and $E(2d_1, \gamma, D)$ for CC-HARQ and IR-HARQ, respectively [47, 64]. The additional resource utilization of this case is obtained as

$$N^{I_{2a}} = \left(1 - P_e^{m_1}\right) P_e^{d_1} \left(1 - P_e^{na}\right)(m_2 + d_1). \tag{G.8}$$

Note that P_e^{na} is critical for URLLC services so that high values of P_e^{na} prohibit achieving the desired reliability target within a limited time. Asymmetric detection to protect NACK signals from being decoded as ACK, NACK repetition, and allocating more resources for feedback signals are among the proposed solutions to enhance the reliability of feedback channel [51, 65, 66].

Failure to decode the metadata

This is known as discontinuous transmission (DTX). In this case, the UE does not know if is intended to a transmission. Thus, it does not forward any feedback. This leads to HARQ timeout, which happens when the BS could not receive ACK/NACK signals by listening to the UL feedback channel within a predefined time interval. This leads to a new retransmission by the BS. We assume the timeout duration equals ACK/NACK feedback time, so that one new retransmission can be performed within the maximum URLLC latency budget. Since the control information required to identify the data block in the initial transmission was not correctly decoded, unlike the previous case, there is no possibility of HARQ combining. Accordingly, the probability of

success $P_{succ}^{I_{2b}}$ and resource usage $N^{I_{2b}}$ are driven as

$$\begin{aligned} P_{succ}^{I_{2b}} &= P_e^{m_1}\left(1-P_e^{ma}\right)\left(1-P_e^{m_2}\right)\left(1-P_e^{d_2}\right), \\ N^{I_{2b}} &= P_e^{m_1}\left(1-P_e^{ma}\right)\left(m_2+d_2\right), \end{aligned} \tag{G.9}$$

where P_e^{ma} is the probability that the BS erroneously assumes receiving ACK that leads to outage failure.

Following those three error cases for in-resource control signalling, we derive the overall outage probability P_{out}^{It} assuming initial transmission plus one retransmission (if needed) as

$$\begin{aligned} P_{out}^{It} &= 1 - P_{succ}^{I_1} - P_{succ}^{I_{2a}} - P_{succ}^{I_{2b}} \\ &= P_e^{m_1}P_e^{m_2}\left[1-P_e^{d_1}-P_e^{d_2}\right] + P_e^{m_1}P_e^{d_2} + P_e^{m_2}P_e^{d_1} \\ &\quad + P_e^{d_{12}}\left[1-P_e^{m_1}-P_e^{m_2}+P_e^{m_1}P_e^{m_2}\right] \\ &\quad + P_e^{ma}\left[P_e^{m_1}(1-P_e^{m_2})(1-P_e^{d_2})\right] \\ &\quad + P_e^{na}\left[(1-P_e^{m_1}-P_e^{m_2}+P_e^{m_1}P_e^{m_2})(P_e^{d_1}-P_e^{d_{12}})\right] \\ &\approx P_e^{m_1}P_e^{m_2} + P_e^{m_1}P_e^{d_2} + P_e^{m_2}P_e^{d_1} + P_e^{d_{12}} \\ &\quad + P_e^{ma}P_e^{m_1} + P_e^{na}P_e^{d_1}. \end{aligned} \tag{G.10}$$

Similarly, the average number of resource usage N^{It} is obtained as

$$\begin{aligned} N^{It} &= N^{I_1} + N_{an}^{I_1} + N^{I_{2a}} + N^{I_{2b}} \\ &= m_1 + d_1 + P_e^{an}(1-P_e^{m_1})(1-P_e^{d_1})(m_2+d_1) \\ &\quad + P_e^{d_1}(1-P_e^{m_1})(1-P_e^{na})(m_2+d_1) \\ &\quad + P_e^{m_1}\left(1-P_e^{ma}\right)\left(m_2+d_2\right) \\ &\approx m_1 + d_1 + P_e^{m_1}(m_2+d_2) + (P_e^{d_1}+P_e^{an})(m_2+d_1). \end{aligned} \tag{G.11}$$

As URLLC deals with low values of errors, the relative cross products are reasonably negligible that make the applied approximations in (G.10) and (G.11) accurate. In Section 5, the accuracy of the approximations are numerically validated.

From (G.10), we realize that the error in the control information along with the miss-detection of the feedback signals are not negligible and have a notable impact on the outage reliability. This is unlike most of the literature studies (e.g. [45–47, 49]) which assume ideal metadata transmission and mainly focus on data outage probability (i.e., $P_e^{d_{12}}$). The overhead and the impact of metadata and UL feedback signals on the required resources are highlighted in (G.11), where we observe that metadata plays an important

role on the network resource utilization. Therefore, it is essential to design and optimize the transmission performance of URLLC by taking into account the data, metadata, and impairments in feedback channel.

We formulate a resource allocation problem to minimize the number of allocated resources while satisfying the QoS requirements of URLLC. The optimization is expressed as:

$$\begin{aligned}
&\min_{d_i, m_i} N^{It} \\
&\text{S.t. C1I: } P_{out}^{It} \leq P_{out}^{tar} \\
&\quad \text{C2I: } E\left(m_i, \gamma, M\right) = P_e^{m_i}, \quad i = 1, 2, \\
&\quad \text{C3I: } E\left(d_i, \gamma, D\right) = P_e^{d_i}, \quad i = 1, 2, \\
&\quad \text{C4I: } \begin{cases} E(d_1, 2\gamma, D) = P_e^{d_{12}} & \text{for CC-HARQ,} \\ E(2d_1, \gamma, D) = P_e^{d_{12}} & \text{for IR-HARQ,} \end{cases} \\
&\quad \text{C5I: } d_i, m_i \in \mathbb{N}, \quad i = 1, 2.
\end{aligned} \tag{G.12}$$

Constraints C1I guarantees the reliability requirement. Constraints C2I-C4I are the error probabilities corresponding to the channel allocations of data and metadata in the initial transmission and the retransmission. Finally, C5I indicates that the number of allocated channels are positive integers. In (G.12), the equality constraints C2I-C4I are not affine and the objective function, constraints C1I, and C5I are non-convex. Thus, it belongs to the family of non-convex combinatorial optimization problems difficult to solve with polynomial complexity. In Section 4.1, we present a solution based on consecutive convex optimization to tackle the non-convexity in (G.12).

3.2 Joint Encoding of Data and Metadata

Assuming joint encoding of the metadata and the data to a single codeword of length n_1 channel uses, the probability of success in the first attempt $P_{succ}^{J_1}$ is given by

$$\begin{aligned}
P_{succ}^{J_1} &= 1 - P_e^{J_1}, \\
P_e^{J_1} &= E(n_1, \gamma, M + D).
\end{aligned} \tag{G.13}$$

The UE sends ACK following correct decoding. Since, both metadata and data are encoded (decoded) together, we assume there are no possibilities for sending NACK as well as the HARQ gain of combining data packets after the retransmission. If an ACK is not received within a predefined timeout period, the network assumes failure and retransmits the packet with a blocklength of n_2 and error probability of $P_e^{J_2}$. Thus, the success probability following the

second transmission $P_{succ}^{J_2}$ is given by

$$P_{succ}^{J_2} = P_e^{J_1}\left(1 - P_e^{ma}\right)\left(1 - P_e^{J_2}\right). \tag{G.14}$$

Consequently, for joint encoding of data and metadata, the overall outage probability P_{out}^{Jt} is calculated as

$$\begin{aligned} P_{out}^{Jt} &= 1 - P_{succ}^{J_1} - P_{succ}^{J_2} \\ &= P_e^{J_1}(P_e^{ma} + P_e^{J_2} - P_e^{ma}P_e^{J_2}) \approx P_e^{J_1}P_e^{ma} + P_e^{J_1}P_e^{J_2}, \end{aligned} \tag{G.15}$$

Also, the corresponding average number of allocated resources N^{Jt} equals

$$\begin{aligned} N^{Jt} &= n_1 + P_e^{J_1}\left(1 - P_e^{ma}\right)n_2 + \left(1 - P_e^{J_1}\right)P_e^{am}n_2 \\ &\approx n_1 + P_e^{J_1}n_2 + P_e^{am}n_2, \end{aligned} \tag{G.16}$$

where P_e^{am} is the probability that ACK feedback is not detected correctly by the BS that results in redundant retransmission.

The resource allocation problem for joint encoding of data and metadata is formulated as follows

$$\begin{aligned} &\min_{n_i} N^{Jt} \\ &\text{S.t.: C1J: } P_{out}^{Jt} = P_e^{J_1}P_e^{ma} + P_e^{J_1}P_e^{J_2} \leq P_{out}^{tar}, \\ &\quad \text{C2J: } E\left(n_i, \gamma, M + D\right) = P_e^{J_i}, \quad i = 1, 2, \\ &\quad \text{C3J: } n_i \in \mathbb{N} \quad i = 1, 2. \end{aligned} \tag{G.17}$$

Similar to (G.12), the objective function and constraint C1J are non-convex. Constraint C2J is not affine, and finally C3J is integer.

4 Proposed Low Complexity Near Optimum Solution

4.1 In-Resource Control Signalling

In order to solve the optimization problem (G.12), an efficient near-optimum low-complexity solution is proposed. In the rest of the paper, we assume CC-HARQ. However, similar procedures can be applied to IR-HARQ. We handle the non-convexity of the problem (G.12) by developing an algorithm based on successive solving a convex optimization problem through the following steps.

- **Step-1: Integer relaxation** First, we relax the constraint C5I in (G.12) and assume the number of channel uses can be positive real number (i.e. $d_i, m_i \in \mathbb{R}^+$).

- **Step-2: Convert non-convex functions to convex form** The objective function and constraint C1I in (G.12) are non-convex posynomial functions [67]. To handle the non-convexity, we introduce new variables as: $x_1 \triangleq \ln m_1,\ x_2 \triangleq \ln d_1,\ x_3 \triangleq \ln m_2,\ x_4 \triangleq \ln d_2,\ y_1 \triangleq \ln P_e^{m_1},\ y_2 \triangleq \ln P_e^{d_1},\ y_3 \triangleq \ln P_e^{m_1},\ y_4 \triangleq \ln P_e^{m_2}\ y_5 \triangleq \ln P_e^{d_{12}}$. Revising problem (G.12) with respect to the new variables and substituting the objective function and constraint C1I with their logarithmic form result in

$$\begin{aligned}
&\min_{x_i,y_i} \ln \Big[e^{x_1} + e^{x_2} + e^{y_1+x_3} + e^{y_1+x_4} + e^{y_2+x_2} \\
&\qquad\qquad + e^{y_2+x_3} + P_e^{an}\left(e^{x_2} + e^{x_3}\right)\Big], \\
&\text{S.t. C1a: } \ln \Big[e^{y_1+y_3} + e^{y_1+y_4} + e^{y_2+y_3} \\
&\qquad + e^{y_5} + P_e^{ma} e^{y_1} + P_e^{na} e^{y_2}\Big] - \ln P_{out}^{tar} \leq 0, \\
&\quad \text{C2Ia: } E\left(e^{x_i}, \gamma, M\right) = e^{y_i}, \quad i = 1, 3, \\
&\quad \text{C3Ia: } E\left(e^{x_i}, \gamma, D\right) = e^{y_i}, \quad i = 2, 4, \\
&\quad \text{C4Ia: } E\left(e^{x_1}, 2\gamma, D\right) = e^{y_5}, \\
&\quad \text{C5Ia: } x_i \in \mathbb{R}^+, \quad i \in \{1, ..., 4\}, \\
&\quad \text{C6Ia: } y_i \in \mathbb{R}, \quad i \in \{1, ..., 5\}.
\end{aligned} \tag{G.18}$$

  In the revised optimization problem (G.18), the objective function and constraint C1Ia are standard convex form [67].

