E2E-Aware Multi-Service Radio Resource Management for 5G New Radio

Radio Access and Resource Management Solutions

Abdul-Mawgood Ali Ali Esswie, Ali

Publication date:
2020

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

E2E-AWARE MULTI-SERVICE RADIO RESOURCE MANAGEMENT FOR 5G NEW RADIO

RADIO ACCESS AND RESOURCE MANAGEMENT SOLUTIONS

BY
ALI ABDELMAWGOOD ALI ALI ESSWIE

DISSERTATION SUBMITTED 2020

AALBORG UNIVERSITY
DENMARK

E2E-Aware Multi-Service Radio Resource Management for 5G New Radio

Radio Access and Resource Management Solutions

Ph.D. Dissertation

Ali Abdelmawgood Ali Ali Esswie

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

Dissertation submitted: August 2020

PhD supervisor: Prof. Preben Mogensen
Aalborg University

PhD Co-supervisor: Prof. Klaus Pedersen
Aalborg University

PhD committee: Associate Professor Carles Navarro Manchón (chair)
Aalborg University

Associate Professor Olav Tirkkonen
Aalto University

Associate Professor Henrik Lehrmann Christiansen
Technical University of Denmark

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

Department: Department of Electronic Systems

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-695-3

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2020

This thesis is dedicated

...to the souls of my Father and Mother.
...to my beloved Wife.

Curriculum Vitae

Ali A. Esswie

Ali A. Esswie received his MSc. degree of the electrical and computer engineering from Memorial University, Canada in 2017. Since October 2017, he has been pursuing his PhD degree with the department of electronic systems at Aalborg University, and employed by Nokia Bell Labs, Aalborg. From 2013 to 2016, he acted as a wireless standards engineer with Intel Labs and Huawei Technologies, respectively. He has authored more than 25 publications and 12 patent filings. His main research interests include 5G new radio, wireless machine learning/AI, radio resource management, ultra-reliable and low latency communications, massive MIMO and channel estimation.

Abstract

The fifth generation (5G) of the cellular technology offers greater support for three main service classes; the ultra-reliable and low-latency communications (URLLC), enhanced mobile broadband (eMBB), and the massive machine type communication (mMTC). URLLC services require the transmission of sporadic and small-payload packets with stringent radio latency and reliability targets. The eMBB applications demand wide-band transmis-sions with extreme peak data rates. Finally, for mMTC, the network is required to simultaneously serve a large number of connected devices, each is associated with strict energy consumption constraints. However, there is a fundamental tradeoff between the achievable latency, reliability, and network spectral efficiency. Concurrently optimizing the quality of service (QoS) of those service classes is one of the major challenges of the 5G new ra-dio and neither been addressed for the former wireless standards. Furthermore, the 5G new radio is designed to support both the frequency and time division duplexing (FDD, TDD) modes. And due to the abundantly available bandwidth at the 3.5 GHz unpaired spectrum, most of the early 5G deployments are envisioned with the TDD duplexing technology. However, achieving such an efficient multi-service-aware resource management is further challenging with TDD. The broader scope of this PhD. project is to research and develop novel and multi-service-aware radio resource management algorithms for multi-QoS 5G networks, spanning both FDD and TDD modes.

The first part addresses the multi-QoS (URLLC-eMBB) multiplexing problem. A QoS-aware multi-user multiple-input multiple-output (MU-MIMO) downlink scheduler is developed based subspace projections. The key idea is to eliminate the scheduling queuing delay of the newly-arriving URLLC packets in case the sufficient radio re-sources are not immediately available. The incoming URLLC transmissions are instantly paired with the active eMBB users which spatial signatures are closest possible to a pre-defined subspace. To control the inter-user interference at the critical URLLC users, the co-scheduled eMBB transmissions are spatially projected on-the-fly into an arbitrary spatial sub-space, to which the paired URLLC users align their respective transceivers into the orthonormal subspace, exhibiting substantially-

zero eMBB interference. Moreover, we have developed several variants of the proposed scheduler for eMBB capacity recovering and spectral efficiency optimization. We adopt highly-detailed system level simulations, with a high degree of realism in line with 3GPP NR assumptions, to evaluate the performance of the proposed schemes. Our simulation results demonstrate considerable improve-ments of the URLLC outage latency and the network capacity, e.g., minimizing the URLLC outage latency by 50 percent while enhancing the network capacity by 79 percent, compared to Rel-15 standard URLLC scheduler.

In the second part of the study, we target achieving the stringent URLLC outage targets in TDD 5G networks. We first demonstrate that the URLLC QoS is further harder to achieve in TDD deployments, mainly due to the TDD frame structure, i.e., no simultaneous downlink and uplink transmissions are possible, and the severe cross-link interference (CLI) when neighboring base-stations or users are adopting opposite transmission directions. A diversity of novel inter-cell coordination schemes are developed for mitigation of the critical CLI. Those schemes incorporate a new set of TDD system design improvements such as semi-static frame configuration, sliding frame-book design, joint hybrid frame design and slot-aware user scheduling, and coordinated transceiver design. Accordingly, developed coordination techniques offer a wide variety of the required inter-cell signaling over-head, TDD frame adaptation flexibility, and the achievable URLLC outage performance. Our results show a no-table URLLC outage improvement compared to standard dynamic TDD setups, e.g., 80 percent URLLC outage latency reduction.

Backed by our former conclusions, the last part of the PhD project demonstrates the potential of adopting a machine learning (ML) algorithms for real-time selection of the TDD radio frame structure. A simple, but efficient, Q-reinforcement-learning (QRL) approach for distributed online TDD frame optimization is proposed. First, a QRL network is utilized to estimate the near-optimal numbers of downlink and uplink transmission opportunities for a balanced traffic handling. A secondary QRL instance is selects the corresponding downlink and uplink symbol structure that minimizes the directional URLLC tail latency. The QRL-based solution is evaluated for both macro networks and newly emerging indoor industrial wireless deployments with dense small cell layouts. The proposed solution offers a significant URLLC outage gain in terms of autonomization of the TDD frame design on a real-time basis, URLLC outage latency reduction, and CLI-avoidance.

Resumé

Den femte generation (5G) af den cellulære teknologi tilbyder bedre støtte til tre vigtige serviceklasser; den ultra-pålidelige og low-latency kommunikation (URLLC), forbedret mobilt bredbånd (eMBB) og den massive maskintype-kommunikation (mMTC). URLLC-tjenester kræver transmission af små pakker med strenge radiolatens - og pålidelighedskrav. EMBB-applikationerne kræver bredbåndstransmissioner med ekstrem høj datahastigheder. Der er grundlæggende kompromisser mellem pålidelighed og latenstid, pålidelighed og spektral effektivitet. Samtidigt er en af de største udfordringer med den nye radio 5G at optimere servicekvaliteten (QoS). Desuden er den nye 5G-radio designet til at understøtte både frekvens- og tids-dupleksing (FDD, TDD). Pga den nye tilgængelige båndbredde ved 3,5 GHz de tidlige 5G-installationer forventes at blive med TDD-dupleksteknologien. At opnå en så effektiv multi-service ressourcestyring er dog mere udfordrende med TDD. Dette ph.d. projekt undersøger og udvikler nye og multi-service radio ressourcestyringsalgoritmer til multi-QoS 5G-netværk.

Den første del vedrører multi-QoS multipleksing-problemet (URLLC - eMBB). En QoS- flerbruger multiple-input multiple-output (MU-MIMO) algoritm. Overordnet gælder det om at eliminere tidsforsinkelse af de nyank - omne URLLC-pakker i tilfælde af, at de tilstrækkelige radioressourcer ikke umiddelbart er tilgængelige. De indgående URLLC-transmissioner parres øjeblikkeligt med de aktive eMBB-brugere, hvis rumlige signaturer er det nærmeste på et foruddefineret underrum. For at kontrollere inter-user koblingen hos de kritiske URLLC-brugere, projiceres de co-planlagte eMBB-transmis - sioner uafhængigt af hinanden i et vilkårligt rumligt underrum, hvortil de parrede URLLC-brugere justerer deres respektive transceivers i det vinkelrette underrum. Vi benytter detaljerede systemniveau-simuleringer med en høj grad af realisme i overensstemmelse med 3GPP NR-antagelserne for at evaluere de udviklede løsninger. Simuleringsresultaterne demonstrerer betydelige forbedringer af URLLC-tidsforsinkelsen og netværkskapaciteten, for eksempel minimeres URLLC-tidsforbruget med 50 percent, mens netværksk - apacit - eten forbedres med 79 percent sammenlignet med Rel-15 standard løsningerne.

I den anden del undersøges TDD 5G. Vi demonstrerer først, at URLLC QoS er sværere at opnå i TDD-implementeringer, hovedsageligt på grund af TDD-rammestrukturen, dvs. at ingen samtidige downlink- og uplink-transmissioner er mulige, og den potentielle kryds-links-interferens (CLI), når nabobasis-stationer eller brugere benytter modsatte transmission - sretninger. Et bredt spektrum af nye inter-celle koordinationsløsninger er blevet udviklet. Disse inkorporer et nyt sæt af TDD-systemdesignforbedringer, såsom semistatisk rammekonfiguration, glidende rammebogdesign, fælles hybrid rammedesign og pladsbevidst brugerplanlægning og koordineret trans - ceiver - design. Vores resultater viser en bemærkelsesværdig forbedring af URLLC-afbrydelser sammenlignet med standard dynamiske TDD-opsætninger, fx 80 percent URLLC reduktion af drift.

Maskinlæringsalgoritmer til realtidsvalg af TDD-rammestruktur er også blevet udviklet. For det første bruges et QRL-netværk til at estimere det næsten optimale antal downlink- og uplink-transmissionsmuligheder til en afbalanceret trafikhåndtering. En sekundær QRL-forekomst vælger den tilsva - rende downlink- og uplink-symbolstruktur, der minimerer retningsbestemt URLLC-haletatens. Den foreslåede løsning tilbyder en betydelig URLLC-strømafbryd - elsesforøgelse med hensyn til autonomisering af TDD-ramme - design på realtid, URLLC-reduktion af drift og CLI-undgåelse.

Contents

List of Abbreviations

WCN	Wireless Communication Networks
IFD	Innovation Fund Denmark
0G	Zero-generation
1G	First-generation
NMT	Nordic mobile telephone
AMPS	Advanced mobile phone telephone
JTACS	Total access communications system
FDMA	Frequency division multiple access
QoS	Quality of service
2G	Second-generation
GSM	System for mobile communications
SMS	Short message services
TDMA	Time division multiple access
GPRS	General packet radio service
IS-95	Interim standard 95
CDMA	Code division multiple access
WCDMA	Wide-band CDMA
3G	Third-generation
HSPA	High speed downlink packet access
FDD, TDD	Frequency and time division duplexing modes

MCSs Modulation and coding schemes

CA Carrier aggregation

4G Fourth-generation

LTE Long-term evolution

IP Internet-protocol

eICIC Enhanced inter-cell interference coordination

DC Dual connectivity

CoMP Coordinated multi-point

NR New radio

3GPP Third generation partnership project

EPC Evolved packet core

SPS Semi-persistent scheduling

eMBB Enhanced mobile broadband

URLLC Ultra-reliable low-latency communications

mMTC Massive machine type communications

TTI Transmission time interval

SCS Sub-carrier spacing

OFDM Orthogonal frequency division multiplexing

FR1 Frequency range 1

FR2 Frequency range 2

UEs User-equipments

CRS Common reference signal

SSBs Synchronization signals blocks

CSI-RS Channel state information reference signals

DMRS Demodulation reference signals

CSIT Channel state information at transmitter

CLI Cross-link-interference

DoFs Degrees of freedom

MAC Media access control

MU-MIMO Multi-user MIMO

CQI Channel quality indication

LA Link adaptation

IRC Interference rejection combining

ML Machine learning

RL Reinforcement learning

SLSs System level simulations

SINR Signal-to-interference-noise-ratio

KPIs Key performance indicators

HARQ Hybrid automatic repeat request

BLER Block error rate

NSBPS Null space based preemptive scheduler

CBR Constant bit rate

PAROS Preemption-aware rank offloading scheduling

PRB Physical radio block

MUPS Multi-user preemptive scheduler

WPF Weighted proportional fair

CCDF Complementary cumulative distribution function

BSP BS processing

FA Frame alignment

CG Configured grant

DG Dynamic grant

CF-TDD CLI-free TDD

NC-TDD Non-coordinated TDD

CRFC-TDD Coordinated radio frame configuration based TDD

CSA-TDD BS-BS CLI suppression algorithm based TDD

sTDD Static TDD

dTDD Dynamic TDD

Thesis Details

Thesis Title:	End-to-End-Aware Multi-Service Radio Resource Management for 5G.
PhD Student:	Ali Abdelmawgood Ali Ali Esswie
Supervisors:	Prof. Preben Mogensen, Aalborg University. Prof. Klaus Pedersen, Aalborg University.

This industrial PhD thesis is the outcome of three years of fruitful research at the Wireless Communication Networks (WCN) section (Department of Electronic Systems, Aalborg University, Denmark), jointly with Nokia Bell Labs. In addition to the presented research outcomes, dissemination activities, through conference attendance, team workshops, and local presentations, in addition to the mandatory external collaboration and courses are fulfilled, as part of the requirements for obtaining the PhD degree.

The main contributions of this thesis are presented by the following publications:

Paper A: **A. A. Esswie** and K. I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5G networks," *in Proc. IEEE ISCC*, Natal, May 2018, pp. 00136-00141.

Paper B: **A. A. Esswie** and K. I. Pedersen, "Null space based preemptive scheduling for joint URLLC and eMBB traffic in 5G networks," *in Proc. IEEE Globecom*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1-6.

Paper C: **A. A. Esswie** and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *in IEEE Access*, vol. 6, pp. 38451-38463, July 2018.

Paper D: **A. A. Esswie** and K. I. Pedersen, "Capacity optimization of spatial preemptive scheduling for joint URLLC-eMBB traffic in 5G new radio," *in Proc. IEEE Globecom*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1-6.

Paper E: **A. A. Esswie**, K. I. Pedersen and P. E. Mogensen, "Preemption-aware rank offloading scheduling for latency critical communications in 5G networks," *in Proc. IEEE VTC-Spring*, Kuala Lumpur, Malaysia, April 2019, pp. 1-6.

Paper F: Guillermo Pocovi, **A. A. Esswie**, and Klaus I. Pedersen, "Channel quality feed-back enhancements for accurate URLLC link adaptation in 5G systems," *in Proc. IEEE VTC-Spring*, Antwerp, April 2020.

Paper G: **A. A. Esswie**, and K.I. Pedersen, "On the ultra-reliable and low-latency communications in flexible TDD/FDD 5G networks," *in Proc. IEEE CCNC*, Las Vegas, Jan. 2020, pp. 1-6.

Paper H: **A. A. Esswie**, K.I. Pedersen, and P. Mogensen, "Semi-static radio frame configuration for URLLC deployments in 5G macro TDD networks," *in Proc. IEEE WCNC*, virtual conference, May 2020, pp. 1-6.

Paper I: **A. A. Esswie**, and K.I. Pedersen, "Inter-cell radio frame coordination scheme based on sliding codebook for 5G TDD systems," *in Proc. IEEE VTC-spring*, Kuala Lumpur, Malaysia, April 2019, pp. 1-6.

Paper J: **A. A. Esswie**, K.I. Pedersen, and P. Mogensen, "Quasi-dynamic frame coordination for ultra- reliability and low-latency in 5G TDD systems," *in Proc. IEEE ICC*, Shanghai, China, May 2019, pp. 1-6.

Paper K: **A. A. Esswie**, and K.I. Pedersen, "Cross link interference suppression by orthogonal projector in 5G dynamic-TDD URLLC systems," *in Proc. IEEE WCNC*, virtual conference, May 2020, pp. 1-6.

Paper L: **A. A. Esswie**, K.I. Pedersen, and P. Mogensen, "Online radio pattern optimization based on dual reinforcement-learning approach for URLLC 5G networks," *IEEE Access*, vol. 8, pp. 132922-132936, 2020.

Paper M: **A. A. Esswie**, and K.I. Pedersen, "Analysis of outage latency and throughput performance in industrial factory 5G TDD deployments," *Submitted to VTC-spring, 2021.*

This thesis has been submitted for assessment in partial fulfillment 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

This dissertation is the outcome of a fruitful PhD research for three years, conducted jointly between Nokia Bell Labs and the Wireless Communication Networks section, Department of Electronic Systems, Aalborg University, Denmark. Moreover, the PhD research study was partly sponsored by the Innovation Fund Denmark (IFD) and the European ONE5G project.

I would like to express my sincere appreciation to my supervisors: Klaus I. Pedersen and Preben Mogensen, without whom it may not be possible to distingiously complete this research project. Their valuable support, advice, guidance, and experience that I have always received all kept me highly motivated and challenged to excel beyond my own expectations. The open-mindset research environment they have offered me is exceptional and a unique personal experience that shall immensely impact my future career. I would like also to give my special thanks to Gilberto Berardinelli who was my first contact point at Aalborg University.

In addition, my sincere thanks are given to Drothe Sparre and Linda Villadsen for their persistent administrative support during my PhD project. Last but certainly not the least, I offer my deepest gratitude to my colleagues at Aalborg University and Nokia Bell Labs. Thank You Jens Steiner and Mads Brix for making my life a lot easier, through your key assistance in the development of the system level simulations. Thank you Guillermo, Ali Karimi, Roberto, Melisa, Pilar, Renato, Erika, and Thomas for being valuable friends before colleagues. Thank you Claudio, Istvan, Jeroen, Daniela, Frank, Troels, Mad L. for the valuable discussions we always had.

My greatest gratitude goes to my wife, Niveen Ozcan, who has been on my side all the time with consistent support, encouragement and understanding. I can not thank you enough for the hard times we passed together during my PhD project.

Finally, I dedicate this PhD work to my parents, who have always been dreaming of me getting a PhD degree.

Ali A. Esswie
Aalborg University, August 2020

Part I

Introduction

Introduction

1 Evolution of Cellular Communications

Nowadays, the cellular communication technology has been a vital element of our day-to-day life. Derived by our communication needs [1], the cellular standards have been rapidly evolving over the past four decades to offer significantly improved capacity, latency, connection density, and energy efficiency, respectively, as depicted by Fig. I.1 [2] (CommScope, 2017). Back in 1920s [3], the German railway first offered mobile telephony services onboard of several national trains. After the world war II, the international developments of portable communication devices have been exponentially progressed. Those were not based on a cellular concept yet. This is refereed to as the zero-generation (0G) of the wireless standards.

Over the early 1980s, the first-generation (1G) of the cellular technology was first triggered by the Nordic mobile telephone (NMT), in Nordic countries (Finland, Sweden, Denmark, and Norway), advanced mobile phone telephone (AMPS) in USA, and total access communications system (JTACS) in Japan [3]. The 1G enabled voice communications for mobile handsets, and was developed based on an analogue system design, with the frequency division multiple access (FDMA) as the baseline radio access technique. Although the 1G has pioneered the cellular standards, it offered a limited capacity and voice quality of service (QoS).

Transiting to the digital domain, the second-generation (2G), referred to as the global system for mobile communications (GSM), is developed by early 1990s, introducing a better voice quality, communication security, and enabling the short message services (SMS). The GSM implementation was based on the time division multiple access (TDMA) [4]. As the consumer needs for mobile data services started to considerably grow, the general packet radio service (GPRS), referred to as 2.5G, was first introduced as an enhanced GSM in order to offer limited data services over mobile networks. Later, the interim standard 95 (IS-95) has been introduced as a new cellular system based on the code division multiple access (CDMA) [5] to offer a further improved data capacity. By the late 1990s, the 2G deployments have dominated the

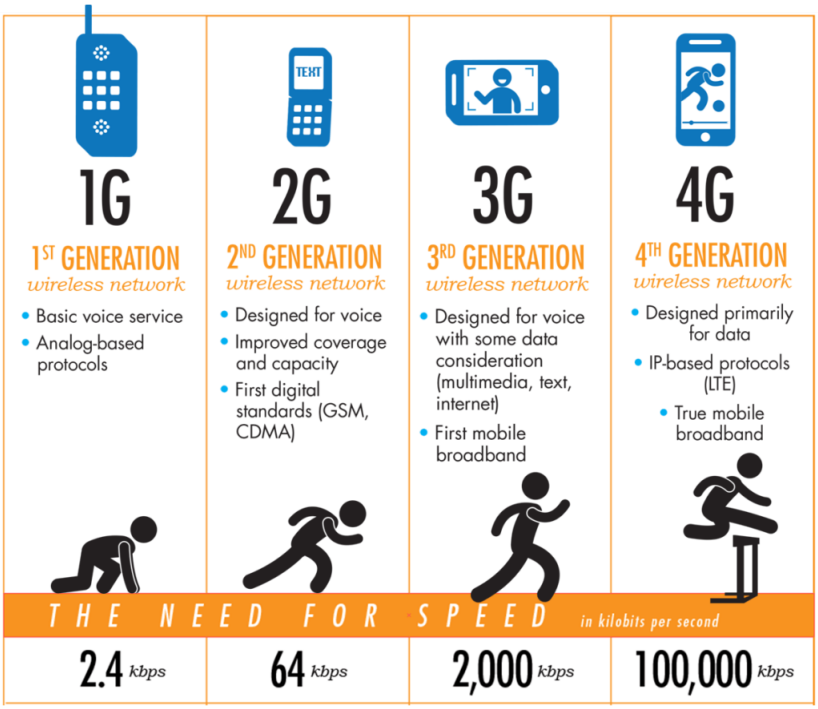


Fig. I.1: Evolution of the cellular communication technology [2] (CommScope, 2017).

international mobile markets by more than an 80% share [6].

However, the consistent growth of the mobile data demands and associated use cases and applications has motivated for a more improved third-generation (3G) of the mobile networks, based on wide-band CDMA (WCDMA) [7]. The 3.5G system variant, referred to as the high speed downlink packet access (HSPA), featured channel bandwidths of 1.25 MHz and 5 MHz, respectively, and supported both the frequency and time division duplexing modes (FDD, TDD). Thereafter, several system design enhancements had been utilized with the development of the HSPA+ to offer improved QoS for voice and data applications. This includes adopting higher modulation and coding schemes (MCSs), carrier aggregation (CA), multi-input multiple-out (MIMO) based transmissions, and partially internet-protocol (IP) based core infrastructure.

Thereafter, the fourth-generation (4G) of the cellular standards, referred to as the long-term evolution (LTE), has come into light, featuring an all-IP based core structure and highly optimized radio interface [8, 9]. LTE systems offered a peak data rate up to 1 Gbps with an improved spectral efficiency of 3 bits/Hz/cell [8, 10], i.e., ITU-R requirements for 3GPP release-8.

Moreover, the LTE radio technology incorporated the orthogonal frequency division multiple access (OFDMA) as a new multiple access technique which better supports MIMO and advanced receiver design due to its flexible frequency domain processing. LTE networks support a wide variety of per-carrier communication bandwidths, i.e., from 1.4 up to 20 MHz. Several major enhancements of the baseline LTE standard have been recently developed to achieve a further enhanced latency performance and spectral efficiency – often called as LTE-advanced networks [11]. Those system improvements include enhanced inter-cell interference coordination (eICIC), improved CA, dual connectivity (DC), improved MIMO transmissions with a larger number of supported spatial layers, enhanced coordinated multi-point (CoMP) transmissions, and multi-carrier aggregation, respectively.

2 5G Introduction

The research activities of the 5G new radio (NR) have been first started by 2010 [12], and accordingly, the first wave of the 5G-NR specifications is completed by the third generation partnership project (3GPP) in 2017, as release-15. Those early specifications define the baseline radio and core components of the 5G-NR system. Generally, two deployment options are defined as non-standalone and standalone roll-outs, respectively [13]. The former acts as the initial 5G deployment option and denotes that the 5G radio interface is integrated with the LTE evolved packet core (EPC). This implies that initial call setups of users in idle modes are communicated over the LTE. The standalone deployment utilizes the full potential of the 5G-NR performance merits by integrating the 5G radio and core interfaces. Subsequently, staring from 2018, the 3GPP groups have started progressing release-16 with further 5G system enhancements, towards a better system optimization of the emerging industrial wireless automation deployments, e.g., Industry 4.0. Those include enhance semi-persistent scheduling (SPS), and improved HARQ retransmissions. As depicted by Fig. I.2 [3GPP 2020], the completion of release-16 is expected by 2020 Q3, while the content definition and early research activities of release-17 are ongoing.

The 5G-NR supports the coexistence of three main service classes: enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine type communications (mMTC), respectively [14]. Fig. I.3 [Nokia] depicts the major performance requirements of each of those classes as follows:

- **eMBB**: eMBB applications demand extreme data rates with large bandwidth allocations. Those were the main driver of the LTE radio standards and accordingly, the 5G-NR is expected to offer a highly improved eMBB capacity. With the 5G-NR, the IMT-2020 visions for eMBB

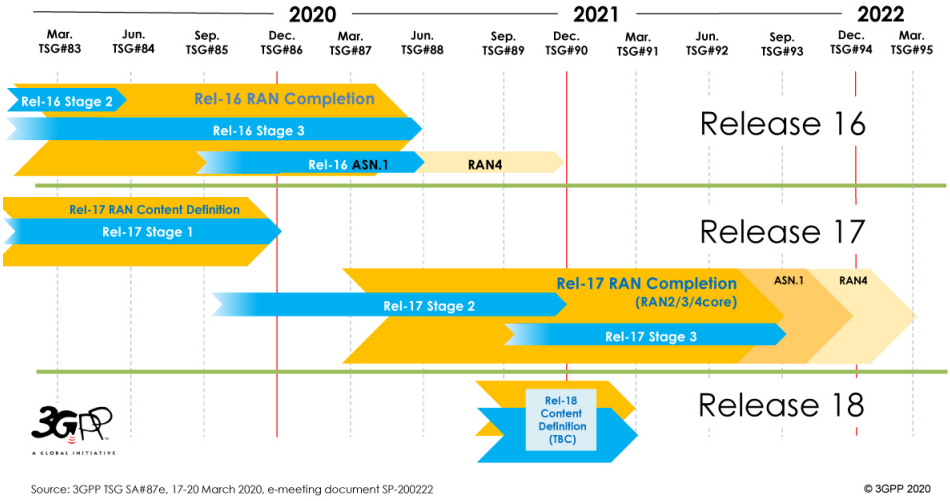


Fig. I.2: 5G new radio standardization time-line [3GPP].

services are 10 Gbit/second as the achievable peak data rate and with a mobility support up to 500 Km/hour.

- **URLLC**: URLLC services denote the transmission of small-payload and sporadically-arriving packets, with a stringent set of radio latency and reliability targets, respectively. Some URLLC use cases demand a 1-ms of radio latency with 99.999% success probability. Such unprecedented latency and reliability requirements were not handled by former radio standards, and thus, those impose a real challenge of the 5G-NR system design.

- **mMTC**: mMTC deployments require the support of a large connection density, i.e., 1 million device/km^2. Furthermore, mMTC devices are designed for optimum energy efficiency and longer battery life, e.g., 10 years of battery life, for which the 5G-NR standards should be optimized towards.

However, there is a fundamental trade-off among the achievable capacity, latency, and reliability, respectively, over the same spectrum [15]. For instance, achieving ultra-reliable wireless transmissions typically require a large radio latency performance though. Hence, the 5G-NR standard introduces a set of system design improvements to offer an agile radio performance, meeting the diverse requirements of the eMBB, URLLC, and mMTC QoSs, respectively, mainly highlighted by:

Agile radio frame structure and numerology [16]: the 5G-NR supports an

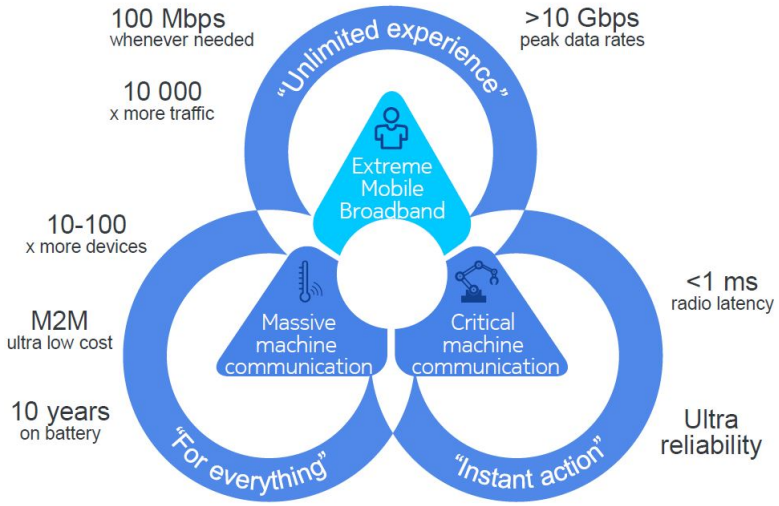


Fig. I.3: Major service classes of the 5G new radio [Nokia].

agile frame design with a variable transmission time interval (TTI) duration and sub-carrier spacing (SCS), respectively. Hence, a 5G-NR radio frame is 10 ms, and consists of 10 sub-frames, each of 1 ms duration. Sub-frames are flexibly divided into $2^n, n = 0, 1, 2...$, slots. Accordingly, a radio slot is of 14 orthogonal frequency division multiplexing (OFDM) symbols duration when a normal cyclic prefix design is used. Within each slot, there can be several transmission opportunities - so called as a mini-slot based transmissions, e.g., transmissions based on 4-OFDM symbol mini-slot. Accordingly, the latency-critical URLLC services are dynamically served with a shorter TTI duration (based on the mini-slot duration) and larger SCS, respectively. Alongside the smarter pipeline PHY processing, and the improved processing capabilities available for the 5G era, this fundamentally reduces both the transmission and processing delays, respectively, at the expense of the increased control overhead, due to the shorter transmissions. However, the latency-tolerant eMBB applications are dynamically scheduled with a larger TTI duration to increase the achievable spectral efficiency.

Higher spectrum utilization [17]: the 5G-NR utilizes the conventional cellular spectrum below 6 GHz (frequency range 1 (FR1)), as well as the higher spectrum over 20 GHz, i.e., FR2. The latter offers abundantly available bandwidth allocations up to 400 MHz. With carrier component aggregation, the serving bandwidth could further be extended up to 1 GHz. With such large available communication bandwidths, enhanced broadband data rates could be provided. However, the higher operating spectrum imposes several propagation challenges of the 5G-NR, e.g., rain absorption issues.

Flexible bandwidth parts operation [18]: as the 5G-NR utilizes much larger communication bandwidths than those supported by former 4G radio standards, it imposes a challenge for user-equipments (UEs) to frequency-scan such large spectrum. Thus, UEs are solely assigned a subset of the total available bandwidth, i.e., a bandwidth part. Accordingly, a bandwidth part specifies a certain set of the radio configurations (bandwidth, SCS, TTI duration, etc) which matches the requested service QoS requirement.

User-centric Radio Design: the 5G-NR adopts a leaner carrier design to allow for UE-centric radio transmissions. Particularly, the 5G-NR does not adopt the common reference signal (CRS) concept, - so called as the always-on signals, except for the transmission of the synchronization signals blocks (SSBs). However, the 5G-NR integrates the transmission of UE-specific downlink modulation reference signals (DMRS) and channel state information reference signals (CSI-RS), respectively. The intuition is that the radio interface transmits control channels and signals only when needed, and accordingly, adapting the radio interface to the UE-specific conditions.

Massive MIMO and Advanced beamforming [19]: advanced MIMO communications are greatly utilized with the 5G-NR standards, where the design of large antenna arrays becomes more feasible in practical deployments due to the much shorter wavelengths, and hence, the shorter antenna spacing requirement. Hence, larger antenna arrays alongside with hybrid beamforming and sufficient channel state information at transmitter (CSIT) acquisition techniques enable multiple simultaneous transmissions over the same time and frequency resources, where they are efficiently separated on the spatial domain, boosting the achievable spectral efficiency.

Multi-QoS Dynamic User Scheduling and Radio Resource Management (**RRM**) [20-23]: the state-of-the-art proposals introduce agile RRM and dynamic user scheduling techniques for the 5G-NR multi-QoS deployments. Those contributions typically adopt a multi-objective optimization techniques towards achieving the diverse, and sometimes conflicting, QoS requirements of active UEs. 3GPP release-15 specifications consider the multi-QoS pre-emptive scheduling [22] as the baseline MAC technique for achieving the stringent radio latency requirements of the latency-critical traffic. It always prioritizes such traffic over other active latency-tolerant traffic by means of immediate preemption; however, recovering the capacity of the latter traffic QoS by smarter re-transmission and coding techniques [24].

3 Scope and Objectives of the PhD Thesis

Dynamic UE scheduling and RRM techniques are of a significant importance for the 5G-NR to achieve the target diverse QoS requirements. Particularly, dynamic UE scheduling denotes how the radio interface is dynamically se-

lecting the UEs to serve at each transmission instant, where the RRM implies the adopted strategy to allocate the actual radio resources to selected UEs, in terms of the time, frequency, and spatial resources. For the 5G-NR deployments with multi-QoS classes, this is specifically a challenging and non-trivial problem to achieve those diverse QoS requirements. State-of-the-art scheduling techniques mainly rely on a long-term network-centric approach for maximizing the achievable network spectral efficiency, although; those are not appropriate for URLLC UEs with stringent latency and reliability bounds. Thus, the broader scope of this thesis is the dynamic UE scheduling and RRM for multi-QoS 5G-NR deployments. It spans both the FDD and TDD 5G deployments. The main considered QoS classes are the URLLC and eMBB, where the corresponding performance targets are optimized by the developed RRM and coordination schemes.

In the following, we present the major research questions and hypothesis addressed by this PhD dissertation as follows:

Q1 How the spatial degrees of freedom (DoFs) of the BS antenna array can be utilized for an enhanced URLLC-eMBB multiplexing performance?

H1 The spatial DoFs of the BS antenna array offer greater multiplexing flexibility in the spatial domain. This implies that multiple traffic streams of different UEs can be simultaneously served over the same time and frequency resources, separated in the spatial domain, and accordingly, enhancing the achievable network spectral efficiency. For the URLLC-eMBB coexistence deployments, such spatial DoFs could be flexibly utilized to reduce the queuing delay of the incoming urgent URLLC packets without significantly impacting the eMBB capacity.

Q2 How sensitive is the achievable URLLC outage latency performance to the TDD frame settings for various 5G-NR system configurations?

H2 Dynamic TDD deployments offer a flexible adaptation of the network resources to the varying traffic demands. In particular, the performance merits of the dynamic-TDD systems appear within the scenarios where the uplink and downlink traffic demands are highly variant in time. Achieving the URLLC stringent outage targets are challenging due to the non-simultaneous availability of the uplink and downlink transmission opportunities and the BS-BS and UE-UE cross link interference (CLI) , respectively. Our hypothesis is that combining an optimized set of the 5G radio configurations, e.g., larger SCS with a faster downlink and uplink link switching, associated with efficient inter-BS CLI avoidance or suppression techniques could offer an attractive URLLC outage performance in dynamic TDD networks.

Q3 How to design flexible and computationally-efficient CLI control mechanisms?

H3 In dynamic TDD systems, controlling the CLI intensity, and especially the BS-BS CLI, should be controlled to achieve a decent URLLC outage performance. A smarter design of the TDD radio frames, combined with CLI-aware dynamic user scheduling, could be a step towards minimizing the occurrences of the CLI, and restricting those to the UEs with the best channel conditions or those of the most relaxed latency targets. Furthermore, efficient inter-BS coordination schemes become vital to ensure coordinated CLI-free resources for critical traffic streams. Those include techniques for partly avoiding the CLI occurrence on a best-effort basis or efficiently suppressing the CLI effect prior to the packet decoding.

Q4 What is the ML learning potential to offer a better TDD radio frame adaptation?

H4 Our hypothesis is that efficiently integrating an ML model to continuously learn and optimize the TDD radio pattern structure could be promising to improve the achievable URLLC outage latency performance. In particular, a sufficient ML modeling could remove the need for pre-defining the TDD radio patterns, in terms of the downlink and uplink symbol placement, rather than dynamically adjusting it on a real-time basis to reduce the URLLC radio latency.

To pursue our research directions and hypotheses, and as shown by Fig. I.4, Part I of the thesis tackles the URLLC-eMBB QoS coexistence problem. For those multi-QoS scenarios, achieving the stringent URLLC radio and reliability targets are highly challenging. Standard RRM techniques typically adopt a predefined QoS-based prioritization approach, where the radio interface is pre-engineered towards achieving the latency-critical URLLC targets, while serving the latency-and-reliability tolerant eMBB applications on a best-effort basis. However, such methodology could inflict a significant degradation of the network spectral efficiency. In this part, we take one step further and utilize the spatial degrees of freedom (DoFs), offered by the BS antenna array, in order to achieve more agile URLLC-eMBB RRM solutions with moderate computational complexity.

As a first step, we develop a multi-stage media access control (MAC) scheduling strategy, where a standard multi-user MIMO (MU-MIMO) scheduler among URLLC-eMBB traffic is combined in cascade with the URLLC preemptive scheduler, i.e., standardized as part of the 3GPP release-15 specifications. It offers moderate performance merits in terms of the achievable URLLC latency and reliability in addition to the network spectral efficiency,

mainly due to the successful best-effort URLLC-eMBB MU-MIMO pairings. Next, we design an enhanced and non-transparent spatial scheduler where the urgent URLLC traffic is immediately served regardless of the active eMBB load, through *on-the-go* eMBB sub-space projections. Our objective is to preserve a minimal queuing delay for incoming URLLC packets while avoiding the significant eMBB capacity losses due to the abrupt URLLC transmissions. Hence, several recovery mechanisms of the eMBB capacity are researched and developed to further enhance the achievable network spectral efficiency.

Moreover, the URLLC traffic modeling implies the fast transmission of small-payload packets, which are arriving sporadically at the transmitter end. Thus, it imposes a novel challenge of having an efficient link adaptation (LA) at the BS, due to the fast variations of the interference statistics. In particular, in those deployments, existing channel quality indication (CQI) estimation and reporting techniques could mislead the BS by indicating either highly optimistic or pessimistic view of the channel and interference conditions, and subsequently, BS applies an insufficient modulation and coding scheme (MCS). Thus, we develop an enhanced CQI feedback for more accurate URLLC LA, within both URLLC and URLLC-eMBB coexistence scenarios.

In the second part of this dissertation, and as the early 5G-NR deployments are envisioned over the TDD spectrum, due to is large available bandwidth, we consider the 5G-NR TDD as the baseline duplexing mode. With dynamic TDD in place, the URLLC targets are highly challenging to achieve because of the additional cross-link interference (CLI), i.e., BS-BS and UE-UE CLI, respectively, and the non-concurrent availability of the uplink and downlink transmission opportunities. Therefore, we first perform a comprehensive evaluation of the achievable URLLC performance boundaries in TDD systems under the key 5G-NR radio design configurations. Based on our obtainable conclusions, we identify the CLI, and in particular the BS-BS CLI, as the main dominant setback of achieving the URLLC performance targets in dynamic TDD macro deployments. Hence, we develop several inter-BS coordination schemes for CLI control, and with different levels of the TDD frame flexibility, computational complexity, and the coordination signaling overhead. We start by a simple; though, performance efficient, semi-static strategy of adapting a network-wide TDD radio frame to the average offered traffic capacity, i.e., entirely eliminating the CLI problem. Furthermore, to boost the achievable TDD frame agility, we utilize a fully dynamic TDD system design with newly-introduced heuristic BS-BS CLI control mechanisms, offering a sufficient trade-off between the CLI intensity and the TDD frame flexibility. That is, we combine rotated frame-book structures, hybrid frame design, and CLI-aware dynamic UE scheduling in order to offer sufficient and dynamic CLI avoidance for the critical URLLC transmissions. Furthermore, we utilize the spatial DoFs of the BS antenna array to near-optimally suppress

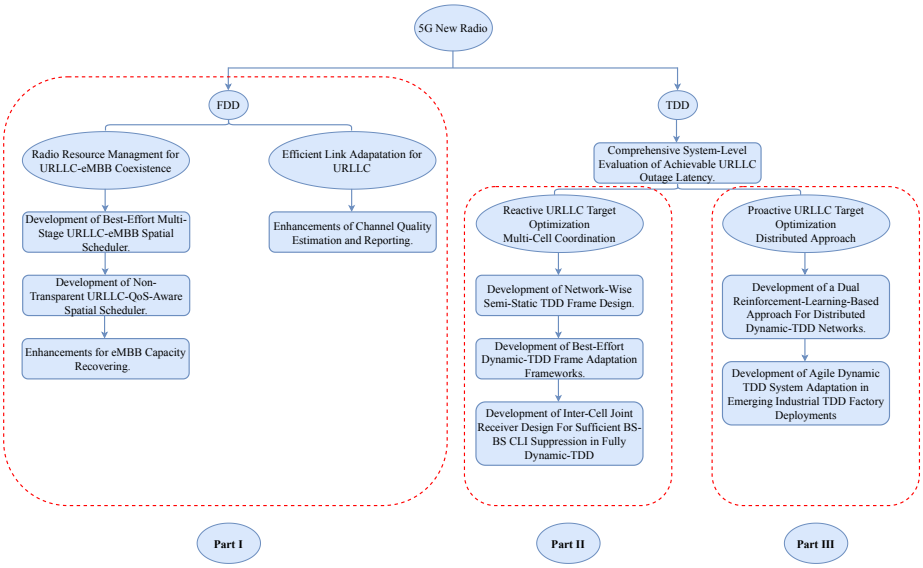


Fig. I.4: Thesis overall scope.

the severe BS-BS CLI, hitting the vulnerable uplink transmissions, thorough a newly-developed joint inter-cell transceiver design. Particularly, BSs seek to isolate the BS-BS CLI spatial subspace from the subspace of the desired uplink transmissions, using a modified design of the interference rejection combining (IRC) receiver.

Finally, Part III of the thesis adopts a machine learning (ML) approach as a viable solution to dynamically determine the optimal TDD pattern, in terms of the number and structure of the downlink and uplink symbols, respectively, which offers the best achievable URLLC outage performance. Unlike former solutions of Part II, the ML approach offers an autonomous optimization of the TDD radio pattern in a distributed manner without the need for inter-BS signaling. Specifically, we develop a dual reinforcement learning (RL) scheme for online pattern optimization, where the learning model is continuously refined on a real-time basis to learn the effective capacity of the uplink and downlink directions as well as the corresponding foreseen radio latency performance. Accordingly, the proposed framework tackles a joint URLLC capacity-latency optimization problem, where it is solved iteratively in time. The performance of the proposed RL-based solution is evaluated within the standard macro 5G-NR TDD deployments in addition to the emerging industrial factory roll-outs, respectively, where insightful conclusions are drawn.

4 Research Methodology

To pursue the research directions of this dissertation, we adopt the classical research methodology as follows.

1. **Formulation of the research problem, objectives, and hypothesis**: prior to the start of each research direction of the PhD project, an extensive review of the state-of-the-art literature is performed. This includes recent academic publications, standard reports and specifications, and recent intellectual properties (IPs), respectively. Then, a series of brainstorming and scouting discussions are triggered with the research supervisors as well as the office industry experts in order to define the major set-backs of the current solutions from achieving our research objectives. Accordingly, we formulate a set of hypothesis to tackle those identified literature limitations.

2. **Compile and develop the research solutions**: based on the defined hypothesis of each research problem, a candidate solution is formulated to reach our performance targets. Typically, identified solutions are analytically obtained since those offer a good view of the performance boundaries. However, due to the complexity of the addressed problems and the 5G-NR design components, sole analytical solutions may require a large set of system simplifications which may jeopardize the realistic implementation of the proposed solutions.

3. **Implementation and performance validation of the solutions**: to assess the performance of the proposed solutions, we evaluate the performance of the proposed solutions using extensive and highly-detailed system level simulations (SLSs), supported by analytical modeling of the problem and solution, when possible. We mainly perform a large set of Monte-Carlo SLSs for each developed feature. The setup of each simulation scenario is ensured to include a realistic implementation of the key 5G-NR design aspects affecting the achievable end performance. Examples include the dynamic link adaptation, 3D spatial channel modeling and propagation conditions, packet decoding based on realistic receiver design implementation, effective signal-to-interference-noise-ratio (SINR) combining, and dynamic traffic arrivals, respectively. To ensure the correctness and integrity of our simulator and the respective simulation results, the simulations are consistently calibrated with other 3GPP partners. Furthermore, to guarantee statistically-reliable results, the length of the adopted simulations is ensured to be sufficiently long enough to cover the rare packet events, e.g., packet drops, which highly impact the achievable URLLC outage performance. The key set

of the simulation design parameters are typically presented to maintain the reproducibility of the results.

4. **Analysis of the performance results and drawing insightful conclusions**: after the simulation results are readily available, post-processing activities are triggered on a diversity of the relevant system key performance indicators (KPIs). Accordingly, an extensive sensitivity analysis is typically performed under a diversity of system settings, where our results are compared to other relevant schemes from other sources such that we do not over-conclude our results. Based on the obtained results and conclusions, the initial hypothesis and solutions could be revisited and/or refined if needed.

5. **Knowledge dissemination**: for each part of the PhD project, the research outcomes and conclusions are documented in the form of journal and conference publications. Furthermore, part of the research outcomes are also presented within the project seminars, technical reports, and team meetings. Finally, the inventive outcomes and ideas of the thesis are protected by Nokia Bell Labs IPs.

5 Contributions

The major contributions of this dissertation are listed as follows:

1. **The development of a low-complexity multi-stage scheduler for URLLC-eMBB coexistence**

   A multi-QoS aware spatial scheduler has been proposed and developed to flexibly trade-off the achievable network spectral efficiency and the critical URLLC outage performance, respectively. The objective is to always prioritize a fast scheduling of the latency-critical URLLC traffic, by utilizing the free available spatial DoFs, while maximizing the eMBB capacity. In case the free available spatial DoFs are not sufficient to accommodate the buffered URLLC traffic, the achievable eMBB capacity is instantly traded-off for the sake of a fast URLLC transmission, i.e., running the cascade URLLC preemptive scheduling. Finally, the proposed scheduling framework demands a simple cascade architecture with a lower processing complexity.

2. **The development of an opportunistic spatial scheduler for joint URLLC-eMBB QoS optimization**

   Derived by the observations and conclusions of work 1., the proposed scheduler efficiently enforces the sufficient spatial DoFs on-the-fly, which are required to instantly accommodate the incoming URLLC packets.

This ensures fast URLLC packet scheduling and transmission regardless of the corresponding eMBB load and the originally available spatial DoFs. The sub-space projection theory is utilized to always allow for an interference-free spatial sub-space for urgent URLLC traffic, when paired with an active victim eMBB traffic. Accordingly, the queuing delay of the URLLC traffic is further minimized without introducing an additional inter-user interference. As a result, using our extensive SLSs, the developed scheduling solution shows a consistently improved URLLC outage performance regardless of the different offered eMBB loads. Moreover, the proposed solution is analytically evaluated to demonstrate that the inflicted eMBB capacity loss is bounded at a minimal level, compared to standard URLLC-eMBB scheduling techniques.

3. **The development and integration of eMBB capacity recovering mechanisms and spectral efficiency maximization enhancements for joint URLLC-eMBB deployments**

   To overcome the eMBB capacity loss, due to the URLLC traffic prioritization, several eMBB capacity recovering techniques have been developed. Proposed solutions combine a multi-bit radio signaling from BS to eMBB UEs, along with an adaptive MIMO rank offloading, in order to compromise the achievable eMBB capacity, and hence, the network spectral efficiency only when needed, i.e., URLLC traffic queuing is foreseen. Proposed solutions have demonstrated greater agility of the dynamic UE selection and scheduling along with an improved ergodic capacity while preserving a decent URLLC outage performance.

4. **A comprehensive evaluation of the achievable URLLC outage performance in dynamic TDD 5G-NR macro networks**

   The key 5G radio design settings and the corresponding dynamic TDD configurations, which immensely affect the achievable URLLC outage performance, are identified and analyzed. Those studies are performed by highly detailed SLSs and designed in a smart way such that to isolate the performance contribution or degradation of each individual system aspect. Thereafter, valuable conclusions are drawn on the baseline dynamic TDD settings in order to achieve a decent URLLC outage performance.

5. **The development of low-complexity inter-BS coordination schemes for CLI avoidance and suppression**

   CLI has been identified as the major set-back of the dynamic TDD macro systems, and thus, the development of inter-BS CLI avoidance schemes is vital for a better URLLC performance. Accordingly, several

coordination frameworks have been developed for inter-BS CLI avoidance with different levels of TDD frame adaptation flexibility, processing complexity, and inter-BS coordination overhead, respectively. Proposed solutions introduce a smarter design of the radio frame-books combined with CLI-aware dynamic UE scheduling, and hybrid TDD frame structures, respectively. Proposed solutions demonstrate a lower coordination complexity while offering agile frameworks for avoiding the severe BS-BS and UE-UE CLI in dynamic TDD deployments. Furthermore, we develop an inter-BS joint IRC receiver design for BSs to suppress the severe impact of the BS-BS CLI on victim uplink receptions. The high-level objective is that aggressor downlink BSs share the spatial signatures of their active downlink UEs, such that victim uplink BSs, with uplink receptions, can utilize such information towards a better uplink IRC receiver design. Unlike the former best-effort CLI avoidance schemes, the proposed solution offers a near-optimal and more reliable URLLC uplink outage performance, however, at the expense of larger coordination signaling overhead size and processing complexity, respectively.

6. **The development of a reinforcement-learning based online pattern optimization scheme for dynamic TDD deployments**

   The ML can be a viable solution for determining the optimum TDD pattern structure, given a certain state of the achievable directional capacity and the radio latency, respectively. Thus, we developed a dual RL approach to periodically determine the best possible TDD pattern structure, in terms of the number and placement of the downlink and uplink symbols across the pattern duration. The RL objective is to select the pattern that best matches the varying uplink and downlink effective capacity as well as the radio latency performance, such that the TDD pattern is continuously optimized for a balanced uplink and downlink traffic handling. The proposed solution neither requires inter-BS coordination overhead nor high processing complexity.

The following list presents the journal and conference publications which represent the main contribution of this PhD thesis as follows:

Paper A: **A. A. Esswie** and K. I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5G networks," *in Proc. IEEE ISCC*, Natal, May 2018, pp. 00136-00141.

Paper B: **A. A. Esswie** and K. I. Pedersen, "Null space based preemptive scheduling for joint URLLC and eMBB traffic in 5G networks," *in Proc. IEEE Globecom*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1-6.

Paper C: **A. A. Esswie** and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *in IEEE Access*, vol. 6, pp. 38451-38463, July 2018.

Paper D: **A. A. Esswie** and K. I. Pedersen, "Capacity optimization of spatial preemptive scheduling for joint URLLC-eMBB traffic in 5G new radio," *in Proc. IEEE Globecom*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1-6.

Paper E: **A. A. Esswie**, K. I. Pedersen and P. E. Mogensen, "Preemption-aware rank offloading scheduling for latency critical communications in 5G networks," *in Proc. IEEE VTC-Spring*, Kuala Lumpur, Malaysia, April 2019, pp. 1-6.

Paper F: Guillermo Pocovi, **A. A. Esswie**, and Klaus I. Pedersen, "Channel quality feed-back enhancements for accurate URLLC link adaptation in 5G systems," *in Proc. IEEE VTC-Spring*, Antwerp, April 2020.

Paper G: **A. A. Esswie**, and K.I. Pedersen, "On the ultra-reliable and low-latency communications in flexible TDD/FDD 5G networks," *in Proc. IEEE CCNC*, Las Vegas, Jan. 2020, pp. 1-6.

Paper H: **A. A. Esswie**, K.I. Pedersen, and P. Mogensen, "Semi-static radio frame configuration for URLLC deployments in 5G macro TDD networks," *in Proc. IEEE WCNC*, virtual conference, May 2020, pp. 1-6.