- **Step-3: Modify equality constraints** Constraints C2Ia-C4Ia are non-affine equality constraints. Without loss of optimality, we can modify them to inequality constraints as

$$\begin{aligned}
&\text{C2Ib: } E\left(e^{x_i}, \gamma, M\right) - e^{y_i} \leq 0, \quad i = 1, 3, \\
&\text{C3Ib: } E\left(e^{x_i}, \gamma, D\right) - e^{y_i} \leq 0, \quad i = 2, 4, \\
&\text{C4Ib: } E\left(e^{x_2}, 2\gamma, D\right) - e^{y_5} \leq 0.
\end{aligned} \tag{G.19}$$

  Proof. Please refer to Appendix A.

- **Step-4: Difference of convex functions.** Constraints C2Ib-C4Ib are the difference of two convex functions.
  Proof. Please refer to Appendix B.
  To handle the non-convexity of constraints C2Ib-C4Ib, we employ successive convex approximation [27, 68]. Applying Taylor expansion for convex functions, the first order approximation of e^{y_i} with respect to a initial point y_i^0 is expressed as

$$e^{y_i^0} + e^{y_i^0}(y_i - y_i^0) \leq e^{y_i}, \tag{G.20}$$

which is an affine function of y_i. Substituting (G.20) into (G.19), we obtain the following problem

$$\begin{aligned}
&\min_{x_i,y_i} \ \ln\Big[e^{x_1}+e^{x_2}+e^{y_1+x_3}+e^{y_1+x_4}+e^{y_2+x_2} \\
&\qquad\qquad +e^{y_2+x_3}+P_e^{an}\left(e^{x_2}+e^{x_3}\right)\Big], \\
&\text{S.t. C1Ia: } \ln\Big[e^{y_1+y_3}+e^{y_1+y_4}+e^{y_2+y_3} \\
&\qquad +e^{y_5}+P_e^{ma}e^{y_1}+P_e^{na}e^{y_2}\Big]-\ln P_{out}^{tar}\leq 0, \\
&\text{C2Ic: } E\left(e^{x_i},\gamma,M\right)-e^{y_i^0}(1+y_i-y_i^0)\leq 0, i=1,3, \\
&\text{C3Ic: } E\left(e^{x_i},\gamma,D\right)-e^{y_i^0}(1+y_i-y_i^0)\leq 0, i=2,4, \\
&\text{C4Ic: } E\left(e^{x_2},2\gamma,D\right)-e^{y_5^0}(1+y_5-y_5^0)\leq 0, \\
&\text{C5Ic: } x_i\in\mathbb{R}^+ \quad i\in\{1,...,4\}, \\
&\text{C6Ic: } y_i\in\mathbb{R} \quad\ \ i\in\{1,...,5\}.
\end{aligned} \tag{G.21}$$

In (G.21), the objective and constraints are convex functions forming a standard convex optimization problem that can be solved via optimization toolbox with polynomial time. Based on above analyses, we apply an iterative algorithm to find a sub-optimal solution for problem (G.12). The convex optimization problem in (G.21) is solved with the initial points y_i^0. The initial points are then updated with the optimal solutions of the previous iteration. Algorithm 1 summarizes the steps of the proposed solution.

Algorithm 1 Proposed Solution for the Optimization Problem (G.12)

1: **Initialize:** The initial points $e^{y_i^0}$, the iteration number $k=1$, and the maximum number of iterations $K_{\max}$.
2: **Repeat:**
3: Solve the convex optimization problem (G.21) with $e^{y_i^{k-1}}$.
4: Update $e^{y_i^k}=e^{y_i}$.
5: Update $k=k+1$.
6: **Until:** $k=K_{\max}$.
7: **Return:** e^{x_i}.

4.2 Joint Encoding of Data and Metadata

Problem (G.17) has similar structure as that of (G.12). Thereby, applying the same approaches as in Section 4.1 yields a standard convex optimization

$$\begin{aligned}
&\min_{w_i, z_i} \ \ln\left[e^{w_1} + e^{z_1+w_2} + P_e^{am} e^{w_2}\right], \\
&\text{S.t.: C1Ja: } \ln\left[e^{z_1} P_e^{ma} + e^{z_1+z_2}\right] - \ln P_{out}^{tar} \leq 0, \\
&\quad \text{C2Ja: } E\left(e^{w_i}, \gamma, D+M\right) - e^{z_i^0}(1 + z_i - z_i^0) \leq 0, \\
&\quad \text{C3Ja: } w_i \in \mathbb{R}^+, \quad i = 1, 2, \\
&\quad \text{C4Ja: } z_i \in \mathbb{R}, \quad\ \ i = 1, 2,
\end{aligned} \tag{G.22}$$

where $w_i \triangleq \ln n_i, z_i \triangleq \ln P_e^{J_i}$. Algorithm 2 presents the steps toward solving problem (G.17).

Algorithm 2 Proposed Solution for the Optimization Problem (G.17)

1: **Initialize:** The initial points $e^{z_i^0}$, the iteration number $k = 1$, and the maximum number of iterations $K_{\max}$.
2: **Repeat:**
3: Solve the convex optimization problem (G.22) with $e^{z_i^{k-1}}$.
4: Update $e^{z_i^k} = e^{z_i}$.
5: $k = k + 1$.
6: **Until:** $k = K_{\max}$.
7: **Return:** e^{w_i}.

5 Numerical Results

This section presents numerical results evaluating the performance of different multiplexing schemes for URLLC. For simulation parameters, we assume equal false-alarm probability P_e^{fa} for P_e^{na} and P_e^{ma} (i.e. $P_e^{na} = P_e^{ma} = P_e^{fa} = 10^{-5}$). Also, to cover the asymmetric detection of feedback signals in URLLC we assume $P_e^{an} = P_e^{am} = 5P_e^{fa} = 5 \times 10^{-5}$. We set the number of maximum iterations $K_{\max} = 5$ for Algorithms 1 and 2.

5.1 In-Resource Control Signalling

Figure G.1 depicts the resource usage performance gains of sending URLLC payloads assuming in-resource control channel multiplexing with respect to outage reliability targets ranging from $P_{out}^{tar} = 10^{-4}$ to $P_{out}^{tar} = 10^{-7}$ and for different channel conditions. We assume data and metadata size are $D = 32$

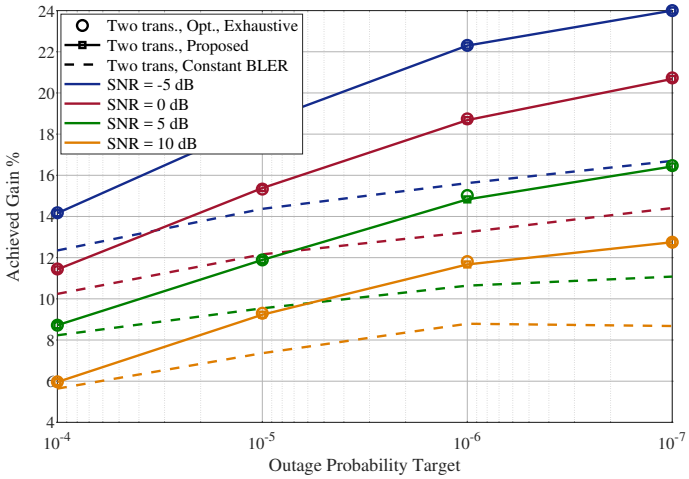

Fig. G.1: Performance analysis of in-resource control signalling for different outage probabilities and channel conditions with $M = 16$ and $D = 32$ bytes. $P_e^{na} = P_e^{ma} = 10^{-5}$ and $P_e^{an} = P_e^{am} = 5 \times 10^{-5}$.

and $M = 16$ bytes, respectively [53, 64]. The gains are compared against the baseline scenario with a single transmission in which both data metadata are encoded with the same error probability equal $P_e^{d_1} = P_e^{m_1} = \frac{1}{2}P_{out}^{tar}$. Considering cases with two transmissions, we present results for three scheduling schemes i) transmission with constant BLER (i.e. $P_e^{d_i} = P_e^{m_i}$, $i = 1, 2$), ii) The proposed solution provided by Algorithm 1, iii) The optimal solution obtained by performing an exhaustive search over the feasible set of points and without approximations in (G.10) and (G.11).

Assuming two transmissions with equal BLER P_e for both metadata and data, the expression of total outage probability in equation (G.10) is further simplified to

$$P_{out}^{St} = 3{P_e}^2 + P_e(P_e^{na} + P_e^{ma}) + P_e^{12} \leq P_{out}^{tar}. \tag{G.23}$$

Solving (G.23) with equality constraint to minimize the number of resources, results

$$P_e^{d_i} = P_e^{m_i} = P_e \approx -\frac{(P_e^{na} + P_e^{ma})}{6} + \frac{\left((P_e^{na} + P_e^{ma})^2 + 12P_{out}^{tar}\right)^{\frac{1}{2}}}{6}, \quad i = 1, 2. \tag{G.24}$$

Accordingly, the resources can be calculated using (G.3). Figure G.1 reveals significant resource efficiency enhancements with two transmissions. As we

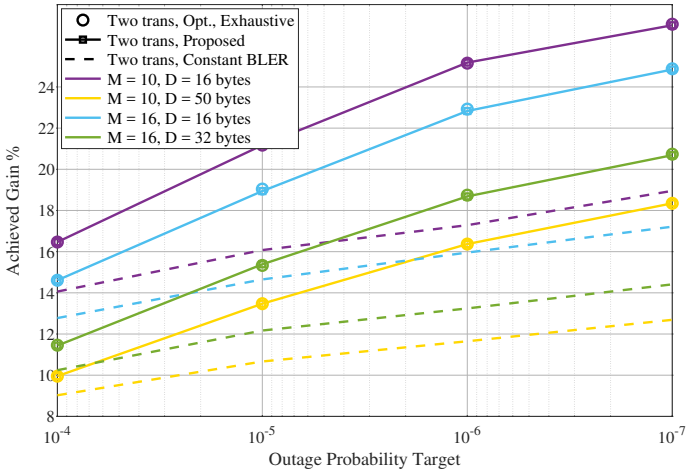

Fig. G.2: Performance analysis of in-resource control signalling for different metadata and data sizes with $\gamma = 0$ dB. $P_e^{na} = P_e^{ma} = 10^{-5}$ and $P_e^{an} = P_e^{am} = 5 \times 10^{-5}$.

observe, the provided gain increases with grows in reliability requirements. For instance, at the outage of $P_{out}^{tar} = 10^{-7}$ (99.99999% reliability) and -5 dB SNR, constants BLER is 16% more resource-efficient as compared to single-shot scheduling. Moreover, the proposed variable error-rate scheduling provides better performance and achieves 24% gain in resource usage. This is because the proposed algorithm schedules the initial transmission with higher rates to minimize the resources. Failure in decoding control information in the first transmission, both data and metadata are scheduled with lower coding-rate to ensure reliability. On the other hand if the UE could not decode the data, HARQ retransmission is scheduled with robust metadata and the data signal quality is enhanced by the HARQ combining of the received packets. Comparing the results from the proposed algorithm with those of optimal solutions by exhaustive search solution, we observe that our solution performs well and approaches similar performance close to the optimal point solution. As the exhaustive search solution is driven by investigating the original resource allocation problem (without approximations), it confirms that the applied approximations in (G.10) and (G.11) are accurate and valid.

Figure G.2 shows the performance gain for different metadata and data set sizes assuming SNR of $\gamma = 0$ dB. For each set, the achieved gain is calculated by comparing against the results of single transmission with the same metadata data sizes. We observe that the gain is higher for short payload sizes. Assuming $M = 10$ and $D = 16$ bytes, the proposed solution is 27% more resource efficient at 10^{-7} outage reliability. Increasing the data size to

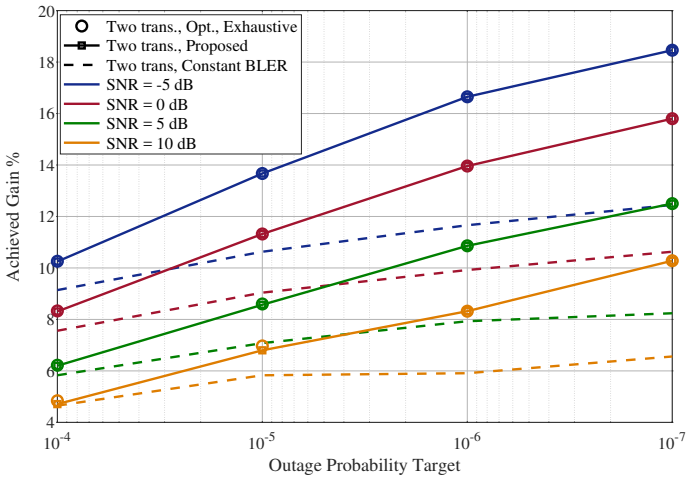

Fig. G.3: Performance analysis of joint encoding of data and metadata for different outage probabilities and channel conditions with $M = 16$ and $D = 32$ bytes. $P_e^{ma} = 10^{-5}$ and $P_e^{am} = 5 \times 10^{-5}$.

$D = 50$ bytes, the gain reduces to 18%. This is due to the fact that the impairment of short packet transmission is decreased as the blocklength grows and the achievable rate converges to the Shannon Capacity. This in turn facilitates low-error scheduling with single transmission.

5.2 Joint Encoding of Data and Metadata

In Figure G.3 and G.4 we evaluate the performance of joint encoding of data and metadata. Figure G.3 plots the achieved gain versus the outage probability for different transmission schemes and SNR values. For the baseline single transmission, the minimum number of required resources n_1 is calculated based on (G.3) such that $E(n_1, \gamma, M + D) = P_{out}^{tar}$. For equal constant BLER with retransmission, the error probabilities are obtained from (G.15) as

$$P_e^{J_1} = P_e^{J_2} \approx \frac{-P_e^{ma}}{2} + \frac{\left((P_e^{ma})^2 + 4P_{out}^{tar}\right)^{\frac{1}{2}}}{2}. \tag{G.25}$$

As can be observed, the proposed solution improves the performance by reducing the number of resources required to guarantee the desired reliability targets. At SNR of $\gamma = 0$ dB and for 99.99999% reliability, it provides 16% gain in comparison to the baseline single transmission. Moreover, the performance is very close to that of the optimal exhaustive search solution. The results in Figures G.1 and G.3 show that retransmission is more favourable at

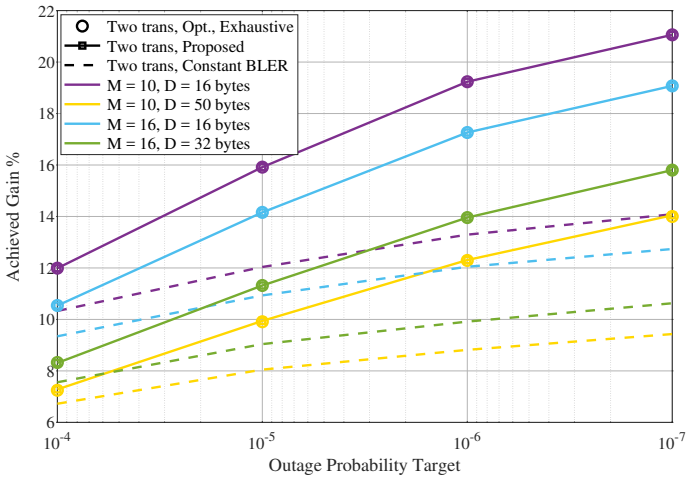

Fig. G.4: Performance analysis of joint encoding for different metadata and data sizes with $\gamma = 0$ dB. $P_e^{na} = P_e^{ma} = 10^{-5}$ and $P_e^{an} = P_e^{am} = 5 \times 10^{-5}$.

low-SNR regimes. The achieved gain of retransmission decreases with an increase in channel quality. The reason is that reliability significantly enhances at high SNRs. Therefore, it is also possible to achieve the reliability target with a relatively low number of resources in a single allocation.

Fig. G.4 shows the performance for different metadata and data sizes. Similar to the results of Fig. G.2, we observe retransmission is more beneficial for small payloads and the gain decreases for larger packets. Comparing in-resource control signalling with joint encoding, we observe that the relative improvements of enabling retransmission are higher for in-resource control transmission. Assuming $M = D = 16$ bytes and at the outage reliability of 10^{-6}, the proposed algorithm results in 23% gain as compared to baseline case for in-resource control signalling. While for joint encoding, 17% improvement is achieved. This is because of the capability of the joint transmission scheme to encode both the metadata and data with a larger codeword that diminishes short blocklength inefficiency.

5.3 Performance Comparison

In Fig. G.5, we provide a comparison between in-resource control scheduling and joint encoding of data and metadata.

The relationship between resource efficiency and feedback errors is also further investigated. The figure plots the minimum number of required resource elements versus operating SNRs for different scheduling schemes. Benefiting from sending information with a larger codeword and in compar-

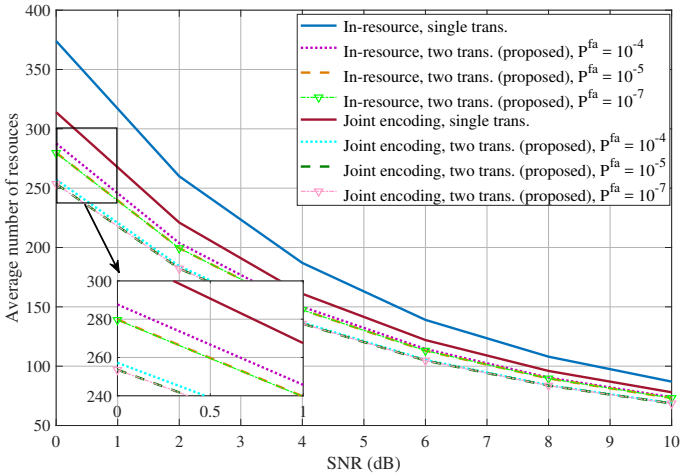

Fig. G.5: Performance comparison of in-resource control signalling and joint encoding of data and metadata with different feedback errors, assuming $P_{out}^{tar} = 10^{-6}$, $M = 10$ and $D = 16$ bytes. $P_e^{na} = P_e^{ma} = P_e^{fa}$, $P_e^{an} = P_e^{am} = 5P_e^{fa}$.

ison in-resource control signalling, joint encoding shows superior resource efficiency for both single and two-transmission schemes. Recall that, though joint encoding is more efficient, its relative gain with respect to single transmission is lower than in-resource control signalling, as highlighted by the previous results. Assuming single transmission and at 5 dB SNR, it provides around 14% gain over the in-resource control transmission. With two transmissions and by applying the proposed optimization, the difference decreases to 8%. However, the gain comes at the expense of higher complexity (more required energy and processing time) at the receiver. Moreover, the plot shows that the performance difference decreases with improving channel conditions. Additionally, we observe that feedback errors have lower (in comparison to data and metadata) impact on the resource usage. Reducing feedback error from 10^{-4} to 10^{-7} results around 1% improvement of the resource efficiency at 5 dB SNR.

6 Conclusion

This paper studied downlink radio resource allocation for URLLC in 5G NR. Two multiplexing methodologies namely as *in-resource control signalling* and *joint encoding (decoding) of data and metadata* are investigated. We proposed an analytical framework to evaluate the allocated resources and the outage probability of URLLC scheduling. It is shown that non-ideal transmission of

control information has a significant impact on the system overhead and reliability of URLLC. For each scheme, we formulated an optimization problem comprising joint link adaptation for data and metadata to minimize resource allocation while guaranteeing URLLC requirements. Since the problems are integer non-convex optimizations, solutions based on successive convex optimizations are proposed. Numerical analyses showed that the proposed algorithms perform close to the optimal solutions, significantly reduce the resource usage and achieve up to 27% resource utilization improvement. Future studies could examine the impact of imperfect (delayed) channel knowledge at transceivers and analyse the performance of multi-node joint transmission.

7 Appendix A

We show that constraints C2Ia-C4Ia are hold with equality at the optimum point. To this end, we first prove that the decoding error probability is always a decreasing function with respect to x. The partial derivative of the function $E(e^x, \gamma, b)$ with respect to x is given by

$$\frac{\partial E(e^x, \gamma, b)}{\partial x} = -\frac{e^x C(\gamma) + b}{2\sqrt{2\pi e^x V(\gamma)}} e^{\frac{-1}{2}\left(\frac{e^x C(\gamma) - b}{\sqrt{e^x V(\gamma)}}\right)^2} \leq 0,$$

meaning that $E(\cdot)$ is monotonically decreasing function of x.

Applying the contradiction theory, let us assume $\{x_i^\star, y_i^\star\}$ are the optimal solutions of (G.18) satisfying at least one of the constraints with non-equality (i.e. $E(e^{x_j^\star}, \gamma, b) < e^{y_j^\star}$). In this case, the achieved minimum number of channel uses is denoted by $N^\star$. We denote $x_j^{\star\star}$ as the solution of $E(e^{x_j^{\star\star}}, \gamma, b) = e^{y_j^\star}$ can be obtained using (G.3). Since $E(\cdot)$ is always decreasing with respect to x, we have $x_j^{\star\star} < x_j^\star$. Suppose a set of points as $\{x_{i,i\neq j}^\star, x_j^{\star\star}, y_i^\star\}$ resulting $N^{\star\star}$ channel uses. This leads to $N^{\star\star} < N^\star$ which is in contradiction with the optimality assumption of $N^\star$. Following similar proofs for other cases, we conclude that modifying equality constraints to inequality does not change the optimal solution.