Paper I: **A. A. Esswie**, and K.I. Pedersen, "Inter-cell radio frame coordination scheme based on sliding codebook for 5G TDD systems," *in Proc. IEEE VTC-spring*, Kuala Lumpur, Malaysia, April 2019, pp. 1-6.

Paper J: **A. A. Esswie**, K.I. Pedersen, and P. Mogensen, "Quasi-dynamic frame coordination for ultra- reliability and low-latency in 5G TDD systems," *in Proc. IEEE ICC*, Shanghai, China, May 2019, pp. 1-6.

Paper K: **A. A. Esswie**, and K.I. Pedersen, "Cross link interference suppression by orthogonal projector in 5G dynamic-TDD URLLC systems," *in Proc. IEEE WCNC*, virtual conference, May 2020, pp. 1-6.

Paper L: **A. A. Esswie**, K.I. Pedersen, and P. Mogensen, "Online radio pattern optimization based on dual reinforcement-learning approach for URLLC 5G networks," *IEEE Access*, vol. 8, pp. 132922-132936, 2020.

Paper M: **A. A. Esswie**, and K.I. Pedersen, "Analysis of outage latency and throughput performance in industrial factory 5G TDD deployments," *Submitted to VTC-spring, 2021.*

Furthermore, during the course of the PhD project, the solutions, which integrate a sufficient inventiveness, are protected by Nokia Bell Labs patent applications. The list of filed patent applications is as follows:

Patent Application 1: Null space based interference pre-cancellation for multi-user ultra low latency communications in 5g networks

Patent Application 2: Enhanced spatial preemptive scheduling for multi-traffic coexistence in 5g new radio

Patent Application 3: Sliding radio frame configuration for dynamic coordinated TDD with limited signaling overhead in 5g new radio

Patent Application 4: BS-BS cross link interference suppression using orthogonal projector in 5G dynamic TDD systems

Patent Application 5: BS cross channel measurement control and configuration

Patent Application 6: Coordinated inter base station measurement procedures based on XN-interface signaling exchange

Patent Application 7: Inter-cell proactive coordination by means of exchange of time-domain outage tables

Patent Application 8: Coordinated interference avoidance and configured grant collisions for reliable uplink HARQ transmission

Patent Application 9: Enhanced procedures for conveying timing information of uplink packets from the UE to the network

Patent Application 10: ML-driven UE antenna panel switching solution

Patent Application 11: Inter-gNB coordination to mitigate PBO limitation in dynamic TDD systems

Furthermore, a considerable share of the PhD research activities is devoted to the development and setting the scenes of the respective system

level simulations. The extensive performance evaluation of all the publications included in this thesis are performed using the Nokia Bell Labs proprietary system-level simulator. The simulator incorporates a highly-detailed modeling of the 5G key system design aspects. This includes the major functionalities of the physical (PHY) and MAC layers, e.g., realistic IRC design and decoding, SINR estimation, 3D spatial channel modeling, adaptive MCS selection, and Chase combining hybrid automatic repeat request (HARQ) retransmissions, respectively.

Particularity, to develop and implement a new feature within the simulator environment, an accurate investigation and tracing of the simulation environment followed by an optimized implementation in object-oriented C++ are performed, respectively. Then, a careful debugging phase is initiated where the new implementation is tested using a simplified set of system configurations. Accordingly, a large set of different simulation scenes are prepared and run to evaluate the performance of the new feature. Finally, the major relevant KPIs are analyzed and accurate conclusions are drawn. Once the feature implementation is finalized, a set of regression tests is applied to guarantee that the new feature implementation does not unnecessarily alter the former verified results of the other features in the simulation environment. Thereafter, a regression test for the new implementation is generated to maintain the result sustainability.

The major feature implementation of this PhD thesis is listed as follows:

- **Multi-stage multi-user preemptive scheduler**: the implementation of this feature enables a cascaded layer design of the dynamic UE scheduling for joint URLLC-eMBB coexistence deployments.
- **Opportunistic spatial URLLC-eMBB scheduler**: the implementation of this feature enables the immediate scheduling of the incoming URLLC packets, over the resources monopolized by active eMBB transmissions using *on-the-go* sub-space projections.
- **eMBB capacity recovering techniques**: the implementation of this feature enables several ergodic capacity recovering techniques.
- **Inter-BS CLI coordination schemes**: those are set of newly implemented features which enable a diversity of inter-BS coordination schemes for CLI avoidance and suppression.
- **ML-based online pattern optimization for TDD deployments**: the implementation of this feature enables the online optimization of the TDD radio patterns based on a dual reinforcement learning approach. The learning as well as the inference are both implemented and executed withing the simulation environment.

6 Thesis Outline

This dissertation mainly consists of five main chapters, and is in the form of a paper collection. Accordingly, the detailed answers to the research questions and hypotheses presented earlier are included in the listed publications. As the dissertation tackles several topics, the several listed publications are sub-grouped, based on the topic relevance, within dedicated chapters. Finally, each chapter starts by an introductory description of the chapter objectives, problem formulations, and final recommendations, where some figures and text phrases from the relevant papers are included.

As depicted by Fig. I.5, the main outline of this thesis is as follows:

- **Part I - Introduction**: It implies the current chapter. It first introduces a brief history of the cellular communications evolution alongside the 5G-NR major advances and current standardization status, respectively. Finally, it presents the main research questions of this dissertation followed by our initial research hypothesis.

- **Part II - Scheduling enhancements for URLLC-eMBB service coexistence**: It tackles the novel coexistence problem among the URLLC and eMBB QoS classes. This chapter spans the papers A, B, C, D, E, and F, respectively, where the first research question Q1 and the corresponding hypothesis H1 are addressed. In this chapter, multiple novel dynamic UE scheduling frameworks are proposed and carefully evaluated using extensive SLSs.

- **Part III - Novel coordination techniques for dynamic-TDD URLLC networks**: This chapter addresses the key challenges of the emerging URLLC dynamic-TDD deployments. In particular, several inter-BS dynamic coordination schemes along with a smarter design of the TDD radio frames and CLI-aware UE scheduling are proposed and developed in order to efficiently combat the severe BS-BS and UE-UE CLI, respectively. This chapter includes the publications G, H, I, J, and K, respectively, and addresses the research questions Q2 and Q3, as well as the corresponding hypotheses.

- **Part IV - Machine learning potential towards improved dynamic-TDD operation**: This chapter investigates the potential of the machine learning in achieving a further flexible and efficient dynamic TDD operation for latency-critical URLLC traffic. A dual reinforcement learning approach is developed to autonomously estimate the best TDD radio pattern which contributes the lowest possible URLLC outage latency. The adopted learning model refines its prediction precision iteratively in time each radio pattern, and neither requires inter-BS coordination

signaling nor high processing complexity. This chapters answers the research question Q4 and the respective hypothesis. It includes the publications L and M, respectively.

- **Part I - Conclusions**: This chapter concludes the thesis by introducing the main findings and suggested future directions accordingly.

References

[1] J. Schiller, *Mobile communications*. Boston, Mass.: Addison-Wesley, 2003.

[2] "*Cellular Wireless 1G, 2G, 3G, 4G, 5G – Watch The Evolution*", James Donovan, CommScope, 2017. [Online]. Available: https://blog.commscopetraining.com/cellular-wireless-watch-the-evolution/.

[3] A. Huurdeman, *The worldwide history of telecommunications*. Hoboken: Wiley-IEEE Press [Imprint], 2005.

[4] H. Omar and W. Zhuang, *Time division multiple access for vehicular communications.* Springer, 2014 .

[5] M. Abu-Rgheff, *Introduction to CDMA wireless communications*. Amsterdam: Elsevier, 2007.

[6] 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.

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

[8] 3GPP TR 25.913, "Requirements for evolved UTRA (EUTRA) and evolved UTRAN (E-UTRAN),", v8.0.0, Release 8, Jan. 2009.

[9] D. Astely, E. Dahlman, A. Furuskr, Y. Jading, M. Lindstrm, and S. Parkvall, "LTE: the evolution of mobile broadband," *IEEE Commun. Mag.*, vol. 47, no. 4, pp. 44–51, April 2009.

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

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

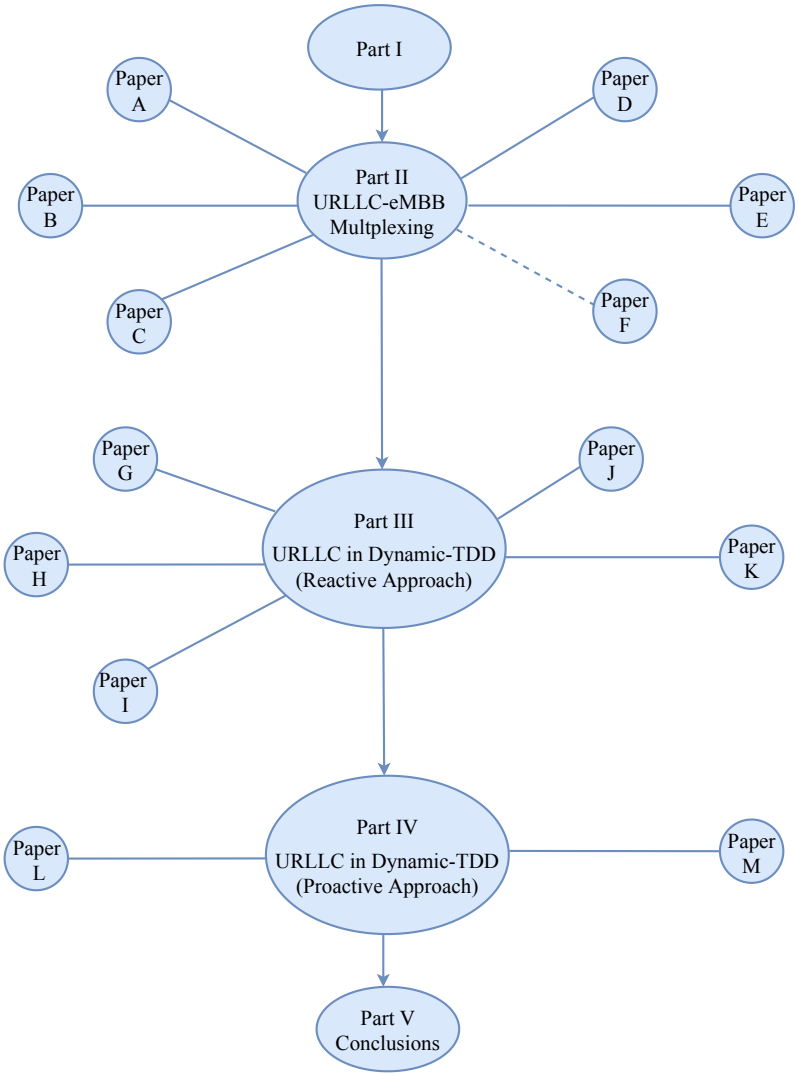


Fig. I.5: Thesis outline.

[12] 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 J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[13] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: a comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, July 2016.

[14] 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 J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, June 2017.

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

[16] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53–59, March 2016.

[17] 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 Netw.*, vol. 21, no. 8, pp. 2657–2676, Nov. 2015.

[18] F. Abinader et al., "Impact of bandwidth part (BWP) switching on 5G NR system performance," *IEEE 5GWF*, Dresden, Germany, 2019, pp. 161-166.

[19] T. Kashima et al., "Large scale massive MIMO field trial for 5G mobile communications system," *IEEE ISAP*, Okinawa, 2016, pp. 602-603.

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

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

[22] K. I. Pedersen, G. Pocovi, J. Steiner and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," *in Proc. IEEE VTC-Fall*, Toronto, ON, 2017, pp. 1-6.

[23] A. Anand, G. De Veciana and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *in Proc. IEEE INFOCOM*, Honolulu, HI, 2018, pp. 1970-1978.

[24] S. R. Khosravirad, L. Mudolo and K. I. Pedersen, "Flexible multi-bit feedback design for HARQ operation of large-Size data Packets in 5G," *in Proc. IEEE VTC-Spring*, Sydney, NSW, 2017, pp. 1-5.

Part II

Scheduling Enhancements For URLLC-eMBB Service Coexistence Over the 5G New Radio

Scheduling Enhancements For URLLC-eMBB Service Coexistence Over the 5G New Radio

This part of the thesis introduces a set of downlink dynamic UE scheduling enhancements for URLLC-eMBB QoS coexistence deployments. Several solutions have been proposed to ensure achieving the URLLC outage performance targets while flexibly optimizing the achievable eMBB ergodic capacity. The introduced solutions have been carefully evaluated through extensive system level simulations, where the major 5G-NR assumptions and best modeling practices are considered.

1 Problem Formulation

As presented in the Part I, the 5G-NR supports three major QoS service classes: URLLC, eMBB, and mMTC [1]. Those require a diversity of various performance targets, which sometimes are conflicting to achieve over the same frequency spectrum. As an example, the URLLC QoS demands a stringent radio latency and reliability performance, while the eMBB applications require broadband communications data rates. There is a fundamental trade-off among the radio latency, reliability, and spectral efficiency over the same communication spectrum [2]. Accordingly, for such multi-QoS coexistence deployments, achieving those diverse performance requirements is a challenging and non-trivial problem, and hence, this is the broader research problem addressed though this part of the thesis.

In order to achieve the stringent URLLC outage latency targets, the incoming URLLC packets should be transmitted without exhibiting a large queuing delay, while adopting proper transmission configurations such the

selected MCS. The latter denotes having a service dependent link adaptation where different first transmission block error rate (BLER) targets for eMBB and URLLC are used. This ensures that the URLLC packets are successfully decoded either from first-time transmission or after the minimal HARQ combining attempts, and hence, reducing the URLLC radio latency accordingly [3]. Therefore, multi-QoS-aware dynamic UE scheduling plays a vital role to achieve such stringent and diverse QoS targets. In the recent literature, the dynamic UE scheduler contributions [4-8] have been tackling the multi-QoS multiplexing problem by the efficient utilization of the time and frequency domain resources through resource pre-allocation, URLLC packet preemption, and coordinated multi-point transmission, respectively.

Inspired by the implementation of large BS antenna arrays with the 5G-NR, the spatial MIMO dynamic UE schedulers have become of a vital importance in order to utilize the offered spatial degrees of freedom (sDoFs), and this is the focus of this PhD part. Recent spatial scheduler proposals typically address a maximization problem of the achievable network spectral efficiency, mainly by means of adopting throughput-based higher-rank multi-user MIMO (MU-MIMO) UE associations [7, 8], and advanced beam-forming techniques [9, 10]. However, this is solely suitable for eMBB services, where the achievable ergodic capacity is the sole optimization objective. Therefore, in this part of the thesis, we address the eMBB-URLLC QoS coexistence problem in the downlink direction, where we propose and develop several novel URLLC-eMBB spatial UE dynamic schedulers for multi-QoS optimization.

2 Objectives

The objective of this part of the PhD thesis are as follows:

- Study the URLLC performance limitations of the standard throughput-based spatial schedulers.
- Design and develop several multi-QoS-aware spatial schedulers for joint latency-capacity optimization.

3 Included Articles

The main relevant papers of this PhD part are listed as follows:

Paper A: Multi-User Preemptive Scheduling For Critical Low Latency Communications in 5G Networks

This paper introduces a multi-stage dynamic UE scheduler for joint URLLC-eMBB macro deployments, where the urgent URLLC packets are always pri-

oritized over the eMBB type for a faster scheduling. In case sufficient radio resources are not immediately available for newly arriving URLLC packets, mainly due to the ongoing eMBB transmissions, the scheduler seeks to fit the URLLC packets in a MU-MIMO transmission with the respective active eMBB traffic flows. The success of the URLLC-eMBB MU-MIMO association is solely based on the maximization of the overall network capacity. In case of a failed MU-MIMO pairing, the urgent URLLC traffic immediately overwrites part of the ongoing eMBB traffic, minimizing the URLLC scheduling queuing delay at the expense of a degraded eMBB capacity. The performance of the proposed solution is evaluated by extensive system level simulations, and compared to the state-of-the-art literature.

Paper B: Null Space Based Preemptive Scheduling For Joint URLLC and eMBB Traffic in 5G Networks

This paper presents a novel null space based preemptive scheduler (NSBPS) for joint URLLC-eMBB networks. The objective of the proposed scheduling framework is to offer an instant URLLC scheduling without further scheduling queuing delays while inflicting a marginal loss of the eMBB spectral efficiency. A predefined subspace is constructed in the spatial domain and assumed known at the URLLC UE ends. In case the sufficient radio resources are not instantly schedulable for arriving URLLC packets, the scheduler pairs the corresponding URLLC UE with the active eMBB UE whose spatial precoder is the closest possible to the defined reference subspace. The scheduler spatially projects the selected eMBB transmission *on-the-go* while the respective URLLC UE shall orient its decoding interference rejection and combining (IRC) matrix into a possible orthogonal sub-space of the reference sub-space. This way, the URLLC traffic is immediately scheduled without inflicting large queuing delays, regardless of the originally available sDoFs and the active eMBB load, and with substantially low inter-UE interference. This paper envisions the importance of the spatial scheduler agility to enforce the needed sDoFs when URLLC packet queuing is foreseen.

Paper C: Opportunistic Spatial Preemptive Scheduling for URLLC and eMBB Coexistence in Multi-User 5G Networks

This paper is an extended version of the Paper B. The paper evaluates the performance of the proposed NSBPS scheduler in a wider and more comprehensive manner with different system settings, various eMBB and URLLC load distributions, respectively. The eMBB traffic is modeled as constant bit rate (CBR), instead of the former full-buffer assumption, to stimulate the broadband live streaming services. Furthermore, due to the *on-the-fly* eMBB precoder projections of the NSBPS scheduler, the victim eMBB UEs exhibit a

capacity loss accordingly. In this work, we analytically evaluate such eMBB loss compared to the state-of-the-art URLLC-eMBB schedulers, where the proposed scheduling framework has demonstrated a marginal eMBB loss.

Paper D: Capacity Optimization of Spatial Preemptive Scheduling for Joint URLLC-eMBB Traffic in 5G New Radio

This paper is based on the proposed scheduling frameworks of Papers A-C. It introduces a capacity recovery mechanism of the victim eMBB UEs, when the former NSBPS scheduler is adopted. The objective is that BSs transfer the knowledge of the performed *on-the-fly* precoder projections to the victim eMBB UEs such that they could re-orient the spatial span of their respective decoders to their original subspace prior to projection. The performance of the proposed capacity enhancement is assessed by extensive system level simulations, where it has been shown that proposed solution offers a considerable improvement of the eMBB capacity and SINR, respectively.

Paper E: Preemption-Aware Rank Offloading Scheduling For Latency Critical Communications in 5G Networks

In this paper, we take on step further and propose an agile URLLC-eMBB UE spatial scheduler. The paper introduces a preemption-aware rank offloading scheduling (PAROS) framework. The high-level idea is that the sDoFs of the antenna arrays are fully utilized to maximize the eMBB spectral efficient; although, those are flexibly traded-off for the sake of the stringent URLLC QoS only when needed. Compared to the scheduling frameworks introduced in papers A-D, the proposed solution herein offers a better scheduling flexibility to offer improved eMBB capacity while achieving a similar URLLC outage performance. In particular, it implies that the scheduler is always attempting eMBB-eMBB MU-MIMO pairings to achieve the maximum possible network spectral efficiency until the maximum MU-MIMO rank is reached on each physical radio block (PRB). In case of incoming URLLC packets and no radio resources are instantly available, the URLLC-eMBB NSBPS scheduler is applied over the PRBs which are monopolized by the ongoing eMBB transmissions and acquire less than the maximum allowable MU-MIMO transmission rank, respectively. Due to the aggressive eMBB-eMBB MU-MIMO transmission, the system PRBs may be overloaded by the maximum MU rank, and such URLLC-eMBB pairing could not be possible. Therefore, the scheduler preemptively offloads the MU rank over a number of PRBs which is sufficiently enough to accommodate the arriving URLLC packets in a URLLC-eMBB MU-MIMO transmission. Proposed scheduler shows a significant improvement of the achievable network capacity while preserving a decent URLLC outage performance.

Paper F: Channel Quality Feedback Enhancements for Accurate URLLC Link Adaptation in 5G Systems

This paper addresses the downlink link adaptation problem of the small-payload URLLC deployments, and is about achieving an accurate link adaptation for URLLC through enhanced CQI feedback methods. Within URLLC deployments, the load fluctuations of the interfering BSs become unpredictable and highly varying in time due to the fast transmissions of the sporadic and small-size URLLC packets. This leads the estimation of the channel quality indication (CQI) at the UE end to be inaccurate. Furthermore, CQI report can be sometimes outdated, from the moment it is measured at the UE until the moment the serving BS considers it in selecting the appropriate downlink MCS. Thus, we propose and develop a novel filtering technique at the UE in order to better estimate and report the lower percentiles of its channel quality distribution. The proposed solutions are proved beneficial, using extensive system level simulations, in terms of achieving a better URLLC outage latency performance. The newly developed *worst-M CQI* report allows the BS to schedule URLLC payloads over the frequency-domain with a random basis while still preserving a high probability of successful decoding despite exhibiting poor channel conditions, e..g, more than 50% reduction of the URLLC outage latency is observed with *worst-M* CQI reporting compared to the standard CQI reports for 8 Mbps of offered traffic load.

4 Main Findings and Recommendations

Main Findings

Paper A introduces an efficient and multi-user preemptive scheduler (MUPS) [15]. Particularity, a standard MU-MIMO pairing among URLLC-eMBB UEs is demonstrated as an attractive solution to reduce the URLLC scheduling queuing delays when the radio resources are monopolized by active eMBB transmissions. The success of the MU-MIMO pairing is highly dependent on the free available sDoFs from the BS antenna array at an arbitrary time. Thus, in case of insufficient sDoFs, a successful MU-MIMO pairing is not possible, thus, the incoming URLLC packets are buffered to the next scheduling opportunity which may jeopardize the achievable URLLC outage performance.

As depicted the by the overall flow diagram in Fig. II.1 [16, Paper C], Papers B, C, and D present a novel NSBPS scheduler, where it artificially enforces the sufficient sDoFs for immediate URLLC packet scheduling in case those sDoFs are not originally available, i.e., a throughput-based MU-MIMO URLLC-eMBB transmission is not possible.

Furthermore, Fig. II.2 [17, Paper B] presents the achievable URLLC outage latency, i.e., radio latency at the 10^{-5} outage probability, for the NSBPS

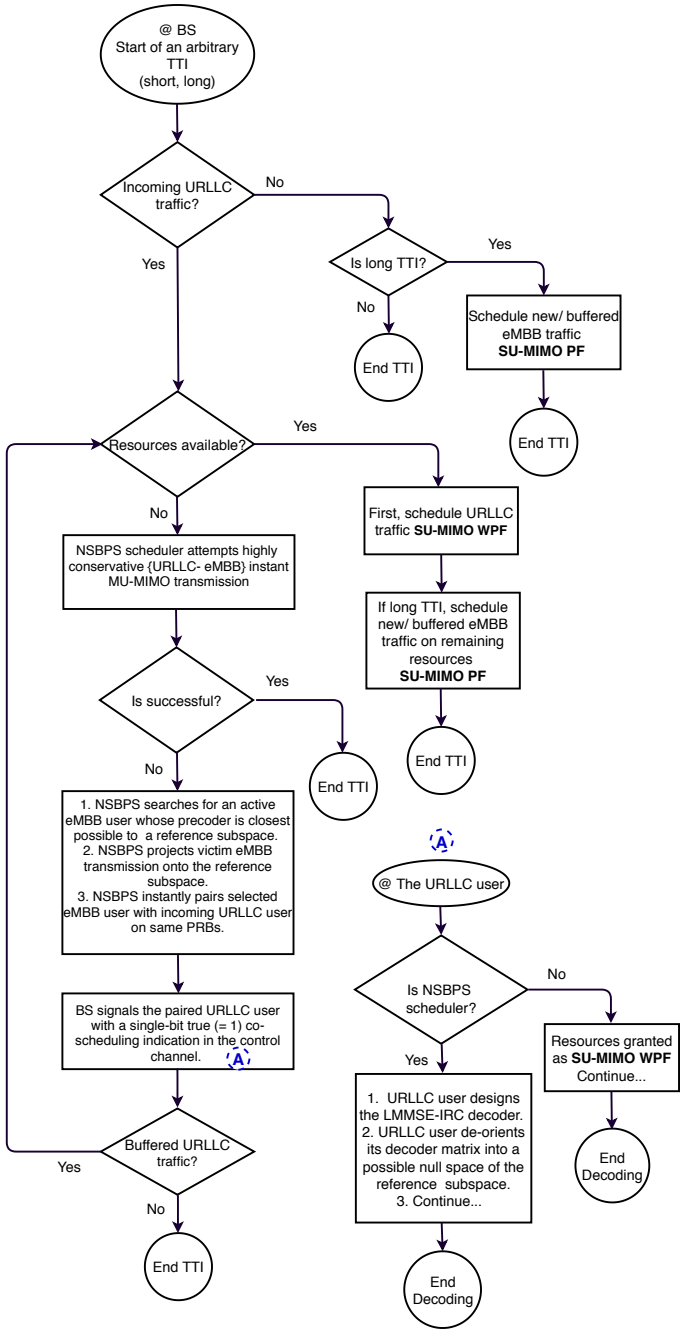


Fig. II.1: Flow diagram of the proposed NSBPS scheduler in papers B-D [16, Paper C].

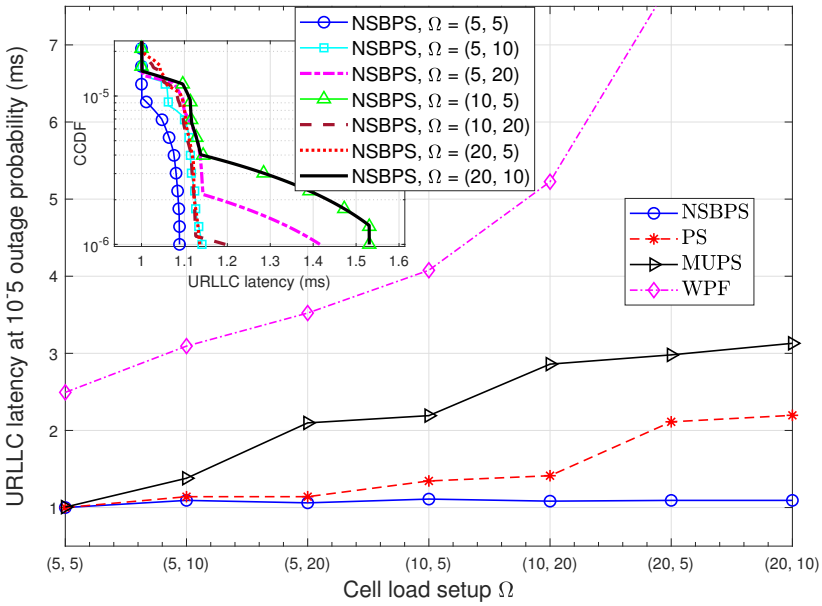


Fig. II.2: Achievable URLLC outage performance of the proposed spatial scheduler [17, Paper B].

[18], MUPS [15], weighted proportional fair (WPF), and preemptive scheduler [11], respectively. Those are evaluated with different load distributions $\Omega = (K_{\text{eMBB}}, K_{\text{URLLC}})$, where K_{eMBB} and K_{URLLC} are the average numbers of the eMBB and URLLC UEs per BS. The key observation is that the NSBPS scheduler preserves a decent and reliable URLLC outage performance, i.e., ~ 1 ms, regardless of the load distribution Ω. This is due to the *on-the-go* enforcement of the needed sDoFs, and hence, achieving efficient URLLC-eMBB MU-MIMO pairings. Furthermore, as can be observed from Fig. II.2, the throughput-based schedulers such as the MUPS and WPF clearly fail to offer a reliable URLLC outage latency.

Moreover, Paper E further extends the MAC scheduler agility by proposing the PAROS framework. The PAROS scheduler seeks to adaptively utilize the full potential of the antenna array sDoFs for maximizing the network spectral efficiency. However, it preemptively trades-off those sDoFs when URLLC queuing delay is foreseen. As depicted by the complementary cumulative distribution function (CCDF) of the URLLC radio latency in Fig. II.3 [18, Paper E], the proposed PAROS scheduling framework preserves a similar decent URLLC outage latency as the baseline NSBPS scheduler, while achieving 80% increase in the achievable MU throughput compared to the

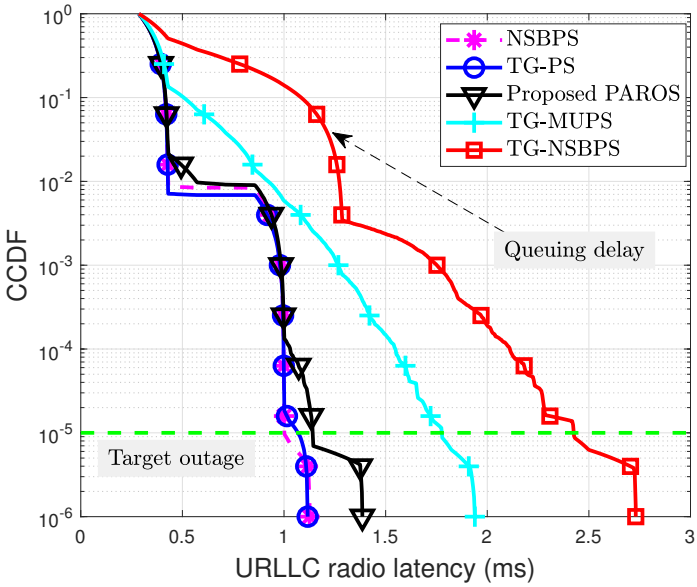


Fig. II.3: Achievable URLLC outage performance of the proposed PAROS scheduler [18, Paper E].

NSBPS scheduler, as depicted by Fig. II.4 [18, Paper E]. This is attributed to the offered scheduler flexibility to trade-off the network spectral efficiency with the stringent URLLC performance targets.

Main recommendations

In the following, we summarize the major research recommendations of this part of the thesis as follows:

1. Utilizing the spatial DoFs of the BS antenna array in the dynamic UE scheduling is an attractive solution to achieve the diverse QoS targets for the URLLC-eMBB service coexistence deployments.

2. UE-centric spatial scheduler are vital to achieve the diverse requirements of the URLLC and eMBB QoS targets as the network-centric, i.e., throughput-based, spatial schedulers fail to satisfy the stringent URLLC outage latency requirements.

3. The proposed multi-QoS spatial scheduling frameworks in this part can not be directly applied to current 3GPP specification of the 5G-NR. Those require additional standard impact by defining the required gNB-UE radio signaling and the URLLC/eMBB UE decoding behavior.

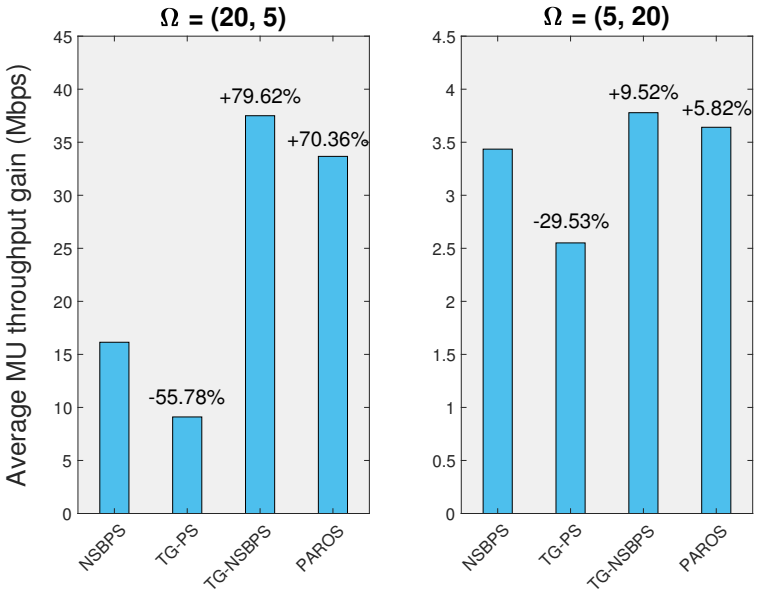


Fig. II.4: Achievable throughput performance of the proposed PAROS scheduler [18, Paper E].

References

[1] *NR and NG-RAN overall description*; Stage-2 (Release 15), 3GPP, TS 38.300, V2.0.0, Dec. 2017.

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

[3] 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. 28912-28922, May 2018.

[4] 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. VTC2019-Spring*, Kuala Lumpur, Malaysia, 2019, pp. 1-6.

[5] A. Karimi, K. I. Pedersen and P. Mogensen, "5G URLLC performance analysis of dynamic-point selection multi-user resource allocation," *in Proc. IEEE ISWCS*, Oulu, Finland, Oct. 2019, pp. 379-383.

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

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

[8] K. I. Pedersen, G. Pocovi and J. Steiner, "Preemptive scheduling of latency critical traffic and its impact on mobile broadband performance," *in Proc. IEEE VTC-Spring*, Porto, July 2018, pp. 1-6.

[9] M. Li, X. Guan, C. Hua, C. Chen and L. Lyu, "Predictive pre-allocation for low-latency uplink access in industrial wireless networks," *in Proc. IEEE INFOCOM*, Honolulu, HI, 2018, pp. 306-314.

[10] G. Gottardi, G. Oliveri and A. Massa, "Capacity-driven design of clustered array architectures for new generation 5G MU-MiMo systems," *in Proc. IEEE USNC-URSI*, Atlanta, GA, USA, 2019, pp. 1483-1484.

[11] K. Uchida, M. Fujimoto, K. Kitao and T. Imai, "Time-varying channel interference reduction by interference channel measurement in MU-MIMO," *in Proc. IEEE iWEM*, Qingdao, China, 2019, pp. 1-2.

[12] A. Kausar, S. Kausar and H. Mehrpouyan, "Hybrid beam-forming smart antenna for 5G networks," *in Proc. IEEE USNC-URSI*, Atlanta, GA, USA, 2019, pp. 1525-1526.

[13] J. M. McKinnis, I. Gresham and R. Becker, "Figures of merit for active antenna enabled 5G communication networks," *in Proc. IEEE GSMM*, Boulder, CO, USA, 2018, pp. 1-7.

[14] R. Kotaba, C. Navarro Manchn, T. Balercia and P. Popovski, "Uplink transmissions in URLLC systems With shared diversity resources," *in IEEE Wireless Commun. Letters*, vol. 7, no. 4, pp. 590-593, Aug. 2018.

[15] A. A. Esswie and K. I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5G networks," *in Proc. IEEE ISCC*, Natal, May 2018, pp. 00136-00141.

[16] 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. 38451-38463, 2018.

[17] A. A. Esswie and K. I. Pedersen, "Null space based preemptive scheduling for joint URLLC and eMBB traffic in 5G networks," *in Proc. IEEE Globecom*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1-6.

[18] A. A. Esswie, K. I. Pedersen and P. E. Mogensen, "Preemption-Aware Rank Offloading Scheduling for Latency Critical Communications in 5G Networks," *in Proc. IEEE VTC2019-Spring*, Kuala Lumpur, Malaysia, April 2019, pp. 1-6.

Paper A

Multi-User Preemptive Scheduling For Critical Low Latency Communications in 5G Networks

Ali A. Esswie and Klaus I. Pedersen

The paper has been published in the
2018 IEEE Symposium on Computers and Communications (ISCC)

Abstract

5G new radio is envisioned to support three major service classes: enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine type communications. Emerging URLLC services require up to one millisecond of communication latency with 99.999% success probability. Though, there is a fundamental trade-off between system spectral efficiency (SE) and achievable latency. This calls for novel scheduling protocols which cross-optimize system performance on user-centric; instead of network-centric basis. In this paper, we develop a joint multi-user preemptive scheduling strategy to simultaneously cross-optimize system SE and URLLC latency. At each scheduling opportunity, available URLLC traffic is always given higher priority. When sporadic URLLC traffic appears during a transmission time interval (TTI), proposed scheduler seeks for fitting the URLLC-eMBB traffic in a multi-user transmission. If the available spatial degrees of freedom are limited within a TTI, the URLLC traffic instantly overwrites part of the ongoing eMBB transmissions to satisfy the URLLC latency requirements, at the expense of minimal eMBB throughput loss. Extensive dynamic system level simulations show that proposed scheduler provides significant performance gain in terms of eMBB SE and URLLC latency.

Index Terms— URLLC; 5G; MU-MIMO; Channel hardening; RRM; Preemptive scheduling.

1 Introduction

The standardization of the fifth generation (5G) new radio (NR) is progressing with big momentum within the 3rd generation partnership project (3GPP) community, to release the first 5G specifications [1-3]. Ultra-reliable and low-latency communications (URLLC) is envisioned as a key requirement of the 5G-type communications, to support broad categories of many new applications from wireless industrial control, autonomous driving, and to tactile internet [4]. URLLC services require stringent latency and reliability levels, e.g., 1 ms at the $1 - 10^{-5}$ reliability level [5]. Such a challenging latency limit denotes that a URLLC packet which can not be transmitted and successfully decoded before the URLLC latency deadline, is considered as information-less and of no-use.

Simultaneously achieving the requirements of extreme spectral efficiency (SE) for enhanced mobile broadband (eMBB) services and ultra-low latency for URLLC applications is a challenging problem [6]. Achieving such URLLC latency demands more radio resources with ultra-low target block error rate (BLER); though, it leads to a significant loss in the network SE. Also, reserving dedicated resources for URLLC traffic is spectrally inefficient due to its sporadic nature.

To meet the stringent URLLC requirements, various studies have been recently presented in the open literature. User-specific scheduling with flexible transmission time intervals (TTIs) [7, 8] is recognized as an enabler to achieve the URLLC latency limit, e.g., URLLC traffic with a short TTI and eMBB with a longer TTI. However, the former increases the aggregate overhead of the control channel. Additionally, different configurations of microscopic and macroscopic diversity [9] are proven beneficial for URLLC to significantly reduce the outage probability of the signal-to-interference-noise-ratio (SINR). Advanced medium access control enhancements [10] are also reported towards optimized scheduling of URLLC traffic, including link adaptation filtering in partly-loaded cells, dynamic and load-dependent BLER optimization. Furthermore, preemptive scheduling [11, 12] is recently studied to instantly schedule URLLC traffic within a shared channel, monopolized by an ongoing eMBB transmission. Compared to existing studies, achieving the URLLC latency requirements comes at the expense of a degraded SE, e.g., high degrees of macroscopic diversity. Needless to say that a flexible and multi-objective scheduling algorithm, which captures the maximal system degrees of freedom (DoFs), is critical to reach the best achievable URLLC-eMBB multiplexing gain.

In this paper, a multi-user preemptive scheduling (MUPS) strategy for densely populated 5G networks is proposed. MUPS aims to simultaneously cross-optimize the network SE and URLLC latency. At each scheduling TTI, MUPS scheduler assigns URLLC traffic a higher priority for immediate scheduling without buffering. If sporadic URLLC traffic arrives at the 5G general NodeB (gNB) during an arbitrary TTI, the gNB first attempts to fit the URLLC packets within an ongoing eMBB transmission. If the spatial DoFs are insufficient, the gNB decides to immediately overwrite, i.e., preemptively schedule (PS), the physical resource blocks (PRBs) over which URLLC users reported the best received SINR. Compared to conventional PS scheduler, proposed MUPS utilizes the spatial DoFs, offered by the transmit antenna array, to extract the best achievable multiplexing gain, satisfying *both*: URLLC latency budget and eMBB throughput requirements.

Due to the complexity of the 5G NR system and the addressed problems, performance evaluation is validated using advanced system level simulations which offer high degree of realism and ensure reliable statistical results. Those simulations are based on widely accepted models and being calibrated with the 3GPP 5G NR assumptions [1-3].

This paper is organized as follows. Section 2 presents the system model. Section 3 outlines the problem formulation and proposed MUPS scheduler. Performance analysis appears in Section 4 and the paper is concluded in Section 5.

2 System Model

We consider a downlink (DL) multi-user multiple-input multiple-output (MU-MIMO) system, with C cells. Each cell is equipped with N_t transmit antennas while there are K-uniformly-distributed users per cell, each with M_r receive antennas. Users are dynamically multiplexed through orthogonal frequency division multiple access (OFDMA), and with 15 KHz sub-carrier spacing. There are two types of DL traffic under evaluation: (1) URLLC time-sporadic traffic of Z-bit finite payload per user with a Poisson point arrival process λ, and (2) eMBB full buffer traffic with infinite payload. The cell loading condition is described by $K_{URLLC} + K_{eMBB} = K$, where K_{URLLC} and K_{eMBB} denote the average number of URLLC and eMBB users per cell, respectively. URLLC traffic is scheduled with a short TTI of 2 OFDM symbols (mini-slot of 0.143 ms) to meet the URLLC latency budget [1]. However, eMBB users are scheduled with a long TTI of 14 OFDM symbols (slot of 1 ms) to maximize system SE.

A maximum MU subset $G \in K$, where $G_c \leq N_t$ is allowed per PRB per cell, with equal power sharing. Thus, the received DL signal at the k^{th} user from the c^{th} cell is given by

$$\mathrm{y}_{k,c} = \mathbf{H}_{k,c}\mathbf{V}_{k,c}s_{k,c} + \sum_{g\in G_c, g\neq k} \mathbf{H}_{k,c}\mathbf{V}_{g,c}s_{g,c} + \sum_{j=1, j\neq c}^{C} \sum_{g\in G_j} \mathbf{H}_{g,j}\mathbf{V}_{g,j}s_{g,j} + \mathbf{n}_k, \tag{A.1}$$

where $\mathbf{H}_{k,c} \in \mathcal{C}^{M_r \times N_t}, \forall k \in \{1,\ldots,K\}, \forall c \in \{1,\ldots,C\}$ is the 3GPP spatial channel matrix seen by the k^{th} user from the c^{th} cell, $\mathbf{V}_{k,c} \in \mathcal{C}^{N_t \times 1}$ and $s_{k,c}$ are the precoding vector (assuming a single stream transmission) and the transmitted symbol, respectively. $\mathbf{n}_k$ is the additive Gaussian white noise at the k^{th} user. The first summation in eq. (A.1) stands for the inter-user interference and the second considers the inter-cell interference. The received signal after applying the antenna combining vector $\mathbf{U}_{k,c} \in \mathcal{C}^{M_r \times 1}$ is given by

$$\mathrm{y}^*_{k,c} = (\mathbf{U}_{k,c})^{\mathrm{H}}\,\mathrm{y}_{k,c}, \tag{A.2}$$

where $(.)^{\mathrm{H}}$ indicates the Hermitian transpose. The antenna combining vector is designed based on the linear minimum mean square error interference rejection combining (LMMSE-IRC) criteria [13], in order to project the received signal on a signal subspace which minimizes the MSE, given by

$$\mathbf{U}_{k,c} = \left(\mathbf{H}_{k,c}\mathbf{V}_{k,c}\,(\mathbf{H}_{k,c}\mathbf{V}_{k,c})^{\mathrm{H}} + \mathrm{W}\right)^{-1} \mathbf{H}_{k,c}\mathbf{V}_{k,c}, \tag{A.3}$$

where $\mathrm{W} = \mathbb{E}\left(\mathbf{H}_{k,c}\mathbf{V}_{k,c}\left(\mathbf{H}_{k,c}\mathbf{V}_{k,c}\right)^{\mathrm{H}}\right) + \sigma^2\mathrm{I}_{M_r}$ is the interference covariance matrix, $\mathbb{E}(.)$ denotes the statistical expectation, and I_{M_r} is $M_r \times M_r$ identity matrix. The received SINR at the k^{th} user can be expressed as

$$\Upsilon_{k,c} = \frac{p_k^c \left|\mathbf{H}_{k,c}\mathbf{V}_{k,c}\right|^2}{1 + \sum\limits_{g\in G_c, g\neq k} p_g^c \left|\mathbf{H}_{k,c}\mathbf{V}_{g,c}\right|^2 + \sum\limits_{j\in C, j\neq c}\sum\limits_{g\in G_j} p_g^j \left|\mathbf{H}_{g,j}\mathbf{V}_{g,j}\right|^2}, \tag{A.4}$$

where p_k^c is the transmission power of the k^{th} user in the c^{th} cell. The per-user per-PRB data rate can then be calculated as,

$$r_{k,rb} = \log_2\left(1 + \frac{1}{\eta_c}\Upsilon_{k,c}\right), \tag{A.5}$$

where $\eta_c = \mathbf{card}(G_c)$ is the MU rank on this PRB.

Moreover, the link adaptation of the data transmission is based on the frequency-selective channel quality indication (CQI) reports to satisfy a target BLER. However, the CQI reports from the MU pairs can be misleading since the calculation of the inter-user interference and power sharing are not considered in the CQI estimation. Hence, to stabilize the link adaptation process against MU variance, an offset of δ dB is applied to the single-user (SU) CQI values before the modulation and coding scheme (MCS) level is selected,

$$\Gamma_{\mathrm{MU}} = \Gamma_{\mathrm{SU}} - \delta, \tag{A.6}$$

where Γ_{MU} and Γ_{SU} are the updated MU and reported CQI levels, respectively. Additionally, due to the bursty nature of the URLLC traffic, it sporadically destabilizes the reported CQI levels [10], especially when an MU transmission is not possible due to the fast varying interference patterns; otherwise, the interference from the co-scheduled users contributes to stabilizing the URLLC CQI levels. Thus, we further apply a sliding filter, e.g., a low pass filter, in order to smooth the instantaneous variation rate of the CQI levels as follows,

$$\partial(t) = \xi\Gamma_{\mathrm{MU}} + (1-\xi)\partial(t-1), \tag{A.7}$$

where $\partial(t)$ is the MU CQI value to be considered for link adaptation and MCS selection at the t^{th} TTI, and $\xi \leq 1$ is a tunable coefficient to specify how much weight should be given to current reported CQI value.

3 Proposed Multi-User Preemptive Scheduling

In this section, the concept of the proposed MUPS scheduler is introduced. Under the 5G umbrella, there are multi user-specific, instead of network-

specific, objectives which need to be fulfilled simultaneously, e.g., eMBB SE maximization, URLLC latency and BLER minimization as follows,

$$\forall k_{eMBB} \in \mathcal{K}_{eMBB} : \arg \max_{\mathcal{K}_{eMBB}} \sum_{k_{eMBB}=1}^{K_{eMBB}} \sum_{rb \in RB_k} r_{k,rb}, \tag{A.8}$$

$$\forall k_{URLLC} \in \mathcal{K}_{URLLC} : \arg \min_{\mathcal{K}_{URLLC}} (\beta)\,, \beta \leq 1\,\text{ms}, \tag{A.9}$$

$$\forall k \in \mathcal{K} : \arg \min_{\mathcal{K}} (\psi), \tag{A.10}$$

where $\mathcal{K}_{eMBB}$ and $\mathcal{K}_{URLLC}$ denote the set of active eMBB and URLLC users, respectively. β and ψ indicate the URLLC latency at the $1 - 10^{-5}$ reliability level and user BLER, respectively. This is a challenging and non-trivial optimization problem, e.g., achieving Shannon SE requires infinite latency budget. The proposed MUPS aims at achieving the maximum possible system SE, while at the same time preserving the URLLC required latency.

As shown in Fig. A.1, if there is no incoming URLLC traffic at an arbitrary TTI, MUPS assigns SU dedicated resources to incoming or buffered eMBB traffic based on the proportional fair (PF) criteria as

$$\Theta_{\text{PF}} = \frac{r_{k,rb}}{\overline{r_{k,rb}}}, \tag{A.11}$$

$$k^*_{eMBB} = \arg \max_{\mathcal{K}_{eMBB}} \Theta_{\text{PF}}, \tag{A.12}$$

where $\overline{r_{k,rb}}$ is the average delivered data rate of the k^{th} user. If incoming URLLC traffic is aligned at the start of the current TTI, e.g., either it is a short URLLC or long eMBB TTI, MUPS applies the weighted PF (WPF) criteria to instantly schedule URLLC traffic with a higher priority on available resources as given by

$$\Theta_{\text{WPF}} = \frac{r_{k,rb}}{\overline{r_{k,rb}}}\alpha, \tag{A.13}$$

where α is the scheduling coefficient and $\alpha_{URLLC} \gg \alpha_{eMBB}$. Afterwards, MUPS schedules pending or new eMBB traffic on remaining resources.

If URLLC traffic arrives at the gNB during an eMBB TTI transmission while scheduling resources are not available, gNB attempts to dynamically multiplex the incoming short-TTI URLLC users within the ongoing long-TTI eMBB transmissions, if there are sufficient spatial DoFs on this TTI. The spatial DoFs represent the ability to jointly process several signals between different sets of transmitters and receivers, if corresponding channels are highly uncorrelated. Accordingly, URLLC users experience no buffering overhead and then the URLLC latency budget can be satisfied. If a successful pairing, i.e., MU URLLC-eMBB transmission over an arbitrary PRB, is not possible,

gNB will instantly overwrite the best reported PRBs, known from the URLLC CQI reports, with the incoming URLLC traffic. Thus, victim eMBB transmissions will exhibit a throughput loss.