8 Appendix B

To prove the convexity of decoding error probability, it is sufficient to show that the second derivative of $E(e^x, \gamma, b)$ presented in (G.26) is positive.

$$\frac{\partial^2 E(e^x,\gamma,b)}{\partial^2 x} = \frac{e^{\frac{-1}{2}\left(\frac{e^x C(\gamma)-b}{\sqrt{e^x V(\gamma)}}\right)^2}}{4\sqrt{2\pi}}$$
$$\times \left[\frac{C^3(\gamma)e^{3x} - b^3 - b^2 C(\gamma)e^x + bC^2(\gamma)e^{2x} + 3bV(\gamma)e^x + C(\gamma)V(\gamma)e^{2x}}{e^x V(\gamma)\sqrt{e^x V(\gamma)}}\right]$$
$$- \frac{e^{\frac{-1}{2}\left(\frac{e^x C(\gamma)-b}{\sqrt{e^x V(\gamma)}}\right)^2}(e^x C(\gamma)+b)}{4\sqrt{2\pi e^x V(\gamma)}}$$
$$= \frac{e^{\frac{-1}{2}\left(\frac{e^x C(\gamma)-b}{\sqrt{e^x V(\gamma)}}\right)^2}}{4e^x V(\gamma)\sqrt{2\pi e^x V(\gamma)}}(e^x C(\gamma)-b)\left[C^2(\gamma)e^{2x} + 2C(\gamma)e^x b + b^2 - V(\gamma)e^x\right]$$
$$= \frac{e^{\frac{-1}{2}\left(\frac{e^x C(\gamma)-b}{\sqrt{e^x V(\gamma)}}\right)^2}}{4e^x V(\gamma)\sqrt{2\pi e^x V(\gamma)}}\left(C(\gamma)e^x + b + \sqrt{V(\gamma)e^x}\right)$$
$$\times \underbrace{(e^x C(\gamma)-b)}_{=\Pi}\underbrace{\left(C(\gamma)e^x + b - \sqrt{V(\gamma)e^x}\right)}_{=\Phi}. \tag{G.26}$$

In (G.26), positivity of the term Π holds when $e^x C(\gamma) \geq b$ implying that $0 < E(e^x,\gamma,b) \leq \frac{1}{2}$ which is a valid assumption for URLLC. Also term Φ has quadratic form and is convex with respect to $e^{\frac{x}{2}}$ ($\frac{\partial^2\Phi}{\partial^2 e^{\frac{x}{2}}} = 2C(\gamma) \geq 0$). Therefore, Φ is minimized setting $\frac{\partial\Phi}{\partial e^{\frac{x}{2}}} = 0$, that results in $e^{\frac{x}{2}} = \sqrt{V(\gamma)}/2C(\gamma)$. The minimum value of Φ is given by

$$\Phi^{\min} = b - \frac{V(\gamma)}{4C(\gamma)}. \tag{G.27}$$

Taking the first derivative of $\Phi^{\min}$ with respect to γ, we have

$$\frac{\partial\Phi^{\min}}{\partial\gamma} = \frac{1}{4\ln 2}\frac{\gamma^2 + 2\left(\gamma - \ln(1+\gamma)\right)}{(1+\gamma)^3\ln^2(1+\gamma)}. \tag{G.28}$$

Given that $\gamma \geq \ln(1+\gamma)$ we conclude $\frac{\partial\Phi^{\min}}{\partial\gamma} \geq 0$, indicating that $\Phi^{\min}$ is monotonically increasing function of γ. As $\Phi^{\min}$ is also increasing with b, it is sufficient to show that $\Phi^{\min} \geq 0$ for few number of information bits and low-value of SNR. Assuming $b = 1$ bit and $\gamma = -100$ dB, we have $\Phi^{\min} = 0.77$. We therefore conclude that $E(e^x,\gamma,b)$ is convex function of x. It is straightforward to show that e^y is also a convex function with respect to y. This completes the proof.

References

[1] 3GPP Technical Specification 38.300, "NR and NG-RAN overall description; stage-2," Version 15.5.0, March 2019.

[2] 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.

[3] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 124–130, June 2018.

[4] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication," *IEEE Network*, vol. 32, no. 2, pp. 24–31, March 2018.

[5] B. Chang, L. Zhang, L. Li, G. Zhao, and Z. Chen, "Optimizing resource allocation in URLLC for real-time wireless control systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 8916–8927, September. 2019.

[6] 3GPP Technical Report 38.913, "Study on scenarios and requirements for next generation access technologies," Version 14.1.0, March 2017.

[7] N. H. Mahmood, M. Lauridsen, G. Berardinelli, D. Catania, and P. Mogensen, "Radio resource management techniques for eMBB and mMTC services in 5G dense small cell scenarios," in *Proc. 84th IEEE VTC Fall*, Montreal, Canada, September 2016.

[8] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, 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.

[9] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28 912–28 922, 2018.

[10] R. Abreu, T. Jacobsen, K. Pedersen, G. Berardinelli, and P. Mogensen, "System level analysis of eMBB and Grant-Free URLLC multiplexing in uplink," in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*, April 2019, pp. 1–5.

[11] A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Pocovi, and P. Mogensen, "Efficient low-complexity packet scheduling algorithm for mixed URLLC and eMBB traffic in 5G," *In Proc. 2019 IEEE 89th Vehicular Technology Conference - VTC2019-Spring*, May, 2019.

[12] A. Anand, G. D. Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, April 2018, pp. 1970–1978.

[13] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38 451–38 463, 2018.

[14] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen, "5G centralized multi-cell scheduling for URLLC: Algorithms and system-level performance," *IEEE Access*, vol. 6, pp. 72 253–72 262, 2018.

[15] A. Karimi, K. I. Pedersen, and P. Mogensen, "5G URLLC performance analysis of dynamic-point selection multi-user resource allocation," *in Proc. 2019 International Symposium on Wireless Communication Systems (ISWCS)*, August, 2019.

[16] N. H. Mahmood, M. Lopez, D. Laselva, K. Pedersen, and G. Berardinelli, "Reliability oriented dual connectivity for URLLC services in 5G New Radio," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, August 2018, pp. 1–6.

[17] G. Berardinelli, N. Huda Mahmood, R. Abreu, T. Jacobsen, K. Pedersen, I. Z. Kovács, and P. Mogensen, "Reliability analysis of uplink grant-free transmission over shared resources," *IEEE Access*, vol. 6, pp. 23 602–23 611, 2018.

[18] Z. Ding, R. Schober, P. Fan, and H. V. Poor, "Simple semi-grant-free transmission strategies assisted by non-orthogonal multiple access," *IEEE Transactions on Communications*, pp. 1–1, 2019.

[19] T. H. Jacobsen, R. Abreu, G. Berardinelli, K. I. Pedersen, I. Z. Kovács, and P. Mogensen, "Multi-cell reception for uplink grant-free ultra-reliable low-latency communications," *IEEE Access*, vol. 7, pp. 80 208–80 218, 2019.

[20] R. M. Cuevas, C. Rosa, F. Frederiksen, and K. I. Pedersen, "Uplink ultra-reliable low latency communications assessment in unlicensed spectrum," in *2018 IEEE Globecom Workshops*, December 2018, pp. 1–6.

[21] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultra-reliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, September. 2016.

[22] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 72–78, June 2017.

[23] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[24] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, July 2014.

[25] P. Popovski, J. J. Nielsen, C. Stefanovic, E. d. Carvalho, E. Strom, K. F. Trillingsgaard, A. Bana, D. M. Kim, R. Kotaba, J. Park, and R. B. Sorensen, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Network*, vol. 32, no. 2, pp. 16–23, March 2018.

[26] Y. Hu, M. Ozmen, M. C. Gursoy, and A. Schmeink, "Optimal power allocation for QoS-constrained downlink multi-user networks in the finite blocklength regime," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 5827–5840, September 2018.

[27] W. R. Ghanem, V. Jamali, Y. Sun, and R. Schober, "Resource allocation for multi-user downlink URLLC-OFDMA systems," *CoRR*, vol. abs/1901.05825, 2019. [Online]. Available: http://arxiv.org/abs/1901.05825

[28] A. Anand and G. de Veciana, "Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2411–2421, November 2018.

[29] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 127–141, January 2018.

[30] ——, "Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications," *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 2266–2280, May 2018.

[31] C. Xiao, J. Zeng, W. Ni, X. Su, R. P. Liu, T. Lv, and J. Wang, "Downlink MIMO-NOMA for ultra-reliable low-latency communications," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 780–794, April 2019.

[32] M. Amjad and L. Musavian, "Performance analysis of NOMA for ultra-reliable and low-latency communications," in *2018 IEEE Globecom Workshops (GC Wkshps)*, December 2018, pp. 1–5.

[33] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, "Short-packet downlink transmission with non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4550–4564, July 2018.

[34] A. Azari, M. Ozger, and C. Cavdar, "Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning," *IEEE Communications Magazine*, vol. 57, no. 3, pp. 42–48, March 2019.

[35] C. Liu, M. Bennis, M. Debbah, and H. V. Poor, "Dynamic task offloading and resource allocation for ultra-reliable low-latency edge computing," *IEEE Transactions on Communications*, pp. 1–1, 2019.

[36] L. Zhang and Y. Liang, "Average throughput analysis and optimization in cooperative IoT networks with short packet communication," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 11 549–11 562, December 2018.

[37] Y. Hu, M. Serror, K. Wehrle, and J. Gross, "Finite blocklength performance of cooperative multi-terminal wireless industrial networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 5778–5792, July 2018.

[38] Y. Hu, A. Schmeink, and J. Gross, "Optimal scheduling of reliability-constrained relaying system under outdated CSI in the finite blocklength regime," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6146–6155, July 2018.

[39] P. Nouri, H. Alves, and M. Latva-aho, "Performance analysis of ultra-reliable short message decode and forward relaying protocols," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, August 2018.

[40] Y. Gu, H. Chen, Y. Li, and B. Vucetic, "Ultra-reliable short-packet communications: Half-duplex or full-duplex relaying?" *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 348–351, June 2018.

[41] C. Guo, L. Liang, and G. Y. Li, "Resource allocation for low-latency vehicular communications: An effective capacity perspective," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 905–917, April 2019.

[42] C. F. Liu and M. Bennis, "Ultra-reliable and low-latency vehicular transmission: An extreme value theory approach," *IEEE Communications Letters*, vol. 22, no. 6, pp. 1292–1295, June 2018.

[43] C. She, C. Liu, T. Q. S. Quek, C. Yang, and Y. Li, "Ultra-reliable and low-latency communications in unmanned aerial vehicle communication systems," *IEEE Transactions on Communications*, pp. 1–1, 2019.

[44] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy HARQ," *IEEE Wireless Communications Letters*, vol. 3, no. 5, pp. 529–532, October 2014.

[45] E. Dosti, M. Shehab, H. Alves, and M. Latva-aho, "Ultra reliable communication via CC-HARQ in finite block-length," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–5.