For $N_t = 8$ transmit antennas at the gNB, dual codebooks are defined in LTE-Pro standards [14] for DL channel quantization at the user's side, and are given by

$$\boldsymbol{\Lambda}_1 = \left\{ \boldsymbol{v}_{1,1}, \boldsymbol{v}_{1,2} \ldots, \boldsymbol{v}_{1,2^{B_1}} \right\}, \tag{A.14}$$

$$\boldsymbol{\Lambda}_2 = \left\{ \boldsymbol{v}_{2,1}, \boldsymbol{v}_{2,2} \ldots, \boldsymbol{v}_{2,2^{B_2}} \right\}, \tag{A.15}$$

where $\boldsymbol{v}_{i,j}$ denotes the j^{th} codeword of the i^{th} codebook, B_1 and B_2 are the numbers of bits of the two precoding matrix indices, reported from each user for the gNB to select one codeword from each codebook. Each user projects its estimated DL channel on both codebooks to select the closest possible codewords as

$$\widehat{\boldsymbol{v}}_1 = \arg \max_{\boldsymbol{v}_1 \in \boldsymbol{\Lambda}_1} \left\| \hat{\mathbf{H}} \boldsymbol{\Lambda}_1 \right\|^2, \tag{A.16}$$

$$\widehat{\boldsymbol{v}}_2 = \arg \max_{\boldsymbol{v}_2 \in \boldsymbol{\Lambda}_2} \left\| \hat{\mathbf{H}} \boldsymbol{\Lambda}_2 \right\|^2, \tag{A.17}$$

where $\|.\|$ denotes the 2-norm operation. The final precoding vector at the gNB is obtained by the spatial multiplication of both precoders, and is given by

$$\mathbf{V} = \widehat{\boldsymbol{v}}_1 \times \widehat{\boldsymbol{v}}_2. \tag{A.18}$$

For a MU transmission on a given PRB, the zero-forcing (ZF) beamforming is used to null the inter-user interference between the co-scheduled pairs as expressed by

$$\mathbf{V}_{\mathrm{MU}} = \left[\mathbf{V}_1 \ldots \mathbf{V}_{\mathrm{G}} \right], \tag{A.19}$$

$$\mathbf{V}_{\mathrm{zf}} = \mathbf{V}_{\mathrm{MU}} \left(\mathbf{V}_{\mathrm{MU}}^{\mathrm{H}} \mathbf{V}_{\mathrm{MU}} \right)^{-1} \mathrm{diag} \left(\sqrt{P} \right), \tag{A.20}$$

where $\mathbf{V}_{\mathrm{G}}$ and $\mathbf{V}_{\mathrm{zf}}$ present the precoder of the g^{th} user enrolled in a MU-MIMO transmission and the ZF beamforming matrix, where its column vectors are the data beamforming vectors of the MU pairs. The MU transmission success is based on the maximization of the Chordal distance between the ZF beamformers of the co-scheduled users as follows,

$$\arg \max_{\mathrm{V_{eMBB}} \in \boldsymbol{\mathcal{V}}_{\mathrm{eMBB}}} \mathbf{d} \left(\mathrm{V_{URLLC}}, \mathrm{V_{eMBB}} \right), \tag{A.21}$$

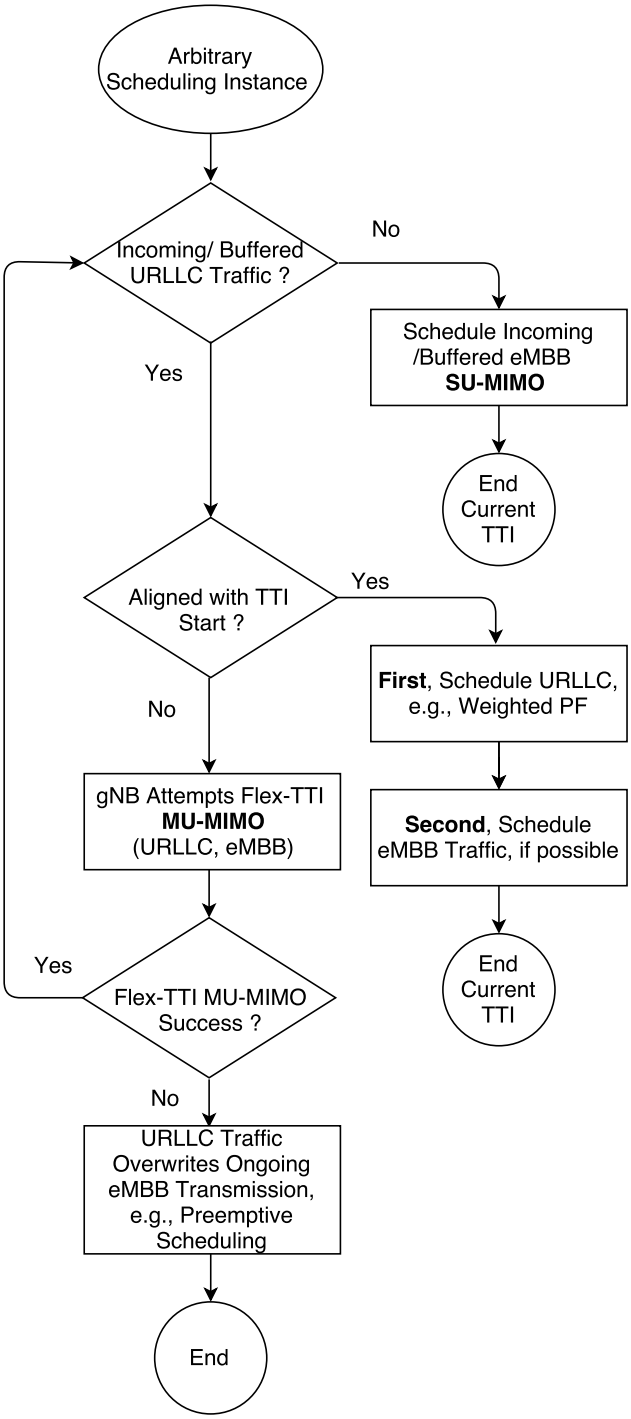


Fig. A.1. Flow diagram of proposed MUPS scheduler.

where $\boldsymbol{\mathcal{V}}_{\text{eMBB}}$ represents the set of ZF precoders of the eMBB active user set. The Chordal distance is calculated as

$$\mathbf{d}\left(\mathrm{V}_{\text{URLLC}}, \mathrm{V}_{\text{eMBB}}\right) = \frac{1}{\sqrt{2}} \left\| \mathrm{V}_{\text{URLLC}} \mathrm{V}_{\text{URLLC}}^{\mathrm{H}} - \mathrm{V}_{\text{eMBB}} \mathrm{V}_{\text{eMBB}}^{\mathrm{H}} \right\|. \tag{A.22}$$

Upon MU pairing success, the aggregate achievable data rate on a given PRB r_{rb} is expressed by the sum rate of both co-scheduled URLLC and eMBB users as

$$r_{rb} = \left(r_{eMBB} + r_{URLLC} - \Delta\right), \tag{A.23}$$

where Δ represents the eMBB and URLLC SU rate loss due to the MU inter-user interference. If a MU pairing is not possible, due to either insufficient spatial DoFs or low number of active eMBB users, the URLLC traffic immediately overwrites the PRBs over which it experiences the best CQI levels. Thus, the eMBB users which have ongoing transmissions on these PRBs suffer from throughput degradation. However, recovery mechanisms can be arbitrarily considered not to include these PRBs as part of the HARQ chase combining process and propagate errors, e.g., consider these PRBs as information-less. Then, the sum rate on victim PRBs can be expressed only by the achievable URLLC rate as

$$r_{rb} = r_{URLLC}. \tag{A.24}$$

For the sake of a fair URLLC latency evaluation, we compare the MUPS performance with the preemptive-only scheduling (PS) [11], where incoming URLLC traffic always overwrites ongoing eMBB transmissions without buffering, at the expense of the system SE. As it will be discussed in Section 4, we demonstrate that a conservative multi-TTI MU-MIMO transmission can be an attractive solution to approach both URLLC latency and eMBB SE requirements.

4 Simulation Results

Extensive dynamic system level simulations have been conducted, following the 5G NR specifications in 3GPP [3]. The major simulation parameters are listed in Table A.1, where the baseline antenna setup is 8×2 unless otherwise mentioned.

Fig. A.2 shows the empirical complementary cumulative distribution function (CCDF) of the URLLC latency statistics. We define the cell loading state by $\Omega = (K_{eMBB}, K_{URLLC})$, where the aggregate URLLC offered load per cell in bits/s is calculated as: $K_{URLLC} \times \lambda \times Z$. Looking at the URLLC latency

Table A.1: Simulation Parameters.

Parameter	Value
Environment	3GPP-UMA,7 gNBs, 21 cells, 500 meters inter-site distance
Channel bandwidth	10 MHz, FDD
gNB antennas	8, 16 and 64 Tx, 0.5λ
User antennas	2, 8, 16 and 64 Rx, 0.5λ
User dropping	uniformly distributed URLLC: 5 and 10 users/cell eMBB: 5 , 10 and 20 users/cell
User receiver	LMMSE-IRC
TTI configuration	URLLC: 0.143 ms (2 OFDM symbols) eMBB: 1 ms (14 OFDM symbols)
MAC scheduler(s)	URLLC: WPF, SU/MU-MIMO and PS eMBB: PF, and SU/MU-MIMO
CQI	periodicity: 5 ms, with 2 ms latency, $\xi = 0.01$
HARQ	asynchronous HARQ, chase combining HARQ round trip time = 4 TTIs
Link adaptation	dynamic MCS target URLLC BLER : 1% target eMBB BLER : 10%
Traffic model	URLLC: bursty, Z=50 bytes, $\lambda = 250$ eMBB: full buffer
MU-MIMO setup	MU beamforming : ZF MU rank (η) : 2 CQI offset (δ) : 3 dB
Link to system mapping	Mean mutual information per coded bit [11]

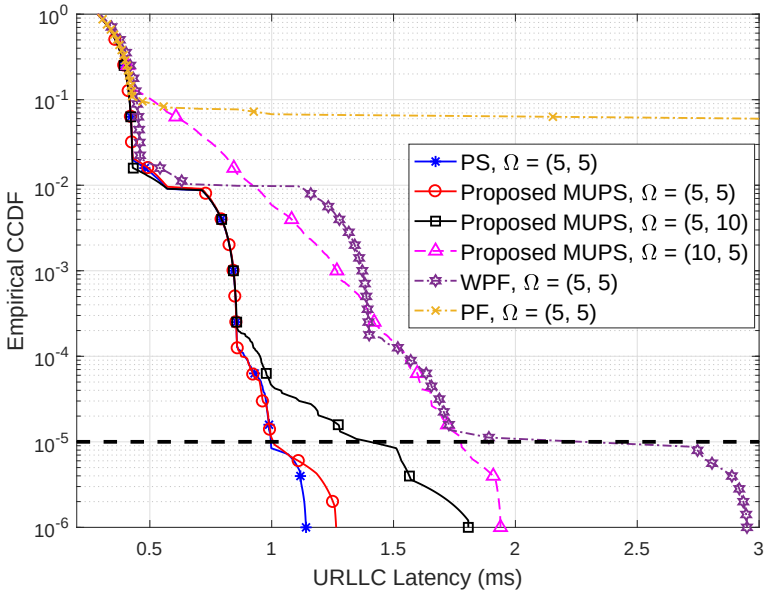


Fig. A.2. URLLC latency of MUPS, PS, PF and WPF schedulers.

at the 10^{-5} level, both proposed MUPS and PS schedulers achieve the 1-ms limit with $\Omega = (5,5)$. By increasing the system loading, e.g., $K_{eMBB} = 10$ and $K_{URLLC} = 10$, the inter-cell interference becomes a dominant component and hence, all schedulers suffer from throughput and latency degradation. Though, MUPS scheduler still shows a decent URLLC latency performance, e.g., 1.7 ms at 10^{-5} level.

PF scheduler suffers from URLLC latency error floor since both URLLC and eMBB users have the same scheduling priority, thus, URLLC large queuing delays occur. WPF shows optimized URLLC latency; however, it doesn't achieve the 1-ms limit since the sporadic URLLC traffic, which is available during an eMBB TTI transmission, is buffered, i.e., not scheduled instantly, until the next available TTI opportunity.

Fig. A.3 shows the empirical CDF of the average cell throughput in Mbps of the proposed MUPS and PS schedulers under different loading conditions. Under all cell loading states, the MUPS scheduler shows significant gain over PS scheduler, e.g., ~ 26.54% gain with $\Omega = (20,5)$. MUPS scheduler exhibits a better system SE due to: (1) the successful multi-TTI MU transmissions, and (2) reduction in the number of the experienced PS scheduling events. For the same number of the URLLC users K_{URLLC}, increasing the number of eMBB users K_{eMBB} significantly enhances the MU DoFs, hence, an incoming URLLC user has higher probability to experience an immediate MU pairing

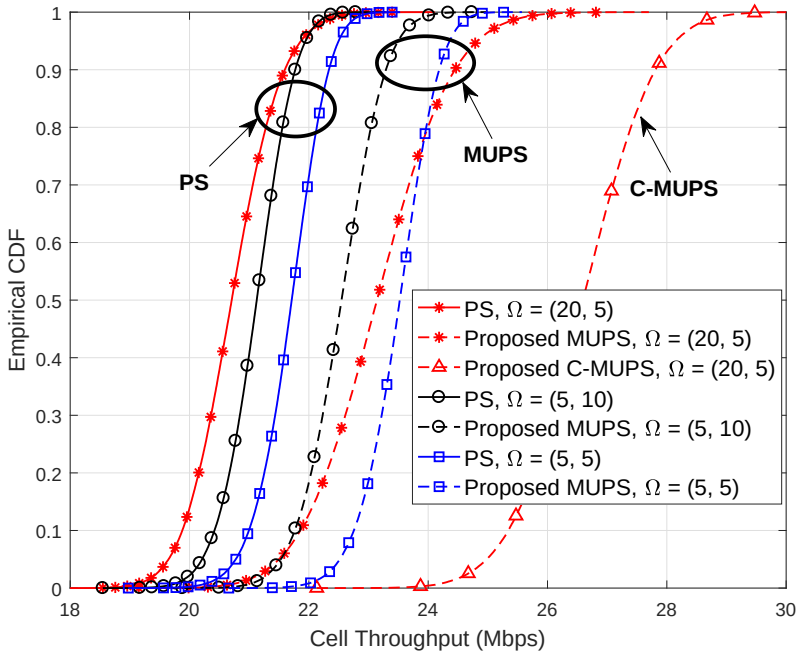


Fig. A.3. Cell throughput of MUPS, C-MUPS and PS schedulers.

success, without falling back to SE-less-efficient PS scheduling. Under such high K_{eMBB} loading, MUPS scheduler attempts many MU pairing success events; however, with limited MU gain due to the aggregate level of inter-cell interference and the higher buffering time. Thus, we also consider a modified version of the MUPS scheduler, denoted as conservative MUPS (C-MUPS), where the URLLC-eMBB pairing success becomes more restricted by the user spatial separation as

$$|\angle(\mathrm{V}_{\mathrm{URLLC}}) - \angle(\mathrm{V}_{\mathrm{eMBB}})|^{o} \geq \theta, \tag{A.25}$$

where θ is a predefined spatial separation threshold. Thus, C-MUPS achieves lower number of MU attempts with further significant MU gain, e.g., ~ 62% gain in average cell throughput with $\Omega = (20, 5)$ and $\theta = 60^{o}$, as shown in Fig. A.3.

As depicted in Fig. A.4, it shows the average achievable MU throughput increase with respect to average SU throughput. As can be noticed, increasing K_{URLLC} offers limited DoFs due to the short TTI length of the URLLC users. Furthermore, increasing the URLLC load results in more sporadic packet arrivals and hence, destabilizing the link adaptation. Increasing the eMBB load offers great spatial DoFs per each URLLC user. With C-MUPS, it shows

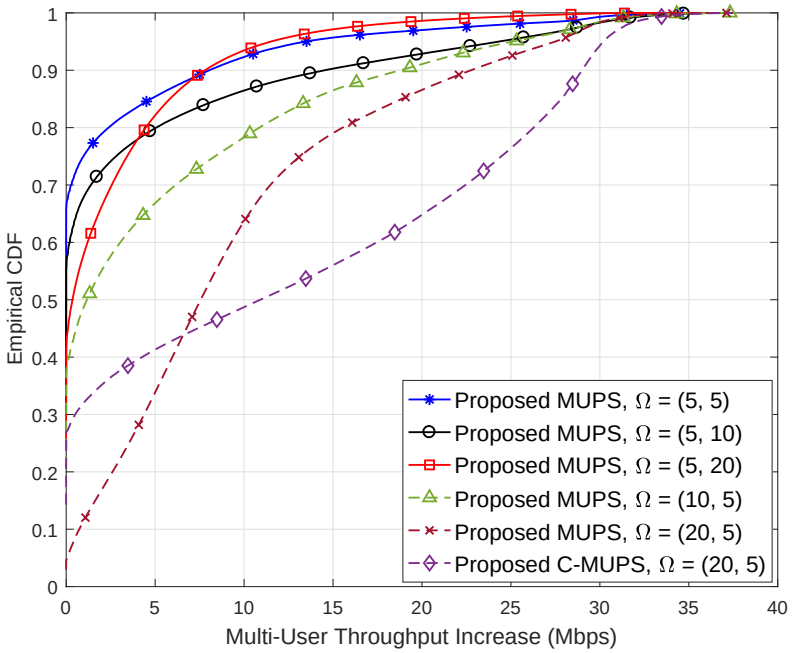


Fig. A.4. MU throughput of the MUPS and C-MUPS schedulers.

that less MU success events are experienced, e.g., 72% instead of 95% for MUPS with $\Omega = (20, 5)$; however, further higher MU throughput is achieved.

Examining the eMBB user performance, Fig. A.5 presents a comparison of the eMBB average user throughput. Proposed scheduler shows improved eMBB user throughput, under all loading conditions. The gain in the eMBB user throughput is strongly dependent on the levels of inter-cell and inter-user interference. With light loading conditions, e.g., $\Omega = (5, 5)$, the MUPS scheduler experiences few successful pairings with sub-optimal MU gain because of the insufficient available spatial DoFs, e.g., due to the low value of K_{eMBB} . On the opposite, under heavy loading conditions, e.g., $\Omega = (20, 5)$, MUPS achieves a higher number of successful MU pairings with higher MU gain as the quality of the MU transmission enhances with the number of active eMBB users K_{eMBB}.

Interestingly, the MU performance can be further improved with a larger number of antennas, equipped at both transmitter and receiver. Channel hardening [15, 16] denotes a fundamental channel phenomenon where the variance of the channel mutual information shrinks as the number of antennas grows,

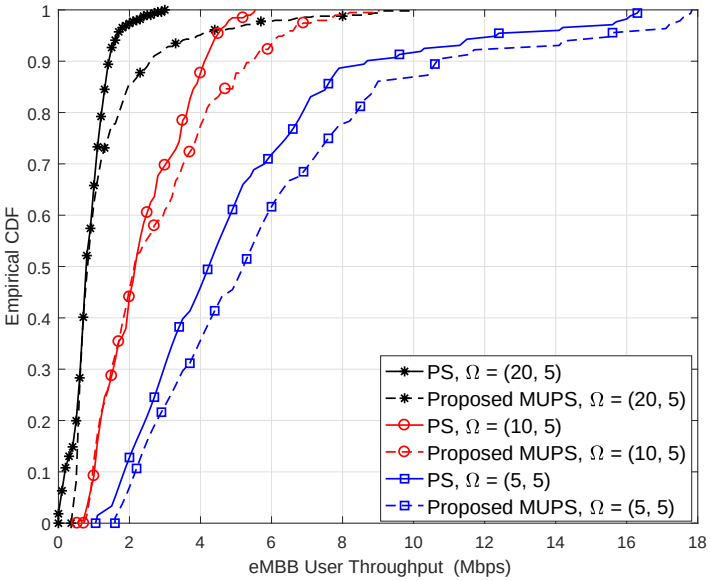


Fig. A.5. eMBB user throughput of the MUPS and PS schedulers.

$$\sigma^2 = \frac{1}{\min\left(N_t, M_r\right)} \left(\frac{\|\mathbf{H}\|^2}{\mathbb{E}\left(\|\mathbf{H}\|^2\right)} \right). \tag{A.26}$$

Consequently, the fading channel starts to act as a non-fading channel where the channel eigenvalues become less sensitive to the actual distribution of the channel entries. Thus, the channel hardens and becomes much more directional on desired paths with less leakage on the interfering paths, as shown in Fig. A.6. As a result, both MU and URLLC performance can be significantly improved.

Fig. A.7 introduces the received user SINR in dB, sampled over both URLLC and eMBB users with $\Omega = (20, 5)$. For a fair performance comparison, each user is assumed to feedback its serving cell with the exact channel entries without quantization, since there is no a standard quantization codebook for $N_t > 8$ and $M_r > 8$. The channel is decomposed and fed-back by the singular value decomposition [17] as: $\mathbf{H} = \boldsymbol{\mathcal{U}}\boldsymbol{\Sigma}\boldsymbol{\mathcal{V}}^{\mathrm{H}}$, where $\boldsymbol{\mathcal{U}} \in \mathcal{C}^{M_r \times M_r}$ and $\boldsymbol{\mathcal{V}} \in \mathcal{C}^{N_t \times N_t}$ are unitary matrices and $\boldsymbol{\Sigma} \in \Re^{M_r \times N_t}$ is the channel singular matrix. The received user SINR levels are significantly enhanced with the number of antennas due to the channel hardening effect. Consequently, further more MU successful pairing events can be achieved with sufficient spatial separation.

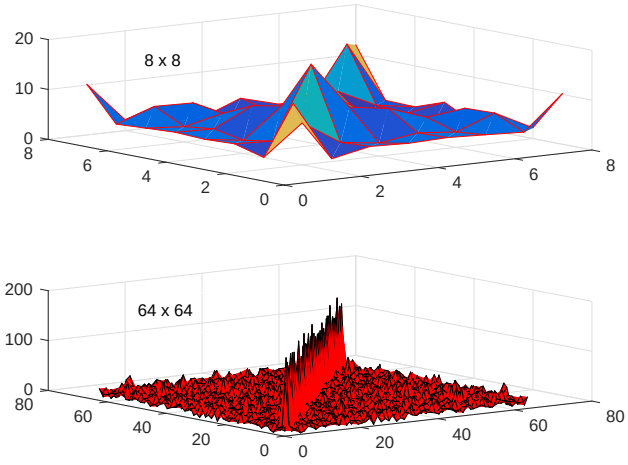


Fig. A.6. Channel hardening of $\mathrm{H}^{\mathrm{H}}\mathrm{H}$ with (N_t, M_r) setup.

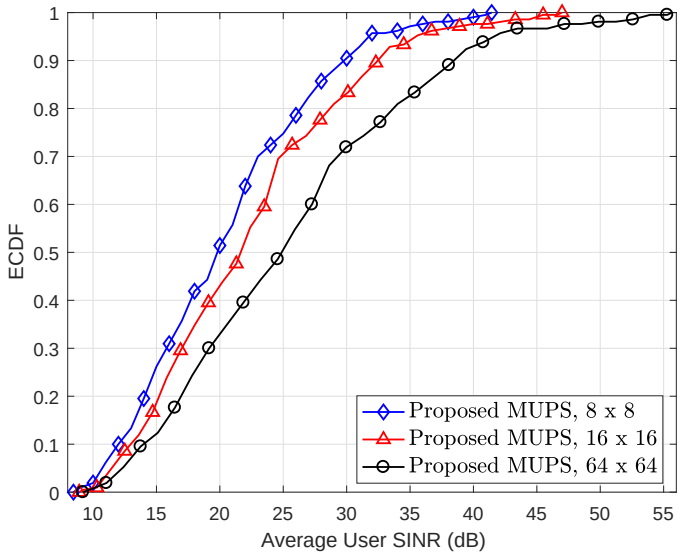


Fig. A.7. User received SINR with (N_t, M_r) setup.

5 Conclusion

In this work, a joint multi-user preemptive scheduler (MUPS) has been proposed for densely populated 5G networks. Proposed scheduler operates efficiently with different traffic types, e.g., full buffer enhanced mobile broadband (eMBB) and sporadic ultra-reliable low-latency communication (URLLC) traffic. MUPS cross-optimizes the network performance such that the maximum possible spectral efficiency and ultra low latency are simultaneously achievable. Using extensive system level simulations, the proposed scheduler provides significant performance gain, e.g., ~ 62% gain in average cell throughput, under different network configurations. The performance of the MUPS scheduler is shown to improve with the number of eMBB users until the interference levels become dominant. Hence, proposed conservative MUPS shows further enhanced MU gain by limiting the inter-user interference. Furthermore, increasing the number of antennas is shown to harden the wireless channel and thus, further improved URLLC performance can be satisfied. A detailed study on the robustness of the URLLC performance under such a scenario will be considered in a future work.

References

[1] NR and NG-RAN overall description; Stage-2 (Release 15), 3GPP, TS 38.300, V2.0.0, Dec. 2017.

[2] Study on new radio access technology; Radio access architecture and interfaces (Release 14), 3GPP, TR 38.801, V14.0.0, March 2017.

[3] Study on scenarios and requirements for next generation access technologies (Release 14), 3GPP, TR 38.913, V14.3.0, June 2016.

[4] IMT vision – "Framework and overall objectives of the future development of IMT for 2020 and beyond", international telecommunication union (ITU), ITU-R M.2083-0, Feb. 2015.

[5] E. Dahlman et al., "5G wireless access: requirements and realization," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 42-47, Dec. 2014.

[6] B. Soret, P. Mogensen, K. I. Pedersen and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," *in Proc. IEEE Globecom*, Austin, TX, 2014, pp. 1391-1396.

[7] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen and A. Szufarska, "A flexible 5G frame structure design for FDD cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53-59, March 2016.

[8] 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 Proc. IEEE Globecom*, Washington, DC, 2016, pp. 1-7.

[9] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen and P. Mogensen, "Signal quality outage analysis for URLLC in cellular networks," *in Proc. IEEE Globecom*, San Diego, CA, 2015, pp. 1-6.

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

[11] K.I. Pedersen, G. Pocovi, J. Steiner, and S. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," *in Proc. IEEE VTC*, Toronto, 2017, pp. 1-6.

[12] G. C. Buttazzo, M. Bertogna and G. Yao, "Limited preemptive scheduling for real-time systems: a survey," *IEEE Trans. Ind. Informat.*, vol. 9, no. 1, pp. 3-15, Feb. 2013.

[13] Y. Ohwatari, N. Miki, Y. Sagae and Y. Okumura, "Investigation on interference rejection combining receiver for space–frequency block code transmit diversity in LTE-advanced downlink," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 191-203, Jan. 2014.

[14] Evolved universal terrestrial radio access (E-UTRA); Physical layer procedures (Release 12), 3GPP, TS 36.213, V12.4.0, Feb. 2015

[15] T. L. Narasimhan and A. Chockalingam, "Channel hardening-exploiting message passing receiver in large-scale MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 847-860, Oct. 2014.

[16] A. A. Esswie, M. El-Absi, O. A. Dobre, S. Ikki and T. Kaiser, "A novel FDD massive MIMO system based on downlink spatial channel estimation without CSIT," *in Proc. IEEE ICC*, Paris, 2017, pp. 1-6.

[17] D. W. Browne, M. W. Browne and M. P. Fitz, "CTH07-4: singular value decomposition of correlated MIMO channels," *in Proc. IEEE Globecom*, San Francisco, CA, 2006, pp. 1-6.

Paper B

Null Space Based Preemptive Scheduling For Joint URLLC and eMBB Traffic in 5G Networks

Ali A. Esswie and Klaus I. Pedersen

The paper has been published in the
2018 IEEE Global Communications Conference (GLOBECOM)

Abstract

In this paper, we propose a null-space-based preemptive scheduling framework for cross-objective optimization to always guarantee robust URLLC performance, while extracting the maximum possible eMBB capacity. The proposed scheduler perpetually grants incoming URLLC traffic a higher priority for instant scheduling. In case that radio resources are not immediately schedulable, proposed scheduler forcibly enforces an artificial spatial user separation, for the URLLC traffic to get instantly scheduled over shared resources with ongoing eMBB transmissions. A pre-defined reference spatial subspace is constructed for which scheduler instantly picks the active eMBB user whose precoder is the closest possible. Then, it projects the eMBB precoder on-the-go onto the reference subspace, in order for its paired URLLC user to orient its decoder matrix into one possible null space of the reference subspace. Hence, a robust decoding ability is always preserved at the URLLC user, while cross-maximizing the ergodic capacity. Compared to the state-of-the-art proposals from industry and academia, proposed scheduler shows extreme URLLC latency robustness with significantly improved overall spectral efficiency. Analytical analysis and extensive system level simulations are presented to support paper conclusions.

Index Terms— URLLC; eMBB; Null space; MU-MIMO; 5G; Preemptive; Puncture scheduling.

1 Introduction

Emerging fifth generation (5G) systems are envisioned to support two major service classes: ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB) [1]. URLLC refer to the future services that demand extremely reliable and low latency data communication, i.e., one-way latency up to 1 ms with 10^{-5} outage probability [2]. That is, the quality of service (QoS) of the URLLC-type applications is infringed if more than one packet out of 10^5 packets are not successfully decoded within the 1 ms deadline. This URLLC QoS is immensely different from that of the current long term evolution (LTE) technology [3], where the overall spectral efficiency (SE) is the prime objective.

To satisfy such stringent latency requirements, the system should be always engineered so that blocking a URLLC packet is a very rare event. Therefore, URLLC services must satisfy their individual outage capacity, instead of the ergodic capacity. That is, by setting an ultra-tight target block error rate (BLER) to always ensure a sufficient URLLC decoding ability. This way, it leads to a significant loss of the network SE due to the fundamental tradeoff between reliability, latency and the achievable SE [4].

In the recent literature, diverse 5G scheduling contributions have been introduced. User-centric scheduling with variable transmission time intervals

(TTIs) [5] is essential to minimize the URLLC frame alignment and queuing delays. Furthermore, URLLC spatial diversity techniques are vital to preserve a sufficient URLLC signal-to-interference-noise-ratio (SINR). For example, the work in [6] demonstrates that a 4 × 4 multi-input multi-output (MIMO) microscopic diversity and two orders of macroscopic diversity are required to reach the URLLC outage SINR level. A recent study [7] further extends the usage of the spatial diversity for URLLC by flexibly allocating coded segments of the URLLC payload message to different interfaces. Thus, a better latency-reliability tradeoff can be achieved by reducing the original payload transmission time. Additionally, URLLC punctured scheduling (PS) [8] is a state-of-the-art scheme to further minimize the queuing delay of the URLLC traffic, where sporadic URLLC traffic is instantly scheduled by overwriting part of the radio resources, monopolized by ongoing eMBB transmissions.

However, the majority of the URLLC scheduling studies consider a monotonic optimization structure of the URLLC outage capacity. Therefore, URLLC requirements can be proportionally satisfied only with the size of the URLLC granted resources or received SINR levels. However, when joint eMBB and URLLC traffic coexists on the same radio spectrum, this approach fails to reach a proper system ergodic capacity.

In this work, a null-space-based preemptive scheduling (NSBPS) for joint eMBB and URLLC traffic is proposed. Proposed scheduler seeks to dynamically fulfill a jointly constrained objective, for which the URLLC QoS is guaranteed, while achieving the best possible eMBB capacity. If the available radio resources are not sufficient to accommodate the URLLC payload, NSBPS forcibly fits the URLLC traffic within an ongoing eMBB transmission in an instant, controlled, semi-transparent and biased multi-user MIMO (MU-MIMO) transmission. The proposed NSBPS instantly selects an active eMBB user whose transmission is most aligned within an arbitrary reference subspace. It spatially projects the selected eMBB transmission onto the reference subspace for which its paired URLLC user de-orients its decoding matrix into one possible null-space. Accordingly, a robust SINR level is preserved at the URLLC user side. Compared to the state-of-the-art studies, proposed NSBPS shows extreme robustness of the URLLC QoS with significantly improved ergodic capacity.

Due to the complexity of the 5G new radio (NR) system model [1-3] and addressed problems therein, the performance of the proposed scheduler is validated by extensive system simulations (SLS), and supported by analytical analysis of the major performance indicators. Those simulations are based on widely accepted models and calibrated against the 5G NR specifications to ensure highly reliable statistical results.

Notations: $(\mathcal{X})^{\mathrm{T}}$, $(\mathcal{X})^{\mathrm{H}}$ and $(\mathcal{X})^{-1}$ stand for the transpose, Hermitian, and inverse operations of $\mathcal{X}$, $\mathcal{X} \cdot \mathcal{Y}$ is the dot product of $\mathcal{X}$ and $\mathcal{Y}$, while

$\overline{\mathcal{X}}$ and $\|\mathcal{X}\|$ represent the mean and 2-norm of $\mathcal{X}$. $\mathcal{X} \sim \mathbb{CN}(0, \sigma^2)$ indicates a complex Gaussian random variable with zero mean and variance σ^2, $\mathcal{X}_\kappa, \kappa \in \{\text{llc}, \text{mbb}\}$ denotes the type of user $\mathcal{X}$, $\mathbb{E}\{\mathcal{X}\}$ and $\mathbf{card}(\mathcal{X})$ are the statistical expectation and cardinality of $\mathcal{X}$.

The paper is organized as follows. Section 2 presents the system and signal models, respectively. Section 3 states the problem formulation and detailed description of the NSBPS scheduler. Extensive system level simulation results are introduced in Section 4, and paper is concluded in Section 5.

2 System Model

We consider a 5G-NR downlink (DL) MU-MIMO system where there are C cells, each equipped with N_t transmit antennas, and K uniformly distributed user equipment's (UEs) per cell, each with M_r receive antennas. Users are multiplexed by the orthogonal frequency division multiple access (OFDMA). Two types of DL traffic are under assessment as: (a) URLLC bursty FTP3 traffic model with a finite B−byte payload and Poisson arrival process λ, and (b) eMBB full buffer traffic with infinite payload size. The total number of UEs per cell is: $K_{\text{mbb}} + K_{\text{llc}} = K$, where K_{mbb} and K_{llc} are the average numbers of eMBB and URLLC UEs per cell, respectively.

The agile 5G-NR frame structure is adopted [5], where the URLLC and eMBB UEs are scheduled with variable TTI periodicity. As depicted in Fig. B.1, eMBB traffic is scheduled with a long TTI of 14-OFDM symbols for SE maximization while URLLC traffic with a shorter TTI of 2-OFDM symbols due to its latency budget. In the frequency domain, the smallest scheduling unit is the physical resource block (PRB), which is 12 sub-carriers and with 15 kHz sub-carrier spacing.

A maximal subset of MU co-scheduled URLLC-eMBB user pairs $\mathtt{G}_c \in \mathcal{K}_c$ is allowed over an arbitrary PRB in the c^{th} cell, where $G_c = \mathbf{card}(\mathtt{G}_c)$, $G_c \leq N_t$ is the number of co-scheduled UEs and $\mathcal{K}_c$ is the set of all active UEs in the c^{th} cell. Since $N_t \leq KM_r$, user selection on top of equal power allocation is assumed for MU pairing. The received DL signal at the k^{th} user from the c^{th} cell can be modeled as

$$\mathrm{y}_{k,c}^{\kappa} = \mathbf{H}_{k,c}^{\kappa}\mathbf{v}_{k,c}^{\kappa}s_{k,c}^{\kappa} + \sum_{g\in\mathtt{G}_c, g\neq k} \mathbf{H}_{k,c}^{\kappa}\mathbf{v}_{g,c}s_{g,c}$$

$$+ \sum_{j=1, j\neq c}^{C} \sum_{g\in\mathtt{G}_j} \mathbf{H}_{g,j}\mathbf{v}_{g,j}s_{g,j} + \mathbf{n}_{k,c}^{\kappa}, \tag{B.1}$$

where $\mathbf{H}_{k,c}^{\kappa} \in \mathcal{C}^{M_r \times N_t}, \forall k \in \{1, \ldots, K\}, \forall c \in \{1, \ldots, C\}$ is the wireless channel observed at the k^{th} user from the c^{th} cell, $\mathbf{v}_{k,c}^{\kappa} \in \mathcal{C}^{N_t \times 1}$ is the zero-forcing precoding vector, assuming a single layer transmission per user, where it is

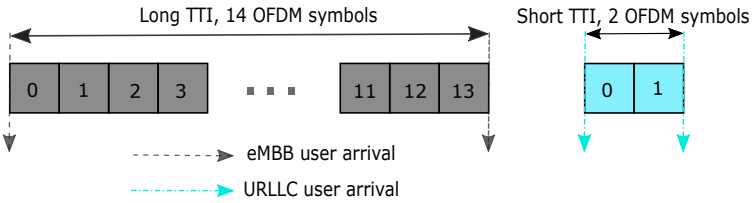


Fig. B.1. Flexible TTI scheduling in 5G NR.

given as: $\mathbf{v}_{k,c}^{\kappa} = \left(\mathbf{H}_{k,c}^{\kappa}\right)^{\mathrm{H}} \left(\mathbf{H}_{k,c}^{\kappa}\left(\mathbf{H}_{k,c}^{\kappa}\right)^{\mathrm{H}}\right)^{-1}$. $s_{k,c}^{\kappa}$ and $\mathbf{n}_{k,c}^{\kappa}$ denote the transmitted symbol and the additive white Gaussian noise at the k^{th} user, respectively. The first and second summation terms represent the intra-cell inter-user and inter-cell interference, generated from either the URLLC or eMBB traffic. In this work, the 3GPP 3D spatial channel model [9] is adopted, where the DL channel coefficient observed by the m^{th} receive antenna from the n^{th} transmit antenna is composed from Q spatial clusters, each with Z rays as

$$h_{(m,n)_k}^{\kappa} = \frac{1}{\sqrt{Q}} \sum_{q=0}^{Q-1} \sqrt{\delta_k}\, \mathcal{G}_{q,k}\, r_{(m,n,q)_k}, \tag{B.2}$$

where $\delta_k = \ell \epsilon_k^{\varrho} \mu_k$, ℓ and μ_k are the propagation and shadow fading coefficients, respectively, and ϵ_k^{ϱ} is the distance, with ϱ as the pathloss factor, and $\mathcal{G}_{q,k} \sim \mathbb{CN}(0,1)$. The steering factor $r_{(m,n,q)_k}$ is given by

$$r_{(m,n,q)_k} = \sqrt{\frac{\xi\psi}{Z}} \sum_{z=0}^{Z-1} \begin{pmatrix} \sqrt{\mathfrak{D}_{\mathrm{BS}}^{m,n,q,z}(\theta_{\mathrm{AoD}}, \varphi_{\mathrm{EoD}})}\, e^{j(\eta d \overline{f} + \Phi_{m,n,q,z})} \\ \times \sqrt{\mathfrak{D}_{\mathrm{UE}}^{m,n,q,z}(\theta_{\mathrm{AoA}}, \varphi_{\mathrm{EoA}})}\, e^{j(\eta d \sin(\theta_{\mathrm{m,n,q,z,AoA}}))} \\ \times e^{j\eta ||s|| \cos(\varphi_{\mathrm{m,n,q,z,EoA}}) \cos(\theta_{\mathrm{m,n,q,z,AoA}} - \theta_s) t} \end{pmatrix}, \tag{B.3}$$

where ξ and ψ are the power and large-scale coefficients, $\mathfrak{D}_{\mathrm{BS}}$ and $\mathfrak{D}_{\mathrm{UE}}$ are the antenna patterns at the BS and UE, respectively, η is the wave number, θ denotes the horizontal angle of arrival θ_{AoA} and departure θ_{AoD}, while φ denotes the elevation angle of arrival φ_{EoA} and departure φ_{EoD}, respectively. s is the speed of the k^{th} user, $\overline{f} = f_x \cos\theta_{\mathrm{AoD}} \cos\varphi_{\mathrm{EoD}}$ is the displacement vector of the uniform linear transmit array.

The received signal at the k^{th} user is decoded by applying the antenna combining as: $\left(\mathrm{y}_{k,c}^{\kappa}\right)^{*} = \left(\mathbf{u}_{k,c}^{\kappa}\right)^{\mathrm{H}} \mathrm{y}_{k,c}^{\kappa}$, where $\mathbf{u}_{k,c}^{\kappa}$ is designed by the linear minimum mean square error interference rejection combining (LMMSE-IRC) receiver [10]. The received SINR level at the k^{th} user is then calculated as

$$\Upsilon_{k,c}^{\kappa} = \frac{p_k^c \left\| \mathbf{H}_{k,c}^{\kappa} \mathbf{v}_{k,c}^{\kappa} \right\|^2}{1 + \sum\limits_{g \in \mathbb{G}_c, g \neq k} p_g^c \left\| \mathbf{H}_{k,c}^{\kappa} \mathbf{v}_{g,c}^{\kappa} \right\|^2 + \sum\limits_{j \in C, j \neq c} \sum\limits_{g \in \mathbb{G}_j} p_g^j \left\| \mathbf{H}_{g,j}^{\kappa} \mathbf{v}_{g,j}^{\kappa} \right\|^2}, \tag{B.4}$$

where p_k^c is the transmit power intended for the k^{th} user. Then, the k^{th} user received rate on a given PRB is given by

$$r_{k,rb}^{\kappa} = \log_2\left(1 + \frac{1}{G_{k,c}}\Upsilon_{k,c}^{\kappa}\right). \tag{B.5}$$

Accordingly, the user SINR levels across different $\mathcal{N}$ sub-carriers are mapped into a single effective SINR using the effective exponential SNR mapping [11] as

$$\left(\Upsilon_{k,c}^{\kappa}\right)^{\text{eff.}} = -\partial \ln\left(\frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}} e^{-\frac{\left(\Upsilon_{k,c}^{\kappa}\right)^{i}}{\partial}}\right), \tag{B.6}$$

with ∂ as a calibration parameter.

3 Proposed NSBPS Scheduler

3.1 Problem Formulation

Under a 5G-NR system, there are user-centric, instead of network-centric, QoS utility functions. These are highly coupled and need to be reliably fulfilled, e.g., eMBB rate maximization and URLLC latency minimization as

$$\forall k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}} : R_{\text{mbb}} = \underset{k_{\text{mbb}}\in\mathcal{K}_{\text{mbb}}}{\arg\max} \sum_{k_{\text{mbb}}=1}^{K_{\text{mbb}}} \sum_{rb\in\Xi_{k_{\text{mbb}}}^{\text{mbb}}} \beta_{k_{\text{mbb}}} r_{k_{\text{mbb}},rb}^{\text{mbb}}, \tag{B.7}$$

$$\forall k_{\text{llc}} \in \mathcal{K}_{\text{llc}} : \underset{k_{\text{llc}}\in\mathcal{K}_{\text{llc}}}{\arg\min}\left(\Psi_{k_{\text{llc}}}\right),\ \Psi_{k_{\text{llc}}} \leq 1\,\text{ms}, \tag{B.8}$$

where R_{mbb} is the overall eMBB ergodic capacity, $\mathcal{K}_{\text{mbb}}$ and $\mathcal{K}_{\text{llc}}$ represent the active sets of eMBB and URLLC users, respectively, $\Xi_{k_{\text{mbb}}}^{\text{mbb}}$ and $\beta_{k_{\text{mbb}}}$ denote the granted set of PRBs and a priority factor of the k^{th} eMBB user. $\Psi_{k_{\text{llc}}}$ is the URLLC target one-way latency, which is expressed as

$$\Psi_{k_{\text{llc}}} = \Lambda_{\text{q}} + \Lambda_{\text{bsp}} + \Lambda_{\text{fa}} + \Lambda_{\text{tx}} + \Lambda_{\text{uep}}, \tag{B.9}$$

where $\Lambda_{\text{q}}, \Lambda_{\text{bsp}}, \Lambda_{\text{fa}}, \Lambda_{\text{tx}}, \Lambda_{\text{uep}}$ are the queuing, BS processing, frame alignment, transmission, and UE processing delays, respectively. Λ_{fa} is upper-bounded by the short TTI interval while Λ_{bsp} and Λ_{uep} are bounded by 3-OFDM symbol duration [12], due to the enhanced processing capabilities with the 5G-NR. Hence, Λ_{tx} and Λ_{q} are the major impediment against achieving the hard URLLC latency budget. Λ_{tx} depends on the outage SINR level as given by

$$\Lambda_{\text{tx}} = \frac{\text{B}}{\left(\Xi_{k_{\text{llc}}}^{\text{llc}}\log_2\left(1 + \frac{\Upsilon_{k_{\text{llc}}}^{\text{llc}}}{I}\right)\right)}, \tag{B.10}$$

where $\digamma$ is the outage gap between the expected and actual received SINR levels. The URLLC queuing delay Λ_{q} can be modeled by the $\mathcal{A}/\mathcal{A}/\mathtt{a}/\phi$ queuing model [13], where the first $\mathcal{A}$ denotes a Poisson packet arrival, second $\mathcal{A}$ means exponential service times out of the queue, notation $\mathtt{a}$ represents the maximum number of the URLLC simultaneous transmissions, and notation ϕ implies that an arriving URLLC packet will be dropped if there are ϕ outstanding packets, worth of more than 1 ms in the queue. Thus, the probability of the URLLC reliability loss, i.e., $\Lambda_{\mathrm{q}} \geq 1$ ms, is given as

$$\rho_{rl} = \left(\rho_0 \frac{\mathtt{a}^{\mathtt{a}}}{\mathtt{a}!}\right) \rho^{\phi}, \tag{B.11}$$

where ρ_0 is the probability of the queue being empty, and $\rho = \left(\frac{\lambda}{\mathtt{a}\mathcal{O}}\right)$, with $\frac{1}{\mathcal{O}}$ as the mean service time. Thus, to achieve the critical URLLC latency, the transmission and queuing delays should be always minimized to provide further allowance for the re-transmission delay. This can be achieved by guaranteeing a sufficient outage SINR level or allocating excessive PRBs to URLLC traffic in order to further minimize ρ_{rl}. In both cases, the eMBB utility function in (B.7) will be ill-optimized, leading to a severe degradation of the network SE.

3.2 Description of The Proposed NSBPS Scheduler

The proposed NSBPS scheduler seeks to simultaneously cross-optimize the joint objectives of the eMBB and URLLC traffic. Thus, the critical URLLC latency deadline is satisfied regardless of the system loading while reaching the best achievable eMBB performance. In the following sub-sections, we describe the proposed NSBPS scheduler in-detail.

At the BS side:

At an arbitrary TTI instance, if there are no newly incoming URLLC packets, NSBPS allocates single-user (SU) dedicated resources to the new/buffered eMBB traffic based on the standard proportional fair (PF) metric as

$$\Theta\left\{\mathrm{PF}_{k_{\mathrm{mbb}}}\right\} = \frac{r^{\mathrm{mbb}}_{k_{\mathrm{mbb}},rb}}{\overline{r}^{\mathrm{mbb}}_{k_{\mathrm{mbb}},rb}}, \tag{B.12}$$

$$k^{*}_{\mathrm{mbb}} = \underset{k_{\mathrm{mbb}} \in \mathcal{K}_{\mathrm{mbb}}}{\arg\max} \ \Theta\left\{\mathrm{PF}_{k_{\mathrm{mbb}}}\right\}, \tag{B.13}$$

where $\overline{r}^{\mathrm{mbb}}_{k_{\mathrm{mbb}},rb}$ is the average delivered data rate of the k^{th} user. If sporadic DL URLLC packets arrive at the BS while sufficient radio resources are instantly available, the NSBPS scheduler immediately overpowers the eMBB traffic SU priority and assigns SU resources to incoming URLLC traffic based on the weighted PF (WPF) criteria instead as: $\Theta\left\{\mathrm{WPF}_{k_{\kappa}}\right\} = \frac{r^{\kappa}_{k,rb}}{\overline{r}^{\kappa}_{k,rb}}\beta_{k_{\kappa}}$, with $\beta_{k_{\mathrm{llc}}} \gg \beta_{k_{\mathrm{mbb}}}$ for immediate URLLC SU scheduling.

However, under a large offered load, which is envisioned with the 5G-NR, schedulable resources may not be instantly available for critical URLLC traffic and accordingly, significant queuing delays are foreseen. In such case, NSBPS scheduler first attempts to fit the URLLC packets within an active eMBB traffic in a normal and non-biased MU transmission, based on a conservative $\gamma-$orthogonality threshold, where $\gamma \rightarrow [0,1]$. Thus, the incoming URLLC traffic can only be paired with an active eMBB user if they satisfy:

$$1 - \left| \left(\mathbf{v}_{k_{\text{mbb}}}^{\text{mbb}} \right)^{\text{H}} \mathbf{v}_{k_{\text{llc}}}^{\text{llc}} \right|^2 \geq \gamma. \tag{B.14}$$

with $\forall k_{\text{mbb}} \in \{1, \ldots, K_{\text{mbb}}\}, \forall k_{\text{llc}} \in \{1, \ldots, K_{\text{llc}}\}$. The conservative orthogonality threshold is enforced to safeguard the URLLC traffic from potential inter-user interference. However, if the spatial degrees of freedom (SDoFs) are limited within a TTI, i.e., system is incapable to jointly process several signals between different transceivers on the same resources, and such orthogonality can not be instantly offered, NSBPS scheduler immediately alters the system optimization objective into a region that satisfies the URLLC outage requirements, while imposing minimal loss to the eMBB performance. Thus, the scheduler enforces an instant, biased and controlled MU transmission between URLLC-eMBB user pair. The URLLC outage is guaranteed by satisfying the following conditions,

$$\textbf{rank} \left\{ \left(\mathbf{u}_k^{\text{llc}} \right)^{\text{H}} \mathbf{H}_k^{\text{llc}} \mathbf{v}_k^{\text{llc}} \right\} \sim \text{full}, \tag{B.15}$$

$$\textbf{rank} \left\{ \left(\mathbf{u}_k^{\text{llc}} \right)^{\text{H}} \mathbf{H}_k^{\text{llc}} \left(\mathbf{v}_{k^\diamond}^{\text{mbb}} \right)' \right\} \sim 0, \tag{B.16}$$

where $\left(\mathbf{v}_{k^\diamond}^{\text{mbb}} \right)'$ denotes the updated precoder of the co-scheduled eMBB user with the incoming URLLC user. Thus, an arbitrary discrete Fourier transform spatial subspace $\mathbf{v}_{\text{ref}}(\theta)$, pointing towards angle θ, is constructed by

$$\mathbf{v}_{\text{ref}}(\theta) = \left(\frac{1}{\sqrt{N_t}} \right) \left[1, e^{-j2\pi\Delta\cos\theta}, \ldots, e^{-j2\pi\Delta(N_t-1)\cos\theta} \right]^{\text{T}}, \tag{B.17}$$

where Δ is the absolute antenna spacing. Next, the NSBPS searches for one active eMBB user whose precoder is closest possible to the reference subspace as

$$k_{\text{mbb}}^\diamond = \arg \min_{\mathcal{K}_{\text{mbb}}} \mathbf{d} \left(\mathbf{v}_k^{\text{mbb}}, \mathbf{v}_{\text{ref}} \right), \tag{B.18}$$

with the Euclidean distance between $\mathbf{v}_k^{\text{mbb}}$ and $\mathbf{v}_{\text{ref}}$ given by

$$\mathbf{d} \left(\mathbf{v}_k^{\text{mbb}}, \mathbf{v}_{\text{ref}} \right) = \frac{1}{\sqrt{2}} \left\| \mathbf{v}_k^{\text{mbb}} \left(\mathbf{v}_k^{\text{mbb}} \right)^{\text{H}} - \mathbf{v}_{\text{ref}} \mathbf{v}_{\text{ref}}^{\text{H}} \right\|. \tag{B.19}$$

Then, scheduler instantly projects the precoder vector of the selected eMBB user $\mathbf{v}_{k^\diamond}^{\text{mbb}}$ onto $\mathbf{v}_{\text{ref}}$ as given by

$$\left(\mathbf{v}_{k^\diamond}^{\mathrm{mbb}}\right)' = \frac{\mathbf{v}_{k^\diamond}^{\mathrm{mbb}} \cdot \mathbf{v}_{\mathrm{ref}}}{\left\|\mathbf{v}_{\mathrm{ref}}\right\|^2} \times \mathbf{v}_{\mathrm{ref}}, \tag{B.20}$$

where $\left(\mathbf{v}_{k^\diamond}^{\mathrm{mbb}}\right)'$ is the updated eMBB user precoder. The NSBPS scheduler then instantly schedules the incoming URLLC traffic over shared resources with the impacted eMBB user. Since the instant precoder projection is transparent to the victim eMBB user, it exhibits a SE projection loss. However, eMBB loss is constrained minimum, especially under high eMBB user load, e.g., NSBPS scheduler has a higher probability to find an eMBB user whose precoder is originally aligned within $\mathbf{v}_{\mathrm{ref}}$, such that the instant projection process would not greatly impact its achievable capacity. Finally, the BS acknowledges the URLLC user by a single-bit Boolean co-scheduling indication $\alpha = 1$, to be instantly transmitted in the user-centric control channel.

At the URLLC user side:

Upon reception of $\alpha = 1$, the URLLC user realizes that its granted resources, from the scheduling grant, are shared with an active eMBB user whose transmission is aligned within the reference subspace $\mathbf{v}_{\mathrm{ref}}$. Thus, the first-stage decoder matrix of the URLLC user is constructed by a standard LMMSE-IRC receiver to reject the inter-cell interference as

$$\left(\mathbf{u}_k^{\mathrm{llc}}\right)^{(1)} = \left(\mathbf{H}_k^{\mathrm{llc}}\mathbf{v}_k^{\mathrm{llc}}\left(\mathbf{H}_k^{\mathrm{llc}}\mathbf{v}_k^{\mathrm{llc}}\right)^{\mathrm{H}} + \mathbf{W}\right)^{-1}\mathbf{H}_k^{\mathrm{llc}}\mathbf{v}_k^{\mathrm{llc}}, \tag{B.21}$$

where the interference covariance matrix is given by

$$\mathbf{W} = \mathbb{E}\left\{\mathbf{H}_k^{\mathrm{llc}}\mathbf{v}_k^{\mathrm{llc}}\left(\mathbf{H}_k^{\mathrm{llc}}\mathbf{v}_k^{\mathrm{llc}}\right)^{\mathrm{H}}\right\} + \sigma^2\mathrm{I}_{M_r}, \tag{B.22}$$

where I_{M_r} is $M_r \times M_r$ identity matrix. The IRC vector $\left(\mathbf{u}_k^{\mathrm{llc}}\right)^{(1)}$ is then de-oriented to be aligned within one possible null space of the effective inter-user interference subspace $\mathbf{H}_k^{\mathrm{llc}}\mathbf{v}_{\mathrm{ref}}$, as expressed by

$$\left(\mathbf{u}_k^{\mathrm{llc}}\right)^{(2)} = \left(\mathbf{u}_k^{\mathrm{llc}}\right)^{(1)} - \frac{\left(\left(\mathbf{u}_k^{\mathrm{llc}}\right)^{(1)} \cdot \mathbf{H}_k^{\mathrm{llc}}\mathbf{v}_{\mathrm{ref}}\right)}{\left\|\mathbf{H}_k^{\mathrm{llc}}\mathbf{v}_{\mathrm{ref}}\right\|^2} \times \mathbf{H}_k^{\mathrm{llc}}\mathbf{v}_{\mathrm{ref}}. \tag{B.23}$$

This way, the final URLLC decoder vector $\left(\mathbf{u}_k^{\mathrm{llc}}\right)^{(2)}$ exhibits no inter-user interference, providing the URLLC user with a robust decoding ability.