[46] J. P. B. Nadas, O. Onireti, R. D. Souza, H. Alves, G. Brante, and M. A. Imran, "Performance analysis of hybrid ARQ for ultra-reliable low latency communications," *IEEE Sensors Journal*, pp. 1–1, 2019.

[47] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2475–2485, November 2018.

[48] H. Shariatmadari, R. Duan, S. Iraji, Z. Li, M. A. Uusitalo, and R. Jäntti, "Resource allocations for ultra-reliable low-latency communications," *International Journal of Wireless Information Networks*, vol. 24, no. 3, pp. 317–327, September 2017.

[49] C. Sun, C. She, and C. Yang, "Retransmission policy with frequency hopping for ultra-reliable and low-latency communications," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.

[50] 3GPP Technical Documents R1-1808449, "IMT-2020 self-evaluation: UP latency analysis for FDD and dynamic TDD with UE processing capability 2 (URLLC)," August 2018.

[51] H. Shariatmadari, S. Iraji, R. Jantti, P. Popovski, Z. Li, and M. A. Uusitalo, "Fifth-Generation control channel design: Achieving ultra-reliable low-latency communications," *IEEE Vehicular Technology Magazine*, vol. 13, no. 2, pp. 84–93, June 2018.

[52] G. Ku and J. M. Walsh, "Resource allocation and link adaptation in LTE and LTE-Advanced: A tutorial," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1605–1633, third-quarter 2015.

[53] P. Popovski, . Stefanović, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Transactions on Communications*, pp. 1–1, 2019.

[54] A. Belogaev, E. Khorov, A. Krasilov, D. Shmelkin, and S. Tang, "Conservative link adaptation for ultra reliable low latency communications," in *2019 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, June 2019, pp. 1–5.

[55] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji, and R. Jantti, "Link adaptation design for ultra-reliable communications," in *2016 IEEE International Conference on Communications*, May 2016, pp. 1–5.

[56] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner, and Z. Li, "Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements," *IEEE Network*, vol. 32, no. 2, pp. 8–15, March 2018.

[57] 3GPP Technical Report 38.802, "Study on new radio access technology physical layer aspects," Version 14.0.0, March 2017.

[58] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 210–217, March 2018.

[59] G. Berardinelli, S. R. Khosravirad, K. I. Pedersen, F. Frederiksen, and P. Mogensen, "On the benefits of early HARQ feedback with non-ideal prediction in 5G networks," in *2016 International Symposium on Wireless Communication Systems (ISWCS)*, September 2016, pp. 11–15.

[60] B. Makki, T. Svensson, G. Caire, and M. Zorzi, "Fast HARQ over finite blocklength codes: A technique for low-latency reliable communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 194–209, January 2019.

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

[62] J. Chen, L. Zhang, Y. Liang, X. Kang, and R. Zhang, "Resource allocation for wireless-powered IoT networks with short packet communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 1447–1461, February 2019.

[63] C. Sun, C. She, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 402–415, January 2019.

[64] N. H. Mahmood, A. Karimi, G. Berardinelli, K. I. Pedersen, and D. Laselva, "On the resource utilization of multi-connectivity transmission for URLLC services in 5G new radio," *in Proc. 2019 IEEE Wireless Communications and Networking Conference (WCNC) Workshops*, April 2019.

[65] K. I. Pedersen, S. R. Khosravirad, G. Berardinelli, and F. Frederiksen, "Rethink hybrid automatic repeat request design for 5G: Five configurable enhancements," *IEEE Wireless Communications Magazine*, vol. 24, no. 6, pp. 154–160, December 2017.

[66] H. Shariatmadari, R. Duan, S. Iraji, R. Jäntti, Z. Li, and M. A. Uusitalo, "Asymmetric ACK/NACK detection for ultra - reliable low - latency communications," in *2018 European Conference on Networks and Communications (EuCNC)*, June 2018, pp. 1–166.

[67] S. Boyd, S. Kim, L. Vandenberghe, and A. Hassibi, "A tutorial on geometric programming," *Optimization and Engineering*, vol. 8, no. 1, p. 67, April 2007.

[68] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and sub-carrier allocation for full-duplex multi-carrier non-orthogonal multiple access systems," *IEEE Transactions on Communications*, vol. 65, no. 3, pp. 1077–1091, March 2017.

Paper H

On the Resource Utilization of Multi-Connectivity Transmission for URLLC Services in 5G New Radio

Nurul Huda Mahmood, Ali Karimi, Gilberto Berardinelli, Klaus I. Pedersen, and Daniela Laselva

The paper has been published in the
IEEE Wireless Communications and Networking Conference (WCNC) Workshops, 2019.

Abstract

Multi-connectivity with packet duplication, where the same data packet is duplicated and transmitted from multiple transmitters, is proposed in 5G New Radio as a reliability enhancement feature. This paper presents an analytical study of the outage probability enhancement with multi-connectivity, and analyses its cost in terms of resource usage. The performance analysis is further compared against conventional single-connectivity transmission. Our analysis shows that, for transmission with a given block error rate target, multi-connectivity results in more than an order of magnitude outage probability improvement over the baseline single-connectivity scheme. However, such gains are achieved at the cost of almost doubling the amount of radio resources used. Multi-connectivity should thus be selectively used such that its benefits can be harnessed for critical users, while the price to pay in terms of resource utilization is simultaneously minimized.

1 Introduction

The fifth generation (5G) New Radio (NR) wireless network standard introduced Ultra Reliable Low Latency Communication (URLLC) service class with the design goals of providing very high reliability and low latency wireless connectivity [1]. Its use cases include Industry 4.0 automation, communication for intelligent transport services and tactile Internet. Different design goals have been identified for different applications, with one of the more stringent targets being a 99.999% reliability (i.e. 10^{-5} outage probability) at a maximum one way data-plane latency of *one* millisecond (ms).

A number of solutions addressing the scheduling and resource allocation aspects of URLLC have been proposed in the literature. These include short transmission time intervals (TTI) [2], faster processing [3] and enhanced URLLC-aware scheduling solution including pre-emption [4].

On the other hand, examples of studies investigating URLLC from an analytical perspective include [5, 6], among others. Reference [5] breaks down URLLC into three major building blocks, namely: *(i) risk* representing decision making uncertainty, *(ii)* extreme values at the *tail* of a distribution affecting high reliability, and *(iii)* the *scale* at which various network elements requiring URLLC services are deployed. The authors then discuss various enablers of URLLC and their inherent tradeoffs, and present several mathematical tools and techniques that can be used to design URLLC solutions. Reference [6] analyse the reliability of uplink grant-free schemes, which have the potential of reducing the latency by avoiding the handshaking procedure for acquiring a dedicated scheduling grant, and demonstrate their latency benefits with respect to conventional grant-based approach.

The stochastic nature of the wireless channel is one of the main constraints

in achieving the stringent URLLC service requirements. Ensuring high reliability requires overcoming variations in the received signal strength caused by the channel. Diversity is a well proven technique in this regard [7]. It is now being revisited as a reliability improvement feature for URLLC services through Multi-Channel Access (MCA) solutions [8–10]. MCA is a promising family of radio resource management approach that allows a user equipment (UE) to be simultaneously served over multiple channels through one or more transmitting nodes. Carrier aggregation is an example of single-node MCA, whereas examples of multi-node MCA include joint transmission, multi-connectivity (MC) and downlink-uplink decoupling [8].

This paper specifically addresses reliability oriented MC with packet duplication, focusing on the downlink transmission direction. MC is a generalization of the dual-connectivity (DC) concept, first standardized in 3GPP release-12 as a throughput enhancement feature [11]. MC with packet duplication involves duplication of a packet destined for a particular UE, which is then transmitted to the UE through multiple transmitting nodes. The current 5G NR release-15 standard specifies that packet duplication has to occur at the Packet Data Convergence Protocol (PDCP) layer [1]. Transmissions from the individual nodes are independent at the lower layers, and thus can be transmitted with different transmission parameters, e.g., modulation and coding schemes (MCS). Reliability improvement with MC introduces transmission diversity that can overcome some of the causes of transmission failures, such as deep fades and/or strong interference.

Data duplication greatly enhances the probability of successfully receiving the data packet, albeit at the cost of increased resource usage. Recently, the 3GPP has acknowledged the resource efficiency challenge of PDCP duplication and is studying how to improve the packet delivery efficiency in future releases [12]. Several solutions have been discussed, comprising selective duplication to minimize the used resources for duplicates, and timely discarding of redundant duplicates. These enhancements are seen even more crucial when envisioning the extension of the number of radio links participating to the packet delivery or to the number of simultaneous duplicates as compare to release-15.

This paper provides a thorough analysis of MC. In particular, the reliability improvement with MC, measured in terms of the outage probability gain, is analytically derived. In addition, the operational cost of MC in terms of resource utilization is also analysed. Due to the limited space, a detailed analysis of the transmission latency could not be included in this work. The latency aspect of URLLC is implicitly covered by the considered 5G numerology, which allows at most a single hybrid automatic repeat request (HARQ) retransmission within the assumed one ms latency budget.

The rest of the paper is organized as follows. The system model is introduced in Section 2. Section 3 and Section 4 present a detailed analysis of the

outage probability enhancement and the corresponding resource usage with baseline SC and reliability-oriented MC, respectively. Numerical results are then presented in Section 5 followed by concluding remarks in Section 6.

2 Setting the Scene

2.1 Latency Components

The downlink one-way latency of a given transmission (Y) is defined from the time a payload arrives at the lower layer of the transmitting base station (BS), until it is successfully decoded at the UE. If the UE correctly decodes the packet in the first transmission, the latency is that of a single transmission, as illustrated in Figure H.1, and is given by

$$Y = t_{fa} + t_{bp} + t_{tx} + t_{up}, \tag{H.1}$$

where t_{fa} is the frame alignment delay. The payload transmission time is denoted by t_{tx}. The processing times at the BS and the UE are represented by t_{bp} and t_{up}, respectively.

The frame alignment delay is a random variable uniformly distributed between *zero* and *one* TTI. Depending on the packet size, channel quality and scheduling strategy, the transmission time t_{tx} can vary from one to multiple TTIs. Considering the small payload of URLLC traffic, we assume $t_{tx} = 1$ TTI in this work. The processing time at the UE (t_{up}) is also assumed to be one TTI.

In the case of failure to successfully decode the data message in the first transmission, a fast HARQ mechanism ensures quick retransmission of the message. In this case, the transmission is subject to additional delay(s), which includes the additional time it takes to transmit a negative acknowledgement (NACK), process it at the BS and schedule the packet for retransmission. The HARQ round trip time (RTT) t_{HARQ}^{RTT}, defined from the start time of the first transmission until the start time of the retransmission, is assumed to be *four* TTIs.

2.2 System Assumptions

In order to meet the stringent latency target of URLLC services, a very flexible frame structure for 5G NR offering different options to shorten the TTI duration, as compared to LTE, is introduced by 3GPP [13]. In particular, the subcarrier spacing (SCS) can be expanded up to 480 kHz (note, SCS of 480 kHz is specified but not supported in release-15 [2]), thereby reducing the minimum scheduling interval considerably. In addition, 'mini-slots' are introduced whereby the number of orthogonal frequency-division multiplexed

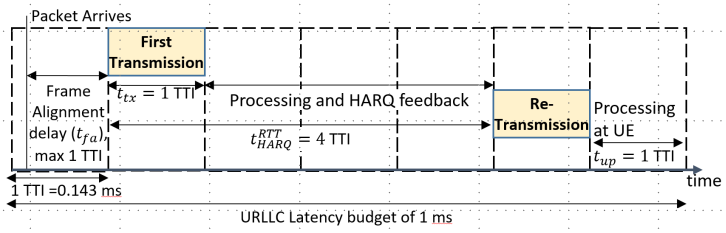

Fig. H.1: URLLC latency budget of one ms can accommodate maximum one HARQ retransmission at four OFDM symbols mini-slot with 30 kHz sub-carrier spacing, corresponding to 0.143 ms TTI.

(OFDM) symbols per TTI can also vary. In contrast with the LTE slot duration of 14 OFDM symbols per TTI, mini-slots in 5G NR can compose of $1-13$ symbols. The recommended mini-slot lengths are two, four and seven symbols, corresponding to TTIs of $0.07, 0.143$ and 0.25 ms at 30 kHz SCS. This allows shorter transmission slots without increasing the SCS, which is particularly suitable for low frequency bands.

In this work, we assume a four symbol mini-slot at 30 kHz SCS, resulting in a transmission duration of 0.143 ms. This leaves sufficient time budget for the first transmission, processing at the UE and a single HARQ retransmission (if needed) within the one ms latency target for URLLC services, as illustrated in Figure H.1. In case of retransmission(s), multiple retransmitted packets are combined using Chase combining, resulting in a boost in the desired signal power [14].

In order to further enhance low latency support, we adopt the *in-resource control signaling* proposed in [15] along with front-loading of the demodulation reference symbol (DMRS) [16]. The main idea is to embed control information on the fly at the start of time-frequency resources allocated to the user in the downlink. This allows the metadata to be processed and decoded as soon as it is received, i.e. while the data is still being received. From the latency perspective, this is advantageous as it allows performing channel estimation earlier and can enable early HARQ feedback as detailed in [17]. HARQ feedbacks are always assumed to be received correctly.

We assume that the metadata (i.e., the control information needed to decode the transmission) and the data for the l^{th} transmission are encoded separately with different target block error rates (BLER) given by $P_e^{m,l}$ and $P_e^{d,l}$, respectively; where $l \in \{1,2\}$. Upon Chase combining following a retransmission, the data outage probability is given by $P_e^{d,c}$. Note that, $P_e^{d,l} > P_e^{d,c}\ \forall l \in \{1,2\}$. Table H.1 provides an overview of the different outage probabilities introduced and derived in this contribution.

In terms of MC operation, we assume that the high priority URLLC packets are scheduled at the secondary node immediately upon arrival at the

PDCP layer. Thus, transmission through each of the secondary nodes can also accommodate a single retransmission within the one ms latency budget, if needed.

Table H.1: Introduced and derived outage probabilities

Introduced Outage Probabilities:	
$P_e^{m,l}$	Metadata BLER target in the l^{th} transmission
$P_e^{d,l}$	Data BLER target in the l^{th} transmission
$P_e^{d,c}$	Data error probability after Chase combining following retransmission
Derived Outage Probabilities	
P_{out}^{SC}	Outage probability of the baseline single-connectivity.
P_{out}^{MC}	Outage probability of MC with data duplication.

3 Reliability Enhancement with Multi-Connectivity

In this section, we present an analytical derivation of the reliability enhancement with MC with PDCP duplication as defined in 3GPP standard [9], and compare it against the baseline single-connectivity (SC) transmission. Ideal link adaptation is assumed, i.e. the achieved outage probabilities after transmission are assumed to be the same as the transmission BLER targets.

3.1 Baseline Outage Probability with Single-Connectivity

The outage probability with SC considering separate BLER targets for the metadata and data is analyzed first. Due to its critical nature, we assume that the metadata is encoded with a lower BLER target, i.e. $P_e^{m,l} < P_e^{d,l}$.

Under conventional eMBB transmissions, the BLER target of the data part is much higher than that of the metadata (i.e., $P_e^{d,l} \gg P_e^{m,l}$). Hence, the impact of the metadata outage on the overall outage probability is negligible, and the outage probability can be readily approximated by the data BLER target $P_e^{d,l}$. However, the same cannot be assumed for URLLC services requiring high reliability where the data is also transmitted with a stringent BLER target.

The events that can occur upon transmission are depicted in Figure H.2. There are three different possible outcomes of processing the first trans-

mission at the receiver: failure to decode the metadata (with probability $P_e^{m,1}$), metadata decoded but failure to decode the data (with probability $\left(1-P_e^{m,1}\right)P_e^{d,1}$) and successful decoding of the data packet in the first attempt. The probability of success in the first transmissions is given by

$$P_{succ}^{SC,1} = \left(1-P_e^{m,1}\right)\left(1-P_e^{d,1}\right). \tag{H.2}$$

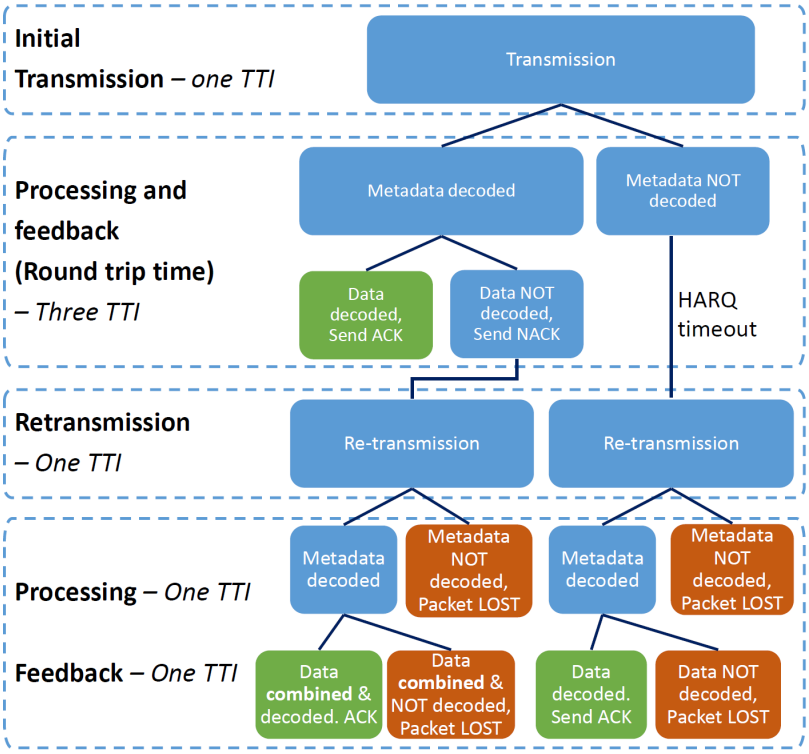

Fig. H.2: Difference possible reception events in the single-connectivity scenario.

A HARQ NACK cannot be transmitted if the metadata is not successfully decoded in the first transmission. This leads to a HARQ *time out*, which occurs when a HARQ feedback (ACK/NACK) is not received within a predefined time interval. The transmitter then retransmits the packet assuming the initial transmission failed. However, there is no possibility of Chase combining in this case since the control information needed to identify the packet in the first transmission was not successfully decoded. Thus, the success probability with retransmission following a HARQ time out is given by

$$P_{succ}^{SC,2-TO} = P_e^{m,1}\left(1-P_e^{m,2}\right)\left(1-P_e^{d,2}\right). \tag{H.3}$$

In this study, we set the HARQ *time out* time and the time it takes to process the retransmission at the BS (t_{bp}) to be equal to three TTIs, thus ensuring the same retransmission latency as that with HARQ retransmission.

In the event of receiving a NACK, the packet is retransmitted. Reception of the retransmitted data can be attempted by Chase combining with the initially received data whose metadata was successfully decoded in the first transmission. In this case, the probability of successful decoding is $\Pr[\gamma_c \geq \gamma_t]$, where $\gamma_c = \gamma_1 + \gamma_2$ is the achieved signal to interference plus noise ratio (SINR) following Chase combining. The SINR of the l^{th} transmission, and the target SINR, are denoted by γ_l and γ_t, respectively. The decoding success probability in this case of HARQ retransmission, $P_{succ}^{SC,2-NR}$, is given by

$$P_{succ}^{SC,2-NR} = \left(1 - P_e^{m,1}\right) P_e^{d,1} \left(1 - P_e^{m,2}\right) \times \left(1 - \Pr[\gamma_c < \gamma_t | \gamma_1 < \gamma_t]\right). \quad \text{(H.4)}$$

Using Baye's rule, $\Pr[\gamma_c < \gamma_t | \gamma_1 < \gamma_t]$ can be expressed as $\Pr[\gamma_c < \gamma_t, \gamma_1 < \gamma_t] / \Pr[\gamma_1 < \gamma_t]$. Since $\gamma_1 < \gamma_c$, we have $\Pr[\gamma_c < \gamma_t, \gamma_1 < \gamma_t] = P_e^{d,c}$. Hence, $P_{succ}^{SC,2-NR}$ can be further simplified to

$$P_{succ}^{SC,2-NR} = (1 - P_e^{m,1})(1 - P_e^{m,2})(P_e^{d,1} - P_e^{d,c}). \quad \text{(H.5)}$$

The success probability following a retransmission is then equal to the sum of the success probability of retransmission after time-out, and that upon retransmission following a NACK. The sum can be expressed as

$$P_{succ}^{SC,2} = \left(1 - P_e^{m,2}\right) \left[P_e^{m,1} \left(1 - P_e^{d,2}\right) + \left(1 - P_e^{m,1}\right) \left(P_e^{d,1} - P_e^{d,c}\right)\right]. \quad \text{(H.6)}$$

Consequently, the total outage probability for the baseline SC scenario is

$$P_{out}^{SC} = 1 - P_{succ}^{SC,1} - P_{succ}^{SC,2}. \quad \text{(H.7)}$$

3.2 Outage Probability Analysis in Multi-Connectivity Scenario

We now analyze the outage probability of MC considering data duplication at the PDCP layer, as defined in 3GPP release-15 [9]. In this MC variant, data packets are duplicated and shared between the master node and the secondary node(s) at the PDCP layer. The packets are then transmitted independently from each node, i.e., they can have different MCS and transmitted over different resource blocks (RB). At the UE end, the lower layers up to the radio link control layer treat each of the packet received from the different nodes as separate packets and attempt to decode them individually. Successfully received packets are then forwarded to the PDCP layer. If multiple copies are successfully decoded, the PDCP layer keeps the first received packet while discarding any later copies.

Assuming independent transmissions of the same data packet over M nodes, the packet is lost if it is not successfully decoded from any of the M nodes. Hence, the outage probability is given by

$$P_{out}^{MC}(M) = \prod_{n=1}^{M} P_{out,n}^{SC}, \tag{H.8}$$

where $P_{out,n}^{SC}$ is the outage probability through the n^{th} node, and can be evaluated using Eq. (H.7). In the case of identical outage probabilities through all links (i.e., transmission with the same BLER targets from all nodes), the outage probability further simplifies to $P_{out}^{MC}(M) = \left(P_{out}^{SC}\right)^{M}$.