3.3 Analytic Analysis Compared to State of The Art

We compare the performance of the proposed NSBPS scheduler against the state-of-the-art schedulers as follows:

1. Punctured scheduler (PS) [8]: the URLLC traffic is always assigned a higher scheduling priority. If radio resources are not available, PS scheduler

instantly overwrites part of the ongoing eMBB transmissions, i.e., immediately stop an ongoing eMBB transmission, for instant URLLC scheduling. PS scheduler shows significant improvement of the URLLC latency performance at the expense of highly degraded SE.

2. Multi-user punctured scheduler (MUPS) [14]: in our past work, we considered a MU scheduler on top of the PS scheduler. MUPS first attempts to achieve a successful MU-MIMO transmission between a URLLC-eMBB user pair; however, it is a transparent, non-biased and non-controlled MU-MIMO. If the SDoFs are limited, MUPS scheduler rolls back to PS scheduler. MUPS has shown an improved performance tradeoff between system SE and URLLC latency; however, with a limited and non-robust gain, due to the non-controlled MU-MIMO and the SE-less efficient PS events.

Accordingly, the aggregate eMBB user rate can be linearly calculated from the individual sub-carrier rates for simplicity, assuming OFDMA flat fading channels, as

$$r^{\text{mbb}}_{k_{\text{mbb}}} = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb}. \tag{B.24}$$

Then, the portion of the radio resources $\Gamma^{\text{llc}}_{k_{\text{mbb}}}$ allocated to the k^{th} eMBB user, and being altered by the sporadic URLLC traffic, can be expressed by a set of random variables, as

$$\mathbf{\Gamma} = \left(\Gamma^{\text{llc}}_{k_{\text{mbb}}} \mid k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}} \right). \tag{B.25}$$

Since URLLC packets are of small payload size, it is reasonably to assume that $\Gamma^{\text{llc}}_{k_{\text{mbb}}} \leq \Xi^{\text{mbb}}_{k_{\text{mbb}}}$ is almost surely satisfied. Hence, the actual eMBB rate is formulated by the joint URLLC-eMBB rate allocation function, given by

$$R_{k_{\text{mbb}}} = \mathcal{F}\left(\Xi^{\text{mbb}}_{k_{\text{mbb}}}, \Gamma^{\text{llc}}_{k_{\text{mbb}}} \right). \tag{B.26}$$

For an instance, if an eMBB user is allocated SU dedicated resources, then $\mathcal{F}\left(\Xi^{\text{mbb}}_{k_{\text{mbb}}}, \Gamma^{\text{llc}}_{k_{\text{mbb}}} \right) = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb}$. with no capacity loss. However, due to the prioritized URLLC traffic, the actual eMBB user rate suffers a loss over a portion of the allocated resources, expressed by the rate loss function Π as

$$\mathcal{F}\left(\Xi^{\text{mbb}}_{k_{\text{mbb}}}, \Gamma^{\text{llc}}_{k_{\text{mbb}}} \right) = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb} \left(1 - \Pi\right), \tag{B.27}$$

where the rate loss function $\Pi : [0,1] \rightarrow [0,1]$ indicates the effective portion of impacted PRBs of the k^{th} eMBB user. Under the proposed NSBPS scheduler, the gain of the updated eMBB effective channel is given as

$$\mathcal{Q}^{\text{mbb}}_{k} = \frac{1}{\left[\left(\mathbf{H}^{\text{mbb}}_{k} \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)' \right) \times \left(\mathbf{H}^{\text{mbb}}_{k} \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)' \right)^{\text{H}} \right]^{-1}}, \tag{B.28}$$

where $\mathcal{Q}_k^{\text{mbb}}$ is the achievable post-projection channel gain of the k^{th} eMBB user, and its magnitude can be rewritten in terms of the precoder projection loss, i.e., the *on-the-fly* eMBB precoder update from $\mathbf{v}_{k^\diamond}^{\text{mbb}}$ to $\left(\mathbf{v}_{k^\diamond}^{\text{mbb}}\right)'$, as

$$\mathcal{Q}_k^{\text{mbb}} = \left\| \mathbf{H}_k^{\text{mbb}} \mathbf{v}_{k^\diamond}^{\text{mbb}} \right\|^2 \times \sin^2\left(\theta_{\left[\mathbf{v}_{k^\diamond}^{\text{mbb}}, \left(\mathbf{v}_{k^\diamond}^{\text{mbb}}\right)'\right]}\right), \tag{B.29}$$

where $\sin^2\left(\theta_{\left[\mathbf{v}_{k^\diamond}^{\text{mbb}}, \left(\mathbf{v}_{k^\diamond}^{\text{mbb}}\right)'\right]}\right)$ introduces the eMBB projection loss, over the shared resources with the URLLC traffic, with $\theta_{\left[\mathbf{v}_{k^\diamond}^{\text{mbb}}, \left(\mathbf{v}_{k^\diamond}^{\text{mbb}}\right)'\right]}$ as the spatial angle deviation between its original and projected precoders. Thus, $\overset{\text{NSBPS}}{\Pi}$ can be expressed as

$$\overset{\text{NSBPS}}{\Pi} = \left(\frac{\Gamma_{k_{\text{mbb}}}^{\text{llc}}}{\Xi_{k_{\text{mbb}}}^{\text{mbb}}}\right) \times \sin^2\left(\theta_{\left[\mathbf{v}_{k^\diamond}^{\text{mbb}}, \left(\mathbf{v}_{k^\diamond}^{\text{mbb}}\right)'\right]}\right). \tag{B.30}$$

Due to the constraints in (14) and (18), the projection loss is always guaranteed to be minimized, i.e., $\sin^2\left(\theta_{\left[\mathbf{v}_{k^\diamond}^{\text{mbb}}, \left(\mathbf{v}_{k^\diamond}^{\text{mbb}}\right)'\right]}\right) \ll 1$. For the PS scheduler, the rate loss function is expressed in terms of the entire URLLC resources inducing the resource allocation of the eMBB user, since the eMBB transmission is instantly stopped over these resources, as

$$\overset{\text{PS}}{\Pi} = \left(\frac{\Gamma_{k_{\text{mbb}}}^{\text{llc}}}{\Xi_{k_{\text{mbb}}}^{\text{mbb}}}\right). \tag{B.31}$$

Finally, the MUPS scheduler exhibits an average eMBB capacity loss due to the persistent PS events, if the normal MU-MIMO scheduler fails; thus, the rate loss can be given as

$$\overset{\text{MUPS}}{\Pi} = \Phi\left(\frac{\Gamma_{k_{\text{mbb}}}^{\text{llc}}}{\Xi_{k_{\text{mbb}}}^{\text{mbb}}}\right), \tag{B.32}$$

where $\Phi \leq 1$ is a fraction to indicate the probability density of rolling back to PS scheduler, under a specific cell loading. Hence, the average eMBB user rate can be calculated as

$$\overline{R}_{k_{\text{mbb}}} = \Xi_{k_{\text{mbb}}}^{\text{mbb}} r_{k_{\text{mbb}},rb}^{\text{mbb}} \left(1 - \mathbb{E}\left\{\Pi\right\}\right). \tag{B.33}$$

Based on (B.26) - (B.33), it can be concluded that the proposed NSBPS scheduler provides the best achievable eMBB and URLLC joint performance against state-of-the-art schedulers.

Table B.1: Simulation Parameters.

Parameter	Value
Environment	3GPP-UMA,7 gNBs, 21 cells, 500 meters inter-site distance
Channel bandwidth	10 MHz, FDD
Antenna setup	BS: 8 Tx, UE: 2 Rx
User dropping	uniformly distributed URLLC: 5, 10 and 20 users/cell eMBB: 5 , 10 and 20 users/cell
User receiver	LMMSE-IRC
TTI configuration	URLLC: 0.143 ms (2 OFDM symbols) eMBB: 1 ms (14 OFDM symbols)
CQI	periodicity: 5 ms, with 2 ms latency
HARQ	asynchronous HARQ, Chase combining HARQ round trip time = 4 TTIs
Link adaptation	dynamic modulation and coding target URLLC BLER : 1% target eMBB BLER : 10%
Traffic model	URLLC: bursty, B=50 bytes, $\lambda = 250$ eMBB: full buffer

4 Simulation Results

In this section, we present the extensive SLS results of the NSBPS scheduler, following the 5G-NR specifications, where the main simulation parameters are listed in Table B.I.

Fig. B.2 shows the URLLC average one-way latency Ψ at the 10^{-5} outage probability, under proposed NSBPS, PS, MUPS, and WPF schedulers. On the top left, a close snap of the complementary cumulative distribution function (CCDF) of the URLLC latency distribution is further presented. We define the cell load setup by: $\Omega = (K_{\text{mbb}}, K_{\text{llc}})$. The proposed NSBPS scheduler clearly provides a significantly robust and steady URLLC latency against different cell load conditions, and hence, independently from the aggregate levels of interference. The overall performance gain of the NSBPS scheduler is due to: 1) the guaranteed instantaneous URLLC scheduling without queuing in a controlled (almost surely occurs), biased (for the sake of the URLLC user), and semi-transparent (URLLC user is aware of it) MU transmission, leading to no inter-user interference at the URLLC user, 2) the constrained-minimum eMBB user rate loss function, and 3) the enforced regularization of the inter-cell interference spatial distribution within a limited span, due to the fixed subspace projection, and hence, the linear MMSE-IRC receiver nulls the av-

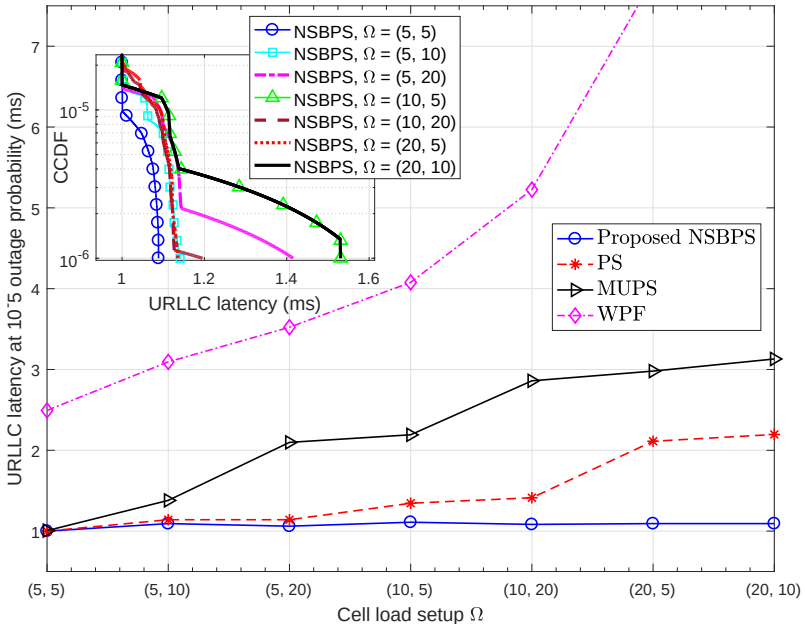


Fig. B.2. URLLC one-way latency Ψ at 10^{-5}outage.

erage inter-cell interference more efficiently and with improved SDoFs.

The PS scheduler shows an optimized URLLC latency in the low load region, at the expense of degraded eMBB performance. However, in the high load region when the inter-cell interference levels are extreme, PS scheduler provides a degraded URLLC latency performance due to the experienced retransmissions and degraded capacity per PRB. The MUPS scheduler shows a fair tradeoff between URLLC latency and the eMBB SE, where the non-controlled URLLC-eMBB MU-MIMO transmissions reduce the URLLC decoding ability. Finally, the WPF scheduler exhibits the worst URLLC latency performance, where the URLLC packets are queued for multiple TTIs if the radio resources are not instantly schedulable.

As shown in Fig. B.3, the empirical CDF (ECDF) of the average cell throughput in Mbps is presented. The NSBPS scheduler provides the best achievable cell throughput compared to other schedulers, due to the always constrained-minimum rate loss function of the victim eMBB users. The PS scheduler exhibits severe loss of the network SE due to the punctured eMBB transmissions. However, the WPF scheduler achieves an improved capacity since no puncture-events are allowed; however, at the expense of the worst URLLC latency. Finally, the MUPS scheduler shows further improved capacity, due to the successful MU events; however, with limited MU gain since

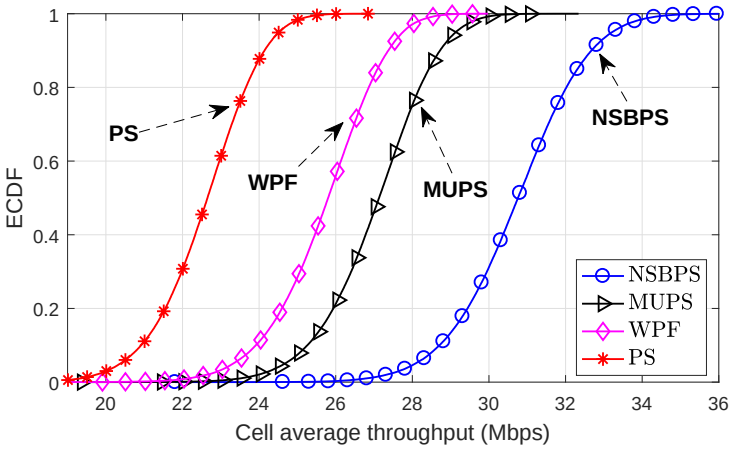


Fig. B.3. Cell average throughput for $\Omega = (5, 5)$.

when a successful MU pairing is not possible, MUPS falls back to SE-less-efficient PS scheduler.

Examining the eMBB performance, Fig. B.4. shows the average eMBB user throughput in Mbps, for all schedulers under evaluation, where similar conclusions can be clearly obtained. For instance, with $\Omega = (5, 5)$, where the system SDoFs are limited by the small number of active eMBB users, i.e., $K_{\text{mbb}} = 5$, the proposed NSBPS shows a gain $\sim$ 28.9% in the eMBB user throughput than the MUPS scheduler. Under such SDoF-limited state, the MUPS scheduler is highly likely to roll back to PS scheduler, i.e., $\Phi \sim 1$, while the NSBPS forcibly enforces these missing SDoFs, sufficient enough to instantly fit the URLLC traffic within an eMBB transmission.

Finally, Table B.II presents the achievable MU throughput gain of the NSBPS and MUPS schedulers. The best achievable MU gain of the NSBPS over the MUPS scheduler is obtained when the system is originally SDoF-limited, i.e., $\Omega = (5, 5)$. With SDoF-rich loading states such as $\Omega = (20, 5)$, the MUPS scheduler rarely falls back to PS scheduler, i.e., $\Phi \sim 0$, and hence, an improved MU gain is achieved.

5 Conclusion

A null space based preemptive scheduler (NSBPS) has been proposed for joint 5G URLLC and eMBB traffic. The proposed NSBPS scheduler aims to fulfill a constraint-coupled objective, for which the URLLC quality of service is almost surely guaranteed while achieving the maximum possible ergodic capacity. Extensive system level simulations and analytic gain analysis have been conducted for performance evaluation. Compared to the state-of-the-art

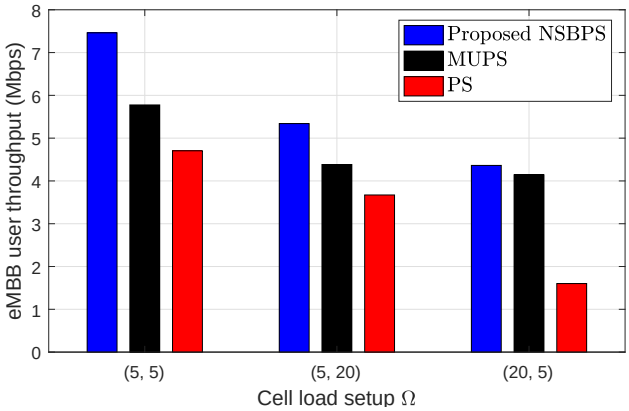


Fig. B.4. eMBB average user throughput.

Table B.2: Average MU gain of the NSBPS and MUPS schedulers.

Scheduler	$\Omega = (5,5)$	$\Omega = (5,20)$	$\Omega = (20,5)$
MUPS (Mbps)	7.69	12.13	23.05
NSBPS (Mbps)	22.92	24.91	27.78
Gain (%)	+198.04	+105.35	+20.52

scheduler proposals from academia and industry, the proposed NSBPS shows extreme robustness of the URLLC latency performance, i.e., regardless of the cell loading, and aggregate interference levels, while providing significantly improved eMBB performance. A comprehensive study on the performance of the proposed scheduler will be considered in a future work.

6 Acknowledgments

This work is partly funded by the Innovation Fund Denmark (IFD) -- case number: 7038-00009B. Also, part of this work has been performed in the framework of the Horizon 2020 project ONE5G (ICT-760809) receiving funds from the European Union.

References

[1] NR and NG-RAN overall description; Stage-2 (Release 15), 3GPP, TS 38.300, V2.0.0, Dec. 2017.

[2] Service requirements for the 5G system; Stage-1 (Release 16), 3GPP, TS 22.261, V16.2.0, Dec. 2017.

[3] Study on new radio access technology; Radio access architecture and interfaces (Release 14), 3GPP, TR 38.801, V14.0.0, March 2017.

[4] B. Soret, P. Mogensen, K. I. Pedersen and M. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," *in Proc. IEEE Globecom*, Austin, TX, 2014, pp. 1391-1396.

[5] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen and A. Szufarska, "A flexible 5G frame structure design for FDD cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53-59, March 2016.

[6] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen and P. Mogensen, "Signal quality outage analysis for URLLC in cellular networks," *in Proc. IEEE Globecom*, San Diego, CA, 2015, pp. 1-6.

[7] J. J. Nielsen, R. Liu and P. Popovski, "Ultra-reliable low latency communication using interface diversity," *IEEE Trans. Commun,* vol. 66, no. 3, pp. 1322-1334, March 2018.

[8] K.I. Pedersen, G. Pocovi, J. Steiner, and S. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," *in Proc. IEEE VTC*, Toronto, 2017, pp. 1-6.

[9] Study on 3D channel model for LTE; Release 12, 3GPP, TR 36.873, V12.7.0, Dec. 2014

[10] Y. Ohwatari, N. Miki, Y. Sagae and Y. Okumura, "Investigation on interference rejection combining receiver for space–frequency block code transmit diversity in LTE-advanced downlink," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 191-203, Jan. 2014

[11] S. N. Donthi and N. B. Mehta, "An accurate model for EESM and its application to analysis of CQI feedback schemes," *IEEE Trans. Wireless Commun.*, vol. 10, no. 10, pp. 3436-3448, Oct. 2011.

[12] Physical layer procedures; Evolved universal terrestrial radio access (Release 15), 3GPP, TS 36.213, V15.1.0, March. 2018.

[13] Bertsekas, D. and Gallager, R. (1992). *Data Networks*. 2nd ed. Michigan: Prentice Hall.

[14] Ali A. Esswie, and K.I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5G networks," *in Proc. IEEE ISCC*, Natal, 2018, pp. 1-6.

Paper C

Opportunistic Spatial Preemptive Scheduling For URLLC and eMBB Coexistence in Multi-User 5G Networks

Ali A. Esswie and Klaus I. Pedersen

The paper has been published in the
IEEE Access

Abstract

The fifth generation (5G) of the mobile networks is envisioned to feature two major service classes: ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB). URLLC applications require a stringent one-way radio latency of 1 ms with 99.999% *success probability while eMBB services demand extreme data rates. The coexistence of the URLLC and eMBB quality of service (QoS) on the same radio spectrum leads to a challenging scheduling optimization problem, that is vastly different from that of the current cellular technology. This calls for novel scheduling solutions which cross-optimize the system performance on a user-centric, instead of network-centric basis. In this paper, a null-space-based spatial preemptive scheduler for joint URLLC and eMBB traffic is proposed for densely populated 5G networks. Proposed scheduler framework seeks for cross-objective optimization, where critical URLLC QoS is guaranteed while extracting the maximum possible eMBB ergodic capacity. It utilizes the system spatial degrees of freedom in order to instantly offer an interference-free subspace for critical URLLC traffic. Thus, a sufficient URLLC decoding ability is always preserved, and with the minimal impact on the eMBB performance. Analytical analysis and extensive system level simulations are conducted to evaluate the performance of the proposed scheduler against the state-of-the-art scheduler proposals from industry and academia. Simulation results show that proposed scheduler offers extremely robust URLLC latency performance with a significantly improved ergodic capacity.*

Index Terms—5G; Radio resource management; Scheduling; Ultra-reliable low-latency communications (URLLC); Enhanced mobile broadband (eMBB); MU-MIMO; Preemptive; Null space

1 Introduction

The 3rd generation partnership project (3GPP) is progressing the standardization of the fifth generation (5G) standards with a big momentum [1 - 4]. The first 5G specifications support two major service classes: ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB) [5], respectively. The URLLC denote the future applications which demand extremely reliable and low latency radio transmissions, i.e., one-way radio latency of 1 ms, associated with $1 - 10^{-5}$ success probability [6, 7]. That is, a URLLC packet is of no-use if it can not be successfully decoded within the 1 ms latency deadline. Accordingly, supporting such stringent URLLC latency specifications enables many novel use cases [8], including smart grids, tactile internet, wireless industrial control, and real time vehicle-to-vehicle communications.

However, due to the limited available spectrum in the centimeter-wave region, both eMBB and URLLC applications shall coexist on the same car-

rier. Thus, achieving such extreme spectral efficiency (SE) for eMBB applications and the ultra reliability and low latency for URLLC services becomes a challenging scheduling task, due to the fundamental trade-off between latency, reliability and SE [9]. For instance, to satisfy such unprecedented URLLC requirements, the system should be forcibly engineered such that blocking URLLC packets is a rare event. This can be achieved by setting an extremely tight block error rate (BLER) to preserve a sufficient URLLC signal-to-interference-noise-ratio (SINR) [10]. Consequently, URLLC users must fulfill their outage capacity of interest [11] at the expense of the overall ergodic capacity, leading to a severe loss of the network SE.

1.1 State of The Art URLLC Scheduling Studies

Recently, the multiplexing of coexistent URLLC and eMBB traffic on the same radio spectrum is gaining progressive research attention in both industry and academia. The agile 5G frame structure design is shown to be of great significance to satisfy the URLLC latency [12 - 15], where users can be scheduled on transmission time intervals (TTIs) of different durations. For instance, eMBB traffic is scheduled with a long TTI duration to meet its extreme SE requirements while URLLC traffic can be scheduled on a shorter TTI duration for its tight latency deadline. Nevertheless, the latter case induces an increased control signaling overhead, which in turn degrades the control channel (CCH) capacity.

Moreover, spatial diversity techniques are considered as enablers for the URLLC by preserving a sufficient received SINR point. The study in [16] demonstrates that a 4×4 multi-input multi-output (MIMO) microscopic diversity along with two orders of macroscopic diversity are essential to reach the outage SINR point, required to achieve the URLLC latency limit at the 10^{-5} outage in 3GPP macro networks. These conclusions are also supported by URLLC realistic measurement campaigns [17]. Hence, the URLLC latency budget can be achieved by enhancing the decoding ability.

The recent work in [18] further broadens the adoption of the spatial diversity for URLLC communications. It flexibly assigns different coded segments of the URLLC payload to several active interfaces, i.e., transmitters, based on the associated latency, reliability, and bit rate properties. This is a substitute of transmitting duplicate versions of the URLLC packets from different transmitters at the same time. Thus, a better latency-reliability trade-off can be achieved by reducing the original payload transmission time. Additionally, the work in [19] considers a semi-shared resource allocation algorithm for the URLLC-type communications. It avoids preserving an exclusive set of the radio resources for the URLLC traffic due to its sporadic nature; however, it splits the URLLC resource allocation into two chunks as: 1) shared resources with other eMBB traffic, and 2) dedicated single-user (SU) resources.

The overall SE is enhanced; yet, with employing non-linear transceivers to compensate for the inter-user interference across the shared resources.

Furthermore, system-level packet duplication (PD) with the dual connectivity architecture in the 5G new radio (NR) [20], where users are simultaneously connected to a primary and secondary cell, is envisioned to offer great reliability levels to address such URLLC outage requirements. However, in order not to excessively consume the radio resources by redundant packets, the benefit of the URLLC PD is relevant to specific scenarios, where channels are highly unfavorable.

Additionally, the study in [21] reports advanced scheduling enhancements for optimized URLLC latency performance, including dynamic and load-dependent BLER optimization, refined hybrid automatic repeat request (HARQ) and link adaptation filtering in partly loaded cells. On another side, punctured scheduling (PS) [22] is a state-of-the-art study which aims at eliminating the scheduling queuing delay component of the stochastic URLLC traffic. If URLLC queuing is foreseen, due to resource shortage, PS scheduler instantly overwrites part of the ongoing eMBB transmissions for immediate URLLC scheduling, at the expense of a highly degraded eMBB SE. Subsequently, enhanced PS (E-PS) scheduler [23] is recently introduced to provide an improved ergodic capacity by informing the victim eMBB users of which physical resource blocks (PRBs) have been punctured by URLLC transmissions, in order to avoid erroneous Chase combing HARQ process, i.e., punctured resources are considered information-less. Code-block (CB) based HARQ re-transmission [24, 25] schemes are also proposed to reduce the overhead size of the punctured eMBB re-transmissions; however, a multi-bit HARQ ACK/NACK is required.

Finally, a multi-user-punctured scheduler (MU-PS) [26] is recently demonstrated to offer an attractive tradeoff between system ergodic capacity and URLLC (outage) performance. MU-PS first attempts to fit the sporadically incoming URLLC traffic within an ongoing eMBB traffic in a standard MU-MIMO transmission. If the MU pairing can not be satisfied at an arbitrary TTI, MU-PS scheduler falls back to PS scheduler for instant URLLC scheduling without queuing. Despite the achievable enhanced SE, MU-PS has shown a non-robust URLLC latency performance since the standard MU pairing constraint is only dependent on the rate maximization. Thus, it may lead to a further degraded SINR level of the URLLC traffic, due to the power sharing and the resulting inter-user interference.

Compared to the state-of-the-art schedulers, the URLLC outage capacity is monotonically satisfied, only with the associated dedicated resource allocation size or the provided decoding SINR level. When eMBB and URLLC traffic coexists on same spectrum, such approach results in severe degradation of the overall SE. Needless to say, a flexible scheduling framework for cross-objective optimization is still critical in scenarios where an efficient

multiplexing of the eMBB and URLLC traffic is mandated.

1.2 Paper Contribution

In this work, we propose a null-space-based preemptive scheduler (NSBPS) for densely populated 5G networks. The proposed NSBPS aims to dynamically cross optimize a jointly constrained system utility, where the URLLC quality of service (QoS) is always guaranteed while achieving the maximum possible ergodic capacity. If the instantaneous schedulable radio resources are not sufficient to contain the incoming URLLC traffic, NSBPS scheduler forcibly fits the URLLC traffic within an ongoing eMBB transmission in a controlled, biased, and semi-transparent MU-MIMO transmission. Proposed scheduler pre-defines a reference spatial subspace, pointing to an arbitrary direction. Then, it instantly searches for an active eMBB transmission which is most aligned within the reference subspace. Next, NSBPS scheduler spatially projects the selected eMBB transmission onto the reference subspace, in order for its paired URLLC user to orient its decoding vector within one possible null-space, thus, no residual inter-user interference is experienced at the URLLC user. Compared to the state-of-the-art scheduling studies from industry and academia, proposed NSBPS shows extreme robustness of the URLLC QoS with significantly enhanced ergodic capacity. The major framework of this work is summarized as follows:

- We extend our recent studies [11, 26] to propose a comprehensive performance analysis of the NSBPS scheduler under diversity of traffic and network settings.
- Compared to the state-of-the-art scheduler proposals from latest 3GPP standards, the derived NSBPS scheduler shows extreme URLLC latency robustness while approaching the network ergodic capacity.
- Proposed NSBPS scheduler is compliant with the 5G-NR standardization and requires neither excessive control overhead nor higher processing complexity.

Due to the complexity of the 5G-NR and addressed problems therein [1 - 3], the performance of the proposed NSBPS scheduler is evaluated by highly-detailed system level simulations (SLSs), and supported by analytical analysis of the key performance indicators. Following the same methodology as in [11, 26], these simulations are based on widely accepted mathematical models and calibrated against the 3GPP 5G-NR assumptions of the majority of the resource management functionalities, e.g., HARQ, link-to-system mapping, and adaptive link adaptation. Furthermore, simulation results are ensured to be statistically reliable by preserving an extremely sufficient simulation confidence interval.

Notations: $(\mathcal{X})^{\mathrm{T}}$, $(\mathcal{X})^{\mathrm{H}}$ and $(\mathcal{X})^{-1}$ stand for the transpose, Hermitian, and inverse operations of $\mathcal{X}$, $\mathcal{X} \cdot \mathcal{Y}$ is the dot product of $\mathcal{X}$ and $\mathcal{Y}$, while $\overline{\mathcal{X}}$ and $\|\mathcal{X}\|$ represent the mean and 2-norm of $\mathcal{X}$. $\mathcal{X} \sim \mathbb{CN}(0, \sigma^2)$ presents a complex Gaussian random variable with zero mean and variance σ^2, $\mathcal{X}^{\kappa}, \kappa \in \{\mathrm{llc}, \mathrm{mbb}\}$ denotes the type of user $\mathcal{X}$, $\mathbb{E}\{\mathcal{X}\}$ and $\mathbf{card}(\mathcal{X})$ are the statistical expectation and cardinality of $\mathcal{X}$.

The rest of this paper is organized as follows. Section 2 introduces the system and signal models. Section 3 presents the addressed problem formulation. Section 4 discusses the proposed NSBPS scheduler in detail. Section 5 describes an analytical gain analysis compared to the state-of-the-art studies, and extensive system level performance evaluation is drawn in Section 6. Section 7 concludes the paper.

2 Setting the Scene

2.1 System Model

We consider a downlink (DL) 5G-NR network where the URLLC and eMBB service classes coexist [11, 26]. There are C cells, each equipped with N_t transmit antennas, and K uniformly-distributed user equipment's (UEs) per cell, each equipped with M_r receive antennas. Users are dynamically multiplexed by the orthogonal frequency division multiple access (OFDMA) [27]. We assess three types of DL traffic as: (1) URLLC sporadic FTP3 traffic with finite $\mathrm{B}_{\mathrm{llc}}-$byte payload size and Poisson arrival process λ, (2) eMBB full buffer traffic model with infinite payload size, and (3) eMBB constant bit rate (CBR) traffic model [28], i.e., broadband video streaming, with a predetermined number of packets $\check{n}$, each is $\mathrm{B}_{\mathrm{mbb}}-$byte, and packet inter-arrival rate $\check{\imath}$.

The average number of UEs per cell is expressed as: $K_{\mathrm{mbb}} + K_{\mathrm{llc}} = K$, where K_{mbb} and K_{llc} are the average numbers of eMBB and URLLC UEs per cell, respectively. Hence, the offered URLLC load per cell is given by: $K_{\mathrm{llc}} \times \mathrm{B}_{\mathrm{llc}} \times \lambda$, while the eMBB full buffer load is infinite and the CBR load per cell is: $K_{\mathrm{mbb}} \times \left(\frac{\mathrm{B}_{\mathrm{mbb}}}{(\check{n}-1)\check{\imath}}\right)$, respectively. The flexible frame structure of the 5G-NR is adopted in this work [12], where the URLLC and eMBB UEs are scheduled with variable TTI duration. As depicted in Fig. C.1, eMBB traffic is scheduled per a long TTI of 14-OFDM symbols for maximizing its perceived SE while the URLLC traffic is scheduled per a shorter TTI of 2-OFDM symbols, i.e., mini-slot, due to its latency requirements. In the frequency domain, the minimum schedulable unit is the PRB, each is 12 sub-carriers of 15 kHz spacing. In line with [12, 13], the scheduling grant is transmitted within the resources assigned to each user, i.e., in-resource CCH. Thus, the minimum resource allocation per user should be sufficiently large to accommodate the

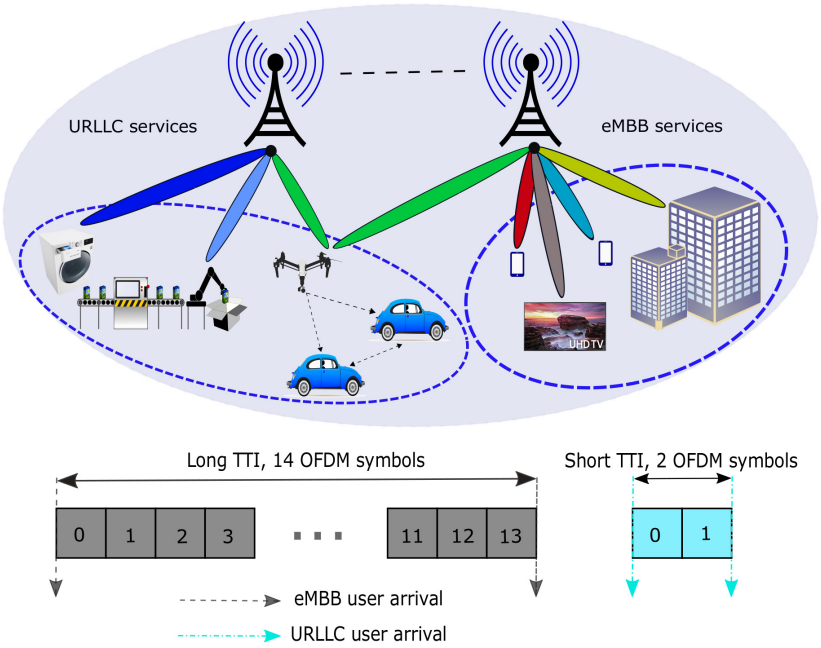


Fig. C.1. Agile 5G system model and frame structure.

in-resource CCH in addition to its desired payload.

Dynamic link adaptation with adaptive selection of the modulation and coding schemes (MCS) is assumed [29], based on the frequency-selective channel quality indication (CQI) user reports. Due to the bursty nature of the FTP3 URLLC and CBR eMBB traffic, the set of active interferers in the system changes sporadically in return, leading to a highly varying interference pattern. Thus, a sliding low pass filter is applied on the instantaneous CQI reports [21] to smooth out the variance of the interference pattern as

$$\partial(t) = \tilde{a}\,\mathring{A} + (1-\tilde{a})\partial(t-1), \tag{C.1}$$

where $\partial(t)$ is the final CQI value based on the averaged interference covariance, to be considered for MCS selection at the t^{th} TTI, $\mathring{A}$ is the CQI value calculated based on the instantaneous interference pattern, and $\tilde{a} \leq 1$ is the filter coefficient to indicate how much confidence should be given to the current reported CQI value. Finally, the Chase combining HARQ re-transmissions [30] are implemented to relax the target BLER transmission requirements, upon the reception of an associated NACK feedback.

2.2 Signal Model

A MU-MIMO signal modeling is adopted in this work, where a maximum subset of MU co-scheduled URLLC-eMBB user pairs $\mathtt{G}_c \in \mathcal{K}_c$ is allowed, where

$G_c = \mathbf{card}(\mathtt{G}_c)$, $G_c \leq N_t$ is the number of co-scheduled users and $\mathcal{K}_c$ is the set of active UEs in the c^{th} cell. Thus, the DL signal, received by the k^{th} user from the c^{th} cell is given by

$$\mathrm{y}^{\kappa}_{k,c} = \mathbf{H}^{\kappa}_{k,c}\mathbf{v}^{\kappa}_{k,c}s^{\kappa}_{k,c} + \sum_{g\in\mathtt{G}_c, g\neq k}\mathbf{H}^{\kappa}_{k,c}\mathbf{v}_{g,c}s_{g,c}$$

$$+ \sum_{j=1, j\neq c}^{C}\sum_{g\in\mathtt{G}_j}\mathbf{H}_{g,j}\mathbf{v}_{g,j}s_{g,j} + \mathbf{n}^{\kappa}_{k,c}, \tag{C.2}$$

where $\mathbf{H}^{\kappa}_{k,c} \in \mathcal{C}^{M_r\times N_t}, \forall k \in \{1,\ldots,K\}, \forall c \in \{1,\ldots,C\}$ is the 3D channel seen at the k^{th} user from the c^{th} cell, $\mathbf{v}^{\kappa}_{k,c} \in \mathcal{C}^{N_t\times 1}$ is the zero-forcing precoding vector, with the assumption of a single layer transmission per user, where $\mathbf{v}^{\kappa}_{k,c} = \left(\mathbf{H}^{\kappa}_{k,c}\right)^{\mathrm{H}}\left(\mathbf{H}^{\kappa}_{k,c}\left(\mathbf{H}^{\kappa}_{k,c}\right)^{\mathrm{H}}\right)^{-1}$. $s^{\kappa}_{k,c}$ and $\mathbf{n}^{\kappa}_{k,c}$ are the transmitted symbol and the additive white Gaussian noise at the k^{th} user, respectively. The first summation indicates the intra-cell interference while the second presents the inter-cell interference, resulted from either the URLLC or eMBB traffic. The 3GPP 3D spatial channel model [31] is adopted, where the DL channel spatial coefficient seen by the m^{th} receive antenna from the n^{th} transmit antenna is composed from Q spatial paths, each with Z rays, and is expressed by

$$h^{\kappa}_{(m,n)_k} = \frac{1}{\sqrt{Q}}\sum_{q=0}^{Q-1}\sqrt{\delta_k}\,\mathcal{G}_{q,k}\,r_{(m,n,q)_k}, \tag{C.3}$$

where $\delta_k = \ell\epsilon_k^{\varrho}\mu_k$ is a constant, ℓ and μ_k are the propagation and shadow fading factors, respectively, ϵ_k^{ϱ} is the physical distance between transceivers, with ϱ as the pathloss exponent, $\mathcal{G}_{q,k} \sim \mathbb{CN}(0,1)$ is a randomness source per channel path. Hence, the channel steering coefficient $r_{(m,n,q)_k}$ is calculated as

$$r_{(m,n,q)_k} =$$

$$\sqrt{\frac{\xi\psi}{Z}}\sum_{z=0}^{Z-1}\begin{pmatrix} \sqrt{\mathfrak{D}^{m,n,q,z}_{\mathrm{BS}}(\theta_{\mathrm{AoD}},\varphi_{\mathrm{EoD}})}\,e^{j(\eta d\overline{f}+\Phi_{m,n,q,z})} \\ \sqrt{\mathfrak{D}^{m,n,q,z}_{\mathrm{UE}}(\theta_{\mathrm{AoA}},\varphi_{\mathrm{EoA}})}\,e^{j(\eta d\sin(\theta_{\mathrm{m,n,q,z,AoA}}))} \\ e^{j\eta||s||\cos(\varphi_{\mathrm{m,n,q,z,EoA}})\cos(\theta_{\mathrm{m,n,q,z,AoA}}-\theta_s)t} \end{pmatrix}, \tag{C.4}$$

where ξ and ψ are the power and large-scale coefficients, $\mathfrak{D}_{\mathrm{BS}}$ and $\mathfrak{D}_{\mathrm{UE}}$ are the antenna patterns at the base-station (BS) and UE, respectively, η is the wave number, θ denotes the horizontal angle of arrival θ_{AoA} and departure θ_{AoD}, while φ implies the elevation angle of arrival φ_{EoA} and departure φ_{EoD}, respectively. s is the user speed, $\overline{f} = f_x\cos\theta_{\mathrm{AoD}}\cos\varphi_{\mathrm{EoD}} + f_y\cos\varphi_{\mathrm{EoD}}\sin\theta_{\mathrm{AoD}} + f_z\sin\varphi_{\mathrm{EoD}}$ is the displacement vector of the transmit antenna array (for a uniform linear array, $f_y = f_z = 0$). Accordingly, the

received signal at the k^{th} user is decoded by applying the receiver vector $\mathbf{u}_{k,c}^{\kappa}$, given by

$$\left(\mathrm{y}_{k,c}^{\kappa}\right)^{*} = \left(\mathbf{u}_{k,c}^{\kappa}\right)^{\mathrm{H}} \mathrm{y}_{k,c}^{\kappa}, \tag{C.5}$$

where $\mathbf{u}_{k,c}^{\kappa}$ is the antenna combining vector, designed by the linear minimum mean square error interference rejection combining (LMMSE-IRC) receiver [32]. Hence, the received SINR at the k^{th} user, assuming an error-free link adaptation process, is expressed by

$$\Upsilon_{k,c}^{\kappa} = \frac{p_k^c \left\| \mathbf{H}_{k,c}^{\kappa} \mathbf{v}_{k,c}^{\kappa} \right\|^2}{1 + \sum\limits_{g \in \mathtt{G}_c, g \neq k} p_g^c \left\| \mathbf{H}_{k,c}^{\kappa} \mathbf{v}_{g,c}^{\kappa} \right\|^2 + \sum\limits_{j \in C, j \neq c} \sum\limits_{g \in \mathtt{G}_j} p_g^j \left\| \mathbf{H}_{g,j} \mathbf{v}_{g,j}^{\kappa} \right\|^2}, \tag{C.6}$$

where p_k^c is the k^{th} user receive power. Then, the received per-PRB data rate of the k^{th} user is expressed as

$$r_{k,rb}^{\kappa} = \log_2 \left(1 + \frac{1}{G_c} \Upsilon_{k,c}^{\kappa}\right). \tag{C.7}$$

Finally, the effective exponential SNR mapping [33] is applied to map the received SINR levels across $\mathcal{N}$ allocated sub-carriers into one effective SINR as

$$\left(\Upsilon_{k,c}^{\kappa}\right)^{\text{eff.}} = -\mathcal{O} \ln \left(\frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} e^{-\frac{\left(\Upsilon_{k,c}^{\kappa}\right)^{i}}{\mathcal{O}}} \right), \tag{C.8}$$

where $\mathcal{O}$ is a calibration parameter.

3 Problem Formulation

The 5G-NR system performance should be continuously optimized per user-centric, instead of network-centric basis. However, the individual user utility functions are highly correlated and need to be reliably fulfilled, e.g., eMBB rate maximization and URLLC latency minimization as

$$\forall k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}} : \arg\max_{\mathcal{K}_{\text{mbb}}} \sum_{k_{\text{mbb}}=1}^{K_{\text{mbb}}} \sum_{rb \in \Xi_{k_{\text{mbb}}}^{\text{mbb}}} \beta_{k_{\text{mbb}}} r_{k_{\text{mbb}},rb}^{\text{mbb}}, \tag{C.9}$$

$$\forall k_{\text{llc}} \in \mathcal{K}_{\text{llc}} : \arg\min_{\mathcal{K}_{\text{llc}}} (\Psi), \tag{C.10}$$

$$\text{s.t.} \left\| \mathbf{v}_k^{\kappa} \sqrt{\mathrm{P}} \right\|^2, \Psi \leq 1\,\text{ms},$$

where $\mathcal{K}_{\text{mbb}}$ and $\mathcal{K}_{\text{llc}}$ represent the active sets of eMBB and URLLC users, respectively, $\Xi^{\text{mbb}}_{k_{\text{mbb}}}$ and $\beta_{k_{\text{mbb}}}$ imply the granted set of PRBs and a priority factor of the k^{th} eMBB user. Ψ is the URLLC target one-way latency, assuming a successful first transmission, which can be given by

$$\Psi = \Lambda_{\text{q}} + \Lambda_{\text{bsp}} + \Lambda_{\text{fa}} + \Lambda_{\text{tx}} + \Lambda_{\text{uep}}, \tag{C.11}$$

where $\Lambda_{\text{q}}, \Lambda_{\text{bsp}}, \Lambda_{\text{fa}}, \Lambda_{\text{tx}}, \Lambda_{\text{uep}}$ are the queuing, BS processing, frame alignment, transmission, and UE processing delays, respectively. Λ_{fa} is upper-bounded by the short TTI interval while Λ_{bsp} & Λ_{uep} are each bounded by 3-OFDM symbol duration [34], due to the enhanced processing capabilities which come with the 5G-NR. Hence, Λ_{q} and Λ_{tx} become the main obstruction against reaching out the hard URLLC latency budget. Λ_{tx} depends on the URLLC outage SINR as

$$\Lambda_{\text{tx}} = \frac{\text{B}_{\text{llc}}}{\left(\Xi^{\text{llc}}_{k_{\text{llc}}} \log_2\left(1 + \frac{\gamma^{\text{llc}}_{k_{\text{llc}}}}{F}\right)\right)}, \tag{C.12}$$

where F is an outage SINR gap to represent a non-ideal link adaptation process. The URLLC queuing delay Λ_{q} can be mathematically represented by an arbitrary queuing model. For instance, we adopt the $\mathcal{A}/\mathcal{A}/1$ queuing model from data networks theory [35], where the first $\mathcal{A}$ implies a Poisson packet arrival, second $\mathcal{A}$ denotes exponential service times, and notation '1' represents a single layer URLLC transmission. Thus, the mean queuing delay $\overline{\Lambda}_{\text{q}}$, can be expressed as

$$\overline{\Lambda}_{\text{q}} = \frac{1}{\overline{\Lambda}_{\text{tx}}\,(1-\rho)}, \tag{C.13}$$

where $\rho = \left(\frac{\lambda}{\overline{\Lambda}_{\text{tx}}}\right)$ is the URLLC traffic intensity, with $\overline{\Lambda}_{\text{tx}}$ as the mean transmission time. Thus, in order to achieve the critical URLLC latency, the transmission and queuing delays should be always minimized to provide further allowance for the HARQ re-transmission delay, if the first transmission is not successful.

Fig. C.2 depicts the URLLC transmission delay versus the received SINR level for different URLLC payload sizes B_{llc} while Fig. C.3 describes the associated URLLC queuing delay. As can be observed, with a larger URLLC payload size, a higher SINR point should be always guaranteed to the URLLC UEs in order to reduce the transmission delay. However, the corresponding queuing delay is shown to significantly depend on the URLLC packet arrival rate, e.g., a larger arrival rate with a degraded mean transmission time results in immensely higher queuing delays. This requires allocating excessive radio resources to URLLC traffic or adopting conservative URLLC transmissions. Consequently, the eMBB utility function in (C.9) is severely under optimized, leading to a significant degradation of the network SE.

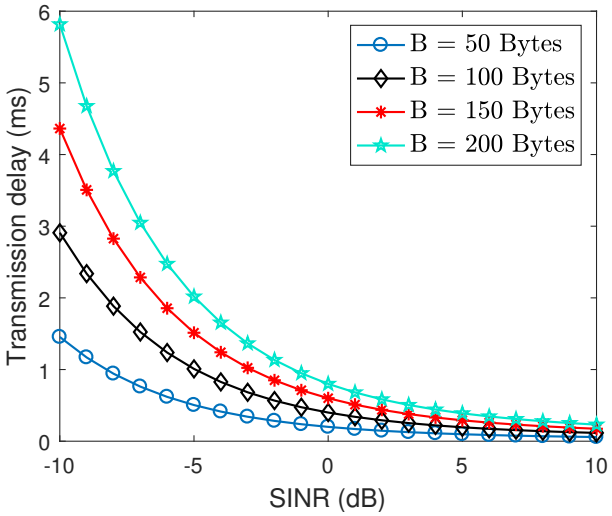


Fig. C.2. URLLC transmission delay with B_{llc}, $\Xi_k^{llc} = 10$ MHz.

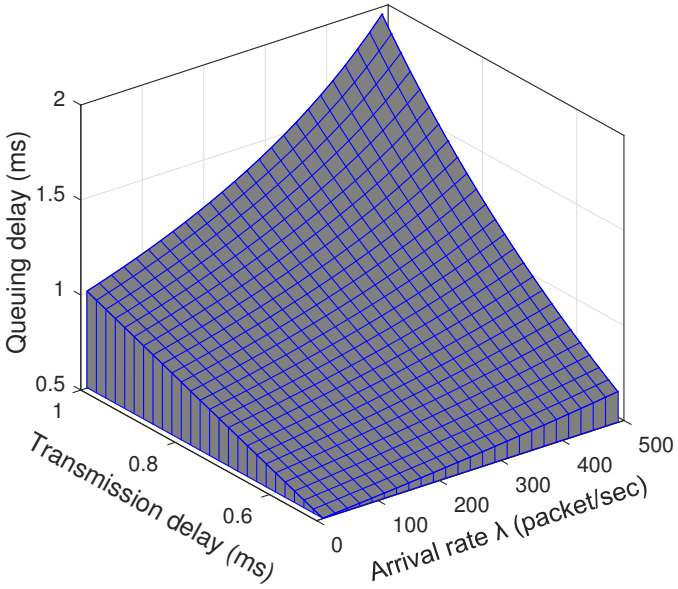


Fig. C.3. URLLC queuing delay with λ and $\overline{\Lambda}_{tx}$.

4 Proposed Spatial Preemptive Scheduling For URLLC and eMBB Coexistence

The proposed NSBPS scheduler seeks to simultaneously cross-optimize the joint performance objectives of the eMBB and URLLC traffic. Thus, the critical URLLC latency deadline is satisfied regardless of the system load while providing the best achievable eMBB performance. When radio resources are not instantly schedulable for incoming URLLC traffic, NSBPS scheduler immediately searches for an ongoing eMBB transmission, that is spatially closest possible to a predefined spatial subspace, i.e., reference subspace. The scheduler instantly projects the selected eMBB transmission onto the reference subspace *on-the-fly*, and accordingly, it assigns the bursty URLLC traffic a portion of the victim eMBB radio resources. At the URLLC user side, it de-orients its decoding vector into one possible null space of the reference subspace; hence, experiencing no inter-user interference, as depicted in Fig. C.4. In the following sub-sections, we describe the proposed NSBPS scheduler in-detail.

4.1 Proposed NSBPS – At the BS Side

Starting at an arbitrary TTI instance, the newly arrived or buffered eMBB traffic is scheduled over single-user (SU) dedicated resources, if there are no pending URLLC arrivals. To dynamically multiplex the active eMBB user allocations across available resources, the proportional fair (PF) scheduling criterion [36] is applied as

$$\Theta_{\text{PF}} = \frac{r^{\text{mbb}}_{k_{\text{mbb}},rb}}{\bar{r}^{\text{mbb}}_{k_{\text{mbb}},rb}}, \tag{C.14}$$

$$k^{*}_{\text{mbb}} = \arg\max_{\mathcal{K}_{\text{mbb}}} \ \Theta_{\text{PF}}, \tag{C.15}$$

where $\bar{r}^{\text{mbb}}_{k_{\text{mbb}},rb}$ is the average delivered data rate of the k^{th} eMBB user. However, in case of URLLC new DL arrivals at the BS while sufficient schedulable resources are instantly available, the NSBPS scheduler overwrites the eMBB user SU scheduling priority for the sake of the newly arrived URLLC traffic, by the weighted PF scheduling criteria (WPF) as

$$\Theta_{\text{WPF}} = \frac{r^{\kappa}_{k_{\kappa},rb}}{\bar{r}^{\kappa}_{k_{\kappa},rb}} \beta_{k_{\kappa}}, \tag{C.16}$$

with $\beta_{k_{\text{llc}}} \gg \beta_{k_{\text{mbb}}}$ for immediate URLLC SU scheduling.

Nonetheless, with a large offered loading level, which is foreseen with the 5G-NR, sufficient resource allocation may not be instantly available for the incoming URLLC traffic. For example, URLLC packets may arrive at the BS

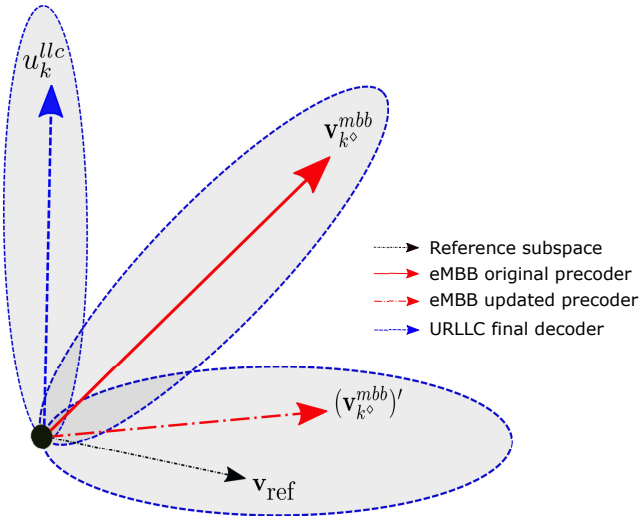


Fig. C.4. NSBPS scheduler: eMBB precoder projection and URLLC decoder orientation.

during an eMBB transmission slot (14-OFDM symbol). Hence, larger scheduling delays, i.e., queuing and/or segmentation delays, are experienced. The URLLC segmentation delay indicates that arrived URLLC payload is segmented and transmitted over multiple TTIs, due to insufficient instant resource allocation or degraded capacity per PRB. For such case, the proposed NSBPS scheduler first attempts fitting the URLLC traffic within one active eMBB transmission using a standard and non-biased MU-MIMO transmission, and based on a highly conservative $\gamma-$orthogonality threshold, with $\gamma \rightarrow [0,1]$. Thus, incoming URLLC traffic can only be paired with an active eMBB transmission if:

$$1 - \left| \left(\mathbf{v}_{k_{\text{mbb}}}^{\text{mbb}} \right)^{\text{H}} \mathbf{v}_{k_{\text{llc}}}^{\text{llc}} \right|^2 \geq \gamma. \tag{C.17}$$

The conservative, i.e., large, orthogonality threshold is forcibly applied to protect the URLLC traffic against potential inter-user interference. If the system spatial degrees of freedom (SDoFs) are restrained during an arbitrary TTI and such large orthogonality requirements can not be satisfied, the NSBPS scheduler instantly enforces a semi-transparent, i.e., URLLC-aware transmission, controlled, i.e., independently from the available SDoFs, and biased, i.e., for the sake of URLLC user end, MU-MIMO transmission. The URLLC outage requirements are then achieved by satisfying:

$$\textbf{rank} \left\{ \left(\mathbf{u}_k^{\text{llc}} \right)^{\text{H}} \mathbf{H}_k^{\text{llc}} \, \mathbf{v}_k^{\text{llc}} \right\} \sim \text{full}, \tag{C.18}$$

$$\mathbf{rank}\left\{\left(\mathbf{u}_k^{\mathrm{llc}}\right)^{\mathrm{H}}\mathbf{H}_k^{\mathrm{llc}}\left(\mathbf{v}_{k^\diamond}^{\mathrm{mbb}}\right)'\right\}\sim 0, \tag{C.19}$$

where $\left(\mathbf{v}_{k^\diamond}^{\mathrm{mbb}}\right)'$ is the actual precoder of the co-scheduled eMBB user with the incoming URLLC user. Then, an arbitrary spatial subspace is pre-defined in the discrete Fourier transform beamforming domain [37] as

$$\mathbf{v}_{\mathrm{ref}}(\theta)=\left(\frac{1}{\sqrt{N_t}}\right)\left[1,e^{-j2\pi\Delta\cos\theta},\ldots,e^{-j2\pi\Delta(N_t-1)\cos\theta}\right]^{\mathrm{T}}, \tag{C.20}$$

where Δ is the absolute antenna spacing and θ is an arbitrary spatial angle. Accordingly, the NSBPS scheduler searches for one active eMBB user whose transmission is most aligned within the reference subspace $\mathbf{v}_{\mathrm{ref}}(\theta)$ as

$$k_{\mathrm{mbb}}^{\diamond}=\arg\min_{\mathcal{K}_{\mathrm{mbb}}}\mathbf{d}\left(\mathbf{v}_k^{\mathrm{mbb}},\mathbf{v}_{\mathrm{ref}}\right), \tag{C.21}$$

where the Chordal distance $\mathbf{d}$ between $\mathbf{v}_k^{\mathrm{mbb}}$ and $\mathbf{v}_{\mathrm{ref}}$ is expressed by

$$\mathbf{d}\left(\mathbf{v}_k^{\mathrm{mbb}},\mathbf{v}_{\mathrm{ref}}\right)=\frac{1}{\sqrt{2}}\left\|\mathbf{v}_k^{\mathrm{mbb}}\left(\mathbf{v}_k^{\mathrm{mbb}}\right)^{\mathrm{H}}-\mathbf{v}_{\mathrm{ref}}\,\mathbf{v}_{\mathrm{ref}}^{\mathrm{H}}\right\|. \tag{C.22}$$

Next, the NSBPS scheduler applies an instant precoder projection of the selected victim eMBB user $\mathbf{v}_{k^\diamond}^{\mathrm{mbb}}$ onto $\mathbf{v}_{\mathrm{ref}}$ as given by

$$\left(\mathbf{v}_{k^\diamond}^{\mathrm{mbb}}\right)'=\frac{\mathbf{v}_{k^\diamond}^{\mathrm{mbb}}\cdot\,\mathbf{v}_{\mathrm{ref}}}{\left\|\mathbf{v}_{\mathrm{ref}}\right\|^2}\times\mathbf{v}_{\mathrm{ref}}, \tag{C.23}$$

wherein $\left(\mathbf{v}_{k^\diamond}^{\mathrm{mbb}}\right)'$ is the post-projection updated eMBB user precoder. This way, the NSBPS scheduler immediately schedules the sporadic URLLC traffic over partial or full shared resource allocation with the victim eMBB transmission. Thus, in principal, no URLLC queuing delays are experienced. On another side, due to the instant projection of the victim eMBB user precoder, it exhibits a capacity loss; however, it is highly constrained and only limited by the spatial projection loss over the shared resources with the URLLC traffic. Furthermore, under larger eMBB user loading, the NSBPS scheduler is highly likely to find an active eMBB user whose transmission is originally aligned within the reference spatial subspace; hence, the instant spatial projection would not significantly impact its achievable capacity. Finally, the BS transmits a single-bit co-scheduling true indication, i.e., $\alpha=1$, to the intended URLLC user, which is transmitted in the user-centric CCH.

4.2 Proposed NSBPS – At the URLLC User Side

When a true co-scheduling indication $\alpha=1$ is detected, the URLLC user acknowledges that its resource allocation is shared with an active eMBB transmission, whose interference is limited within the reference subspace. Thus,

the URLLC user first designs its decoder vector using a standard LMMSE-IRC receiver, to reject inter-cell interference as

$$\left(\mathbf{u}_k^{\text{llc}}\right)^{(1)} = \left(\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}\left(\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}\right)^{\text{H}} + \mathbf{W}\right)^{-1}\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}, \tag{C.24}$$

where the interference covariance matrix is given by

$$\mathbf{W} = \mathbb{E}\left(\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}\left(\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}\right)^{\text{H}}\right) + \sigma^2\text{I}_{M_r}, \tag{C.25}$$

where I_{M_r} is $M_r \times M_r$ identity matrix. The decoder vector statistics $\left(\mathbf{u}_k^{\text{llc}}\right)^{(1)}$ are then transferred to one possible null space of the observed effective inter-user interference subspace $\mathbf{H}_k^{\text{llc}}\mathbf{v}_{\text{ref}}$, as given by

$$\left(\mathbf{u}_k^{\text{llc}}\right)^{(2)} = \left(\mathbf{u}_k^{\text{llc}}\right)^{(1)} - \frac{\left(\left(\mathbf{u}_k^{\text{llc}}\right)^{(1)} \cdot \mathbf{H}_k^{\text{llc}}\mathbf{v}_{\text{ref}}\right)}{\left\|\mathbf{H}_k^{\text{llc}}\mathbf{v}_{\text{ref}}\right\|^2} \times \mathbf{H}_k^{\text{llc}}\mathbf{v}_{\text{ref}}. \tag{C.26}$$

Accordingly, the final URLLC decoder vector $\left(\mathbf{u}_k^{\text{llc}}\right)^{(2)}$ experiences no inter-user interference, providing the URLLC user with a robust decoding ability. To summarize the major concept of the proposed NSBPS scheduler, Fig. C.5 shows a high level flow diagram of the NSBPS scheduler at the BS and intended URLLC user, respectively.

5 Analytical Analysis Compared to State of The Art URLLC Schedulers

In this section, we introduce an analytical performance comparison of the proposed NSBPS scheduler versus the state-of-the-art schedulers from industry and academia as follows:

1. Punctured scheduler (PS) [22]: in case that sufficient radio resources are not instantly available for the sporadic URLLC traffic, the PS scheduler immediately overwrites part of the ongoing eMBB transmissions by the incoming URLLC traffic. Thus, in principal, the URLLC queuing delay component is significantly minimized. PS scheduler has shown sound improvement of the URLLC latency performance; however, with a highly degraded SE, due to the eMBB unrealizable punctured transmissions.

2. Enhanced punctured scheduler (E-PS) [23]: E-PS scheduler is an improved version of the conventional PS scheduler, which is recently proposed to partially recover the lost eMBB capacity due to puncturing. Punctured eMBB UEs are presumed to be aware of which resources are being punctured

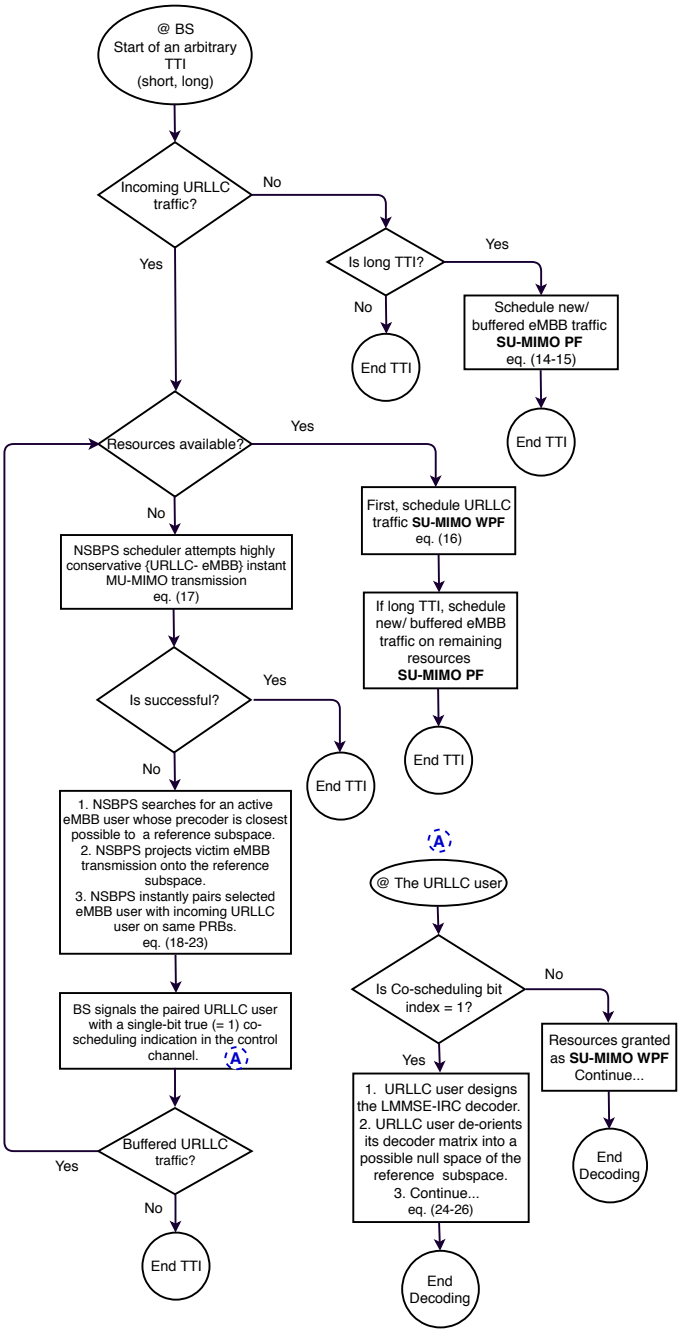


Fig. C.5. Flow diagram of the NSBPS scheduler, at the BS and the intended URLLC user, respectively.

by URLLC traffic. Thus, victim eMBB UEs disregard the punctured PRBs from the Chase combining HARQ process in order not to spread the decoding errors. Furthermore, two code-block (CB) mapping layouts [23 - 25] are evaluated as: fully interleaved (FI), and frequency first (FF) layouts, respectively. The former indicates that CBs associated with an eMBB transport block (TB) are fully interleaved over the time and frequency resources, however, the latter means that CBs are spread over the frequency domain and condensed over the time domain. Moreover, CB-based HARQ feedback is adopted in order for the impacted eMBB UEs to feedback the BS of which punctured CBs could not be successfully decoded, hence, only re-transmitting the victim CBs instead of the full TB, reducing the aggregate HARQ overhead.

3. Multi-user punctured scheduler (MU-PS) [26]: in our recent work, we considered a MU transmission on top of the PS scheduler. The proposed MU-PS scheduler first attempts a non-biased and transparent MU transmission of an URLLC-eMBB user pair. If the system offered SDoFs during an arbitrary TTI are not sufficient, the MU-PS scheduler rolls back to PS scheduler, where the URLLC traffic immediately punctures part of the radio resources, monopolized by ongoing eMBB transmissions. The MU-PS exhibits a fair tradeoff between URLLC latency and overall SE. However, the achievable MU gain is shown to be very restrained with the SDoF-limited conditions, where the MU-PS scheduler is highly likely to fall back to PS scheduler. Furthermore, it has been demonstrated that MU-PS scheduler leads to a degradation of the URLLC decoding ability, due to the potential inter-user interference. Thus, a conservative MU-PS (CMU-PS) scheduler is introduced to further safeguard the URLLC traffic against potentially strong inter-user interference, even if the pairing sum capacity constraint is satisfied. Thus, users can only be paired in a MU-MIMO transmission if their precoders satisfy larger spatial separation as given by

$$\left| \angle \left(\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}} \right) - \angle \left(\mathbf{v}^{\text{llc}}_{k_{\text{llc}}} \right) \right| \geq \vartheta, \tag{C.27}$$

where ϑ is a predefined spatial separation threshold.

Accordingly, the aggregate eMBB user rate is calculated from the individual sub-carrier rates, assuming OFDMA flat fading channels, as

$$r^{\text{mbb}}_{k_{\text{mbb}}} = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb}. \tag{C.28}$$

Next, the fraction of the resources $\Gamma^{\text{llc}}_{k_{\text{mbb}}}$, allocated to the k^{th} eMBB user and being altered by the incoming URLLC traffic, is expressed as a set of random variables, given by

$$\mathbf{\Gamma} = \left(\Gamma^{\text{llc}}_{k_{\text{mbb}}} \mid k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}} \right). \tag{C.29}$$

Due to the small size of URLLC packets, it is reasonable to assume that $\Gamma^{\text{llc}}_{k_{\text{mbb}}} \leq \Xi^{\text{mbb}}_{k_{\text{mbb}}}$ is satisfied. The achievable eMBB user rate can then be formulated by the joint eMBB and URLLC rate allocation function, as expressed by

$$R_{k_{\text{mbb}}} = \mathcal{F}\left(\Xi^{\text{mbb}}_{k_{\text{mbb}}}, \Gamma^{\text{llc}}_{k_{\text{mbb}}}\right). \tag{C.30}$$

For example, an eMBB user exhibits no capacity loss if its associated resource allocation is not induced by incoming URLLC traffic, $\mathcal{F}\left(\Xi^{\text{mbb}}_{k_{\text{mbb}}}, \Gamma^{\text{llc}}_{k_{\text{mbb}}}\right) = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb}$. However, since the URLLC traffic is always prioritized, victim eMBB users exhibit a rate loss over a fraction of the impacted PRBs, where it can be formulated by the rate loss function Π as

$$\mathcal{F}\left(\Xi^{\text{mbb}}_{k_{\text{mbb}}}, \Gamma^{\text{llc}}_{k_{\text{mbb}}}\right) = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb} \left(1 - \Pi\right), \tag{C.31}$$

where the rate loss function $\Pi : [0,1] \rightarrow [0,1]$ represents the effective portion of impacted PRBs of the k^{th} eMBB user. Under the proposed NSBPS framework, the updated eMBB effective channel gain is expressed as

$$\mathcal{Q}^{\text{mbb}}_{k} = \frac{1}{\left[\left(\mathbf{H}^{\text{mbb}}_{k}\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'\right) \times \left(\mathbf{H}^{\text{mbb}}_{k}\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'\right)^{\text{H}}\right]^{-1}}, \tag{C.32}$$

where $\mathcal{Q}^{\text{mbb}}_{k}$ is the post-projection channel gain of the k^{th} eMBB user. The magnitude of $\mathcal{Q}^{\text{mbb}}_{k}$ can be reformulated in terms of the eMBB projection loss, due to the immediate change of the eMBB precoder from $\mathbf{v}^{\text{mbb}}_{k^{\diamond}}$ to $\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'$, as

$$\mathcal{Q}^{\text{mbb}}_{k} = \left\|\mathbf{H}^{\text{mbb}}_{k}\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right\|^{2} \times \sin^{2}\left(\theta_{\left[\mathbf{v}^{\text{mbb}}_{k^{\diamond}}, \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'\right]}\right), \tag{C.33}$$

where $\sin^{2}\left(\theta_{\left[\mathbf{v}^{\text{mbb}}_{k^{\diamond}}, \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'\right]}\right)$ denotes the eMBB precoder projection loss, over the shared resources with the URLLC traffic, and $\theta_{\left[\mathbf{v}^{\text{mbb}}_{k^{\diamond}}, \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'\right]}$ is the spatial angle discrepancy between its original and projected precoders, respectively. Thus, $\overset{\text{NSBPS}}{\Pi}$ is estimated as

$$\overset{\text{NSBPS}}{\Pi} = \left(\frac{\Gamma^{\text{llc}}_{k_{\text{mbb}}}}{\Xi^{\text{mbb}}_{k_{\text{mbb}}}}\right) \times \sin^{2}\left(\theta_{\left[\mathbf{v}^{\text{mbb}}_{k^{\diamond}}, \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'\right]}\right). \tag{C.34}$$

Due to the constraints in (17) and (21), the eMBB projection loss is guaranteed minimum at all times since:

$$\sin^{2}\left(\theta_{\left[\mathbf{v}^{\text{mbb}}_{k^{\diamond}}, \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'\right]}\right) \ll 1. \tag{C.35}$$

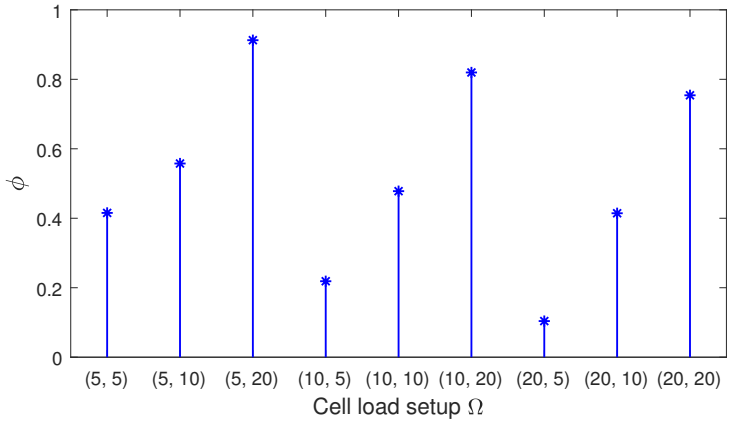


Fig. C.6. MU-PS scheduler: discrete probabilities ϕ of falling back to PS scheduler with Ω.

On another side, the rate loss function of the PS scheduler is expressed by the full URLLC resources altering the eMBB user resources, since the eMBB transmission is instantly stopped over these resources, and it is given by

$$\overset{\text{PS}}{\Pi} = \left(\frac{\Gamma^{\text{llc}}_{k_{\text{mbb}}}}{\Xi^{\text{mbb}}_{k_{\text{mbb}}}} \right). \tag{C.36}$$

The MU-PS scheduler provides an optimized average of the achievable eMBB user rate; however, the MU gain is constrained by the available SDoFs, due to the persistent PS events, if the standard MU-MIMO scheduler fails. Hence, the MU-PS rate loss can be given by

$$\overset{\text{MU-PS}}{\Pi} = \phi \left(\frac{\Gamma^{\text{llc}}_{k_{\text{mbb}}}}{\Xi^{\text{mbb}}_{k_{\text{mbb}}}} \right), \tag{C.37}$$

where $\phi \leq 1$ is the probability of rolling back to PS scheduler under a given cell loading state. Fig. C.6 presents the discrete values of ϕ under different loading conditions, where we define the cell loading as: $\Omega = (K_{\text{mbb}}, K_{\text{llc}})$, and the eMBB full buffer traffic is adopted. As can be observed, with a small number of eMBB users per cell, the system overall SDoFs are highly limited and hence, the MU-PS scheduler is highly likely to roll back to PS scheduler, i.e., $\phi \sim 1$ in order to instantly schedule the offered URLLC traffic. Finally, the average achievable eMBB user rate is expressed by

$$\overline{R}_{k_{\text{mbb}}} = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb} \left(1 - \mathbb{E}\left(\Pi\right)\right). \tag{C.38}$$

Based on (C.28) - (C.38), it can be further observed that the proposed

NSBPS scheduler exhibits the highest ergodic capacity, due to the constrained eMBB rate loss function.

6 Performance Evaluation

The performance of the NSBPS scheduler is validated by extensive SLSs, where the major 5G-NR and radio resource management functionalities are implemented, e.g., agile frame structure, HARQ re-transmission, dynamic link adaptation, and control channel overhead, as described in the subsection II-A. The major simulation parameters are listed in Table C.I. The baseline antenna configuration is 8×2 and the default eMBB traffic is full buffer unless otherwise mentioned.

6.1 Major Performance Comparison

Fig. C.7 depicts the one-way latency of the URLLC traffic at the 10^{-5} outage probability under different cell loading conditions Ω, for the proposed NSBPS, PS, MU-PS, and time-domain WPF (TD-WPF) schedulers. As can be noticed, the NSBPS scheduler offers significant robustness of the URLLC latency performance, independently from the cell loading conditions, and hence, the aggregate interference levels. The performance gain of the NSBPS scheduler is attributed to: a) the elimination of the scheduling queuing delays of the URLLC sporadic traffic, i.e., guaranteed instant URLLC scheduling, b) safeguarding the URLLC traffic from the potential inter-user interference through controlled (almost surely occurs), biased (in favor of the URLLC user), and semi-transparent MU-MIMO transmission, c) compressing the interference spatial dimension, leading to a better LMMSE receiver interference rejection ability, as will be presented in subsection 6.2, and d) the always constrained minimum eMBB cost function.

The PS scheduler provides an optimized URLLC latency performance, especially over the low load region; however, it comes at the expense of a degraded SE. Moreover, it exhibits URLLC performance degradation as the cell load increases, due to the resulting extreme levels of inter-cell interference. Accordingly, a degraded capacity per PRB is experienced. The MU-PS scheduler provides a decent tradeoff between URLLC latency and overall SE due to the achievable MU gain. However, the non-controlled MU interference degrades the URLLC decoding point, especially when the inter-cell interference levels are originally significant. Finally, the TD-WPF scheduler exhibits the worst latency performance since instant URLLC scheduling is not guaranteed, e.g., the URLLC packets are queued for multiple TTIs if the instant schedulable radio resources are not sufficient to accommodate these payloads.

Table C.1: Simulation setup and major parameters

Parameter	Value
Network environment	3GPP Urban Macro (UMa) network with 21 cells and 500 meter inter-site distance
Carrier configuration	10 MHz carrier bandwidth at 2 GHz
Propagation	$128.1 + 37.6 \log(D[km])$ dB; Log-Normal shadowing with 8 dB standard deviation
PHY numerology	15 kHz subcarrier spacing; 12 subcarriers per PRB; 2-OFDM symbols TTI (0.143 ms)
Control channel	Error-free in-resource scheduling grants with dynamic link adaptation
Data channel MCS	QPSK to 64QAM, LTE coding rates
CSI	LTE-alike CQI and PMI, $\tilde{a} = 0.01$, reported every 5 ms; Sub-band size: 8 PRBs
Antenna configuration	8 x 2 MU-MIMO with LTE-alike precoding and LMMSE-IRC receiver
Packet scheduler	Weighted Proportional Fair with priority for URLLC traffic
BLER target	eMBB: 10 percent; URLLC: 1 percent
HARQ	Asynchronous HARQ with Chase combining and 4 TTI round trip time; Max. 6 HARQ retransmissions
RLC setup	Transparent mode
Traffic composition	URLLC: 5, 10, and 20 users / cell; eMBB: 5, 10, and 20 users / cell
UE distribution	Uniformly distributed; 3 km/h UE speed
Traffic model	URLLC: FTP3 downlink traffic, $\lambda = 250$ and B = 50 bytes; eMBB: full buffer and CBR traffic

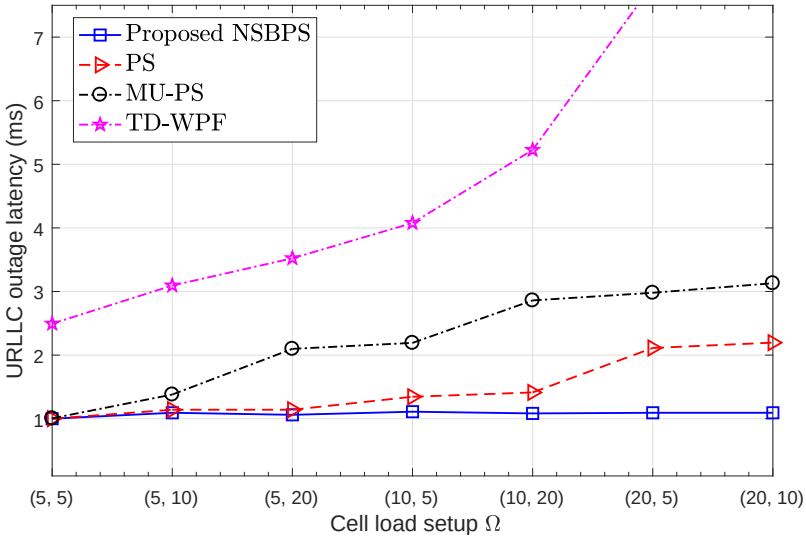


Fig. C.7. URLLC outage latency of the NSBPS, PS, MU-PS, and TD-WPF schedulers, with Ω.

Fig. C.8 shows the average cell throughput in Mbps with the cell loading condition Ω. The NSBPS scheduler achieves the best cell throughput performance because the eMBB cost function is limited by the spatial projection loss, and thus, it is always constrained minimum, compared to the PS, MU-PS, and TD-WPF schedulers. The PS scheduler clearly suffers from severe degradation in the cell ergodic capacity due to the eMBB punctured transmissions. However, the TD-WPF scheduler exhibits an improved cell performance since punctured eMBB transmissions are not allowed; however, at the expense of significant URLLC queuing delays. Finally, the MU-PS scheduler provides a better cell capacity than TD-WPF and PS schedulers, due to the achieved MU gain; however, gain is highly limited by the available system SDoFs, and hence, dependent on the cell loading condition, and aggregate interference levels, e.g., MU-PS scheduler is highly likely to roll back to SE-less-efficient PS scheduler when the system SDoFs are limited within a TTI. In Fig. C.9, we compare the empirical cumulative distribution function (ECDF) of the achievable cell throughput of the proposed NSBPS scheduler against the state-of-the-art E-PS and CMU-PS schedulers, respectively, for $\Omega = (5, 5)$. As can be clearly identified, the NSBPS scheduler still outperforms all schedulers under assessment due to the guaranteed minimum projection loss of the victim eMBB UEs. On the other hand, the CMU-PS scheduler provides an optimized cell throughput performance due to enforcing a conservative MU pairing constraint; thus, the CMU-PS scheduler performs less MU pairings;

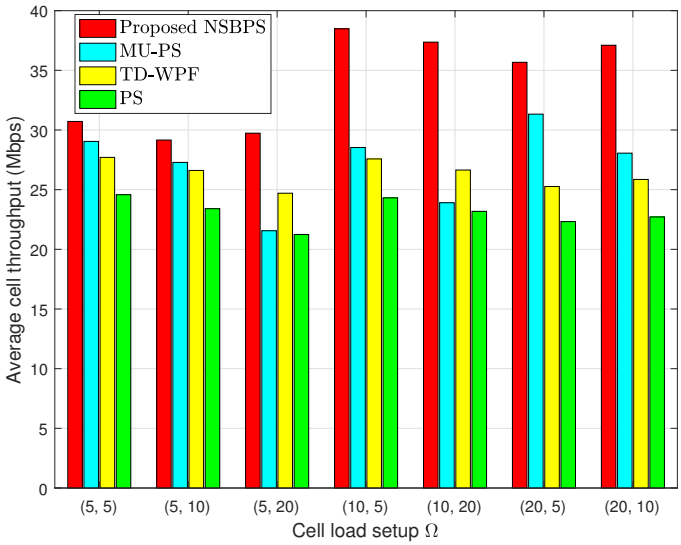


Fig. C.8. Average cell throughput performance of the NSBPS, PS, MU-PS, and TD-WPF schedulers, with Ω.

however, with a higher MU gain. The conventional PS scheduler shows the worst SE because the puncturing events severely degrade the eMBB capacity. Finally, the E-PS scheduler shows an improved cell throughput than the PS, for both the FI and FF CB layouts, respectively. The E-PS scheduler with FI CB layout is shown to slightly outperform that is of the FF CB, since a modest and equal puncturing impact on all CBs minimizes the error probability of the entire TB compared to the case of the FF CB, where only a few CBs, i.e., condensed in the time-domain, are completely damaged due to puncturing.

6.2 Performance Drivers of The Proposed NSBPS Scheduler

Examining the performance drivers of the proposed NSBPS scheduler, Fig. C.10 shows the average achievable capacity per scheduled eMBB/URLLC allocations in bits. The proposed scheduler clearly enhances the allocation average capacity due to the controlled MU pairing, and the limited eMBB projection loss. The MU-PS scheduler shows an improved capacity, however, it depends on the available system SDoFs, e.g., with SDoF-limited condition ($\Omega = (5, 20)$), the MU-PS scheduler exhibits a similar allocation capacity as of the PS scheduler. The PS scheduler provides the worst performance due to the punctured eMBB transmissions and the hard priority of the URLLC traffic. Similar conclusions can be also reached from Fig. C.11, where the average number of the TD queued users is depicted, i.e., the average num-

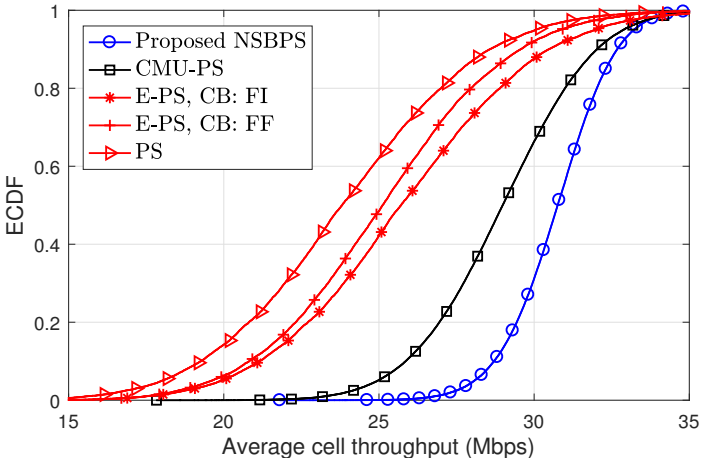


Fig. C.9. Average cell throughput performance of the NSBPS, CMU-PS, E-PS, and PS schedulers, with $\Omega = (5, 5)$.

ber of active users which are queued in the TD scheduler for multiple TTIs until sufficient resources are released. Due to its achievable higher allocation capacity, the NSBPS scheduler shows the lowest number of the TD-queued users against the MU-PS and PS schedulers, respectively. However, under a large offered load, e.g., $\Omega = (20, 5)$, all schedulers under evaluation suffer from a larger queuing delay due to the extreme interference levels, and hence, PRB degraded capacity. Furthermore, Fig. C.12 depicts the URLLC per packet effective SINR in dB, as in eq. (C.8), for $\Omega = (5, 20)$. The NSBPS scheduler provides ~ 1 dB gain in the average FTP3 packet SINR over the PS scheduler. The fixed subspace projection of the victim eMBB transmissions leads to regularizing the inter-cell interference statistics from different cells into a compressed spatial span. Thus, the LMMSE-IRC receiver has better SDoFs to reject and null the interference statistics from the received signal, leading to a better SINR performance with the NSBPS scheduler. However, the MU-PS scheduler exhibits the worst SINR level per FTP3 packet due to the residual inter-user interference from the standard MU transmissions.

6.3 eMBB Realistic Traffic Model

Examining the end-to-end eMBB performance, we also consider a more realistic traffic modeling in order to emulate the coexistence of the broadband video streaming services with the URLLC applications. Under this assumption, a constant bit rate (CBR) traffic modeling is adopted for the eMBB users, where $\check{n} = 10$, $\mathrm{B}_{\mathrm{mbb}} = 320$ KBytes, and $\check{\imath} = 0.6864$ sec. This implies a clip time of ~ 6.1776 sec and CBR load of ~ 4 Mbps per eMBB user. When an arbitrary

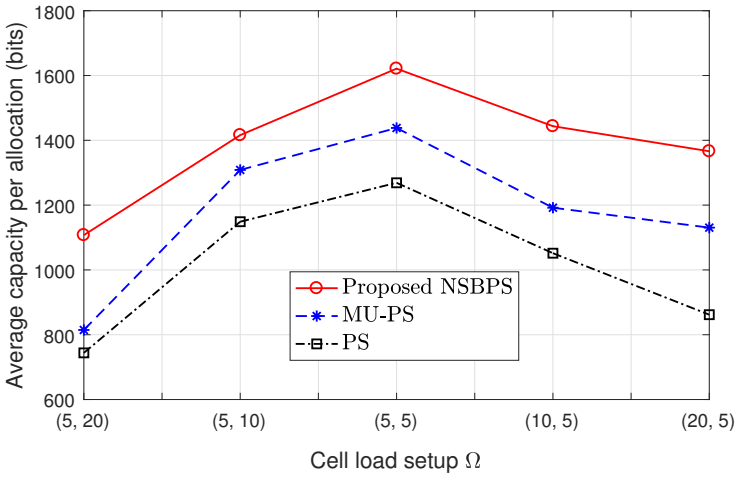


Fig. C.10. Average capacity per scheduled allocation size of the NSBPS, MU-PS, and PS schedulers, with Ω.

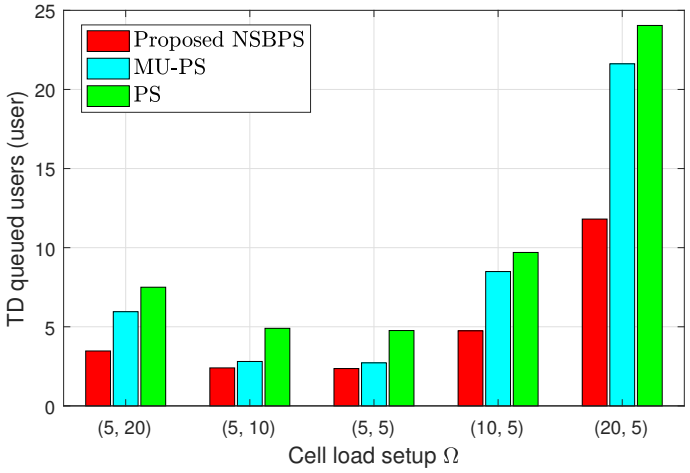


Fig. C.11. TD user queuing performance of the NSBPS, MU-PS, and PS schedulers, with Ω.

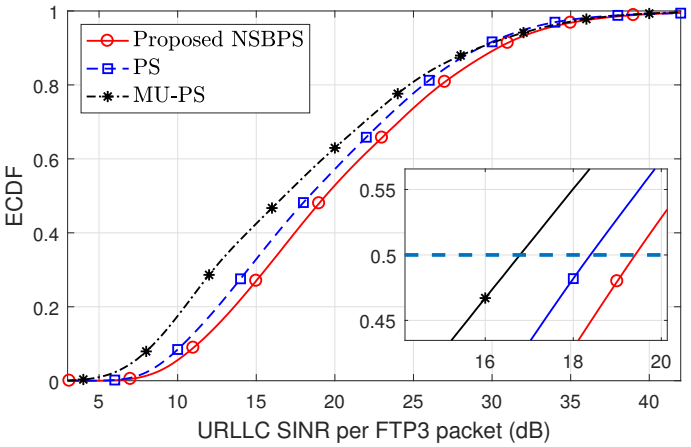


Fig. C.12. URLLC per packet SINR performance of the NSBPS, MU-PS, and PS schedulers, with $\Omega = (5, 20)$.

eMBB user finishes its corresponding streaming session, another eMBB user is generated with a random position in the simulation.

Fig. C.13 depicts the complementary CDF (CCDF) of the URLLC one-way latency, for different antenna configurations, i.e., 8×2 and 8×8, respectively. As can be seen, with 8×2 antenna setup, the URLLC latency performance of both NSBPS and PS schedulers is significantly degraded, where the URLLC 1 ms outage latency can not be satisfied. This is due to the highly varying set of active interferers, resulting from the bursty eMBB CBR traffic. Hence, the resultant fast varying interference pattern disrupts the URLLC link adaptation process, leading to several HARQ re-transmissions before a successful decoding. One possible suggestion is to utilize the channel hardening phenomenon [38] by increasing the size of the transmit and receive antenna arrays, for the same transceiver complexity. With larger antenna arrays, the spatial channel becomes more directive on the desired paths with much less energy leakage on interference paths, leading to a better decoding ability of the LMMSE-IRC receiver. Hence, with 8×8 antenna setup, the URLLC latency performance of both schedulers is clearly improved, achieving the URLLC latency target with the NSBPS scheduler, due to the significantly reduced interference leakage. Finally, Fig. C.14 depicts the ECDF of the achievable eMBB user CBR, where similar conclusions can be drawn.

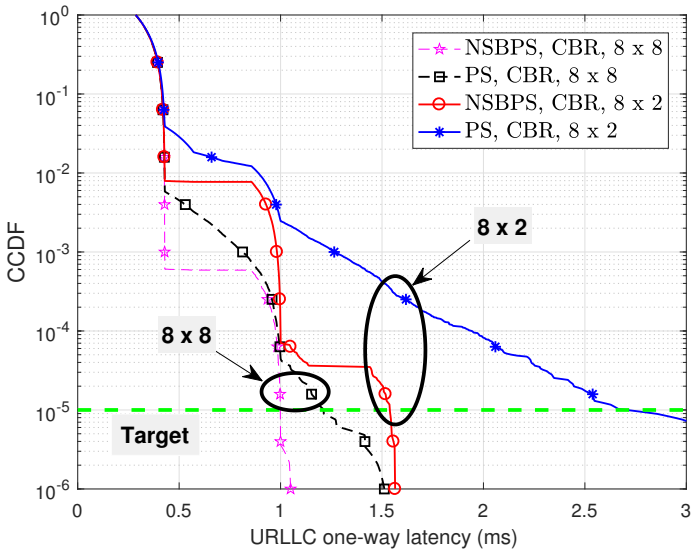


Fig. C.13. URLLC latency CCDF of the NSBPS, and PS schedulers, with eMBB CBR traffic and $\Omega = (5, 10)$.

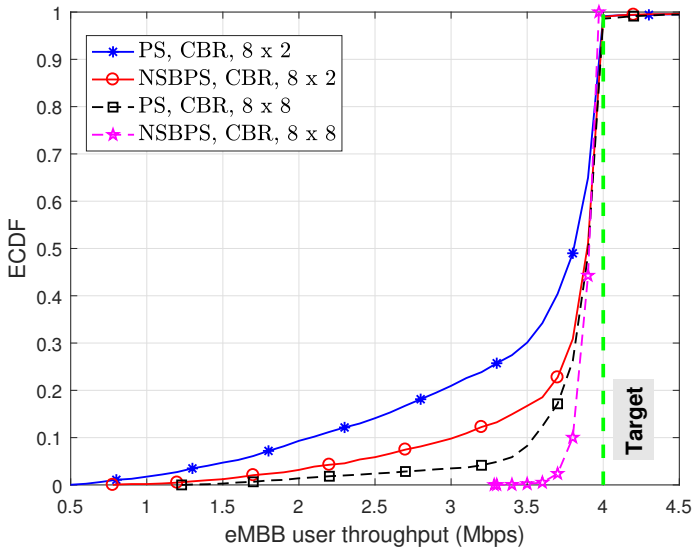


Fig. C.14. eMBB CBR throughput performance of the NSBPS, and PS schedulers, with $\Omega = (5, 10)$.

7 Concluding remarks

An attractive null-space-based preemptive scheduler (NSBPS) for joint eMBB and URLLC traffic is introduced. Proposed NSBPS scheduler guarantees an instant scheduling for the sporadic URLLC traffic, and with the minimal impact on the overall ergodic capacity. Thus, the sporadic URLLC traffic experiences no further queuing delays in order to achieve its critical one-way latency budget. A variety of dynamic system level simulations in addition to an analytic analysis of the major performance indicators are carried out to validate the performance of the proposed scheduler. Compared to the state-of-the-art scheduling proposals from industry and academia, the proposed NSBPS shows extreme URLLC latency robustness with significantly improved eMBB performance.

The major conclusions brought by this paper can be summarized as follows: (1) the transmission and queuing delay components are the major obstacles against achieving the URLLC hard latency, and those are highly correlated and dependent on the URLLC payload size and the mean packet arrival rate, (2) thus, URLLC users must satisfy their outage capacity of interest instead of the overall ergodic capacity, leading to a severe degradation of the network spectral efficiency, (3) proposed NSBPS scheduler instantly schedules the sporadic URLLC traffic regardless of the network loading state, reducing the URLLC queuing delays, and (4) NSBPS scheduler safeguards the URLLC traffic from potential inter-user interference by enforcing sufficient spatial separation through subspace projection. A detailed study on recovering the eMBB capacity will be considered in a future work.

8 Acknowledgments

This work is partly funded by the Innovation Fund Denmark (IFD) – case number: 7038-00009B. Also, 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] NR and NG-RAN overall description; Stage-2 (Release 15), 3GPP, TS 38.300, V2.0.0, Dec. 2017.

[2] Study on new radio access technology (Release 14), 3GPP, TR 38.801, V14.0.0, March 2017.

[3] Study on scenarios and requirements for next generation access technologies (Release 14), 3GPP, TR 38.913, V14.3.0, June 2016.

[4] Study on new radio access technology physical layer aspects (Release 14), 3GPP, TR 38.802, V14.2.0, Sep. 2017.

[5] IMT vision – "Framework and overall objectives of the future development of IMT for 2020 and beyond", international telecommunication union (ITU), ITU-R M.2083-0, Feb. 2015.

[6] P. Popovski, "Ultra-reliable communication in 5G wireless systems," *in Proc. IEEE International Conference on 5G for Ubiquitous Connectivity*, Akaslompolo, 2014, pp. 146-151.

[7] E. Dahlman et al., "5G wireless access: requirements and realization," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 42-47, Dec. 2014.

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

[9] B. Soret, P. Mogensen, K. I. Pedersen and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," *in Proc. IEEE Globecom*, Austin, TX, 2014, pp. 1391-1396.

[10] 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 Netw.*, vol. 32, no. 2, pp. 8-15, Mar. 2018.

[11] Ali A. Esswie, and K.I. Pedersen, "Null space based preemptive scheduling for joint URLLC and eMBB traffic in 5G networks," *Submitted to IEEE Globecom*, Abu Dhabi, 2018.

[12] K. Pedersen, G. Pocovi, J. Steiner and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 210-217, Mar. 2018.

[13] G. Pocovi, K. I. Pedersen and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Netw.*, vol. 6, pp. 28912-28922, May 2018.

[14] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen and A. Szufarska, "A flexible 5G frame structure design for FDD cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53-59, March 2016.

[15] 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 Proc. IEEE Globecom*, Washington, DC, 2016, pp. 1-7.

[16] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen and P. Mogensen, "Signal quality outage analysis for ultra-reliable communications in cellular networks," *in Proc. IEEE Globecom*, San Diego, CA, 2015, pp. 1-6.

[17] G. Pocovi, M. Lauridsen, B. Soret, K. I. Pedersen and P. Mogensen, "Ultra-reliable communications in failure-prone realistic networks," *in Proc. IEEE ISWCS*, Poznan, 2016, pp. 414-418.

[18] J. J. Nielsen, R. Liu and P. Popovski, "Ultra-reliable low latency communication using interface diversity," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1322-1334, March 2018.

[19] R. Kotaba, R. Kotaba, C. N. Manchon, T. Balercia and P. Popovski, "Uplink transmissions in URLLC systems with shared diversity resources," *IEEE Commun. Lett.*, pp. 1-4, Jan. 2018.

[20] J. Rao and S. Vrzic, "Packet duplication for URLLC in 5G: architectural enhancements and performance analysis," *IEEE Netw.*, vol. 32, no. 2, pp. 32-40, March-April 2018.

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

[22] K.I. Pedersen, G. Pocovi, J. Steiner, and S. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," *in Proc. IEEE VTC*, Toronto, 2017, pp. 1-6.

[23] K. I. Pedersen, G. Pocovi, and J. Steiner, "Preemptive scheduling of latency critical traffic and its impact on mobile broadband performance," *in Proc. VTC,* Porto, 2018, pp. 1-6.

[24] K. I. Pedersen, S. R. Khosravirad, G. Berardinelli and F. Frederiksen, "Rethink hybrid automatic repeat request design for 5G: five configurable enhancements," *IEEE Wireless Commun.,* vol. 24, no. 6, pp. 154-160, Dec. 2017.

[25] S. R. Khosravirad, L. Mudolo and K. I. Pedersen, "Flexible multi-bit feedback design for HARQ operation of large-size data packets in 5G," *in Proc. VTC*, Sydney, NSW, 2017, pp. 1-5.

[26] Ali A. Esswie, and K.I. Pedersen, “Multi-user preemptive scheduling for critical low latency communications in 5G networks,” *in Proc. IEEE ISCC*, Natal, 2018, pp. 1-6.

[27] S. Sesia, I. Toufik, and M. Baker, "Orthogonal frequency division multiple access (OFDMA)," in LTE - The UMTS Long Term Evolution: From Theory to Practice , 1, Wiley Telecom, 2011, pp.123-143.

[28] K. Xu, D. Tipper, Y. Qian, P. Krishnamurthy and S. Tipmongkonsilp, "Time-varying performance analysis of multi-hop wireless networks with CBR traffic," *IEEE Trans. Veh. Technol.*, vol. 63, no. 7, pp. 3397-3409, Sept. 2014.

[29] J. P. Singh, Y. Li, N. Bambos, A. Bahai, B. Xu and G. Zimmermann, "TCP performance dynamics and link-layer adaptation based optimization methods for wireless networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1864-1879, May 2007.

[30] E. W. Jang, J. Lee, H. L. Lou and J. M. Cioffi, "On the combining schemes for MIMO systems with hybrid ARQ," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 836-842, Feb. 2009.

[31] Study on 3D channel model for LTE; Release 12, 3GPP, TR 36.873, V12.7.0, Dec. 2014.

[32] Y. Ohwatari, N. Miki, Y. Sagae and Y. Okumura, "Investigation on interference rejection combining receiver for space–frequency block code transmit diversity in LTE-advanced downlink," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 191-203, Jan. 2014.

[33] S. N. Donthi and N. B. Mehta, "An accurate model for EESM and its application to analysis of CQI feedback schemes," *IEEE Trans. Wireless Commun.*, vol. 10, no. 10, pp. 3436-3448, Oct. 2011.

[34] Physical layer procedures; Evolved universal terrestrial radio access (Release 15), 3GPP, TS 36.213, V15.1.0, March. 2018.

[35] Bertsekas, D. and Gallager, R. (1992). Data Networks. 2nd ed. Michigan: Prentice Hall.

[36] D. Parruca and J. Gross, "Throughput analysis of proportional fair scheduling for sparse and ultra-dense interference-limited OFDMA/LTE networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6857-6870, Oct. 2016.

[37] Y. Han, S. Jin, J. Zhang, J. Zhang and K. K. Wong, "DFT-based hybrid beamforming multiuser systems: rate analysis and beam selection," *IEEE J. Sel. Topics Signal Process.*, pp. 1-15, March 2018.

[38] T. L. Narasimhan and A. Chockalingam, "Channel hardening-exploiting message passing receiver in large-scale MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 847-860, Oct. 2014.

Paper D

Capacity Optimization of Spatial Preemptive Scheduling For Joint URLLC-eMBB Traffic in 5G New Radio

Ali A. Esswie and Klaus I. Pedersen

The paper has been published in the
2018 IEEE Global Communications Conference (GLOBECOM)

Abstract

Ultra-reliable and low-latency communication (URLLC) is envisioned as a primary service class of the fifth generation mobile networks. URLLC applications demand stringent radio latency requirements of 1 millisecond with 99.999% confidence. Obviously, the coexistence of the URLLC services and enhanced mobile broadband (eMBB) applications on the same spectrum imposes a challenging scheduling problem. In this paper, we propose an enhanced spatial preemptive scheduling framework for URLLC-eMBB traffic coexistence. The proposed scheduler ensures an instant and interference-free signal subspace for critical URLLC transmissions, while achieving best-effort eMBB performance. Furthermore, the impacted eMBB capacity is then recovered by limited network-assisted signaling. The performance of the proposed scheduler is evaluateeed by highly detailed system level simulations of the major performance indicators. Compared to the state-of-the-art multi-traffic schedulers from industry and academia, the proposed scheduler meets the stringent URLLC latency requirements, while significantly improving the achievable ergodic capacity.

Index Terms— URLLC; eMBB; 5G; Preemptive scheduling, MU-MIMO, Latency.

1 Introduction

The fifth generation (5G) of the mobile communications features two major service classes: ultra-reliable and low-latency communications (URLLC) and enhanced mobile broadband (eMBB) [1, 2]. eMBB applications support stable and delay-tolerant connections with extremely high data rates. However, URLLC critical services demand very low radio latency of 1 millisecond with 10^{-5} outage probability [3]. This category of the URLLC quality of service (QoS) is vastly different from that of the current 4G technology, where the spectral efficiency (SE) is the prime objective. Hence, the support of URLLC is envisioned to enable many future real-time applications such as virtual reality, self-driving vehicles, and tactile internet [4].

However, in pursuit of such extreme SE requirements for eMBB services and tight latency & reliability targets for URLLC, a prime scheduling challenge is how to strategically multiplex such diverse requirements on same spectrum [5]. For instance, to satisfy the URLLC latency and reliability budgets, the system must be forcibly engineered such that blocking a URLLC packet at an arbitrary transmission time interval (TTI) is a rare event. Such scheduling behavior imposes a severe degradation of the overall ergodic capacity, due to the fundamental trade-off between reliability, latency and SE [6].