4 Resource Usage Analysis

This section evaluates the resource usage of the baseline SC and MC with PDCP duplication using results from finite blocklength theory [18]. The number of information bits L that can be transmitted with decoding error probability P_e in R channel use in an additive white Gaussian noise (AWGN) channel with a given SINR γ is

$$L = RC(\gamma) - Q^{-1}(P_e)\sqrt{RV(\gamma)} + \mathcal{O}(\log_2 R), \tag{H.9}$$

where $C(\gamma) = \log_2(1+\gamma)$ is the Shannon capacity of AWGN channels under infinite blocklength regime, $V(\gamma) = \frac{1}{\ln(2)^2}\left(1 - \frac{1}{(1+\gamma)^2}\right)$ is the channel dispersion (measured in squared information units per channel use) and $Q^{-1}(\cdot)$ is the inverse of the Q-function. Using the above, the channel usage R can be approximated as [19]

$$R \approx \frac{L}{C(\gamma)} + \frac{Q^{-1}(P_e)^2 V(\gamma)}{2C(\gamma)^2} \times \left[1 + \sqrt{1 + \frac{4LC(\gamma)}{Q^{-1}(P_e)^2 V(\gamma)}}\right]. \tag{H.10}$$

For a given transmission schemes $\Pi \in \{SC, MC\}$, we first evaluate the BLER target P_e^{Π} that can achieve the desired 10^{-5} URLLC outage probability using the equations derived in Sections 3.1 and 3.2. The channel use per single transmission, R_{Π}, is then calculated by inserting the corresponding values of P_e^{Π} into Eq. (H.10). Finally, the total resource usage including the effect of retransmission is evaluated henceforth.

Single-Connectivity

For single connectivity, the resource usage is R_{SC} with probability $P_{succ}^{SC,1}$ and $2R_{SC}$ with probability $1 - P_{succ}^{SC,1}$. Hence, the total resource usage, $\mathcal{U}_{SC}$, is

straightforwardly obtained as

$$\mathcal{U}_{SC} = \left(2 - P_{succ}^{SC,1}\right) R_{SC}. \tag{H.11}$$

Multi-Connectivity

For MC with PDCP duplication through M nodes, each transmissions are independent with a retransmission occurring in the case of failure of that transmission. Hence, the transmission (or retransmission) through a given node is not cancelled even if the packet has already been correctly decoded from the transmission through other nodes. Thus, MR_{MC} channel uses are used if the initial transmissions through all M nodes are successful, $(M+1)R_{MC}$ channel uses are used if one initial transmission fails, and so on. (Here, we assume, for the ease of presentation, the achieved SINR through all M nodes are the same. This can happen, e.g., when considering a UE at the cell edge with equal received power from multiple BSs. In general, the channel use for each node can be obtained by inserting the appropriate SINR value in Equation (H.10). The corresponding resource usage can then be calculated easily using any numerical computing software.)

In other words, the channel usage is a random variable that can take the values $(M+n)R_{MC}$, for $n = 0 \dots M$, with probability $\binom{M}{n} \left(P_{succ}^{SC,1}\right)^{M-n} \left(1 - P_{succ}^{SC,1}\right)^{n}$. The total resource usage can then be calculated by summing the above for $n = 0 \dots M$. After some algebraic manipulation, this yields

$$\mathcal{U}_{MC} = M \left(2 - P_{succ}^{SC,1}\right) R_{MC}. \tag{H.12}$$

5 Numerical Results

This section presents numerical validation of the derived analytical results. For simplicity, we assume that the outage probabilities remain unchanged over the initial transmission and the retransmission, i.e. $P_e^{d,1} = P_e^{d,2} = P_e^d$ and $P_e^{m,1} = P_e^{m,2} = P_e^m$. This is a reasonable assumption as the time between retransmissions is very short. Moreover, the assumed payload and metadata size is 32 and 16 Bytes, respectively.

5.1 Outage Probability as a Function of P_e^d

The derived outage probabilities with SC and MC are presented as a function of the BLER target on the data channel $\left(P_e^d\right)$ in Figure H.3. Two different metadata BLER targets are considered, namely $P_e^m = {}^{P_e^d}/_2$ and $P_e^m = 1\%$. MC results are evaluated with $M = 2$ and $M = 3$.

With SC, the outage probability remains above the URLLC target of 10^{-5} even for P_e^d as low as 1%. In fact, 10^{-5} outage is only achieved with P_e^m and P_e^d at about 0.15% and 0.3%, respectively. However, the targeted reliability can be met at much higher BLER targets with MC transmission, even when only two nodes are involved (i.e., $M = 2$).

Comparing the outage probabilities, we can observe a large gap between the performance of SC and the MC schemes. This indicates that there is a clear advantage in terms of the outage probability in transmitting multiple copies of the packet, especially at the levels targeted for URLLC applications. Since decoding the metadata is more critical than the data itself, we observe clear advantage in having a lower BLER target for the metadata.

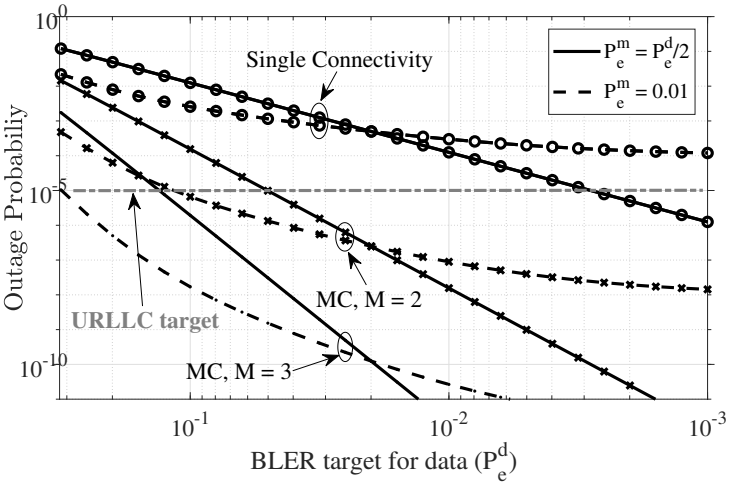

Fig. H.3: Outage Probabilities with single-connectivity and multi-connectivity transmission for $P_e^m = [P_e^d/2, 0.01]$ and $M = [2, 3]$.

5.2 Resource Usage Analysis

Figure H.4 presents the resource usage and corresponding outage probabilities for SC and MC transmission scheme with $M = 2$, where the BLER targets for metadata and data are fixed at 1% and 10% respectively. Since the same BLER targets are assumed for all schemes, the resource usage is normalized by the resource utilization for a single transmission. Please note that we assume the same SINR is achieved through the master and the secondary node, as discussed earlier. This is only to facilitate the analytical derivations and obtain meaningful insights into the performance trend. Performance results with different SINRs can easily be evaluated numerically.

Single-connectivity is expectedly the most resource efficient, though this

comes at an outage probability that is several orders of magnitude higher than the MC scheme. Thus, the price to pay for the higher reliability with MC is the almost doubling of the resource usage and additional signalling overhead. Note however that, resource efficiency is not the main performance indicator in many applications requiring high reliability. Nonetheless, this provides a strong motivation for investigating more resource efficient MC schemes.

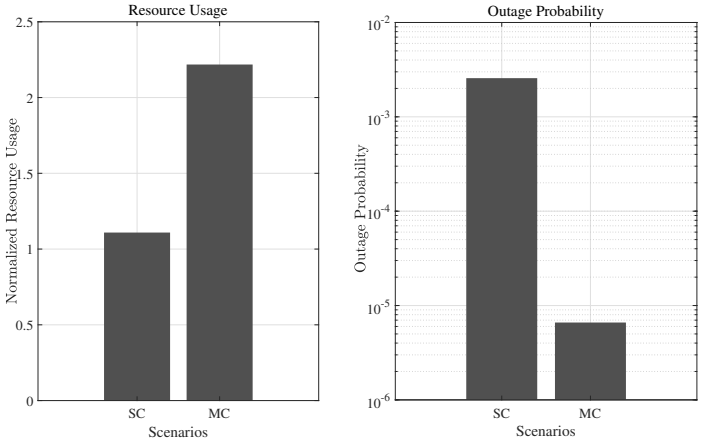

Fig. H.4: Resource usage and corresponding outage probabilities with SC and MC, $P_e^m = 1\%$, $P_e^d = 10\%$ and $M = 2$.

In the rest of this section, we calculate the resource utilization at $1 - 10^{-5}$ reliability target using the results derived in Section 4. Table H.2 shows the required BLER targets for achieving $1 - 10^{-5}$ reliability target for SC and MC transmission scheme with $M = 2$, assuming $P_e^m = P_e^d$. The corresponding channel use per transmission at an SINR of 10 dB derived using results from finite blocklength theory and the final resource usage with retransmission are also listed. We observe that the required BLER target with SC is more than an order of magnitude lower than that with MC.

Figure H.5 presents the total resource usage in terms of the channel use with a BLER target set to achieve 10^{-5} outage probability for SINRs of 0 and 10 dB. Even after taking into account the lower BLER targets required for achieving 10^{-5} outage probability, SC is found to be more resource efficient compared to the considered MC scheme. In fact, 46% to 48% less resources are required with SC, depending on the SINR value. However, the resourced required to achieve a given reliability target with SC may not always be available at a given node. Furthermore, the success probability following the single transmission is much higher with MC, meaning that it has clear advantages in applications with a tight latency budget where even

a single retransmission cannot be accommodated [20].

Table H.2: Resource usage at $1 - 10^{-5}$ reliability target

Tx. scheme (Π)	*BLER target*	R_{Π}	$\mathcal{U}_{\Pi}$
Single-Connectivity	0.183%	85.14	85.44
Multi-Connectivity	3.28%	80.88	166.12

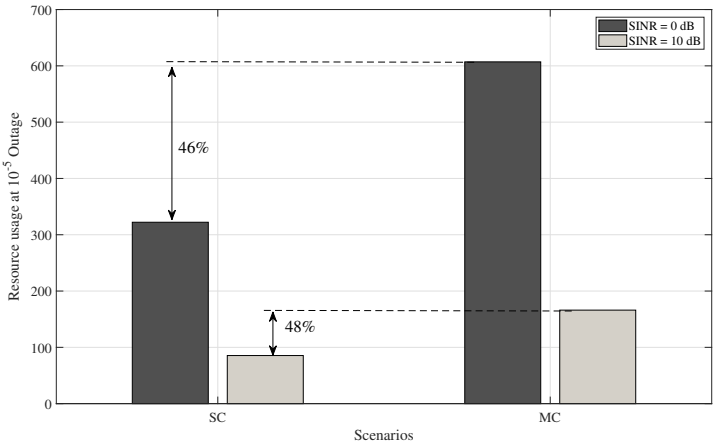

Fig. H.5: Resource usage of the single-connectivity and multi-connectivity scheme at $1 - 10^{-5}$ reliability.