In the recent open literature, eMBB and URLLC service coexistence in 5G new radio (NR) has gained progressive research attention from industry

and academia. Such multi-service scheduling problem is the dominant study item of the upcoming 3GPP release-16 [7]. Furthermore, user-centric TTI scheduling is demonstrated as essential to achieve the URLLC latency and reliability targets [8, 9], i.e., URLLC users are scheduled on a short TTI duration; however, eMBB users on a longer TTI duration. Spatial diversity techniques are also considered key enablers for URLLC, to enhance the URLLC decoding ability by preserving a sufficient signal-to-interference-noise-ratio (SINR) level [10, 11]. Moreover, URLLC preemptive scheduling (PS) [12] is a state-of-the-art technique to instantly schedule sporadic URLLC traffic with minimum queuing delay. If the radio resources are monopolized by ongoing eMBB transmissions, PS scheduler immediately overwrites part of eMBB physical resource blocks (PRBs) for the sake of the incoming URLLC traffic.

In [5], we demonstrated that a standard multi-user multiple-input multiple-output (MU-MIMO) URLLC-eMBB transmission on top of PS scheduler (MUPS) is an attractive solution to provide a fair trade-off between URLLC performance and overall SE. That is, the MUPS scheduler first attempts a URLLC-eMBB MU-MIMO transmission. If the MU pairing is not possible at an arbitrary TTI, MUPS scheduler rolls back to PS scheduler for instant URLLC scheduling. However, when the system spatial degrees of freedom (SDoFs) are limited, MUPS scheduler offers a limited MU gain and degraded URLLC latency and reliability, since the standard MU-MIMO pairing condition is only constrained by the achievable sum rate. In our recent studies [13, 14], we have introduced a null space based preemptive scheduler (NSBPS), altering the MU pairing condition to instantly offer an interference-free signal subspace for sporadic URLLC traffic, through subspace projection, where the loss in the ergodic capacity is upper-bounded by the eMBB projection loss.

In this paper, an enhanced NSBPS (eNSBPS) scheduling framework for downlink (DL) 5G-NR is proposed. When incoming URLLC traffic can not be immediately scheduled, i.e., without queuing or segmentation, the eNSBPS scheduler immediately alters the system optimization to a region where the URLLC QoS is instantly guaranteed, and delay-tolerant eMBB QoS is recovered through limited network-assisted signaling. eNSBPS searches for an active eMBB transmission whose transmission is most aligned within a pre-defined reference spatial subspace. Next, eNSBPS projects the selected eMBB transmission onto the reference subspace for which its instantly paired URLLC user, on the same resources, aligns its decoding matrix into a possible null space of the reference subspace; thus, experiencing an interference-free transmission. Then. the base-station (BS) signals the victim eMBB users with limited signaling components in the control channel to recover the inflicted capacity loss due to the instant spatial projection, hence, achieving the maximum possible ergodic capacity of a multi-traffic MU system. Compared to the state-of-the-art scheduler proposals, eNSBPS scheduling framework shows a robust URLLC performance with a significantly improved ergodic

capacity.

Due to the complexity of the 5G-NR system model and addressed problems herein [1-3], the performance of the proposed eNSBPS scheduler is validated using highly detailed system level simulations (SLSs), where the majority of the 5G-NR protocol stack is implemented and calibrated against the 3GPP 5G-NR assumptions, including but not limited to: dynamic link adaptation & user scheduling, hybrid automatic repeat request (HARQ), 3D channel modeling and estimation.

Notations: $(\mathcal{X})^{\mathrm{T}}$, $(\mathcal{X})^{\mathrm{H}}$ and $(\mathcal{X})^{-1}$ stand for the transpose, Hermitian, and inverse operations of $\mathcal{X}$, $\mathcal{X} \cdot \mathcal{Y}$ is the dot product of $\mathcal{X}$ and $\mathcal{Y}$, while $\overline{\mathcal{X}}$ and $\|\mathcal{X}\|$ are the mean and 2-norm of $\mathcal{X}$. $\angle\mathcal{X}$ denotes the principal phase direction of $\mathcal{X}$. $\mathcal{X}^{\kappa}, \kappa \in \{\mathrm{llc}, \mathrm{mbb}\}$ denotes the type of user $\mathcal{X}$, $\mathbb{E}\{\mathcal{X}\}$ and $\mathbf{card}(\mathcal{X})$ are the statistical expectation and cardinality of $\mathcal{X}$.

This paper is organized as follows. Section 2 presents the system model of this work. Section 3 introduces the problem formulation and detailed description of the proposed scheduler framework. Section 4 discusses the performance evaluation results. The paper is concluded in Section 5 while work acknowledgments are presented in Section 6.

2 System Model

We adopt a DL MU-MIMO transmission in 5G-NR [13, 14], where there are C cells, each with N_t transmit antennas. Each cell serves $K_{\mathrm{mbb}} + K_{\mathrm{llc}} = K$ users on average, each with M_r receive antennas, where K_{mbb} and K_{llc} are the average numbers of eMBB and URLLC users per cell. We asses two types of DL traffic as: a) URLLC sporadic FTP3 traffic with B−byte payload size and a Poisson point arrival λ, and b) eMBB full buffer traffic with infinite payload size. As depicted in Fig. D.1, the agile 5G-NR frame structure is considered in this work, where URLLC traffic is scheduled on a short TTI duration to satisfy its stringent latency targets, i.e., 2-symbol TTI, while eMBB users can be scheduled on a longer TTI duration, i.e., 14-symbol TTI, to maximize the achievable SE. In the frequency domain, users are dynamically multiplexed using orthogonal frequency division multiple access, where the smallest schedulable unit is the PRB, i.e., 12 sub-carriers of 15 kHz sub-carrier spacing.

We assume a maximum subset $\mathtt{G}_c \in \mathcal{K}_c$ of MU URLLC-eMBB co-scheduled users over an arbitrary PRB in the c^{th} cell, where $G_c = \mathbf{card}(\mathtt{G}_c)$, $G_c \leq N_t$ is the MU user rank per PRB and $\mathcal{K}_c$ is the set of active eMBB/URLLC users in the c^{th} cell. The DL received signal at the k^{th} user from the c^{th} cell is expressed by

$$\mathrm{y}^{\kappa}_{k,c} = \mathbf{H}^{\kappa}_{k,c}\mathbf{v}^{\kappa}_{k,c}s^{\kappa}_{k,c} + \sum_{g \in \mathtt{G}_c, g \neq k} \mathbf{H}^{\kappa}_{k,c}\mathbf{v}_{g,c}s_{g,c}$$

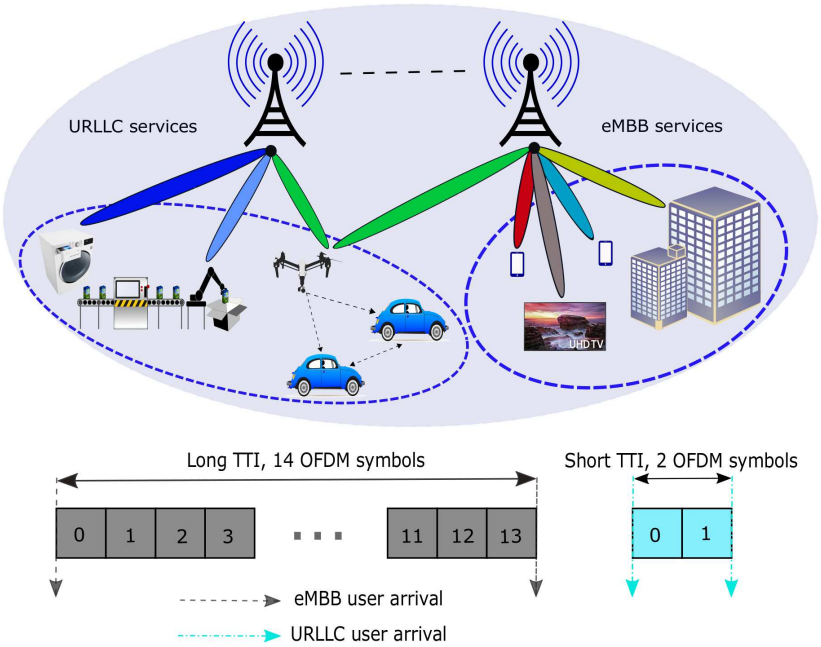


Fig. D.1. Agile TTI structure in 5G-NR.

$$+ \sum_{j=1, j\neq c}^{C} \sum_{g \in \mathbb{G}_j} \mathbf{H}_{g,j} \mathbf{v}_{g,j} s_{g,j} + \mathbf{n}_{k,c}^{\kappa}, \tag{D.1}$$

where $\mathbf{H}_{k,c}^{\kappa} \in \mathcal{C}^{M_r \times N_t}, \forall k \in \{1, \ldots, K\}, \forall c \in \{1, \ldots, C\}$ is the 3D spatial channel matrix [15]. $\mathbf{v}_{k,c}^{\kappa} \in \mathcal{C}^{N_t \times 1}$ is the zero-forcing beamforming vector, with a single spatial stream per user, calculated as: $\mathbf{v}_{k,c}^{\kappa} = \left(\mathbf{H}_{k,c}^{\kappa}\right)^{\mathrm{H}} \left(\mathbf{H}_{k,c}^{\kappa} \left(\mathbf{H}_{k,c}^{\kappa}\right)^{\mathrm{H}}\right)^{-1}$. Finally, $s_{k,c}^{\kappa}$ and $\mathbf{n}_{k,c}^{\kappa}$ indicate the transmitted symbol and the additive white Gaussian noise, respectively. The first summation represents the intra-cell inter-user interference while the latter introduces the inter-cell interference, resulting from the URLLC and eMBB traffic. The received signal is then decoded by the linear minimum mean square interference rejection combining (LMMSE-IRC) [16] vector $\mathbf{u}_{k,c}^{\kappa}$ as

$$\left(\mathrm{y}_{k,c}^{\kappa}\right)^{*} = \left(\mathbf{u}_{k,c}^{\kappa}\right)^{\mathrm{H}} \mathrm{y}_{k,c}^{\kappa}, \tag{D.2}$$

where $\left(\mathrm{y}_{k,c}^{\kappa}\right)^{*}$ is the post-combining received signal. Then, the received SINR at the k^{th} user can be represented by:

$$\Upsilon_{k,c}^{\kappa} = \frac{p_k^c \left\| \mathbf{H}_{k,c}^{\kappa} \mathbf{v}_{k,c}^{\kappa} \right\|^2}{1 + \sum\limits_{g \in \mathtt{G}_c, g \neq k} p_g^c \left\| \mathbf{H}_{k,c}^{\kappa} \mathbf{v}_{g,c}^{\kappa} \right\|^2 + \sum\limits_{j \in C, j \neq c} \sum\limits_{g \in \mathtt{G}_j} p_g^j \left\| \mathbf{H}_{g,j}^{\kappa} \mathbf{v}_{g,j}^{\kappa} \right\|^2}, \tag{D.3}$$

where p_k^c represents the transmit power towards the k^{th} user.

3 Proposed eNSBPS Scheduler

3.1 Problem Formulation

Inline with the 5G-NR targets in the upcoming 3GPP release-16 [7], the eMBB and URLLC QoS classes have to be efficiently multiplexed on the same spectrum. Such requirement implies that the QoS objective functions of the MAC scheduler should be user-centric, instead of network-centric [14]. However, these QoS classes are highly correlated and need to be reliably satisfied, e.g., eMBB SE maximization and URLLC latency minimization as

$$\forall k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}} : R_{\text{mbb}} = \underset{k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}}}{\arg\max} \sum_{k_{\text{mbb}}=1}^{K_{\text{mbb}}} \sum_{rb \in \Xi_{k_{\text{mbb}}}^{\text{mbb}}} \beta_{k_{\text{mbb}}} r_{k_{\text{mbb}},rb}^{\text{mbb}}, \tag{D.4}$$

$$\forall k_{\text{llc}} \in \mathcal{K}_{\text{llc}} : \underset{k_{\text{llc}} \in \mathcal{K}_{\text{llc}}}{\arg\min} \left(\Psi_{k_{\text{llc}}} \right), \ \Psi_{k_{\text{llc}}} \leq 1\,\text{ms}, \tag{D.5}$$

where R_{mbb} is the overall eMBB ergodic capacity, $\mathcal{K}_{\text{mbb}}$ and $\mathcal{K}_{\text{llc}}$ are the active sets of eMBB and URLLC users, respectively. $\Xi_{k_{\text{mbb}}}^{\text{mbb}}$ and $\gamma_{k_{\text{mbb}}}$ are the set of granted PRBs and the scheduling priority of the k_{mbb}^{th} user, respectively, $r_{k_{\text{mbb}},rb}^{\text{mbb}}$ is the achievable per-PRB rate of the k_{mbb}^{th} user. Finally, $\Psi_{k_{\text{llc}}}$ denotes the URLLC one-way radio latency, which can be expressed as (assuming a successful first transmission):

$$\Psi_{k_{\text{llc}}} = \Lambda_{\text{q}} + \Lambda_{\text{bsp}} + \Lambda_{\text{fa}} + \Lambda_{\text{tx}} + \Lambda_{\text{uep}}, \tag{D.6}$$

where $\Lambda_{\text{q}}, \Lambda_{\text{bsp}}, \Lambda_{\text{fa}}, \Lambda_{\text{tx}}, \Lambda_{\text{uep}}$ are random variables to present the URLLC queuing, BS processing, frame alignment, transmission, and user processing delays, respectively. Due to the agile 5G-NR frame structure, Λ_{fa} is upper-bounded by a short TTI duration while Λ_{bsp} & Λ_{uep} are each bounded by 3-OFDM symbol duration [17], due to the enhanced processing capabilities that come with the 5G-NR. Thus, the URLLC queuing delay Λ_{q} and transmission delay Λ_{tx} are the major bottleneck against achieving the stringent URLLC latency targets. As reported in our recent studies [13, 14], these delay components are hardly controlled in a dynamic system, and highly correlated to each others. Furthermore, their statistical behavior vastly varies with

the URLLC arrival rate λ, packet size B, SINR level $\Upsilon^{\text{llc}}_{k_{\text{llc}}}$, and the scheduler buffering behavior.

To achieve the URLLC stringent latency and reliability requirements in eq. (D.5), Λ_{q} and Λ_{tx} must be always controlled at minimum to allow for further delay allowance for the re-transmission(s) within the target 1 ms. This can only be achieved by enforcing a hard URLLC priority in the scheduler queues, or allocating URLLC users with excessive PRB sizes to ensure a sufficient outage SINR level. In both cases, the eMBB utility function in eq. (D.4) is severely under-optimized, resulting in a significant degradation of the system ergodic capacity. In this work, we address such challenging multiplexing requirement and propose a scheduling framework that guarantees the URLLC QoS while significantly improving the system SE.

3.2 Proposed eNSBPS Scheduler – At The BS Side

During an arbitrary TTI, eNSBPS scheduler assigns single-user (SU) resources to new/buffered eMBB traffic, if there are no new URLLC arrivals, based on the proportional fair (PF) [18] criterion as

$$\Theta\left\{\text{PF}_{k_{\text{mbb}}}\right\} = \frac{r^{\text{mbb}}_{k_{\text{mbb}},rb}}{\bar{r}^{\text{mbb}}_{k_{\text{mbb}},rb}}, \tag{D.7}$$

$$k^{*}_{\text{mbb}} = \underset{k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}}}{\arg\max} \;\; \Theta\left\{\text{PF}_{k_{\text{mbb}}}\right\}, \tag{D.8}$$

where $\bar{r}^{\text{mbb}}_{k_{\text{mbb}},rb}$ is the mean delivered data rate of the k^{th}_{mbb} user. Though, if there are new/buffered URLLC packets at the BS queues, and the instant schedulable resources are sufficiently enough to accommodate such payloads, the eNSBPS scheduler overwrites the SU eMBB scheduling priority for the sake of the URLLC traffic, by applying the weighted PF (WPF) criterion as

$$\Theta\left\{\text{WPF}_{k_{\kappa}}\right\} = \frac{r^{\kappa}_{k_{\kappa},rb}}{\bar{r}^{\kappa}_{k_{\kappa},rb}}\gamma_{k_{\kappa}}, \tag{D.9}$$

with $\gamma_{k_{\text{llc}}} \gg \gamma_{k_{\text{mbb}}}$ for immediate URLLC SU scheduling. In case radio resource are not immediately sufficient for the incoming URLLC packets, the eNSBPS scheduler first attempts a highly conservative version of a standard MU-MIMO transmission between the URLLC-eMBB user pair. That is, users are only paired if their corresponding transmission subspaces offer high spatial separation [14] as

$$1 - \left| \left(\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}}\right)^{\text{H}} \mathbf{v}^{\text{llc}}_{k_{\text{llc}}} \right|^2 \geq \eta, \tag{D.10}$$

where $\eta \rightarrow [0,1]$ is a conservative orthogonality threshold. However, if such orthogonality can not be offered at an arbitrary TTI, due to limited SDoFs, the

proposed eNSBPS instantly enforces such orthogonality, for the sake of the URLLC traffic. It pre-defines a discrete Fourier transform spatial reference subspace, pointing towards an arbitrary direction θ as given by

$$\mathbf{v}_{\text{ref}}(\theta) = \left(\frac{1}{\sqrt{N_t}}\right)\left[1, e^{-j2\pi\Delta\cos\theta}, \ldots, e^{-j2\pi\Delta(N_t-1)\cos\theta}\right]^{\text{T}}, \tag{D.11}$$

where Δ is the antenna spacing. Then, scheduler instantly searches for an active eMBB whose transmission is most aligned within the reference subspace, using the minimum Chordal distance as

$$k^{\diamond}_{\text{mbb}} = \arg\min_{\mathcal{K}_{\text{mbb}}} \mathbf{d}\left(\mathbf{v}^{\text{mbb}}_{k}, \mathbf{v}_{\text{ref}}\right), \tag{D.12}$$

where the Chordal distance $\mathbf{d}$ between $\mathbf{v}^{\text{mbb}}_{k}$ and $\mathbf{v}_{\text{ref}}$ is expressed by

$$\mathbf{d}\left(\mathbf{v}^{\text{mbb}}_{k}, \mathbf{v}_{\text{ref}}\right) = \frac{1}{\sqrt{2}}\left\|\mathbf{v}^{\text{mbb}}_{k}\left(\mathbf{v}^{\text{mbb}}_{k}\right)^{\text{H}} - \mathbf{v}_{\text{ref}}\,\mathbf{v}^{\text{H}}_{\text{ref}}\right\|. \tag{D.13}$$

Finally, the eNSBPS scheduler instantly projects the selected eMBB transmission onto the reference subspace as:

$$\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}}\right)' = \frac{\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}}\cdot\,\mathbf{v}_{\text{ref}}}{\left\|\mathbf{v}_{\text{ref}}\right\|^{2}} \times \mathbf{v}_{\text{ref}}, \tag{D.14}$$

with $\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}}\right)'$ as the post-projection eMBB precoder. As shown in Fig. D.2, the victim eMBB transmission inflicts a loss in its principal direction and gain, respectively, due to the instant projection at the BS, as it will be discussed in greater detail in Section 3-D. Then, scheduler immediately allocates shared resources between the incoming URLLC user and the victim eMBB transmission. Finally, as depicted by the timing diagram in Fig. D.3, the BS signals the URLLC user by a single-bit true co-scheduling indication, i.e., $\alpha = 1$ in the control channel, for the URLLC user to de-orient its decoding matrix into one possible null space of the reference subspace, hence, experiencing no intra-cell inter-user interference. Furthermore, to recover the eMBB capacity region, being impacted by the instant spatial projection, the BS also signals the victim eMBB user with:

- $\alpha = 1$, AND
- Multi-bit separation angle $\Phi = \left|\angle\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}}\right) - \angle\left(\mathbf{v}_{\text{ref}}\right)\right|$ between its original principal precoder and the reference subspace, AND
- Timing information of the starting symbol when such spatial projection has been applied, AND/OR
- Multi-bit original precoder length $\beta = \left\|\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}}\right\|$.

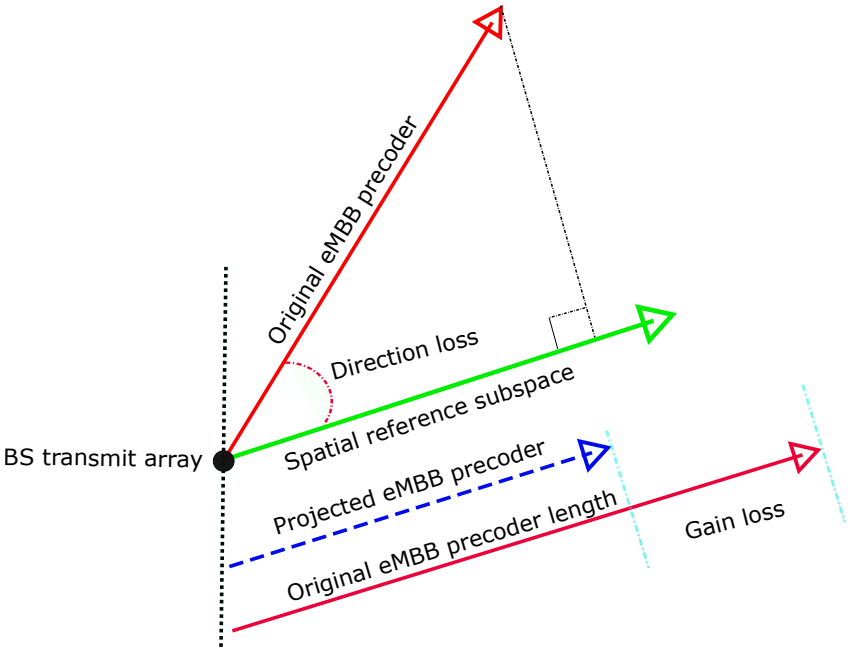


Fig. D.2. Victim eMBB transmission projection.

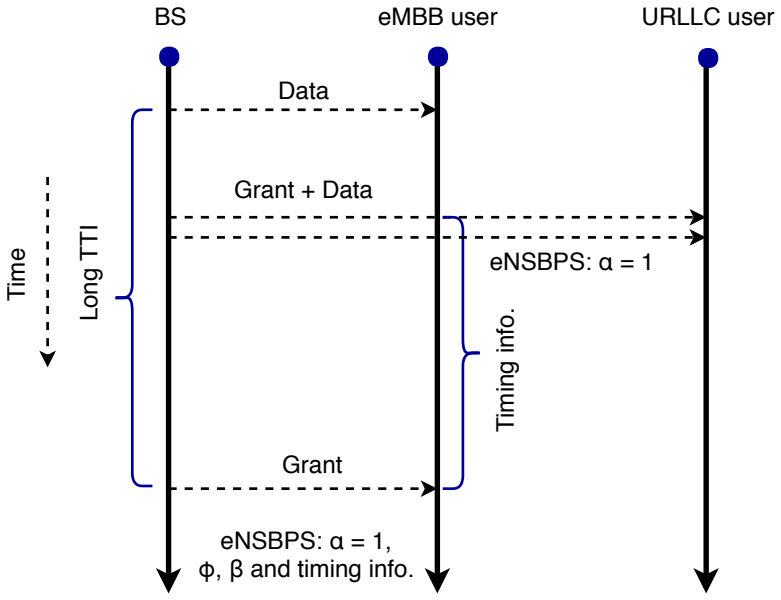


Fig. D.3. Timing diagram of the eNSBPS scheduler.

3.3 Proposed eNSBPS Scheduler – At The URLLC User Side

Upon the reception of a true co-scheduling indication $\alpha = 1$ in the control channel, the URLLC user realizes that its scheduling grant is shared with an active eMBB user, whose transmission is aligned within the pre-known reference subspace. The URLLC user first designs its first-stage LMMSE-IRC decoding matrix in order to reject the inter-cell interference statistics as

$$\left(\mathbf{u}_k^{\text{llc}}\right)^{(1)} = \left(\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}\left(\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}\right)^{\text{H}} + \mathbf{W}\right)^{-1} \mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}, \tag{D.15}$$

with the interference covariance matrix $\mathbf{W}$ given as

$$\mathbf{W} = \mathbb{E}\left(\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}\left(\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}\right)^{\text{H}}\right) + \sigma^2 \text{I}_{M_r}, \tag{D.16}$$

where I_{M_r} is $M_r \times M_r$ identity matrix. Then, $\left(\mathbf{u}_k^{\text{llc}}\right)^{(1)}$ is transferred into one possible null space of the inter-user interference effective channel $\mathbf{H}_k^{\text{llc}}\mathbf{v}_{\text{ref}}$, coming from the paired eMBB user and aligned within the reference subspace as

$$\left(\mathbf{u}_k^{\text{llc}}\right)^{(2)} = \left(\mathbf{u}_k^{\text{llc}}\right)^{(1)} - \frac{\left(\left(\mathbf{u}_k^{\text{llc}}\right)^{(1)} \cdot \mathbf{H}_k^{\text{llc}}\mathbf{v}_{\text{ref}}\right)}{\left\|\mathbf{H}_k^{\text{llc}}\mathbf{v}_{\text{ref}}\right\|^2} \times \mathbf{H}_k^{\text{llc}}\mathbf{v}_{\text{ref}}. \tag{D.17}$$

This way, the second-stage decoder $\left(\mathbf{u}_k^{\text{llc}}\right)^{(2)}$ matrix of the URLLC user experiences no inter-user interference, boosting its received SINR level.

3.4 Proposed eNSBPS Scheduler – At The eMBB User Side

At the eMBB user side, when $\alpha = 1$ is received, it acknowledges that its corresponding transmission is being spatially altered *on-the-fly* to be aligned within the reference subspace. Thus, it inflicts a spatial loss in its spatial gain and principal direction, respectively, e.g., as described in Fig. D.2 and eq. (D.14), the loss in the precoding spatial gain is given by

$$\left\|\left(\mathbf{v}_{k_{\text{mbb}}^{\diamond}}^{\text{mbb}}\right)'\right\| = \|\mathbf{v}_{\text{ref}}\| \left\|\mathbf{v}_{k_{\text{mbb}}^{\diamond}}^{\text{mbb}}\right\| \cos\left(\Phi\right), \tag{D.18}$$

where $\|\mathbf{v}_{\text{ref}}\| = 1$, and the original precoder spatial length $\left\|\mathbf{v}_{k_{\text{mbb}}^{\diamond}}^{\text{mbb}}\right\|$ exhibits a scale-down loss by $\cos\left(\Phi\right)$. Thus, we introduce two setups to recover the eMBB capacity with different signaling overhead as follows.

Setup-1: victim eMBB user attempts to reconstruct its original transmission subspace, that was altered at the BS by the instant spatial projection, and based on the knowledge of the reference subspace, Φ, and β, expressed as

$$\left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}}\right)^{\text{est.}} = \beta\ e^{-j\Phi}\ \mathbf{v}_{\text{ref}}, \tag{D.19}$$

where $\left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}}\right)^{\text{est.}}$ is the estimated original transmission subspace of the victim eMBB user. The first factor β compensates for the loss in the precoder spatial length; however, the second factor $e^{-j\Phi}$ cancels the spatial rotation effect. Then, the eMBB user projects its first-stage LMMSE-IRC decoding matrix $\left(\mathbf{u}_k^{\text{mbb}}\right)^{(1)}$ on its desired estimated effective transmission subspace $\mathbf{H}_k^{\text{mbb}}\left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}}\right)^{\text{est.}}$ as

$$\left(\mathbf{u}_k^{\text{mbb}}\right)^{(2)} = \frac{\left(\left(\mathbf{u}_k^{\text{mbb}}\right)^{(1)} .\ \mathbf{H}_k^{\text{mbb}}\left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}}\right)^{\text{est.}}\right)}{\left\|\mathbf{H}_k^{\text{mbb}}\left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}}\right)^{\text{est.}}\right\|^2} \times \mathbf{H}_k^{\text{mbb}}\left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}}\right)^{\text{est.}}, \tag{D.20}$$

with $\left(\mathbf{u}_k^{\text{mbb}}\right)^{(2)}$ as the second-stage eMBB decoder, that is de-oriented towards its original transmission subspace, thus, maximizing its achievable capacity.

Setup-2: based on the fact that both the length and direction loss of the victim eMBB user depend on the spatial separation angle between its original precoder and the reference subspace, i.e., spatial rotation of Φ, and spatial gain loss factor of $\cos(\Phi)$. Thus, the signaling overhead from the BS to the intended eMBB users can only be limited to Φ, without the need for signaling β. Accordingly, a spatial rotation matrix Γ is constructed and scaled-up by the length loss factor as

$$\Gamma = \left(\frac{1}{\cos(\Phi)}\right)\begin{bmatrix} \left(e^{-j\Phi}\right)_{0,0} & \cdots & \left(e^{-j\Phi}\right)_{0,d-1} \\ \vdots & \ddots & \vdots \\ \left(e^{-j\Phi}\right)_{M_r-1,0} & \cdots & \left(e^{-j\Phi}\right)_{M_r-1,d-1} \end{bmatrix}, \tag{D.21}$$

where d indicates the number of spatial streams per user. Finally, inline with setup-1, the victim eMBB user projects its fist-stage decoding matrix onto the spatial rotation matrix, given by

$$\left(\mathbf{u}_k^{\text{mbb}}\right)^{(2)} = \frac{\left(\left(\mathbf{u}_k^{\text{mbb}}\right)^{(1)} .\ \Gamma\right)}{\|\Gamma\|^2} \times \Gamma. \tag{D.22}$$

4 Simulation Results

In this section, we introduce the SLS results of the proposed eNSBPS scheduler, following the 5G-NR assumptions [5]. The major simulation parameters are listed in Table D.I.

Table D.1: Simulation Parameters.

Parameter	Value
Environment	3GPP-UMA,7 gNBs, 21 cells, 500 meters inter-site distance
Channel bandwidth	10 MHz, FDD
Antenna setup	BS: 8 Tx, UE: 2 Rx
User dropping	uniformly distributed URLLC: 7 users/cell eMBB: 7 users/cell
User receiver	LMMSE-IRC
TTI configuration	URLLC: 0.143 ms (2 OFDM symbols) eMBB: 1 ms (14 OFDM symbols)
CQI	periodicity: 5 ms, with 2 ms latency
HARQ	asynchronous HARQ, Chase combining HARQ round trip time = 4 TTIs
Link adaptation	dynamic modulation and coding target URLLC BLER : 1% target eMBB BLER : 10%
Traffic model	URLLC: bursty, B=50 bytes, $\lambda = 250$ eMBB: full buffer

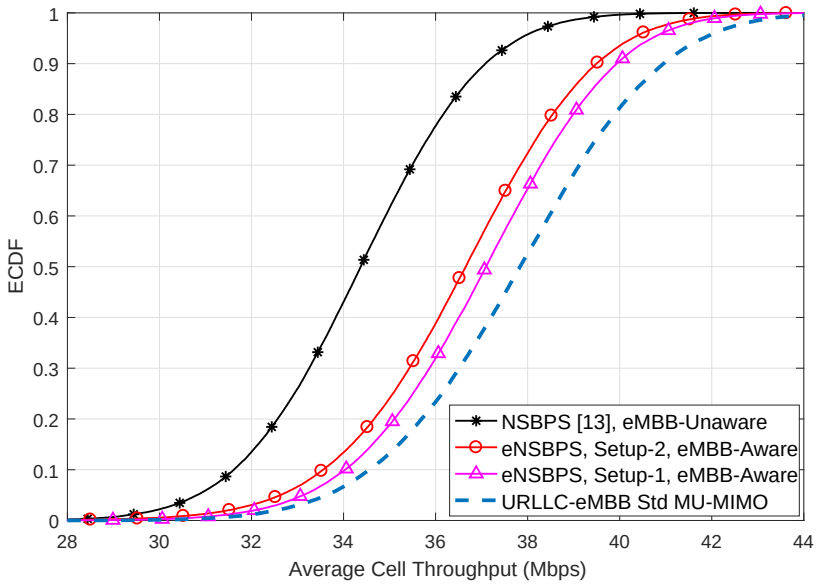


Fig. D.4. Average cell throughput performance (Mbps).

We present a performance comparison of three state-of-the-art schedulers for joint eMBB and URLLC traffic as follows: (1) proposed eNSBPS scheduler with the two techniques to recover the impacted eMBB capacity, (2) our recent NSBPS scheduler [13], where the victim eMBB users are presumed unaware of the spatial projection, hence, a degraded eMBB capacity is exhibited, and (3) a standard (Std) MU-MIMO scheduler between incoming URLLC users and ongoing eMBB transmissions, if the instantly available resources are not sufficient to accommodate the entire URLLC payload.

Fig. D.4 presents the empirical cumulative distribution function (ECDF) of the average achievable cell throughput in Mbps of all three schedulers under evaluation. As can be noticed, the Std URLLC-eMBB MU-MIMO scheduler offers the maximum possible cell throughput since the eMBB transmissions are not biasedly altered for the sake of the URLLC traffic; however, with the worst URLLC latency performance as will be shown in Fig. D.5. The proposed eNSBPS scheduler with the two introduced eMBB recovery techniques, significantly improves the achievable cell throughput against the eMBB-unaware NSBPS scheduler. It approaches the Std MU-MIMO scheduler, while simultaneously preserving the URLLC latency targets. This is because the intentionally lost eMBB capacity at the BS is recovered at the victim eMBB users using BS control signaling. Both setup-1 and setup-2 of the proposed eNSBPS scheduler show a similar cell throughput performance, with further reduced signaling overhead for setup-2, since both the spatial length and direction losses of the victim eMBB users only depend on the separation angle between the eMBB original precoder and the reference subspace at the BS.

Examining the URLLC performance, Fig. D.5 depicts the complementary CDF (CCDF) of the URLLC one-way latency in ms. As can be clearly identified, both proposed eNSBPS and NSBPS schedulers achieve the stringent URLLC latency target of 1 ms at 10^{-5} outage, since under both schedulers, sporadic URLLC traffic is guaranteed an instant and interference-free spatial subspace, hence, improving the URLLC decoding ability and reducing the number of inflicted URLLC re-transmissions. Furthermore, due to the fact that the Std MU-MIMO pairing condition is only constrained by the achievable sum rate, i.e., not a user-centric constraint, a Std URLLC-eMBB MU-MIMO transmission degrades the URLLC decoding SINR level. Additionally, a Std MU-MIMO pairing is not almost surely guaranteed, e.g., if the SDoFs are limited during an arbitrary TTI, MU pairing may not be possible, hence, the incoming URLLC traffic must be queued for multiple TTIs until sufficient radio resources are released. As a result, the Std URLLC-eMBB MU-MIMO scheduler exhibits a significant loss of the URLLC latency performance, not fulfilling its latency targets.

Finally, looking at the individual eMBB performance, Fig. D.6 presents the ECDF of the eMBB user post-detection carrier-to-interference-ratio (CIR) in

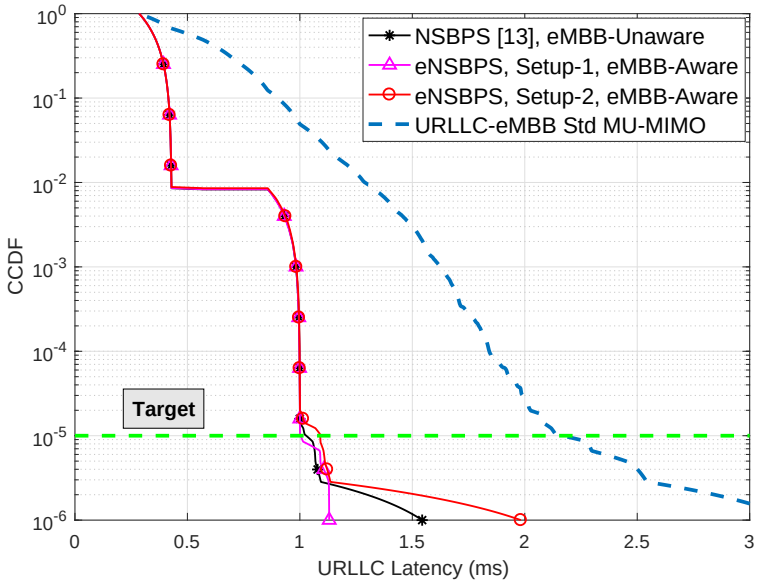


Fig. D.5. URLLC one-way latency performance (ms).

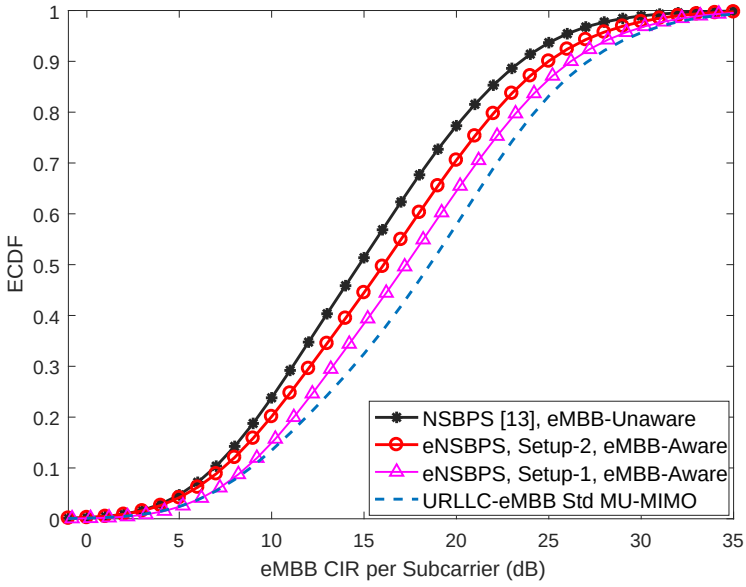


Fig. D.6. Average eMBB CIR per sub-carrier performance (dB).

dB. The Std MU-MIMO scheduler offers the best eMBB CIR since the paired eMBB users are only impacted by the standard MU equal-power sharing and the resultant inter-user interference. That is, eMBB transmissions are not spatially altered for the sake of the paired URLLC traffic, leading to a better cell performance as shown in Fig. D.4. On another side, NSBPS scheduler suffers from the worst eMBB CIR due to the unrealizable eMBB projections. Hence, victim eMBB users exhibit a sub-optimal LMMSE-IRC performance since both the actual and estimated eMBB effective channels are not aligned within the same signal subspace. The proposed eNSBPS, under the two introduced recovery setups, provides a clear enhancement of the end eMBB CIR performance. The eMBB recovery mechanisms of the eNSBPS scheduler re-align the LMMSE-IRC decoding spatial span of the victim eMBB users into its original signal subspace before the inflicted projection at the BS, thus, maximizing their perceived effective channels and SNR levels, respectively.

5 Conclusion

In this work, an enhanced null space based preemptive scheduler (eNSBPS) has been introduced for joint URLLC and eMBB traffic in 5G new radio. Sporadic URLLC traffic is instantly guaranteed an interference-free subspace for immediate and secured transmission without queuing, through eMBB subspace projection. Thus, proposed eNSBPS scheduler offers extreme URLLC latency robustness. The impacted eMBB capacity is then recovered through subspace alignment at the victim eMBB users, hence, maximizing the achievable eMBB capacity. Compared to the state-of-the-art scheduling proposals, extensive system level simulations show that proposed scheduler framework satisfies the stringent URLLC latency targets while significantly improving the overall cell spectral efficiency, by achieving an average gain of $\sim$ 3.2 dB in the eMBB post-detection carrier-to-interference-ratio.

6 Acknowledgments

This work is partly funded by the Innovation Fund Denmark (IFD) – case number: 7038-00009B. Also, 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] Study on new radio access technology (Release 14), 3GPP, TR 38.801, V14.0.0, March 2017.

[2] NR and NG-RAN overall description; Stage-2 (Release 15), 3GPP, TS 38.300, V2.0.0, Dec. 2017.

[3] IMT vision – "Framework and overall objectives of the future development of IMT for 2020 and beyond", international telecommunication union (ITU), ITU-R M.2083-0, Feb. 2015.

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

[5] Ali A. Esswie, and K.I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5G networks," *in Proc. IEEE ISCC*, Natal, 2018, pp. 1-6.

[6] B. Soret, P. Mogensen, K. I. Pedersen and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," *in Proc. IEEE Globecom*, Austin, TX, 2014, pp. 1391-1396.

[7] Study on enhancement of URLLC supporting in 5GC (Work item: release-16); 3GPP, TR *to be specified*, March 2018.

[8] K. Pedersen, G. Pocovi, J. Steiner and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 210-217, Mar. 2018.

[9] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner and Z. Li, "Achieving URLLC: challenges and envisioned system enhancements," *IEEE Netw.*, vol. 32, no. 2, pp. 8-15, April 2018.

[10] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen and P. Mogensen, "Signal quality outage analysis for URLLC in cellular networks," *in Proc. IEEE Globecom*, San Diego, CA, 2015, pp. 1-6.

[11] J. J. Nielsen, R. Liu and P. Popovski, "URLLC using interface diversity," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1322-1334, March 2018.

[12] K. I. Pedersen, G. Pocovi, and J. Steiner, "Preemptive scheduling of latency critical traffic and its impact on mobile broadband performance," *in Proc. VTC,* Porto, 2018, pp. 1-6.

[13] Ali A. Esswie, and K.I. Pedersen, “Null space based preemptive scheduling for joint URLLC and eMBB traffic in 5G networks,” *submitted to IEEE Globecom*, Abu Dhabi, 2018.

[14] Ali A. Esswie, and K.I. Pedersen, “Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks,” *submitted to IEEE Netw.*, June, 2018.

[15] Study on 3D channel model for LTE; Release 12, 3GPP, TR 36.873, V12.7.0, Dec. 2014.

[16] Y. Ohwatari, N. Miki, Y. Sagae and Y. Okumura, "Investigation on interference rejection combining receiver for space–frequency block code transmit diversity in LTE-advanced downlink," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 191-203, Jan. 2014.

[17] Physical layer procedures; Evolved universal terrestrial radio access (Release 15), 3GPP, TS 36.213, V15.1.0, March. 2018.

[18] D. Parruca and J. Gross, "Throughput analysis of proportional fair scheduling for sparse and ultra-dense interference-limited OFDMA/LTE networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6857-6870, Oct. 2016.

Paper E

Preemption-Aware Rank Offloading Scheduling For Latency Critical Communications in 5G Networks

A. A. Esswie, K. I. Pedersen and P. E. Mogensen

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

Abstract

This paper introduces a preemptive rank offloading scheduling framework for joint ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB) traffic in 5G new radio (NR). Proposed scheduler dynamically adapts the overall system optimization among the network-centric ergodic capacity and the user-centric URLLC one-way latency, based on the instantaneous traffic and radio resources availability. The spatial degrees of freedom, offered by the transmit antenna array, are fully exploited to maximize the overall spectral efficiency. However, when URLLC traffic buffering is foreseen, proposed scheduler immediately enforces scheduling pending URLLC payloads through preemption-aware subspace projection. Compared to the state-of-the-art schedulers from industry and academia, proposed scheduler framework shows significant scheduling flexibility in terms of the overall ergodic capacity and URLLC latency performance. The presented results therefore offer valuable insights of how to most efficiently multiplex joint URLLC-eMBB traffic over the 5G NR spectrum.

Index Terms— URLLC; eMBB; 5G; MU-MIMO; New radio; Preemptive; Scheduling.

1 Introduction

The coexistence of conventional human-centric and future machine-centric communications introduces more complex wireless environments [1, 2]. To address such diversified requirements, the standardization of the fifth generation new radio (5G-NR) is readily advancing, with its first specifications issued recently [3, 4]. 5G-NR features two major service classes: ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB). URLLC services require stringent latency and reliability targets, i.e., up to one-way radio latency of 1 ms with 10^{-5} outage probability while eMBB applications seek for broadband data rates [5].

The efficient multiplexing of such diverse quality of service (QoS) classes over a single radio spectrum is a challenging and non-trivial scheduling problem, due to the underlying trade-off between latency, reliability, and aggregated data rate [6]. That is, if the system is forcibly engineered to satisfy the URLLC per-user outage of interest, the eMBB spectral efficiency (SE) will be severely degraded due to the inefficient resource utilization.

Recently, the URLLC and eMBB multiplexing problem has gained growing research attention from academia and industry. Primarily, the variable transmission time interval (TTI) duration with small data payloads is of significant importance to achieve the URLLC targets; however, at the expense of additional signaling overhead [7]. Spatial diversity techniques and dual connectivity [5] are also proved beneficial to improve the URLLC decod-

ing ability by preserving the minimum outage signal-to-interference-noise-ratio (SINR). Furthermore, puncturing scheduler (PS) [8] is a state-of-the-art scheduling technique for joint URLLC-eMBB traffic, where the URLLC scheduling queuing delay becomes independent from the eMBB offered load through disruptive URLLC transmissions over eMBB-monopolized resources.

In our recent study [9], we demonstrated that a standard multi-user multi-input multi-output (MU-MIMO) transmission between URLLC-eMBB pairs is a fair solution to trade-off URLLC latency with overall SE. However, when the system spatial degrees of freedom (SDoFs) are limited, significant URLLC queuing delays are observed since a standard MU-MIMO pairing is only constrained by the achievable sum rate. Hence, in [10], we proposed a biased, and non-transparent version of the standard URLLC-eMBB MU-MIMO to guarantee an immediate and interference-free URLLC scheduling, regardless of the instantaneous system SDoFs and user loading. Thus, the URLLC latency budget is always preserved.

Compared to recent URLLC scheduler proposals, the scheduler operation is monotonically dictated by the URLLC capacity of interest. Examples include URLLC resource pre-allocation, and immediate puncturing. Thus, when URLLC services are multiplexed with eMBB applications on the same spectrum, the maximum system SE becomes infeasible. Needless to say, a multi-QoS-aware scheduling framework, which flexibly adapts the scheduling objectives to the instantaneous traffic state and being able to instantly preempt a particular QoS enforcement, is vital for future 5G-NR use cases.

In this work, we propose a preemption-aware rank offloading scheduling (PAROS) for joint URLLC and eMBB traffic. The proposed scheduler is a multi-objective framework, where both eMBB and URLLC QoS classes are simultaneously optimized on the TTI-level. Proposed PAROS scheduler first targets achieving the maximum possible ergodic capacity by attempting greedy MU eMBB transmissions. However, in case URLLC buffering is foreseen, hence, exceeding the critical URLLC latency budget, the PAROS scheduler enforces an instant subspace-projection for an interference-free URLLC scheduling over shared resources with ongoing eMBB transmissions. If the instantly available SDoFs are limited, the PAROS scheduler enforces an instant SDoF-relaxation through rank offloading, sufficient enough to immediately accommodate the incoming URLLC traffic. Hence, proposed scheduler shows great multiplexing flexibility in terms of the overall ergodic capacity and URLLC latency & reliability targets.

Due to the complexity of the the 5G-NR scheduling problem [3] and addressed issues herein, we assess the performance of the proposed solution using extensive system level simulations, where the major scheduling functionalities are calibrated against the 3GPP 5G-NR assumptions. This includes the 3D channel spatial modeling, dynamic link adaptation, hybrid automatic repeat request (HARQ), dynamic multi-traffic modeling, SINR combining,

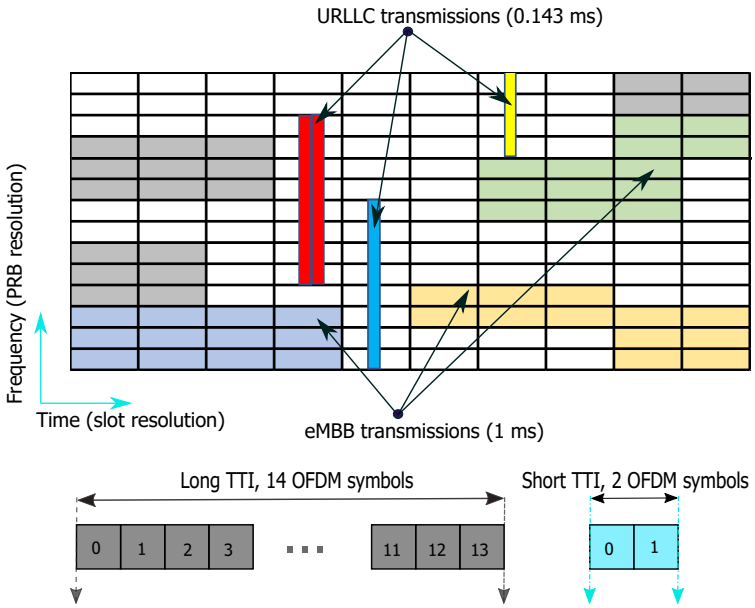


Fig. E.1. Agile 5G-NR frame design and resource allocation.

and dynamic user scheduling.

This paper is organized as follows. System model is presented in Section 2. The proposed scheduler framework is introduced in Section 3 while Section 4 shows the numerical results. Finally, the paper is concluded in Section 5.

2 System Model

We adopt a 5G-NR system with C downlink (DL) base-stations (BSs), each equipped with N_t transmit antennas. Each BS serves an average K uniformly distributed user equipment's (UEs), each with M_r receive antennas and $K = K_{\text{llc}} + K_{\text{mbb}}$, with K_{llc} and K_{mbb} as the average numbers of the URLLC and eMBB UEs per cell. Thus, the average cell loading condition per BS is defined by $\Omega = (K_{\text{mbb}}, K_{\text{llc}})$. The URLLC traffic is characterized by the FTP3 traffic model with a finite B-byte payload size and a Poisson point arrival process λ, while eMBB traffic is full buffer with infinite payload, to offer all-time best effort background load.

The 5G-NR flexible frame design is assumed. As depicted in Fig. E.1, in the time domain, the URLLC traffic is scheduled over short TTI durations of 2-OFDM symbol mini slots, to satisfy its stringent latency budget. The eMBB traffic is scheduled over longer TTI durations of 14-OFDM symbol slots, to

maximize the overall ergodic capacity. Furthermore, in line with [5], the scheduling grant is appended prior to the radio resources of the data payloads, thus, the minimum resource allocation per UE should be sufficiently large to accommodate both data and control symbols. In the frequency domain, the UEs are dynamically multiplexed by the orthogonal frequency division multiple access, where the smallest scheduling unit is the physical resource block (PRB) of 12-subcarriers.

We further assume a throughput-greedy scheduler with controlled, biased and non-transparent MU-MIMO transmissions, where a subset of co-scheduled UEs $\mathtt{G}_c \subseteq \mathcal{K}_c$ is allowed over an arbitrary PRB, where $\mathcal{K}_c$ is the active UE set in the c^{th} cell, $G_c = \mathbf{card}(\mathtt{G}_c)$, $G_c \leq N_t$ is the actual number of co-scheduled UEs and $\mathbf{card}(\cdot)$ indicates the cardinality. The post-decoded DL signal at the k^{th} UE from the c^{th} cell is given by

$$\hat{s}^{\kappa}_{k,c} = \left(\mathbf{u}^{\kappa}_{k,c}\right)^{\mathrm{H}} \mathbf{H}_{k,c}\mathbf{v}^{\kappa}_{k,c}s_{k,c} + \sum_{j=1,j\neq c}^{C} \sum_{g\in\mathtt{G}_j} \left(\mathbf{u}^{\kappa}_{k,c}\right)^{\mathrm{H}} \mathbf{H}_{k,j}\mathbf{v}_{g,j}s_{g,j}$$

$$+ \begin{cases} \sum\limits_{g\in\mathtt{G}_c, g\neq k} \left(\mathbf{u}^{\kappa}_{k,c}\right)^{\mathrm{H}} \mathbf{H}_{k,c}\mathbf{v}^{\{\text{'llc'},\text{'mbb'}\}}_{g,c}s_{g,c}, & \kappa = \{\text{'mbb'}\} \\ \sim 0, & \kappa = \{\text{'llc'}\} \end{cases} + \mathbf{n}^{\kappa}_{k,c}, \tag{E.1}$$

where $\mathcal{X}^{\kappa}, \kappa\in\{\text{'llc'}, \text{'mbb'}\}$ denotes the QoS type requested by UE $\mathcal{X}$, $\mathbf{H}_{k,c} \in \mathcal{C}^{M_r\times N_t}, \forall k \in \{1,\ldots,K\}, \forall c \in \{1,\ldots,C\}$ follows the 3GPP 3D spatial channel [11] from the c^{th} cell to the k^{th} UE, $\mathbf{v}_{k,c} \in \mathcal{C}^{N_t\times 1}$ is the standard zero-forcing precoding vector, assuming a single stream transmission, and is expressed by

$$\mathbf{v}_{k,c} = \left(\mathbf{H}_{k,c}\right)^{\mathrm{H}} \left(\mathbf{H}_{k,c}\left(\mathbf{H}_{k,c}\right)^{\mathrm{H}}\right)^{-1}. \tag{E.2}$$

$s^{\kappa}_{k,c}$, $\hat{s}^{\kappa}_{k,c}$, and $\mathbf{n}^{\kappa}_{k,c} \in \mathcal{C}^{M_r\times 1}$ are the transmitted symbol, decoded symbol and the additive white Gaussian noise, respectively, while $\mathbf{u}^{\kappa}_{k,c}$ is the corresponding linear minimum mean square error interference rejection and combining (LMMSE-IRC) receiver matrix [5], with $(\cdot)^{\mathrm{H}}$ as the Hermitian operation. The first summation in eq. (E.1) models the inter-cell inter-user interference, resulting from either URLLC or eMBB traffic while the second summation represents the intra-cell inter-user interference resulting from the overloaded MU-MIMO transmissions. As will be discussed in Section 3, the URLLC-eMBB MU pairing is biased and altered such that URLLC traffic experiences no inter-user interference, hence, fulfilling its latency and reliability limits.

3 Proposed PAROS Scheduler

3.1 Problem Formulation

Multiplexing of the URLLC and eMBB QoS classes over the same radio spectrum implies a hard scheduling problem. URLLC QoS class must satisfy its outage of interest while eMBB QoS shall align with the network-wide outage. In that sequel, there is a trade-off between the user-centric URLLC and the network-centric eMBB targets. These are highly coupled and must be simultaneously optimized, i.e., eMBB rate maximization, and URLLC latency minimization as

$$\forall k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}} : R_{\text{mbb}} = \underset{k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}}}{\arg\max} \sum_{k_{\text{mbb}}=1}^{K_{\text{mbb}}} \sum_{rb \in \Xi^{\text{mbb}}_{k_{\text{mbb}}}} \beta_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb}, \tag{E.3}$$

$$\forall k_{\text{llc}} \in \mathcal{K}_{\text{llc}} : \underset{k_{\text{llc}} \in \mathcal{K}_{\text{llc}}}{\arg\min} \left(\Psi_{k_{\text{llc}}} \right), \tag{E.4}$$

where $\forall k_{\text{mbb}} \in \{1, \ldots, K_{\text{mbb}}\}, \forall k_{\text{llc}} \in \{1, \ldots, K_{\text{llc}}\}$, R_{mbb} is the overall eMBB ergodic capacity, $\mathcal{K}_{\text{mbb}}$ and $\mathcal{K}_{\text{llc}}$ are the active UE sets of eMBB and URLLC QoS classes, respectively, $\Xi^{\text{mbb}}_{k_{\text{mbb}}}$ and $\beta_{k_{\text{mbb}}}$ imply the allocated set of PRBs and the scheduling priority of the k^{th} eMBB user. $r^{\text{mbb}}_{k_{\text{mbb}},rb}$ is the achievable k^{th} eMBB UE rate per PRB and $\Psi_{k_{\text{llc}}}$ is defined as the URLLC radio latency, as

$$\Psi_{k_{\text{llc}}} = \Lambda_{\text{q}} + \Lambda_{\text{bsp}} + \Lambda_{\text{fa}} + \Lambda_{\text{tx}} + \Lambda_{\text{uep}} + \Lambda_{\text{harq}}, \tag{E.5}$$

where $\Lambda_{\text{q}}, \Lambda_{\text{bsp}}, \Lambda_{\text{fa}}, \Lambda_{\text{tx}}, \Lambda_{\text{uep}}$ and Λ_{harq} are random variables to represent the queuing, BS processing, frame alignment, transmission, UE processing, and HARQ re-transmission delays, respectively. Λ_{fa} is upper bounded by the short TTI duration due to the agile 5G-NR frame structure, while the standardization bodies agreed that Λ_{bsp} and Λ_{uep} are each bounded by 3-OFDM symbol duration [5], because of the enhanced processing capabilities that come with the 5G-NR. Therefore, Λ_{tx}, Λ_{q} and Λ_{harq} are the major delay sources against achieving the URLLC latency deadline.

Therefore, to guarantee the URLLC radio latency limit, the URLLC traffic must fulfill: 1) not being buffered/queued over many TTI instances at the BS scheduler, and 2) one-shot transmissions without segmentation, to further allow for additional Λ_{harq} delay within the 1 ms deadline. This can be achieved by allocating excessive bandwidth for URLLC traffic, and enforcing a hard-coded URLLC higher priority in the scheduling buffers. As a result, the eMBB utility in (E.3) will be severely under-optimized, leading to a significant degradation of the overall SE. In that sequel, we address such multiplexing problem by proposing an efficient and flexibly adaptive scheduling framework.

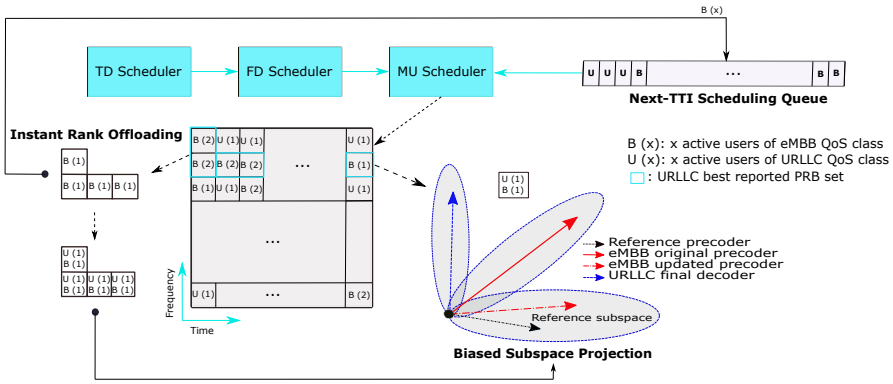


Fig. E.2. Illustration example of the proposed PAROS scheduling framework with $G_c = 2$.

3.2 Proposed Multi-Traffic PAROS Scheduler

The proposed scheduler dynamically alternates the scheduling targets in time such that the network ergodic capacity is maximized at all times by attempting greedy eMBB-eMBB MU-MIMO transmissions. When URLLC traffic buffering is foreseen, i.e., URLLC payload could not get scheduled from the time and frequency domain (TD, FD) schedulers, the proposed scheduler utilizes all system available SDoFs to instantly schedule these URLLC payloads over shared resources with transmitting eMBB UE through interference-free subspace projection based pairing. If the system PRBs are overloaded by eMBB MU transmissions, i.e., the maximum allowed number of per-PRB active users G_c is reached, PAROS scheduler immediately enforces eMBB UE offloading to reach $G_c - 1$ active UEs on the best reported PRBs of these incoming URLLC UEs. Fig. E.2 shows an example of the proposed PAROS scheduler with $G_c = 2$.

At the BS – Time and frequency domain schedulers:

During an arbitrary TTI, if there is no sporadic URLLC traffic, PAROS framework allocates single-user (SU), i.e., rank-1, dedicated resources to newly arrived and/or buffered eMBB traffic, based on the standard proportional fair (PF) criterion over both TD and FD schedulers as

$$\Theta\left\{\mathrm{PF}_{k_{\mathrm{mbb}}}\right\} = \frac{r^{\mathrm{mbb}}_{k_{\mathrm{mbb}},rb}}{\bar{r}^{\mathrm{mbb}}_{k_{\mathrm{mbb}},rb}}, \tag{E.6}$$

$$k^{*}_{\mathrm{mbb}} = \underset{k_{\mathrm{mbb}} \in \mathcal{K}_{\mathrm{mbb}}}{\arg\max} \ \Theta\left\{\mathrm{PF}_{k_{\mathrm{mbb}}}\right\}, \tag{E.7}$$

where $\bar{r}^{\mathrm{mbb}}_{k_{\mathrm{mbb}},rb}$ is the average received rate of the k^{th}_{mbb} UE. If URLLC payloads are available in the TD scheduling buffers, PAROS scheduler instantly overpowers the eMBB TD scheduling priority by the weighted PF criterion

as: $\Theta\left\{\mathrm{WPF}_{k_\kappa}\right\} = \frac{r^{\kappa}_{k,rb}}{\bar{r}^{\kappa}_{k,rb}}\beta_{k_\kappa}$, with $\beta_{k_{\mathrm{llc}}} \gg \beta_{k_{\mathrm{mbb}}}$ for instant URLLC scheduling. Then, the non-biased PF criterion is still applied on the FD scheduler to preserve fairness across the radio PRBs.

At the BS – Multi-user scheduler:

The PAROS scheduler aims to maximize the overall SE by default. Thus, at the MU scheduler, it always attempts greedy eMBB-to-eMBB MU transmissions, where G_c eMBB UEs are co-scheduled on an active PRB if the achievable sum rate is larger than that is of the primary eMBB UE only. In that sequel, the system PRBs are fully utilized with eMBB MU transmissions.

However, under high offered cell load, the schedulable resources may not be instantly available for critical URLLC traffic. Thus, TD and FD schedulers fail to immediately schedule such traffic and it will be queued in the MU scheduling buffers. Then, PAROS first attempts a highly conservative MU transmission between a primary eMBB and secondary URLLC UE pair if their corresponding transmissions satisfy:

$$1 - \left| \left(\mathbf{v}^{\mathrm{mbb}}_{k_{\mathrm{mbb}}}\right)^{\mathrm{H}} \mathbf{v}^{\mathrm{llc}}_{k_{\mathrm{llc}}} \right|^2 \geq \gamma. \tag{E.8}$$

The highly conservative, i.e., large, orthogonality threshold γ is enforced to protect the URLLC traffic against potential inter-user interference from the co-scheduled eMBB UE. If such orthogonality can not be offered at the current TTI, due to limited SDoFs, URLLC traffic shall be queued. Under this scheduling state, PAROS instantly alters the system optimization towards the URLLC latency and reliability targets instead of the ergodic capacity by satisfying the following conditions:

$$\mathbf{rank}\left\{ \left(\mathbf{u}^{\mathrm{llc}}_{k_{\mathrm{llc}}}\right)^{\mathrm{H}} \mathbf{H}_{k_{\mathrm{llc}}} \mathbf{v}^{\mathrm{llc}}_{k_{\mathrm{llc}}} \right\} \sim \text{full}. \tag{E.9}$$

$$\mathbf{rank}\left\{ \left(\mathbf{u}^{\mathrm{llc}}_{k_{\mathrm{llc}}}\right)^{\mathrm{H}} \mathbf{H}_{k_{\mathrm{llc}}} \mathbf{v}^{\mathrm{mbb}}_{k_{\mathrm{mbb}}} \right\} \sim 0. \tag{E.10}$$

Hence, PAROS scheduler instantly applies a biased and user-centric URLLC-eMBB MU transmission for interference-free URLLC scheduling, through subspace projection over the best reported URLLC PRBs with less than G_c active UEs. If such requested PRBs are overloaded with G_c eMBB active UEs, PAROS instantly offloads the eMBB UEs with the lowest achievable rates to preemptively free some SDoFs for URLLC traffic, i.e., it offloads PRBs with MU rank = G_c eMBB UEs down to $G_c - 1$ and biasedly pairs the incoming URLLC UE over these PRBs. Suspended eMBB transmissions are placed in the scheduling buffers according to their respective PF metrics. Furthermore, BS signals these eMBB UEs with a single-bit transmission interruption indication, for them to be aware that prior DL grant is not currently valid.

Towards such biased URLLC-eMBB pairing over an arbitrary PRB, a spatial reference subspace is predefined using the beamformed discrete Fourier transform, pointing to an arbitrary spatial direction θ, given by

$$\mathbf{v}_{\text{ref}}(\theta) = \left(\frac{1}{\sqrt{N_t}}\right)\left[1, e^{-j2\pi\Delta\cos\theta}, \ldots, e^{-j2\pi\Delta(N_t-1)\cos\theta}\right]^{\text{T}}, \tag{E.11}$$

where $(\cdot)^{\text{T}}$ implies the transpose operation and Δ is the antenna inter-distance. Then, PAROS scheduler searches for the active PRBs, from within the best reported PRB set of the incoming URLLC UEs, with at maximum $G_c - 1$ eMBB active UEs and whose active transmissions are closest possible in the spatial domain to the reference subspace as

$$k^{\diamond}_{\text{mbb}} = \arg\min_{\mathcal{K}_{\text{mbb}}} \mathbf{d}\left(\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}}, \mathbf{v}_{\text{ref}}\right), \tag{E.12}$$

where the Chordal distance between $\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}}$ and $\mathbf{v}_{\text{ref}}$ is given by

$$\mathbf{d}\left(\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}}, \mathbf{v}_{\text{ref}}\right) = \frac{1}{\sqrt{2}}\left\|\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}}\left(\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}}\right)^{\text{H}} - \mathbf{v}_{\text{ref}}\mathbf{v}^{\text{H}}_{\text{ref}}\right\|. \tag{E.13}$$

Finally, PAROS spatially projects the transmission of each victim eMBB UE $\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}}$ over selected PRBs onto $\mathbf{v}_{\text{ref}}$ as

$$\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)' = \frac{\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}} \cdot \mathbf{v}_{\text{ref}}}{\|\mathbf{v}_{\text{ref}}\|^2} \times \mathbf{v}_{\text{ref}}, \tag{E.14}$$

where $\mathcal{X} \cdot \mathcal{Y}$ indicates the dot product of $\mathcal{X}$ and $\mathcal{Y}$ and $\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'$ is the post-projection precoder of the victim eMBB UE. Next, PAROS forcibly pairs incoming URLLC UEs over these shared resources with selected eMBB UEs. As the impacted eMBB UEs are not aware of the instant projection, eMBB capacity shall be degraded. However, due to the constraints in (8) and (12), the eMBB capacity is limited specially under high offered eMBB load, i.e., PAROS scheduler has a higher probability to fetch an eMBB UE whose transmission is originally aligned with the reference subspace, hence, the hard-coded spatial projection would not significantly degrade its achievable capacity. Furthermore, in our recent study [5], we have analytically determined that for a generic eMBB transmission, the loss function of the effective channel gain due to such spatial projection is scaled down by $\sin(\Phi)^2 \ll 1$, where Φ is the difference angle between pre-projection $\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}}$ and post-projection $\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}}\right)'$ transmissions, leading to a guaranteed minimum loss rate. The BS scheduler finally signals the intended URLLC UEs with a single-bit true indication $\alpha = 1$.

At the URLLC UE:

When a URLLC UE acknowledges $\alpha = 1$, it realizes that its DL grant is shared with an active eMBB UE and the corresponding interfering trans-

mission is aligned within the reference subspace. Thus, it first designs its first-stage LMMSE-IRC standard decoding matrix as expressed by

$$\left(\mathbf{u}_{k_{\text{llc}}}^{\text{llc}}\right)^{(1)} = \left(\mathbf{H}_{k_{\text{llc}}}\mathbf{v}_{k_{\text{llc}}}^{\text{llc}}\left(\mathbf{H}_{k_{\text{llc}}}\mathbf{v}_{k_{\text{llc}}}^{\text{llc}}\right)^{\text{H}} + \mathbf{W}\right)^{-1}\mathbf{H}_{k_{\text{llc}}}\mathbf{v}_{k_{\text{llc}}}^{\text{llc}}, \tag{E.15}$$

where $(\cdot)^{-1}$ stands for the inverse operation, and the interference covariance matrix $\mathbf{W}$ is given as

$$\mathbf{W} = \mathbb{E}\left\{\mathbf{H}_{k_{\text{llc}}}\mathbf{v}_{k_{\text{llc}}}^{\text{llc}}\left(\mathbf{H}_{k_{\text{llc}}}\mathbf{v}_{k_{\text{llc}}}^{\text{llc}}\right)^{\text{H}}\right\} + \sigma^2\text{I}_{M_r}, \tag{E.16}$$

where $\mathbb{E}\{\cdot\}$ is the statistical expectation, σ^2 is the estimation error variance, and I_{M_r} denotes an identity matrix of size $M_r \times M_r$. Then, the URLLC UE intentionally transfers the statistics of $\left(\mathbf{u}_{k_{\text{llc}}}^{\text{llc}}\right)^{(1)}$ to a possible null space of the inter-user interference effective channel $\mathbf{H}_{k_{\text{llc}}}\mathbf{v}_{\text{ref}}$ as

$$\left(\mathbf{u}_{k_{\text{llc}}}^{\text{llc}}\right)^{(2)} = \left(\mathbf{u}_{k_{\text{llc}}}^{\text{llc}}\right)^{(1)} - \frac{\left(\left(\mathbf{u}_{k_{\text{llc}}}^{\text{llc}}\right)^{(1)} \cdot \mathbf{H}_{k_{\text{llc}}}\mathbf{v}_{\text{ref}}\right)}{\left\|\mathbf{H}_{k_{\text{llc}}}\mathbf{v}_{\text{ref}}\right\|^2} \times \mathbf{H}_{k_{\text{llc}}}\mathbf{v}_{\text{ref}}. \tag{E.17}$$

Hence, the final URLLC decoding matrix $\left(\mathbf{u}_{k_{\text{llc}}}^{\text{llc}}\right)^{(2)}$ shall experience an interference-free transmission, leading to an improved URLLC decoding ability.

3.3 Comparsion to the state of the art URLLC schedulers

In this sub-section, we introduce the state-of-the-art scheduling proposals from both industry and academia, to which we compare the performance of proposed PAROS against.

Null space based preemptive scheduler (NSBPS) [10]: in our previous contribution, we proposed a monotonic scheduling optimization such that when URLLC queuing is inevitable, the MU scheduler enforces a special URLLC-eMBB MU transmission, biased for the sake of the URLLC UEs. Hence, URLLC buffering is further minimized. However, eMBB-eMBB MU transmissions are not allowed to preserve the maximum possible SDoFs for incoming URLLC traffic.

Throughput-greedy NSBPS (TG-NSBPS): an extension of the NSBPS scheduler such that the scheduler always aims to maximizing the overall SE by attempting greedy eMBB-eMBB MU transmissions. When URLLC traffic is about to be buffered, TG-NSBPS instantly applies the NSBPS scheduling for immediate URLLC-eMBB MU pairing, however, only over the URLLC PRB set with less than G_c active eMBB UEs.

Throughput-greedy puncturing scheduler (TG-PS): an extension of the PS scheduler [8] where the MU scheduler always attempts greedy eMBB-eMBB MU transmissions in case there is no buffered URLLC traffic foreseen. Otherwise, to-be-buffered URLLC traffic preemptively overwrites some of the eMBB-monopolized PRBs for immediate scheduling, at the expense of the eMBB capacity degradation.

Throughput-greedy Multi-user PS (TG-MUPS): an extension to the MUPS scheduler in [9], in which the scheduler attempts greedy eMBB-eMBB MU transmissions if there is no URLLC queued traffic. In case URLLC traffic is to be buffered for multiple TTIs, scheduler attempts a *standard and non-biased* URLLC-eMBB MU transmissions based on the achievable sum rate constraint, only over the PRB set with maximum $G_c - 1$ eMBB active UEs. If a successful pairing is not possible, scheduler immediately rolls back to PS scheduler by overwriting several ongoing eMBB transmissions.

4 Numerical Results

The performance evaluation is based on dynamic system level simulations where the 3GPP 5G-NR methdology is followed [5]. We adopt 8×2 antenna setup, with the 3D spatial channel modeling. Dynamic link adaptation and Chase combining HARQ are used to relax the initial block error rate (BLER). The main simulation settings are listed in Table E.I. Herein, we consider the NSBPS scheduler as a reference against other schedulers under evaluation.

Fig. E.3 shows the empirical cumulative distribution function (ECDF) of the average DL cell throughput performance for all assessed schedulers with $\Omega = (5, 5)$. The NSBPS scheduler provides a fair cell throughput performance since all system SDoFs are fully reserved for instant URLLC scheduling, i.e., greedy eMBB-eMBB MU transmissions are not allowed regardless from the URLLC traffic availability. The proposed PAROS scheduler offers a significant improvement of the cell throughput, i.e., an average of 5 Mbps throughput increase compared to the NSBPS scheduler, while the TG-NSBPS scheduler offers the best cell throughput due to the aggressive MU transmissions without rank offloading.

Moreover, the TG-PS scheduler exhibits a severe degradation in the overall throughput due to the puncturing events. Thus, punctured eMBB transmissions suffer from significant capacity loss. Consequently, the SE gain from the greedy eMBB-eMBB MU pairings vanishes due to the puncturing capacity loss, e.g., one URLLC UE may puncture an active PRB with G_c active eMBB UEs, thus, degrading their respective capacity. Finally, the TG-MUPS shows a slightly improved ergodic capacity than the TG-PS due to the successful URLLC-eMBB MU standard pairings, hence, no puncturing is applied. Otherwise, TG-MUPS rolls back to PS scheduler for instant URLLC transmission.

Table E.1: Major simulation parameters.

Parameter	Value
Environment	3GPP-UMA,7 BSs, 21 cells
Channel bandwidth	10 MHz, FDD
Antenna setup	BS: 8 Tx, UE: 2 Rx
User load	$K_{\text{llc}} = 5 \text{ or } 20$,$K_{\text{mbb}} = 5 \text{ or } 20$
User receiver	LMMSE-IRC
TTI configuration	URLLC: 0.143 ms (2 OFDM symbols) eMBB: 1 ms (14 OFDM symbols)
HARQ	asynchronous HARQ, Chase combining HARQ round trip time = 4 TTIs
Link adaptation	dynamic modulation and coding target URLLC BLER : 1% target eMBB BLER : 10%
Traffic model	URLLC: FTP3, B = 50 bytes, $\lambda = 250$ eMBB: full buffer
Multi-user rank	$G_c = 2$

As shown in Fig. E.4, the empirical complementary CDF (ECCDF) of the URLLC radio latency is depicted. Referring to the NSBPS scheduler, the proposed PAROS, and TG-PS schedulers offer a decent URLLC latency performance, approaching its stringent target, i.e., 1 ms at 10^{-5} outage probability. Thus, if there is buffered URLLC traffic at the MU scheduler, which is the last scheduling opportunity for URLLC traffic to get scheduled during the current TTI, both schedulers enforce an *immediate and biased* URLLC transmissions regardless of the scheduler state. Thus, the URLLC queuing delay is significantly minimized. However, the TG-MUPS exhibits an increase of $\sim +43.4\%$ in the URLLC latency than the PAROS scheduler. This is basically due to the *standard and non-biased* URLLC-eMBB MU transmissions, where the resulting inter-user interference degrades the URLLC decoding ability, leading to several re-transmissions prior to a successful decoding. The TG-NSBPS shows the worst URLLC latency since all active PRBs are highly likely to be overloaded with G_c active eMBB UEs. Thus, when URLLC traffic arrives the MU schedulers, it has very limited SDoFs to schedule such critical traffic, resulting in further URLLC queuing delays.

Finally, Fig. E.5 presents a comparison of the achievable MU throughput increase, with respect to the SU case, for two extreme loading states. The MU achievable throughput is defined as the pre-detection sum data rate due to the effective MU pairings at the BS. Thus, for SDoF-rich state, i.e., $\Omega = (20, 5)$, where there is a sufficient number of active eMBB UEs, TG-

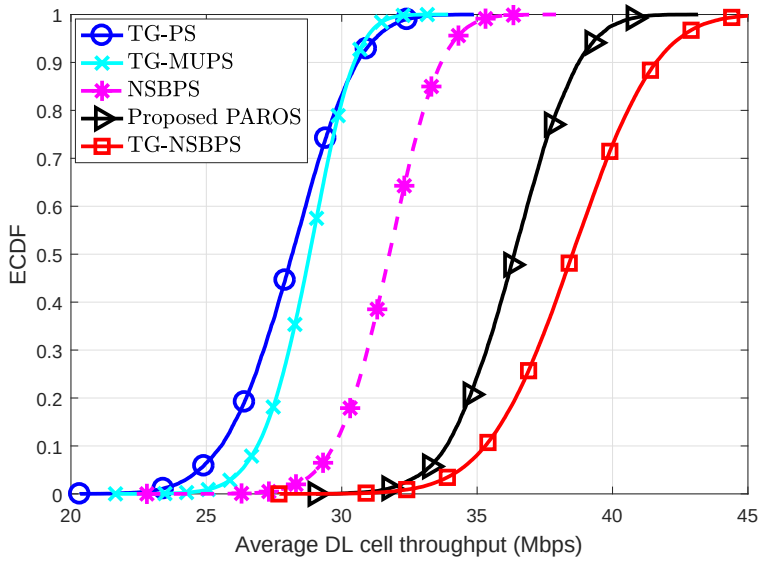


Fig. E.3. Average cell throughput performance (Mbps).

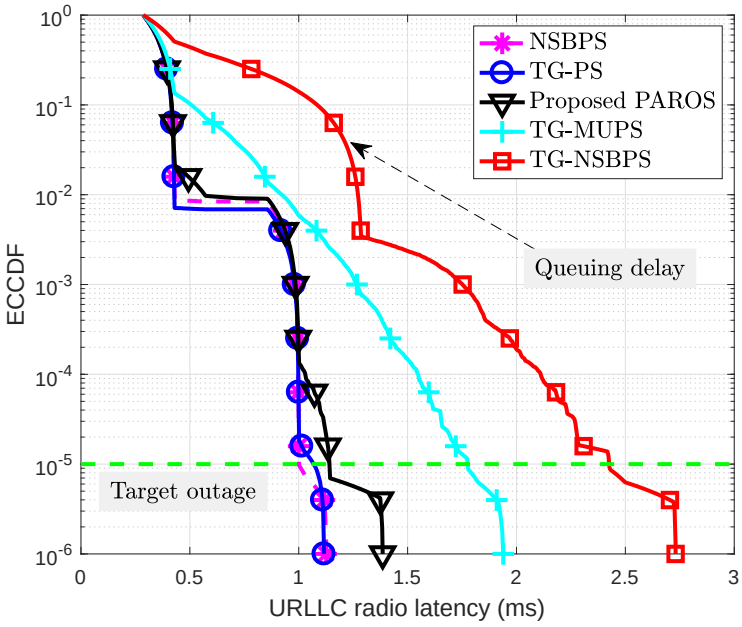


Fig. E.4. URLLC latency performance (ms).

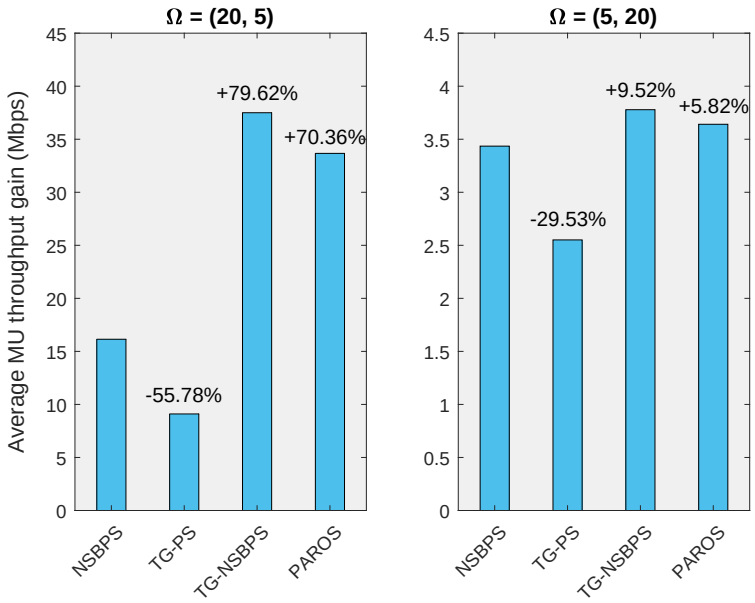


Fig. E.5. MU throughput performance (Mbps), compared to NSBPS.

NSBPS and PAROS schedulers offer a significant enhancement in the achievable MU throughput due to the successful eMBB-eMBB MU pairings. Thus, the ergodic capacity is almost doubled, i.e., $\geq +70\%$ gain. Though, PAROS scheduler exhibits $\sim -9.5\%$ MU loss than TG-NSBPS due to the instant rank offloading when URLLC buffering is envisioned. Finally, the TG-PS scheduler exhibits a severe degradation in the MU throughput since under such loading state, the majority of the system PRBs are overloaded with eMBB MU transmissions. Thus, instant puncturing of these becomes quite costly. With $\Omega = (5, 20)$, the system becomes dictated by URLLC transmissions from the TD and FD schedulers. Hence, all schedulers suffer from MU degradation since URLLC-URLLC MU transmissions are not allowed.

5 Concluding Remarks

We have proposed a preemption-aware rank offloading scheduling (PAROS) framework for 5G new radio. The proposed scheduler shows great scheduling flexibility in multi-traffic scenarios, i.e., URLLC and eMBB. It dynamically adapts the scheduling objectives according to the instantaneous traffic availability and scheduling state. Compared to the state-of-the-art scheduler proposals, the proposed PAROS scheduler offers a significantly improved ergodic capacity of more than 70% gain, while simultaneously satisfying the

URLLC stringent latency and reliability targets, i.e., 1 ms at 10^{-5} outage.

The valuable insights offered by this work are summarized as: (1) for highly loaded cells, multi-traffic spatial schedulers become of a significant importance to trade-off the overall spectral efficiency with the latency and reliability targets, (2) conventional spatial schedulers are not appropriate for latency critical URLLC traffic due to their network-centric, instead of user-centric, scheduling constraints, and (3) these schedulers should be sufficiently flexible to maximize the ergodic capacity by default and be able to preemptively free sufficient degrees of freedom for the sporadic URLLC arrivals. A further flexible URLLC-to-URLLC multi-user scheduling study will be conducted in a future work.

6 Acknowledgments

This work is partly funded by the Innovation Fund Denmark, Grant: 7038-00009B. Also, part of this work is performed in the framework of the Horizon 2020 project ONE5G (ICT-760809) receiving funds from the European Union.

References

[1] R. Drath and A. Horch, "Industrie 4.0: hit or hype?," *IEEE Ind. Electron. Mag.*, vol. 8, no. 2, pp. 56-58, June 2014.

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

[3] Study on new radio access technology (Release 14), 3GPP, TR 38.801, V14.0.0, March 2017.

[4] NR and NG-RAN overall description; Stage-2 (Release 15), 3GPP, TS 38.300, V2.0.0, Dec. 2017.

[5] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Netw.*, vol. 6, pp. 38451-38463, July 2018.

[6] B. Soret, P. Mogensen, K. I. Pedersen and M. C. Aguayo-Torres, "Fundamental trade-offs among reliability, latency and throughput," *in Proc. IEEE Globecom*, Austin, TX, 2014, pp. 1391-1396.

[7] K. Pedersen, G. Pocovi, J. Steiner and A. Maeder, "Agile 5G scheduler for improved E2E performance for different network implementations," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 210-217, Mar. 2018.

[8] K. I. Pedersen, G. Pocovi, and J. Steiner, "Preemptive scheduling of latency critical traffic and its impact on mobile broadband performance," *in Proc. VTC,* Porto, 2018, pp. 1-6.

[9] Ali A. Esswie, and K.I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5G networks," *in Proc. IEEE ISCC*, Natal, 2018, pp. 1-6.

[10] Ali A. Esswie, and K.I. Pedersen, "Null space based preemptive scheduling for joint URLLC and eMBB traffic in 5G networks," *in Proc. IEEE Globecom*, Abu Dhabi, Dec. 2018.

[11] Study on 3D channel model for LTE; Release 12, 3GPP, TR 36.873, V12.7.0, Dec. 2014.

Paper F

Channel Quality Feed-back Enhancements For Accurate URLLC Link Adaptation in 5G Systems

Guillermo Pocovi, A. A. Esswie, and Klaus I. Pedersen

The paper has been published in the
2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)

Abstract

Accurate downlink link adaptation is a major challenge for ultra-reliable and low-latency communications (URLLC) as a consequence of the random and unpredictable load variations at the interfering cells. To address this problem, this paper introduces enhancements to the channel quality indicator (CQI) measurement and reporting procedures for 5G New Radio (NR). The goal is to accurately estimate and report the lower percentiles of the user channel quality distribution. First, a simple and efficient technique is proposed for filtering the channel quality samples collected at the user equipment and, accordingly, estimating tail signal-to-interference-and-noise (SINR) performance. Second, a new CQI reporting format is introduced which better guides downlink scheduling and link adaptation decisions of small URLLC payloads at the gNB. The benefits of the proposed solutions are evaluated via advanced system-level simulations, where it is shown that the proposed solutions significantly outperform existing CQI measurement and reporting schemes. For instance, the 99.999% percentile of the experienced latency is reduced from 1.3 ms to 0.86 ms for the case when URLLC traffic is multiplexed with enhanced mobile broadband (eMBB) traffic.

Index Terms— URLLC; 5G new radio; Channel quality indication (CQI); Link adaptation;

1 Introduction

The 5G New Radio (NR) standard provides enhanced support for enhanced mobile broadband (eMBB), and enables new vertical use cases which demand Ultra-Reliable Low-Latency Communications (URLLC) [1]. In this regard, the 3GPP Release-15 allows the transmission of 32-Byte payloads with a radio latency below 1 ms and 99.999% reliability; whereas the 3GPP community is currently finalizing Release-16 with further enhancements that increase the reliability bound to 99.9999% [2], and address new industrial use cases demanding even lower latencies down to 0.5 ms [3].

To fulfil the stringent URLLC requirements, 5G NR incorporates a wide range of enhancements as compared to preceding technologies. For instance, faster processing times and a flexible frame structure with shorter transmission time intervals (TTI) allow to fulfil the 1 ms latency requirement with up to one Hybrid Automatic Repeat Request (HARQ) retransmission within the latency budget [4]. This enables flexible link adaptation in the sense that a first transmission is scheduled to achieve a moderate block error probability (BLEP) target of, e.g., 10^{-3}, and rely on the HARQ process to ensure a residual BLEP below 10^{-5} for the retransmission [5].

In the downlink (DL) direction, link adaptation for the selection of a modulation and coding scheme (MCS) is based on the channel quality indicator (CQI) feedback information from the User Equipments (UE). For NR, the

BLEP constraint associated with the CQI reports from the UEs is network-configured and can be either 10^{-1} or 10^{-5} [6]. The accuracy and integrity of the CQI reports are of vital importance for fulfilling the strict URLLC reliability requirements [5]. This is challenging in multi-cell cellular networks, where fast and random (unpredictable) interference fluctuations are often experienced, which make the signal to interference-and-noise ratio (SINR) at the UE to also vary rapidly [7, 8]. This problem is especially challenging under fractional-load conditions, as also observed for LTE [9]. In such cases, the MCS selection is typically assisted by adopting outer loop link adaptation (OLLA) mechanisms for fine-tuning the MCS selection at the gNB according to the received HARQ ACK/NACK feedback [10]. The open literature presents several studies on OLLA techniques and related enhancements, see e.g. [11] which proposes a self-optimization algorithm to adjust the OLLA initial offset, and [12] that introduces a dynamic OLLA step size adjustment. However, one of the main challenges of such techniques is their slow convergence time especially when operating with low BLEP targets ($\leq 10^{-3}$), thus, making them unsuitable for URLLC applications.

With the target of supporting even lower latency and/or higher reliability in upcoming NR releases, this paper proposes link adaptation enhancements for URLLC, including the UE reported CQI information. Examples of earlier pioneering studies on CQI design for orthogonal frequency division multiple access (OFDMA) systems to foster radio channel-aware scheduling and link adaptation include [13–18]. Common for those studies is that the objectives were to optimize the user experienced average data rate. However, for URLLC applications, the objective is to accurately control the BLEP for every single transport block transmission in coherence with the ultra-reliability constraint. Here, a CQI report is needed that corresponds to an estimate of the worst-case SINR conditions that the UE is likely to experience until the next received CQI [19]. In pursuit of such solutions, a simple and efficient technique is proposed for filtering the channel quality samples collected at the UEs to estimate tail of the UE-experienced SINR conditions. Secondly, a new CQI reporting format is introduced which better guides the scheduling and link adaptation decisions of small URLLC payloads at the gNB. Similar principles have been studied in [20] and [21]. In [20], a new pilot signals design is proposed such that the CQI accounts for the interference that would be observed if the entire network were actively transmitting; whereas in [21], the gNB collects multiple CQI reports to estimate the maximal degradation from the instant the CQI was measured until it is applied for DL transmission. Contrary to these studies, our proposal requires only minor modifications to the CQI measurement and reporting procedure at the UE side, and does not require changes to the existing NR physical layer reference signals design. The performance and benefits of the proposed techniques are evaluated in a highly-dynamic environment, including the effects of multiple users and cells

and corresponding time-varying traffic and interference. Given the complexity of the system model, the adopted methodology consists of system-level simulations following the Release-16 NR modelling assumptions in 3GPP for URLLC [22]. Good practice is applied in order to generate trustworthy and statistical-reliable results.

The rest of the paper is structured as follows. Section 2 further sets the scene by introducing the system model and problem formulation, respectively. Section 3 provides an overview of the proposed CQI measuring and reporting procedures. Performance results are presented in Section 7, followed by conclusions in Section 5.

2 Setting the Scene

2.1 Network Layout and Traffic Modeling

We consider a macro cellular network with C cells, deployed in a sectorized manner, each with three sectors and 500-meter inter-site distance. Two different traffic compositions are considered: (i), U_u URLLC UEs are deployed in each cell, where the URLLC traffic is modeled as small payloads of B_u Bytes, which arrive at each URLLC UE in the DL direction following a Poisson arrival process with a mean arrival rate λ [packets/s]. The offered load of URLLC traffic per cell is given by $U_u \times B_u \times \lambda$. In case (ii), additional U_e eMBB UEs are deployed in each cell, where the eMBB traffic is modelled with constant-bit-rate (CBR) DL flows, e.g. video streaming, consisting of a predefined number of packets n_e generated per UE, each with payload size of B_e and fixed inter-arrival time of T_e [s]. Once the n_e packets are successfully delivered to the UE, the UE leaves the network and a new one is generated at a random location in the network. The CBR load per cell is $U_e \times \frac{B_e}{(n_e-1)T_e}$.

Users are dynamically scheduled in both the time- and frequency domain using OFDMA. The physical layer configuration consists of 30 kHz sub-carrier spacing (SCS), a physical radio block (PRB) size of 12 sub-carriers (360 kHz), and a TTI duration of 2 OFDM symbols (71.4 μs). Considering the gNB and UE processing capabilities specified in [23], the adopted physical layer configuration allows to fulfil the 1 ms latency target even with one HARQ retransmission.

2.2 URLLC Link Adaptation Challenges

One challenge for accurate link adaptation (and scheduling) of URLLC payloads relates to the tracking of the radio channel and interference variations. Given that URLLC payloads are generally small-sized, they are often scheduled over less PRBs than available within the total carrier bandwidth, offering a weak frequency domain diversity for localized resource allocation,

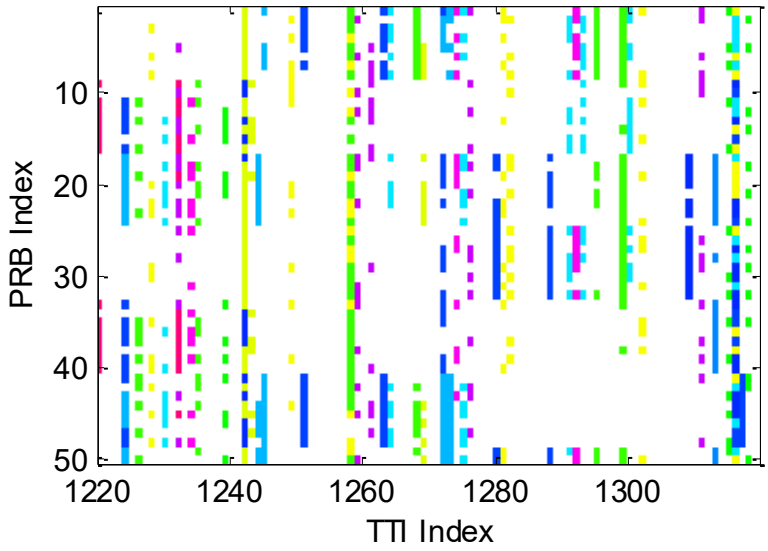


Fig. F.1: Time trace of the downlink PRB allocation in one cell serving URLLC traffic. A color identifies one UE which is served in the downlink direction.

while some frequency diversity can be achieved with distributed resource allocation. In addition, the experienced UE SINR is highly time-variant due to rapid load fluctuations of the neighboring cells. As an example, Fig. F.1 presents a time trace of the allocated PRBs of a cell serving a set of URLLC users (obtained from a dynamic system-level simulation). As can be observed from Fig. F.1, the PRB activity is a highly time-variant random process, which causes the experienced SINR at different UEs to be rapidly time-variant as well (due to variations of the experienced inter-cell interference). This implies that if a UE measures the SINR on certain PRBs at a given time, it might be several dBs different shortly after (say from one TTI to another).

2.3 Objective of the Study

Due to UE SINR estimation imperfections, CQI measuring and reporting delays and the additional latencies such as gNB processing times, it is not considered realistic to accurately track the time- and frequency-variants of the UE experienced SINR. The objective is therefore to design a CQI report that expresses the worst-case SINR conditions that the UE is likely to experience until the next received CQI. One key challenge is to avoid a too-pessimistic CQI estimation, as it reduces the network spectral efficiency, and accordingly, limits the number of URLLC UEs that can be served in the network.

3 Proposed CQI Enhancements

In the following sub-sections, we describe the basic principles of the CQI measuring and reporting procedure as per the 5G NR standard, followed by the introduction of the two proposed CQI enhancements.

3.1 CQI Measuring and Reporting Procedure

The CQI represents the highest supported MCS with which the UE can decode its data with an error probability no larger than a certain constraint. The CQI takes into account the receiver type, number of antennas and potential interference cancellation/suppression capabilities at the UE. The CQI is included in the Channel State Information (CSI) feedback to the gNB, together with the preferred precoding matrix indicator (PMI), rank indicator (RI), among other UE reports [6, Sec. 5.2].

Fig. F.2 shows a flow chart of the CQI measurement and reporting procedure. In the first step, the gNB configures the UE via the Radio Resource Control (RRC) signalling with one or multiple CSI reporting and resource settings. These include, among others, configuration of the time-domain behaviour of the report, e.g., aperiodic or periodic reporting, number of reported frequency sub-bands $\mathcal{S} = \{1, ..., S\}$, the CQI table which shall be used for the report, as well as the Channel State Information Reference Signals (CSI-RS) to be used for desired-signal and interference measurements.

Next, the UE performs channel quality measurements on the specified CSI-RS. Each individual measurement is filtered and used to estimate the UE's experienced SINR with the specified frequency resolution. The estimated SINR on each sub-band s is then mapped to the MCS index m from the specified CQI table that fulfils the following condition:

$$m_s^* = \arg\max_m \{R_{m,s} | P_e(\Gamma_s) \leq P_{target}\}, \tag{F.1}$$

corresponding to the largest data rate $R_{m,s}$, that can be supported with a block error probability P_e not exceeding P_{target} if scheduled over the s-th sub-band (with experienced SINR Γ_s) using MCS index m. For NR, P_{target} can be either 10^{-1} or 10^{-5} and is implicitly derived from the configured CQI table. More details on the MCS entries for each CQI table can be found in [6]. In practice, this is achieved by having the UEs measure the experienced SINR, followed by evaluation of (F.1) given knowledge of the BLEP vs SINR mapping curve for each of the supported MCSs.

Finally, the UE formats the CQI report following the specified granularity. Two report formats are standardized: i) wideband CQI reports ($S = 1$), where the UE reports a single CQI index, and ii) frequency-selective CQI ($S > 1$), where the UE reports both a wideband CQI and the relative offset of each sub-band with respect to the wideband CQI value.

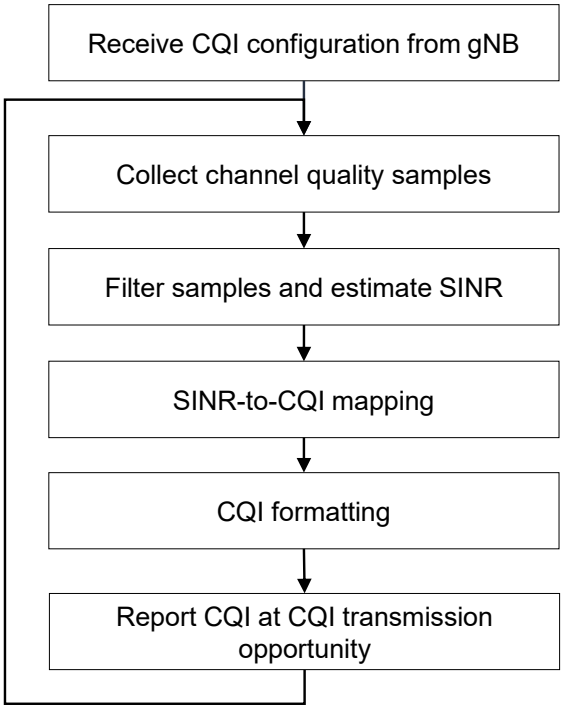


Fig. F.2: CQI measurement and reporting operation at the UE.

3.2 Biased Interference Filtering (BIF)

The first enhancement introduces time-domain filtering of channel quality samples collected at the UE. The UE performs desired-signal and interference measurements on CSI-RS as specified in the CSI resource configuration. As the interference represents one of the main sources of SINR variations, it is proposed that on each measurement instant n, the instantaneous interference measurement on the s-th sub-band, $x_s[n]$, is filtered with a low-pass first-order infinite impulse response (IIR) filter as follows:

$$y_s[n] = \begin{cases} \alpha_u \cdot x_s[n] + (1-\alpha_u) \cdot y_s[n-1], & \text{if } x_s[n] \geq y_s[n-1], \\ \alpha_d \cdot x_s[n] + (1-\alpha_d) \cdot y_s[n-1], & \text{if } x_s[n] < y_s[n-1], \end{cases} \tag{F.2}$$

where $x_s[n]$ and $y_s[n]$ are the instantaneous and filtered interference measurement on the sub-band s over the measurement interval n, and α_u and α_d determine the memory of the filter as well as its bias. As an example, Fig. F.3 shows the filter's output for different settings of α_u and α_d, assuming a zero-mean unit-variance Gaussian distribution at the filter's input. Settings with $\alpha_u = \alpha_d$ correspond to a standard exponentially-weighted moving average filter used for mean value estimation (proposed in [7])

, whereas setting $\alpha_u > \alpha_d$ or $\alpha_u < \alpha_d$ allows to estimate higher or lower percentiles of the input distribution, respectively. Besides, for a fixed α_u/α_d ratio, the value of α_u determines the filter's memory, i.e. how much weight is given to the latest measurement as compared to the previous ones. As the proposed filter is applied to the interference component of the SINR, we consider settings with $\alpha_u > \alpha_d$ for the performance evaluation in Section 7 in order to estimate the worst (highest) interference conditions. Note that the presented filtering procedure is simple in the sense that it only requires storing one $y_s[n]$ sample per sub-band.

A frequency-selective CQI is reported to the gNB containing the filtered interference on each sub-band, $y_s[n]$, together with the latest desired-signal fading information. Note that the latter varies in a much slower time scale and can be generally tracked at the gNB for low UE speeds.

3.3 Worst-M CQI Report

Secondly, a new CQI format is proposed where the UE reports to the gNB: i) a wideband CQI value, that at maximum will result in a BLEP of P_{target} if the gNB schedules a payload with a MCS according to the recently received CQI over the entire band; and ii) a CQI value that at maximum will result in a BLEP of P_{target} if transmitting only over the worst-M subbands, without explicit indication on the position of those subbands.

The worst-M CQI allows the gNB to schedule a small URLLC payload randomly over the frequency-domain (either localized or spread allocation) while still guaranteeing high probability of successful decoding even if it experiences unfavourable conditions of fading and/or interference. Besides, the wideband CQI information can be used for allocations spanning over a larger bandwidth.

The proposed CQI reporting format is similar to the *Best-M* reporting standardized in LTE [1]; however, this scheme applies the opposite criterion when sorting the channel quality measurements, and does not include information on the positions of the M-worst subbands due to the limited benefit of frequency-selective information as observed from Fig. F.1.

4 Performance Evaluation

4.1 Simulation Methodology

A proprietary system-level simulation tool is used to evaluate the performance of the proposed CQI enhancements. The simulation assumptions are summarized in Table F.1. The network layout, UE distribution and traffic follow the description presented in Section 2.1. The network is composed of $C = 21$ cells, with $U_u = 10$ URLLC UEs and optionally $U_e = 10$ eMBB UEs

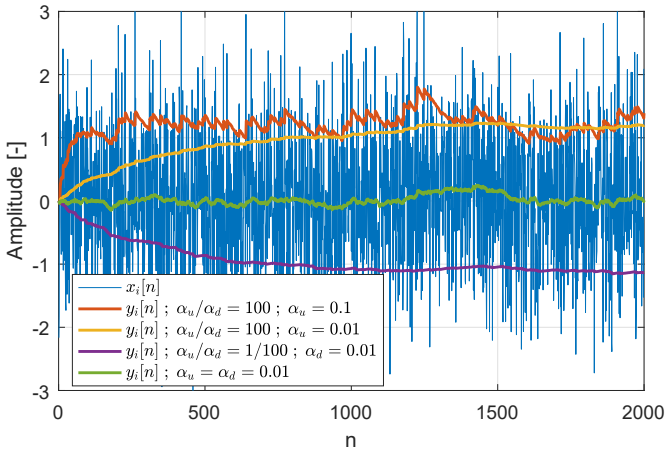


Fig. F.3: Filter's input $x_s[n]$ and output $y_s[n]$ for different settings of α_u and α_d. $x_s[n]$ corresponds to a a zero-mean unit-variance Gaussian distribution.

deployed in each cell.

The simulator's time resolution is one OFDM symbol, and it includes explicit modelling of the majority of radio resource management functionalities such as dynamic packet scheduling and HARQ, as well as time- and frequency-varying inter-cell interference. Closed-loop 4x4 single-user MIMO is assumed for each link and the UE receiver type is minimum mean square error with interference rejection combining (MMSE-IRC). URLLC users are scheduled with a single spatial stream, i.e. benefiting from both transmission and reception diversity against fast fading and radio channel fluctuations, whereas dynamic rank adaptation is assumed for eMBB users allowing multiplexing of up to two spatial streams for favourable SINR conditions.

A frequency- and QoS-aware packet scheduler is assumed, which prioritizes URLLC transmissions and HARQ retransmissions over first transmissions of eMBB traffic. Dynamic link adaptation is applied for both data and the in-resource control channel, which results in varying control overhead depending on the user signal quality and TTI duration (see [7]). The link adaptation is based on the periodical CQI report from the URLLC users. UEs are configured to periodically transmit a CQI report every 1 ms, and a 1 ms processing delay is assumed from the time the CQI is reported until it can be applied for downlink transmissions. Each sub-band consists of 4 PRBs, thus the UE reports CQI for $S = 13$ sub-bands. The proposed measurement and formatting enhancements in Section 3 are presented for different settings of α_u and α_d, and $M = 3$. The latter parameter has been selected in accordance with the average PRB allocation size of URLLC payloads. No outer-loop link adaptation methods are applied.