6 Conclusions

Multi-connectivity is proposed as a potential reliability enhancement solution for URLLC applications. The outage probability considering MC with PDCP duplication as defined by 3GPP is derived and compared against baseline SC scheme in this paper. In contrast with existing works, the reliability of the control channel is specifically considered in the outage probability evaluation. The corresponding resource usages are also derived. Collectively, the derived outage probability and resource usage analysis allow comparing the cost-performance trade-offs of MC as a reliability enhancement solution for URLLC services.

The obtained analytical results show that MC can greatly enhance the outage probability at the expense of increased resource usage. In particular, the outage probability is enhanced by several orders of magnitude, at the

expense of almost doubling of the resource usage. From a resource utilization perspective, single-connectivity is more resource efficient. However, MC is more desirable from a reliability aspect since the reliability levels targeted for URLLC applications may not always be possible with SC. Furthermore, it has clear advantages in scenarios where even a single retransmission cannot be accommodated, for example in certain industrial use cases with less than one ms latency requirements. Our future work includes investigating more resource efficient MC transmission schemes.

References

[1] 3GPP TS 38.300, "NR; Overall description; Stage-2," Jun. 2018, v:15.2.0.

[2] 3GPP TR 38.912, "Study on New Radio (NR) access technology," Jul. 2018, v:15.0.0.

[3] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Comm. Mag.*, vol. 56, no. 3, pp. 210–217, Mar. 2018.

[4] A. Karimi *et al.*, "5G centralized multi-cell scheduling for URLLC: Algorithms and system-level performance," *IEEE Access*, vol. 6, pp. 72 253–72 262, Nov. 2018.

[5] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail, risk and scale," *ArXiv e-prints*, Jan. 2018.

[6] G. Berardinelli *et al.*, "Reliability analysis of uplink grant-free transmission over shared resources," *IEEE Access*, vol. 6, pp. 23 602–23 611, 2018.

[7] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, UK: Cambridge University Press, 2005.

[8] N. H. Mahmood *et al.*, "Multi-channel access solutions for 5G new radio," *CoRR*, vol. abs/1902.01948, 2019.

[9] 3GPP TS 37.340, "E-UTRA and NR; Multi-connectivity; Stage-2," Dec. 2017, v:15.0.0.

[10] N. H. Mahmood *et al.*, "Reliability oriented dual connectivity for URLLC services in 5G New Radio," in *15th International Symposium on Wireless Communication Systems (ISWCS)*, Lisbon, Portugal, Aug. 2018, pp. 1–6.

[11] C. Rosa *et al.*, "Dual connectivity for LTE small cell evolution: functionality and performance aspects," *IEEE Comm. Mag.*, vol. 54, no. 6, pp. 137–143, Jun. 2016.

[12] 3GPP TR 38.825, "Study on NR industrial Internet of Things (IoT)," Nov. 2018, Rel-16, v:0.0.1.

[13] 3GPP TS 38.211, "5G NR; Physical channels and modulation," Jul. 2018, v:15.2.0.

[14] P. Frenger, S. Parkvall, and E. Dahlman, "Performance comparison of HARQ with Chase combining and incremental redundancy for HS-DPA," in *Proc. VTC Fall 2001*, Oct. 2001, pp. 1829–1833.

[15] K. Pedersen *et al.*, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Comm. Mag.*, vol. 54, no. 3, pp. 53–59, Mar. 2016.

[16] N. H. Mahmood *et al.*, "Radio resource management techniques for eMBB and mMTC services in 5G dense small cell scenarios," in *Proc. IEEE VTC-Fall*, Montreal, Canada, Sep. 2016, pp. 1–5.

[17] G. Berardinelli *et al.*, "Enabling early HARQ feedback in 5G networks," in *Proc. IEEE VTC Spring 2016*, Nanjing, China, May.

[18] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[19] A. Anand and G. de Veciana, "Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2411–2421, Nov. 2018.

[20] G. Berardinelli, N. H. Mahmood, I. Rodriguez, and P. Mogensen, "Beyond 5G wireless IRT for Industry 4.0: Design principles and spectrum aspects," in *IEEE Globecom Workshops*, Dec. 2018.

Part V

Conclusions

Conclusions

1 Summary of the Main Findings

A broad set of enhancements and novel ideas are required to accommodate different services in the fifth generation (5G) cellular networks. This PhD dissertation focused on radio resource management (RRM) solutions to serve enhanced Mobile Broadband (eMBB) and Ultra-Reliable Low-Latency Communications (URLLC). We studied an extensive set of RRM strategies, including packet scheduling schemes, through advanced theoretical framework and system-level evaluations.

First, inter-cell interference was studied as one of the main limiting factors towards network densification. A new subspace coordinated scheme motivated by interference alignment (IA) was introduced. The main design concepts of the proposed solution were: distributed implementation, based on only local channel knowledge, and simple minimum mean square error (MMSE) receiver structure. Monte-Carlo based performance evaluations showed that in ultra-dense networks, up to 28% throughput gain is achieved in comparison to the well-known *MaxSNR* precoding. In macro scenario (where the interference level is lower), the benefits decrease. So, it is a less attractive option for practical implementation in interference limited scenarios.

Distributed co-scheduling of URLLC and eMBB for downlink (DL) transmission was studied as the second step. A system model was developed by applying 3GPP Release 15 specifications for New Radio (NR). To reduce URLLC latency and enhance eMBB throughput, a new service-based packet scheduling policy was proposed. It was shown that having elements of channel quality, service type, payload size, and control channel information leads to superior resource allocation decisions. The performance was evaluated by running advanced system-level simulations. The results indicated that the proposed solution significantly improves both URLLC and eMBB performance. At the 99.999% reliability level, it offers up to 98% URLLC latency improvement in comparison to the well-known proportional fair (PF) scheduling. For eMBB traffic, 12% throughput gain is achieved.

System-level simulations results indicate that at low URLLC offered load, transmission time, processing times, and HARQ delays are the main latency components. While, at high load, queuing delay becomes dominant and threatens URLLC. As one solution to address this issue, centralized radio access network (C-RAN) was investigated in the third step. Here, each mobile terminal simultaneously measures and connects to multiple cells. At any given time, the C-RAN dynamically schedules the user from one of the connected cells based on parameters such as channel quality, load condition, etc. spectrum-efficient dynamic point selection (SE-DPS) was studied as one of the possible solution for C-RAN. This feature relies on the achieved multi-site diversity gain. System-level analyses show that SE-DPS reduces the number of required resources for scheduling the cell-edge users. Thereby, it improves queuing delay and URLLC latency when traffic increases.

To achieve the best C-RAN performance for URLLC, carried load per cell should be dynamically balanced. A low-complexity algorithm was proposed to minimize the queuing delay by fast switching users from congested points to lightly loaded secondary cells. Simulation analyses revealed that the proposed solution offers significant URLLC improvement. It was shown that at 99.999% percentile reliability, C-RAN provides up to 99% lower latency in comparison to the distributed solution.

Theoretical studies of the URLLC were conducted next. Fundamentals of URLLC in downlink transmission were investigated in the fourth step. We explored two multiplexing schemes for data and control channel allocation, namely as *in-resource control signalling* and *joint encoding of data and control information*. Expressions for the resource usage and success probability were derived. It was shown that overhead and error of sending control channel are important and have significant impacts on the spectrum efficiency and reliability of transmission. Moreover, it is essential to have reliable feedback channel. Joint link adaptation of data and control information was investigated to optimize the throughput while satisfying URLLC constraints. Solution based on successive convex optimization was proposed. Numerical evaluations showed that the proposed solution performs close to the optimal one, and provides up to 27% resource efficiency gain compared to single shot transmission.

Finally in the last step, the application of PDCP packet duplication was studied for URLLC as a reliability enhancement solution. A theoretical framework was presented. The outage probability and resource usage were driven by taking into account the overhead and reliability of the control channel. The obtained results indicated that packet duplication significantly enhances the outage probability by several orders of magnitude. The gain comes at the price of more resource usage. Comparing the result with single node transmission, it was observed that when the channel and interference knowledge are available at both transceivers, and accurate link adaptation can be

conducted, it is more efficient to adopt single transmission with more conservative modulation and coding scheme (MCS).

2 Recommendations

Based on main findings of this PhD study, the following presents recommendations to answer the research questions stated in Part I:

Q1 How to design a distributed IA solution for 5G?

R1 The proposed *inter-cell interference sub-space coordination (ICISC)* offers remarkable throughput gain for ultra-dense deployment. However, the benefits decrease in urban macro systems.

Q2 How to allocate radio resources to URLLC and eMBB for distributed NR implementation?

R2 Efficient multiplexing of URLLC and eMBB significantly improves the performance for both services. The latency budget needs to be considered as the main factor for time-domain scheduling. Throughout the frequency-domain scheduling, it is essential to achieve the maximum gain of multi-user diversity, while avoiding the segmentation of URLLC payloads. If the available resources are not enough for one URLLC packet, it is recommended to segment a payload with the lowest control channel overhead (i.e., highest throughput).

Q3 How to best utilize C-RAN for URLLC in 5G NR?

R3 Centralized multi-cell scheduling significantly enhances URLLC. The main benefit of C-RAN comes from its capability for fast switching between serving cells and performing instant dynamic load balancing. A proper solution for URLLC should have elements of scheduling budget, multi-cell CQI, and offered load.

Q4 How to multiplex data and control information for URLLC in downlink transmission?

R4 For accurate URLLC evaluation, it is important to take into account the overhead and error of control information, and also impairments in feedback channel. Joint link adaptation of data and control channel provides substantial resource efficiency gain while satisfying URLLC. Joint encoding of data and control channel needs less resources in comparison to in-resource control signalling. However, it suffers from high computational complexity (processing time) at the user-side. Thus, it is recommended to adopt in-resource control signalling for URLLC.

Q5 How data duplication impacts URLLC?

R5 Packet duplication significantly enhances the reliability at the cost of high resource usage. Therefore, its applications and scenarios need to be optimized. When accurate channel and interference information are available at both the base station and mobile terminal, it is recommended to use single-cell transmission and apply dynamic link adaptation.

3 Future Work

There are several RRM techniques and design optimizations to enhance multi-service performance in 5G. Based on the achieved knowledge through this PhD, we list some interesting research topics that could be addressed in future studies.

- Further study of centralized inter-cell interference coordination for co-existence of URLLC and eMBB traffic.
- Packet scheduling and URLLC performance evaluation through millimetre wave frequencies is potentially one interesting research topic.
- Investigating more advanced centralized solutions inspired by machine learning could be considered as one direction for future studies.
- It would be interesting to perform system-level evaluation of centralized scheduling for different configurations (e.g., advanced receiver types or a higher number of antennas) and scenarios (e.g. industrial environments).
- End to end network optimization and packet scheduling for both uplink and downlink in time division duplex (TDD) mode have great potential for future investigations.
- It is highly desirable to evaluate and optimize packet duplication for different channel models (e.g., Rayleigh, Rician fading) and levels of knowledge.
- Finally, exploiting the potential of different multi-node solutions such as coherent joint transmission will help to enhance and optimize future URLLC system design.

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-521-5

AALBORG UNIVERSITY PRESS