For each URLLC payload, the latency is measured from the moment it

arrives at the serving cell until it is successfully received at the UE. This accounts for various constant and variable latency components, namely queuing delay, processing and frame alignment delay, and transmission delay; the latter includes the effects of HARQ retransmissions and payload segmentation over multiple TTIs. An *infinite* delay is assumed for payloads not successfully decoded after 6 HARQ retransmissions. The latency of each received URLLC payload is collected and used to form empirical complementary cumulative distribution functions (CCDF). The key performance indicator (KPI) is the achievable latency with 99.999% probability, i.e., the 10^{-5} percentile of the URLLC latency CCDF. The simulation time corresponds to at least 5.000.000 successfully received URLLC payloads in order to ensure a reasonable confidence level for the considered performance metric.

The obtained performance is compared against the following state-of-the-art schemes [7]: i) CQI based on latest/unfiltered channel quality measurements, which is a special case of the proposed BIF scheme with $\alpha_u = \alpha_d = 1$, and ii) CQI based on mean SINR estimation, which corresponds to $\alpha_u = \alpha_d < 1$.

4.2 Performance Results without eMBB Traffic

Fig. F.4 shows the CCDF of the URLLC latency for different CQI schemes and fixed offered load of 10 Mbps per cell, for the case without eMBB traffic. URLLC transmissions experience a minimum delay of $\sim$ 0.29 ms which is a consequence of the 71.4 μs transmission duration and encoding/decoding processing times at gNB and UE, respectively. At the 10^{-5} percentile, a CQI report based on instantaneous channel quality measurements ($\alpha_u = \alpha_d = 1$) is not sufficient to fulfil the 1 ms latency requirement. This is a consequence of the fast (per-TTI) varying load conditions which results in inaccurate link adaptation and thus a large number of HARQ retransmissions. In contrast, other configurations experience at most one HARQ retransmission at the 10^{-5} percentile, and thus achieve the 1 ms latency target accordingly. For instance, the BIF scheme achieves a retransmission probability between $2 \cdot 10^{-5}$ and $8 \cdot 10^{-6}$ with $\alpha_u/\alpha_d = 10$ and $\alpha_u/\alpha_d = 100$, respectively. That is, the latter parameter setting achieves the target 99.999% reliability with a single transmission, and hence, it can be considered an attractive CQI measurement solution for industrial use cases demanding latencies down to 0.5 ms.

Fig. F.5 summarizes the latency at the 10^{-5} percentile for 8 Mbps and 14 Mbps offered loads of URLLC traffic. The benefits of the BIF scheme are mainly relevant for low offered URLLC loads since there are generally sufficient resources to operate with lower error-rate (lower MCS) without increasing the probability of queuing delay to other users. Worst-3 report also provides good URLLC outage performance, especially if the report is based on the time-averaged interference measurements ($\alpha_u = \alpha_d = 0.01$). At

Table F.1: Simulation assumptions

Parameter	Value
Network env.	3GPP Urban Macro (UMa) network with 21 cells and 500 meter inter-site distance [24]
PHY numerology	30 kHz subcarrier spacing; 12 subcarriers per PRB; TTI size of 2 OFDM symbols (71.4 μs)
Carrier config.	20 MHz carrier bandwidth (50 PRBs) at 4 GHz
Duplexing	Frequency division duplexing (FDD)
Control channel	Error-free in-resource scheduling grants with dynamic link adaptation [7]
CQI/CSI configuration	CQI and PMI, reported every 1 ms with 1 ms processing delay; Sub-band size: 4 PRBs;
Antenna config.	4 x 4 single-user MIMO and MMSE-IRC receiver
Packet scheduler	Proportional Fair; strict priority for URLLC traffic
HARQ	Async. HARQ with Chase combining; Max. 6 HARQ retransmissions. Processing time as in [23]
RLC	RLC Unacknowledged mode
Traffic composition	Case a) 10 URLLC UEs per cell; Case b) 10 URLLC UEs + 10 eMBB UEs per cell
UE distribution	Uniformly distributed in outdoor locations
Traffic model	URLLC: FTP3 DL traffic; $B_u = 50$B; Variable offered load per cell eMBB: CBR DL traffic: $B_e = 160$kB; $n_e = 10$; 5Mbps offered load per cell

higher offered loads, the larger and fast-varying interference makes it difficult to achieve the required reliability with a single transmission, and therefore, both the proposed solution and the state-of-the-art scheme deliver a similar performance.

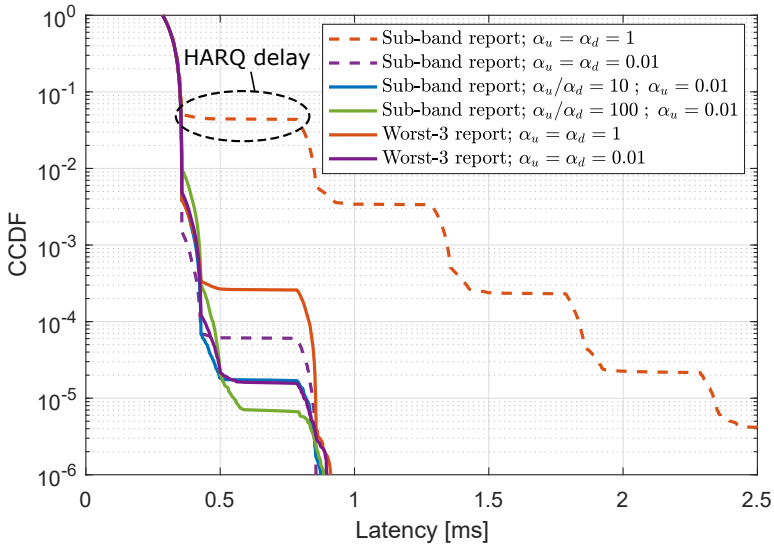


Fig. F.4: URLLC latency distribution for different CQI reporting and measurement schemes. Offered load of URLLC traffic is 10 Mbps per cell.

4.3 Performance Results with eMBB Traffic

Fig. F.6 shows the URLLC performance for cases with a mixture of URLLC and eMBB users, with an offered load of 2 and 5 Mbps for URLLC and eMBB traffic, respectively. Even though URLLC transmissions are fully prioritized by the packet scheduler, the larger inter-cell interference from scheduling eMBB users significantly degrades the URLLC latency performance. For instance, the CQI scheme with $\alpha_u = \alpha_d = 0.01$, which was deemed suitable for the URLLC-only case (Fig. F.4 and F.5), does not longer fulfil the 1 ms latency target with 99.999% reliability when eMBB traffic co-exists in the system. This is a consequence of the significantly different interference pattern with frequent transitions between *low* load and *high* load (up to 100% PRB utilization) when eMBB users arrive or leave the system. In such conditions, there is a substantial benefit of using the proposed Worst-M and BIF technique, as these focus on estimating the tails of the UE's SINR distribution (worst-case interference conditions). The best performance is generally obtained with the BIF technique, whereas Worst-3 offers a slightly worse performance, although, it has the benefit of lower UL signalling overhead due to single sub-band reporting.

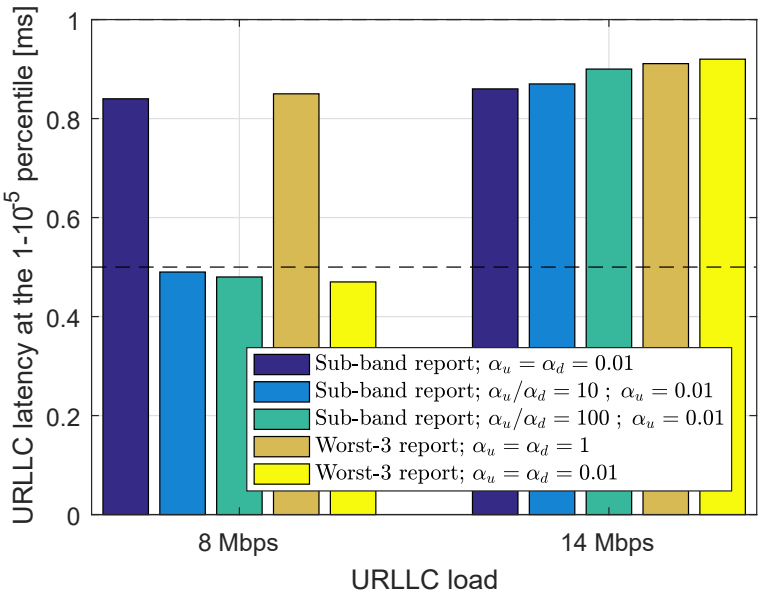


Fig. F.5: Summary of URLLC latency performance at the 99.999% percentile for two offered loads of URLLC traffic. The 0.5 ms latency target in NR Release-16 is indicated with a horizontal dashed line.

5 Conclusions

In this paper we have addressed the problem of link adaptation imperfections for reliable downlink transmissions of URLLC traffic. Two enhancements have been proposed to the Channel Quality Indicator (CQI) measuring and reporting procedure at the UE: *Biased Interference filtering* (BIF) of the collected channel quality measurements, and *Worst-M* CQI reporting format, which target to estimate and report the lower percentiles of the UE's channel quality distribution. Performance results show how the proposed schemes facilitate downlink transmission of small and sporadic URLLC payloads with low BLEP constraints, e.g. $< 10^{-3}$, without relying on traditional outer-loop link adaptation methods. In scenarios with low offered loads of URLLC traffic, BIF and the Worst-M schemes allow to achieve latencies down to 0.5 ms at the 99.999% percentile, which is a new requirement imposed by some industrial vertical applications. In scenarios with a mixture of URLLC and dynamic eMBB traffic, the proposed solutions significantly outperform existing techniques and achieve the 1 ms and 99.999% URLLC requirement.

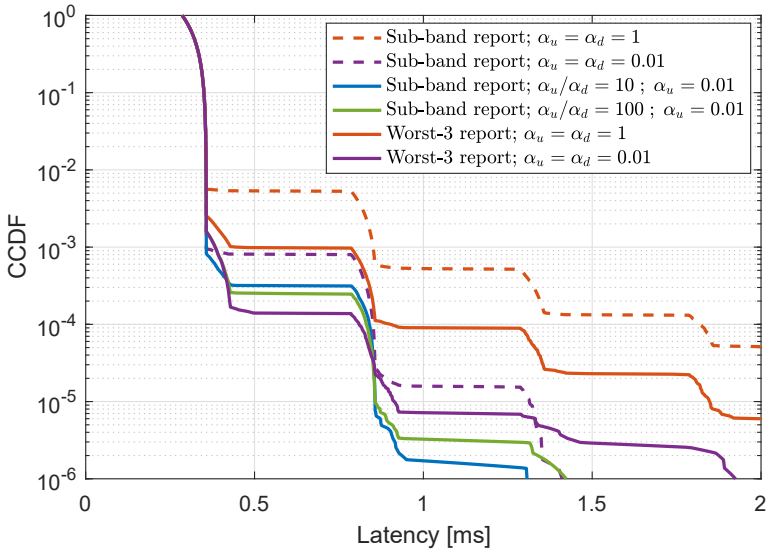


Fig. F.6: URLLC latency distribution for different CQI reporting and measurement schemes in scenarios with a mixture of URLLC and eMBB traffic. The offered load is 2 and 5 Mbps for URLLC and eMBB traffic, respectively.

Acknowledgements

This work is partly funded by the Innovation Fund Denmark (IFD) under File No. 7039-00009B and 7038-00009B.

References

[1] 3GPP Technical Specification TS 38.300 v15.3.1, “NR and NG-RAN Overall Description; Stage 2," Oct. 2018.

[2] 3GPP RP-181477, “New SID on Physical Layer Enhancements for NR URLLC ," June 2018.

[3] 3GPP Technical Specification TS 22.104 v16.0.0, “Service requirements for cyber-physical control applications in vertical domains," Dec. 2018.

[4] 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.

[5] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji, and R. Jäntti, “Link adaptation design for ultra-reliable communications", *in Proc. IEEE ICC*, Kuala Lumpur, 2016, pp. 1-5.

[6] 3GPP Technical Specification TS 38.214 v15.3.0, “Physical layer procedures for data", Sept. 2018.

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

[8] 3GPP R1-1901555, “Baseline performance achievable with Rel-15 URLLC for factory automation", Mar. 2019.

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

[10] K. I. Pedersen et al., “Frequency domain scheduling for OFDMA with limited and noisy channel feedback," *in Proc. IEEE VTC*, Baltimore, MD, 2007, pp. 1792-1796.

[11] A. Durán, M. Toril, F. Ruiz, and A. Mendo, “Self-optimization algorithm for outer loop link adaptation in LTE," *IEEE Commun. Lett.* , vol. 19, no. 11, pp. 2005–2008, Nov. 2015.

[12] F. Blanquez-Casado, G. Gomez, M. Aguayo-Torres, and J. Entrambasaguas, “eOLLA: an enhanced outer loop link adaptation for cellular networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 20, Jan. 2016.

[13] R. Agarwal, V. R. Majjigi, Z. Han, R. Vannithamby, and J. M. Cioffi, “Low complexity resource allocation with opportunistic feedback over downlink OFDMA networks," IEEE J. Sel. Areas Commun., vol. 26, no. 8, pp. 1462-1472, Oct. 2008.

[14] T. S. Kang and H. M. Kim, “Opportunistic feedback assisted scheduling and resource allocation in OFDMA systems," in Proc. IEEE ICCS, Oct. 2006, pp. 1-5.

[15] S. Yoon, O. Somekh, O. Simeone, and Y. Bar-Ness, “A comparison of opportunistic transmission schemes with reduced channel information feedback in OFDMA downlink," in Proc. IEEE PIMRC, Sep. 2007.

[16] J. Leinonen, J. Hamalainen, and M. Juntti, “Performance analysis of downlink OFDMA frequency scheduling with limited feedback," in Proc. IEEE ICC, Jun. 2008, pp. 3318-3322.

[17] Y.-J. Choi and S. Bahk, “Selective channel feedback mechanism for wireless multichannel scheduling," in Proc. IEEE WoWMoM, 2006.

[18] J. Chen, R. A. Berry, and M. L. Honig, “Performance of limited feedback schemes for downlink with finite coherence time," in Proc. IEEE ISIT, Jun. 2007, pp. 2751-2755.

[19] M. Bennis, M. Debbah and H. Vincent Poor, “Ultrareliable and low-Latency wireless communication: tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834-1853, Oct. 2018.

[20] D. Phan-Huy, P. Chauveau, A. Galindo-Serrano, and M. Deghel, “High Data Rate Ultra Reliable and Low Latency Communications in Bursty Interference," in Proc. International Conference on Telecommunications (ICT), Jun. 2018.

[21] A. Belogaev, E. Khorov, A. Krasilov , D. Shmelkin, and S. Tang, “Conservative Link Adaptation for Ultra Reliable Low Latency Communications," in Proc. IEEE BlackSeaCom, Jun. 2019.

[22] 3GPP TR 38.824 v16.0.0, “Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC)", Mar. 2019.

[23] 3GPP R1-1808449, “IMT-2020 self-evaluation: UP latency analysis for FDD and dynamic TDD with UE processing capability 2 (URLLC)", Aug. 2018.

[24] 3GPP Technical Report TR 38.901, “Study on channel model for frequencies from 0.5 to 100 GHz", Dec. 2017.

Part III

Coordination Techniques For Dynamic TDD URLLC Networks

Coordination Techniques For Dynamic TDD URLLC Networks

In this part of the thesis, several novel heuristic inter-BS coordination schemes of the severe cross-link-interference (CLI) have been proposed and developed. The achievable URLLC outage latency is comprehensively studied and evaluated in multi-UE & multi-cell TDD 5G-NR deployments. The performance of the introduced solutions has been compared to the state-of-the-art TDD proposals in the recent literature, through highly detailed system level simulations, with a high degree of modeling realism.

1 Problem Formulation

The early 5G deployments are envisioned over the unpaired 3.5 GHz spectrum due to its abundantly available communication bandwidths [1, 2]. Therefore, the time division multiplexing (TDD) mode has become of a significant importance for the 5G success. For dynamic TDD algorithm, BSs dynamically in time change the structure of their radio frames, in terms of the number and timing of the downlink and uplink transmission opportunities, in order to meet the varying directional traffic demands.

Achieving the stringent URLLC latency and reliability targets [3] are further challenging in TDD networks. This is mainly attributed to:

1. The restriction of the TDD system design of having an exclusive downlink or uplink transmission opportunity at a time. For instance, as depicted by the packet timing example in Fig. III.1 [9, Paper G], the transmission time interval (TTI) is assumed to span two OFDM symbol duration while the delays for the BS processing (bsp), frame alignment (fa) and HARQ re-transmissions are all considered. As clearly shown,

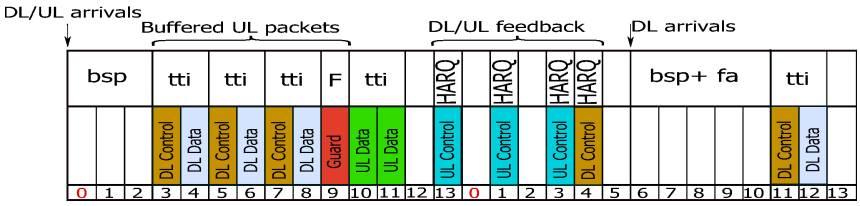


Fig. III.1: An example of the URLLC one-way latency components [9, Paper G].

the urgent arriving uplink packets are further buffered to the next available uplink TTI after three downlink TTIs. Such additional queuing delay makes the stringent URLLC radio latency budget further challenging to achieve. Accordingly, the optimization of the dynamic TDD frame structure is vital to achieve a decent URLLC outage performance in TDD deployments.

2. The BS-BS and UE-UE CLI from the coexistence of BSs and UEs with simultaneous opposite transmission directions [4], as depicted by Fig. III.2 [10, Paper K]. Particularly, in macro TDD deployments, the BS-BS CLI has been demonstrated to be a critical performance limitation [9] due to the large transmit power difference between the downlink interfering transmissions and the corresponding victim uplink receptions. Accordingly, the BS-BS CLI leads to a continuous uplink traffic blockage. Therefore, transmitted uplink packets typically consume several HARQ re-transmission combining attempts before a successful decoding leading to an accumulation of the buffered uplink traffic. Thus, controlling the network CLI through inter-BS coordination schemes, is essential for a proper dynamic TDD operation.

The recent studies from state-of-the-art TDD literature typically consider multi-cell joint UE scheduling, coordinated uplink power control, and dynamic resource muting [5, 6]. The main objective is that the dynamically-identified aggressor CLI sources are either muted or configured to transmit with a lower power level in order to either reduce or avoid the severe CLI. Moreover, advanced beam-forming and joint receiver design techniques [7] have been recently introduced as viable solutions to combat the BS-BS CLI, utilizing the spatial degrees of freedom offered by the BS antenna array. Although those solutions offer clear performance merits, they were not designed to suite the URLLC use cases with stringent radio latency and reliability targets. Therefore, this part of the thesis focuses on the problem of controlling the network CLI for URLLC TDD networks, while preserving the frame flexibility of the TDD deployments through the development of novel inter-BS CLI coordination schemes.

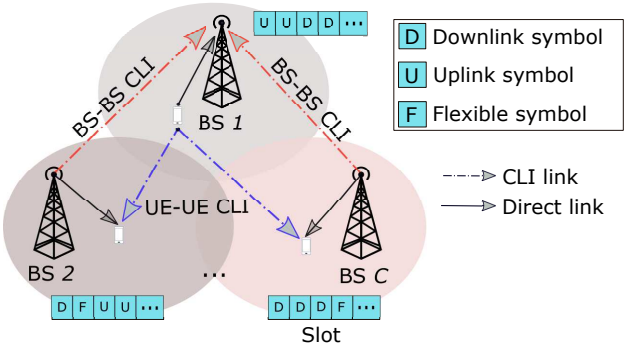


Fig. III.2: BS-BS and UE-UE CLI in dynamic TDD systems [10, Paper K].

2 Objectives

As being indicated by the former research questions Q2 and Q3 and corresponding hypothesis, the objective of this part of the PhD thesis are as follows:

- Study and evaluate the achievable URLLC latency and reliability performance within dynamic TDD macro networks.
- Propose and develop several novel CLI coordination schemes in order to dynamically avoid or suppress the severe CLI.

3 Included Articles

The main relevant papers of this PhD part are listed as follows:

Paper G: On the Ultra-Reliable and Low-Latency Communications in Flexible TDD/FDD 5G Networks

In this paper, we study and evaluate the URLLC outage performance for the 5G-NR TDD deployments. In particular, the feasibility of achieving the URLLC outage targets are identified for the 5G-NR TDD macro networks. The achievable URLLC outage performance is compared to that is of the frequency division multiplexing (FDD) systems, and with a diversity of the 5G-NR system design variants. The individual performance contributions of the 5G-NR sub-carrier spacing (SCS), TTI duration, CLI, and the TDD link switching periodicity are all captured in a system-level setting. For the uplink direction, the dynamic and configured grant (DG and CG) uplink scheduling schemes [8] are considered and corresponding URLLC performance is evaluated. Finally, the paper envisions the 5G-NR radio system settings in order to

achieve a decent URLLC performance within the TDD roll-outs. In particular, the setting of the SCS equals 30 kHz with a TTI duration of 4 OFDM symbols is shown to be suitable to achieve the URLLC 1 ms within an interference-free environment. With increasing the offered load, the queuing delay starts to dominate the achievable URLLC radio latency, hence, adopting larger bandwidths becomes vital. Finally, with considering the BS-BS CLI, the URLLC outage targets are hardly feasible in dynamic TDD systems.

Paper H: Semi-Static Radio Frame Configuration for URLLC Deployments in 5G Macro TDD Networks

This paper introduces a semi-static frame coordination scheme for dynamic TDD macro deployments. Based on the insights brought by Paper G, we first identify the BS-BS CLI as a critical performance bottleneck of the dynamic TDD macro networks. Our objective is to avoid the occurrence of the network CLI while preserving a semi-static frame adaptation to the varying average traffic demands. Each BS continuously monitors its respective offered traffic statistics. Neighboring BSs exchange indications about their current traffic demand, in terms of the how much downlink and uplink traffic is available, over the back-haul interfaces. Those traffic statistics are unequally weighted by Kaiser filter coefficients. The intuition of the applied traffic filtering is that BSs with the largest amount of downlink or uplink traffic size have a higher priority in deciding the next radio frame structure. A network-specific filtered traffic statistic is calculated. Accordingly, the corresponding upcoming TDD radio frame configuration is selected, where it is adopted by all coordinating BSs. This way, the BS-BS and UE-UE CLI is completely avoided through a simple; but, efficient and dynamic TDD frame coordination scheme. Proposed solution is demonstrated to offer 40% reduction of the URLLC outage latency compared to the static TDD scheme, mainly due to the offered semi-static frame adaptation to average network traffic demands.

Paper I: Inter-Cell Radio Frame Coordination Scheme Based on Sliding Codebook for 5G TDD Systems

This paper proposes a fully dynamic inter-BS TDD frame coordination scheme for joint eMBB-URLLC networks. The objective, unlike the proposed scheme in Paper H, is to offer BS-specific TDD frame adaptation to the BS varying traffic demand while heuristically avoiding the network CLI. A sliding phase-offset radio frame book is constructed. The frame book consists of several frame sub-books. Each sub-book contains an arbitrary number of the radio frames, where they are all configured with the same number of downlink and uplink transmission opportunities, however, the downlink and uplink symbol structure is cyclic shifted. Therefore, BSs are able to dynamically se-

lect the radio frame configurations that best satisfy their current traffic needs while minimizing the number of CLI-hit sub-frames across the radio frame. This is achievable by selecting the radio frames with the proper numbers of the downlink and uplink transmission opportunities, which meet the corresponding traffic demands, while selecting the respective symbol structure that allows for the least inter-BS CLI occurrence. Using extensive system level simulations, proposed solution has demonstrated considerable capacity improvements compared to both static and non-coordinated dynamic TDD deployments, respectively. For instance, more than 144% increase of the achievable throughput is observed with the proposed solution compared to the uncoordinated dynamic TDD setup.

Paper J: Quasi-Dynamic Frame Coordination for Ultra-Reliability and Low-Latency in 5G TDD Systems

Based on the conclusions of Paper I, the degrees of freedom which allow for a sufficient CLI reduction mainly depend on the size of the frame-book. Accordingly, the larger the frame-book, the better the CLI avoidance ability, however, it comes at the cost of increased inter-BS coordination signaling overhead. Furthermore, the semi-static coordination approach of Paper H only offers network-specific TDD frame flexibility rather than BS-specific frame adaptation in order to completely avoid the network CLI. Therefore, it is shown to inflict a loss in the URLLC outage performance when the inter-BS traffic fluctuations are high, e.g., for the low offered load region. In this paper, we seek to achieve BS-specific TDD frame adaptation, unlike Paper H, while completely avoiding the requirement of the inter-BS coordination, unlike Papers H and I. First, a hybrid radio frame design is defined. It denotes that a predefined set of the radio slots are statically configured across all neighboring BSs, regardless of the time-variant radio frame configurations. Secondly, we introduce a slot-aware dynamic UE scheduler, where the BSs are aware of the predefined static slot set. Accordingly, they preemptively schedule the UEs with the worst radio conditions, i.e., in terms of the worst channel quality indication (CQI) report, during the predefined static slot set. This is regardless if they are not schedulable during those slots based on the baseline scheduling criterion. BSs independently select the radio frame configurations which best satisfy their respective traffic demands; though, with securing the vulnerable transmissions against the severe CLI. The performance of the proposed solution has been assessed using highly detailed system level simulations, where a significant URLLC outage latency improvement is achieved compared to state-of-the-art dynamic TDD proposals, e.g., 92% outage latency reduction compared to the non-coordinated dynamic TDD.

Paper K: Cross-Link Interference Suppression By Orthogonal Projector For 5G Dynamic TDD URLLC Systems

This paper introduces a novel CLI suppression algorithm for 5G-NR dynamic TDD macro deployments. Unlike Papers H-J, the proposed algorithm utilizes a fully and BS-specific dynamic TDD frame flexibility while efficiently suppressing the severe BS-BS CLI. In dynamic TDD roll-outs, the traditional linear interference rejection and combining (IRC) receivers fail to decently suppress the severe BS-BS CLI. This is attributed to the coexistence of multiple principal BS-BS CLI interferers, sparse in the spatial domain. Therefore, the linear averaging of the linear IRC receiver leads to losing some vital CLI information, thus, degrading the overall decoding performance. In this paper, an inter-BS joint IRC receiver design is introduced, where the principal BS-BS CLI is efficiently suppressed . First, the aggressor BSs, with downlink radio slots, exchange the spatial signatures of the UEs, to be scheduled in the downlink direction, with the neighboring uplink BSs. Thus, the victim uplink BSs are able to estimate the basis of the effective BS-BS CLI sub-space. Accordingly, using the orthogonal projection theory, uplink BSs calculate the projector sub-space of the BS-BS CLI sub-space. Therefore, for victim uplink transmissions, the designed uplink IRC decoding matrix is spatially projected *on-the-go* into the estimated CLI projector sub-space, and thus, substantially suppressing the severe BS-BS CLI from the desired uplink data. The proposed solution shows a significant BS-BS CLI suppression gain compared to the standard IRC design case for the same dynamic TDD setup. However, such gain is obtainable at the expense of a larger inter-BS signaling overhead.

4 Main Findings and Recommendations

Main Findings

Table III.1 presents a high-level comparison of the developed coordination schemes in this part of the thesis, and are described by Papers H-K. As can be observed, the various developed schemes require a different level of the coordination overhead and processing complexity, and subsequently, they are demonstrated to offer the best achievable URLLC outage performance for certain offered load regions. For instance, the semi-static coordination scheme of Paper H requires an inter-BS coordination with very simple processing complexity and infrequent & very low signaling overhead. However, it mainly operates best for moderately and highly loaded deployments. Due to the adoption of a network-wide common TDD radio frame, the proposal in Paper H is not suitable for the lightly loaded cases where the inter-BS traffic fluctuations are high. Similarly to Paper H, the fully-dynamic TDD proposal of Paper I demands low processing and inter-BS signaling overhead, though,

it mainly does not operate the best within the highly loaded networks. In those deployments, the CLI, and particularly the BS-BS CLI, becomes the dominant factor of the overall network performance. Thus, to further combat the increasing CLI, the heuristic CLI avoidance approach of Paper I shall require larger frame-books with many more cyclic-shifted radio frame patterns, which results in increasing the inter-BS signaling overhead.

Finally, the coordinated IRC receiver design of Paper K offers a decent URLLC performance, regardless of the offered load region. This is mainly attributed to the BS-BS CLI suppression, on-the-go with ongoing victim uplink transmissions. However, unlike the former schemes, it demands the a relatively high processing and coordination overhead. Finally, the proposal of Paper J requires no inter-BS coordination, thus, no signaling overhead is exhibited. This is due to the pre-configuration of a static slot set across all possible radio frames that an arbitrary BS can select. Thus, the static slot set offers CLI-free channels for reliable UL new payload and HARQ (re-) transmissions. Although, alike Paper I, with high offered loads, and accordingly, more severe CLI, the size of the required static slot set increases. Thus, the quasi-dynamic TDD proposal of paper J may approach the static TDD scheme under the high load region.

Fig. III.3 [9, 10, 12 - Papers G, H, K] depicts a comparison of the achievable URLLC outage latency for the non-coordinated dynamic TDD, semi-static TDD (Paper H), and the coordinated IRC design for dynamic TDD (Paper K), respectively. As can be seen, for an offered load of 0.25 Mbps with a downlink-to-uplink traffic ratio of 1:1, the non-coordinated dynamic TDD achieves a decent URLLC outage latency, approaching the 1 ms target. This is mainly because of the minimum inflicted queuing delay and network CLI, respectively. The semi-static scheme however exhibits a large URLLC latency degradation, due to the network-specific, rather than the BS-specific, TDD radio frame adaptation. Therefore, with such low offered load, the inter-BS traffic fluctuations are high while the CLI is relaxed. Accordingly, adopting a common network TDD radio frame is sub-optimal. At high load of 1 Mbps, the CLI, and especially the BS-BS CLI, starts to dominate the URLLC outage performance. Thus, the dynamic TDD, without CLI control mechanisms, suffers a clear URLLC latency performance loss. The semi-static approach obviously provides a considerable latency reduction of 73% compared to dynamic TDD, due to the absence of the network CLI. With such a highly loaded scenario, the inter-BS traffic variations converge to a similar average, hence, a common network TDD frame provides a decent inter-BS URLLC performance. Finally, for both load regions, the coordinated IRC receiver design offers a decent URLLC performance due to the achieved CLI suppression gain.

Finally, Table III.2 [10] presents a comprehensive comparison of the achievable URLLC outage latency for some of the developed schemes in this thesis

Table III.1: Overall comparison of the developed TDD radio frame coordination schemes in this part of thesis.

Scheme	Inter-BS Coordination	Coordination overhead	TDD Frame Flexibility	Performance Regions
Semi-static TDD (Paper H)	Yes	Very Low	Network-Specific Semi-Static	Moderate and High Load
RFC-codebook TDD (Paper I)	Yes	Low	BS-Specific Fully-Dynamic	Low and Moderate Load
Quasi-dynamic TDD (Paper J)	No	-	BS-Specific Quasi-Dynamic	Moderate Load
Quasi-dynamic TDD (Paper K)	Yes	Moderate	BS-Specific Fully-Dynamic	High, Moderate and Low Load

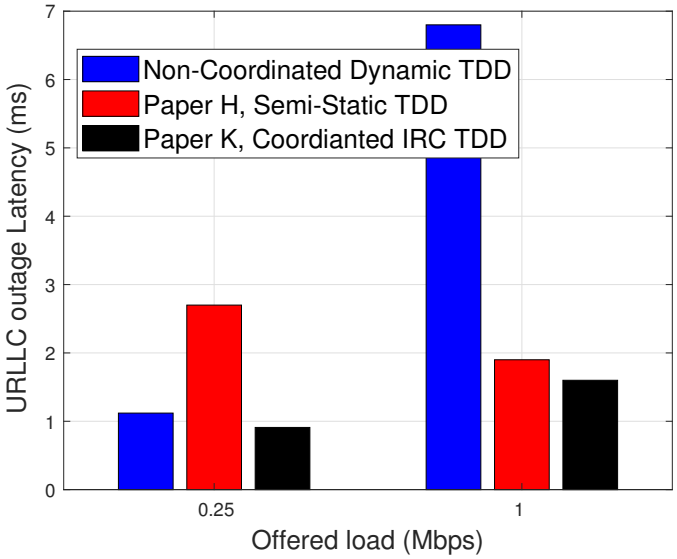


Fig. III.3: The achievable URLLC outage latency with offered load in Mbps, for different TDD coordination schemes [9, 10, 12 - Papers G, H, K].

part (described by Papers I-K) and with different offered loads. Particularity, the schemes under evaluations are listed as follows:

1. **CLI-free TDD** (**CF-TDD**): a fully dynamic and uncoordinated TDD system, although, with the assumption of the BS-BS and UE-UE CLI are perfectly canceled. This hypothetical deployment acts as the reference case to the other schemes under assessment.

2. **Non-coordinated TDD** (**NC-TDD**): a fully non-coordinated dynamic TDD deployments, where the BS-BS and UE-UE CLI can be inflicted among BSs and UEs adopting simultaneously opposite link directions.

3. **Coordinated radio frame configuration based TDD** (**CRFC-TDD**) [11]: it represents Paper J.

4. **BS-BS CLI suppression algorithm based TDD** (**CSA-TDD**) [10]: it represents Paper K.

As can be observed from Table III.2, the CF-TDD scheme offers a steadily URLLC outage latency due to the CLI absolute absence. Thus, the packet buffering delay becomes the sole major source of the radio latency while increasing the offered load size from 4 to 7 Mbps, respectively. On the other hand, the NC-TDD scheme clearly exhibits a progressive increase of the URLLC outage latency with the offered load. This is mainly due to the

stronger BS-BS CLI when increasing the downlink traffic split. Moreover, the CRFC-TDD utilizes an opportunistic way in order for the UEs with the worst channel conditions to dynamically avoid the severe CLI. However, with increasing the offered downlink load, the intensity of the CLI increases (due to the coexistence of more principal interferers), and accordingly, the UL latency is highly degraded due to the several required uplink HARQ combining attempts before uplink packets are successfully decoded. Finally, the CSA-TDD proposal obviously provides a decent URLLC outage latency, approaching the optimal CF-TDD case with all offered loads. This is attributed to the BS-BS CLI cancellation.

Main recommendations

In the following, we summarize the major research recommendations of this part of the thesis as follows:

1. Each of the developed TDD coordination schemes is suitable for a certain load region. For the lightly loaded networks, a fully non-coordinated dynamic TDD is sufficient since the CLI intensity and resource utilization are low. For moderately and highly loaded scenarios, the semi-static TDD scheme is the recommended solution to completely avoid the severe network CLI while offering a semi-static adaptation of the network TDD frame configuration to the network traffic statistics. Fig. III.4 depicts an overall comparison of the developed TDD schemes in Papers H-K with the CLI intensity and the offered load size.

2. The proposed coordination schemes in this part of thesis require newly introduced inter-BS signaling exchange over the back-haul links. The semi-static TDD requires BSs to exchange the introduced buffered traffic indications over the back-haul links, as part of the specified load information. The coordinated IRC design demands the exchange of the downlink UE spatial signatures for a decent BS-BS CLI suppression. Those new information objects and associated signaling procedures are vital to reap the benefits of the developed solutions. Therefore, it is recommended that those should be introduced in the upcoming 3GPP releases.

References

[1] J. Lee et al., "Spectrum for 5G: global status, challenges, and enabling techs," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 12-18, March 2018.

Table III.2: Comparison of the URLLC outage latency of several coordination schemes with offered load per BS, and DL:UL ratio = 2 : 1 [10].

Offered load	CF-TDD		NC-TDD		CRFC-TDD		CSA-TDD	
	DL	UL	DL	UL	DL	UL	DL	UL
4 Mbps	7.15 0.0%	14.76 0.0%	8.47 +16.9%	105.34 +150.8%	7.75 +8.0%	24.12 +48.1%	7.36 +2.89%	17.4 +16.4%
5 Mbps	8.04 0.0%	15.17 0.0%	1663 +198%	6063 +199%	14.24 +55.6%	201.6 +172%	8.43 +4.7%	18.0 +17%
6 Mbps	11.04 0.0%	16.29 0.0%	7394 +199.4%	18390 +199.6%	3150 +198.6%	12540 +199.4%	11.47 +3.82%	19.32 +17%
7 Mbps	17.28 0.0%	18.23 0.0%	12480 +199.4%	25610 +199.7%	6575 +198.9%	19470 +199.6%	19.8 +13.5%	23.07 +23.4%

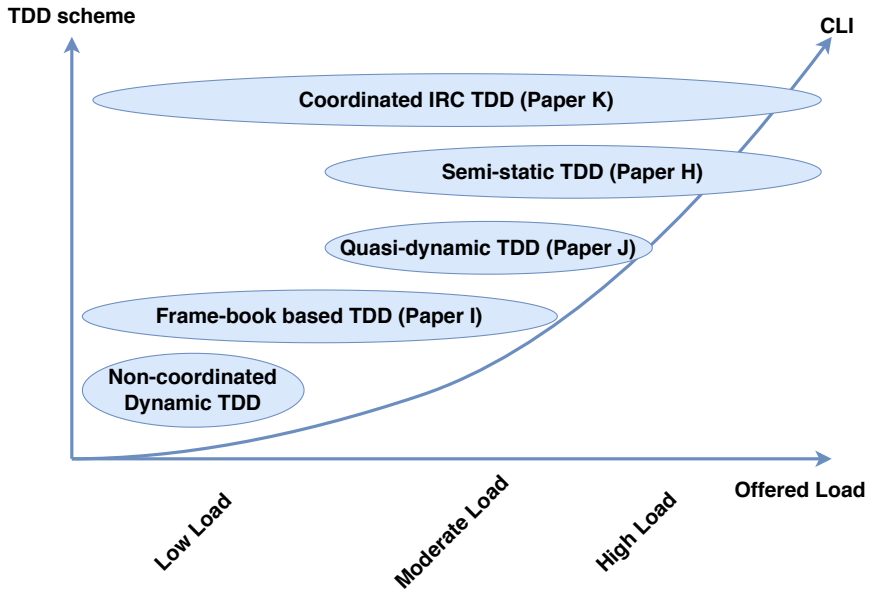


Fig. III.4: Recommended performance regions of the developed TDD coordination schemes.

[2] K. I. Pedersen, G. Berardinelli, F. Frederiksen and P. Mogensen, "A flexible 5G wide area solution for TDD with asymmetric link operation," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 122-128, April 2017

[3] *Service requirements for the 5G system*; Stage-1 (Release 16), 3GPP, TS 22.261, V16.6.0, Dec. 2018.

[4] *Cross link interference handling and remote interference management (RIM) for NR*; (Release 16); 3GPP, TR 38.828, V16.0.0, June 2019.

[5] A. Łukowa and V. Venkatasubramanian, "Performance of strong interference cancellation in flexible UL/DL TDD systems using coordinated muting, scheduling and rate allocation," *in Proc. IEEE WCNC*, Doha, April 2016, pp. 1-7.

[6] Z. Huo, N. Ma and B. Liu, "Joint user scheduling and transceiver design for cross-link interference suppression in MU-MIMO dynamic TDD systems," *in Proc. IEEE ICCC*, Chengdu, Sep. 2017, pp. 962-967.

[7] E. d. O. Cavalcante, G. Fodor, Y. C. B. Silva and W. C. Freitas, "Distributed beamforming in dynamic TDD MIMO networks with cell to cell interference constraints," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 788-791, Oct. 2018.

[8] R. Abreu, T. Jacobsen, K. Pedersen, G. Berardinelli, and P. Mogensen, “System level analysis of eMBB and grant-free URLLC multiplexing in uplink,” in Proc. IEEE VTC-spring, Kuala Lumpur, 2019.

[9] Ali A. Esswie, and K.I. Pedersen, “On the ultra-reliable and low-latency communications in flexible TDD/FDD 5G networks,” *in Proc. IEEE CCNC*, , Las Vegas, NV, USA, Jan. 2020, pp. 1-6.

[10] Ali A. Esswie, and K.I. Pedersen, “Cross link interference suppression by orthogonal projector in 5G dynamic-TDD URLLC systems,” *in Proc. IEEE WCNC*, May 2020.

[11] Ali A. Esswie, K.I. Pedersen, and P. Mogensen, “Quasi-dynamic frame coordination for ultra- reliability and low-latency in 5G TDD systems,” *in Proc. IEEE ICC*, Shanghai, China, May 2019, pp. 1-6.

[12] A. A. Esswie, K. I. Pedersen and P. E. Mogensen, "Semi-static radio frame configuration for URLLC deployments in 5G macro TDD networks," *in Proc. WCNC*, Seoul, Korea (South), 2020, pp. 1-6.

Paper G

On the Ultra-Reliable and Low-Latency Communications in Flexible TDD/FDD 5G Networks

Ali A. Esswie and Klaus I. Pedersen

The paper has been published in the
2020 IEEE Consumer Communications and Networking Conference (CCNC)

Abstract

The ultra-reliable and low-latency communication (URLLC) is the key driver of the current 5G new radio standardization. URLLC encompasses sporadic and small-payload transmissions that should be delivered within extremely tight radio latency and reliability bounds, i.e., a radio latency of 1 ms with 99.999% *success probability. However, such URLLC targets are further challenging in the 5G dynamic time division duplexing (TDD) systems, due to the switching between the uplink and downlink transmission opportunities and the additional inter-cell cross-link interference (CLI). This paper presents a system level analysis of the URLLC outage performance within the 5G new radio flexible TDD systems. Specifically, we study the feasibility of the URLLC outage targets compared to the case with the 5G frequency division duplexing (FDD), and with numerous 5G design variants. The presented results therefore offer valuable observations on the URLLC outage performance in such deployments, and hence, introducing the state-of-the-art flexible-FDD technology.*

Index Terms— Dynamic-TDD; Flexible-FDD; 5G new radio; URLLC; Cross link interference (CLI).

1 Introduction

The fifth generation (5G) new radio (NR) is designed to support a variety of services such as ultra-reliable and low-latency communications (URLLC) [1], industrial time sensitive communications (TSC) [2], and enhanced mobile broadband (eMBB) communications [3]. Those come with challenging requirements for the packet latency, jitter, and aggregated capacity, respectively. On another side, dynamic time division duplexing (TDD) is the major duplexing technology for 5G NR due to the wide spectrum availability of unpaired bands, i.e., the 3.5 GHz band, and spectrum above 6 GHz [4]. Additionally, the frequency division duplexing (FDD) is also supported for 5G NR, and considered especially relevant for deployments at bands below 6 GHz [5]. In this regard, fulfilling the URLLC requirements for FDD systems is obviously more manageable since both base-stations (BSs) and user-equipments (UEs) always have simultaneous uplink (UL) and downlink (DL) transmission opportunities.

Although, for TDD deployments, it is further challenging to fulfill such targets due to the restriction of either having exclusively UL or DL transmissions. Hence, in a multi-cell multi-user scenario, it becomes a hard problem to ensure that the URLLC latency and reliability requirements are met for all active UEs, as the inter-UE timing relations may likely be different. It is therefore a non-trivial problem how to dynamically adjust the UL-DL switching for 5G NR TDD.

The standardization body has accordingly defined a flexible slot format

design [6], where the traffic adaptation could occur per 14-OFDM symbol slots. In principle, such a design allows BSs to dynamically adapt their link directions, i.e., UL or DL symbols, according to a local selection criterion such as the buffered traffic statistics (incl. e.g., the related head of line delay). Although, when different neighboring BSs concurrently adopt opposite transmission link directions, it comes with the cost of potentially severe cross-link interference (CLI) [7]. CLI is highly critical for achieving the URLLC outage requirements, where especially the BS-BS CLI is problematic due to the higher BS transmit power as compared to the UE transmit powers. Accordingly, the majority of the recent TDD studies tackled the CLI issue either by pre-avoidance or post-cancellation techniques. In [8], coordinated inter-cell user scheduling, and advanced UL power control are introduced to minimize the average network CLI. Furthermore, opportunistic frame coordination schemes [7, 9] are proposed to pre-avoid the occurrence of the BS-BS and UE-UE CLI on a best-effort basis. Moreover, perfect BS-BS CLI cancellation using full packet exchange and orthogonal projector estimation are discussed in [10, 11].

In this paper, we study the URLLC outage performance in an advanced system-level setting with high degree of realism. Particularly, how to most efficiently manage the switching between the UL and DL transmission opportunities to best meet the URLLC traffic conditions is investigated, assuming bi-directional random time-variant traffic. The impact of adjusting the TDD switching pattern at different time-resolutions is analyzed, including a sensitivity analysis for other system-level parameter settings and algorithm variants. For dynamic TDD, we isolate the effect of the CLI by presenting both cases where the CLI is realistically modeled, in addition to the case where an optimal CLI cancellation is assumed. To the best of our knowledge, no prior studies have presented such system-level URLLC outage results and related recommendations for 5G NR TDD deployments.

This paper is organized as follows. Section 2 discusses the system modeling. Section 3 introduces the URLLC radio latency analysis in dynamic-TDD systems, while Section 4 presents our adaptation criterion of the dynamic link selection. The URLLC outage latency assessment is introduced in Section 5. Finally, the flexible-FDD duplexing mode is discussed in Section 6, while conclusions appear in Section 7.

2 System Modeling

We consider a 5G-NR dynamic TDD macro network with C BSs, each with N_t antennas, where there are K^{dl} and K^{ul} uniformly-distributed DL and UL active UEs per BS, each with M_r antennas. We assume inter-BS synchronized TDD transmissions, as depicted in Fig. G.1. Additionally, the URLLC-alike

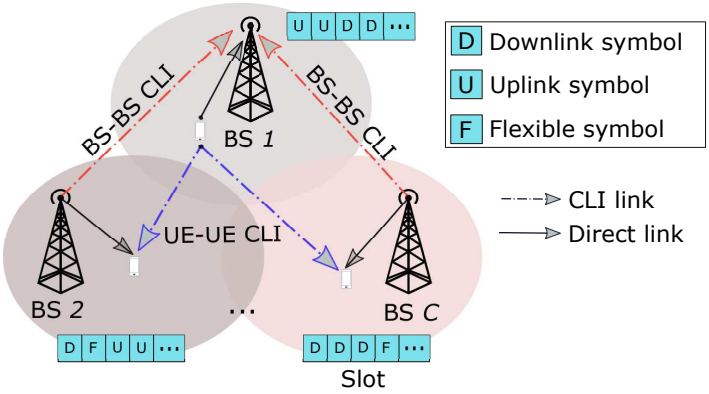


Fig. G.1. Dynamic-TDD deployment per slot periodicity.

sporadic FTP3 traffic model is adopted with the packet sizes of f^{dl} and f^{ul} bits, and Poisson Point Processes, with mean packet arrivals λ^{dl} and λ^{ul}, in the DL and UL directions, respectively. Thus, the average offered traffic load per BS in the DL direction is expressed as: $\Omega^{\text{dl}} = K^{\text{dl}} \times f^{dl} \times \lambda^{\text{dl}}$, and in the UL direction as: $\Omega^{\text{ul}} = K^{\text{ul}} \times f^{ul} \times \lambda^{\text{ul}}$. The total offered load per BS is: $\Omega = \Omega^{\text{dl}} + \Omega^{\text{ul}}$.

We adopt the state-of-the-art 3GPP 5G-NR configurations. UEs are multiplexed using the orthogonal frequency division multiple access (OFDMA), and with 30 kHz sub-carrier spacing (SCS). The smallest resource unit, granted to an active UE, is the physical resource block (PRB) of 12 consecutive SCs. The dynamic user scheduling is applied per a TTI duration of 4-OFDM symbols, for faster URLLC transmissions.

3 URLLC Radio Latency Analysis

The 3GPP 5G-NR release-15 standard has defined several slot format designs [6]. A slot format denotes a certain placement of the DL [D], UL [U], and flexible [F], OFDM symbols within a slot duration of 14 OFDM symbols. The flexible symbols imply that these could be used either for UL/DL transmissions or as guard intervals between consecutive DL and UL symbols. The average one-way URLLC latency in the DL direction Ψ_{dl} is given by

$$\Psi_{\text{dl}} = \Lambda_{\text{bsp}} + \psi_{\text{tq}} + \psi_{\text{fa}} + \psi_{\text{tti}} + \alpha\psi_{\text{harq}} + \Lambda_{\text{uep}}, \tag{G.1}$$

where Λ_{bsp}, ψ_{tq}, ψ_{fa}, ψ_{tti}, ψ_{harq} and Λ_{uep} denote the BS processing, DL total queuing, DL frame alignment, DL packet transmission, DL hybrid automatic repeat request (HARQ) re-transmission, and UE processing delays, respectively. α implies the target block error rate (BLER), e.g., for a URLLC-alike

BLER = 1%, $\alpha = 0.01$, and $\alpha = 0$ if the packet has been successfully decoded from the first transmission. As can be observed, Λ_{bsp}, ψ_{tti} and Λ_{uep} impose a constant delay offset, and are only dependent on the UE/BS processing capabilities, and TTI size, respectively; however, ψ_{tq} and ψ_{harq} are time-varying DL delay components, depending on the DL offered load level, DL and UL link switching delay, and the inflicted DL interference, respectively.

Accordingly, the DL HARQ delay ψ_{harq} is expressed as

$$\psi_{\text{harq}} = \Lambda_{\text{uep}} + \varphi_{\text{fa}} + \varphi_{\text{nack}} + \Lambda_{\text{bsp}} + \psi_{\text{tq}} + \psi_{\text{fa}} + \psi_{\text{tti}}, \tag{G.2}$$

where φ_{fa} implies the alignment delay towards the first UL control channel opportunity for the UE to transmit the HARQ negative acknowledgment (NACK), with φ_{nack} as the NACK transmission time. The summation $\varphi_{\text{fa}} + \varphi_{\text{nack}} + \Lambda_{\text{bsp}}$ represents the total delay from the time a UE has identified a corrupted DL packet until the BS becomes aware of it. Subsequently, the total DL queuing delay ψ_{tq} is calculated by

$$\psi_{\text{tq}} = \psi_{\text{q}} + \psi_{\text{tdd}}, \tag{G.3}$$

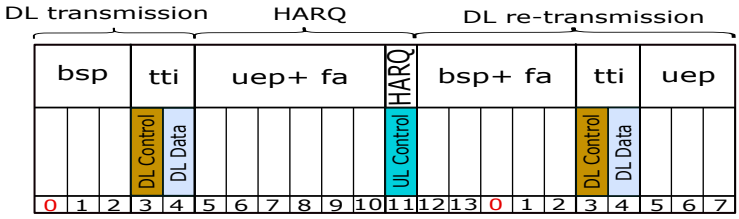
where ψ_{q} implies the packet queuing delay due to the dynamic multi-user scheduling, and ψ_{tdd} is TDD UL-DL link-switching delay, i.e., the additional DL buffering delay towards the first available DL transmission symbol(s) due to the non-concurrent DL and UL transmission availability. For instance, with FDD, $\psi_{\text{tdd}} = 0$ ms. Fig G.2.a shows an example of the factors which contribute to the average one-way DL latency Ψ_{dl}, where a single DL packet associated with one HARQ re-transmission is assumed. As can be observed, the DL packet is decoded at its intended UE after 22 OFDM-symbol duration, i.e., 0.7 ms, from its arrival time at the BS, satisfying the URLLC 1-ms radio latency target; however, with the assumption of immediate DL scheduling and transmission once the packet arrives the BS DL buffer, i.e., $\psi_{\text{td}} = 0$.

Similarly, the one-way URLLC UL latency Ψ_{ul} follows a similar behavior as Ψ_{dl}; however, with a linear delay offset due to the UL scheduling. Specifically, with dynamic-grant (DG) UL scheduling, UEs first align to the first available transmission opportunity of the UL control channel, i.e., φ_{fa}, in order to send the scheduling request (SR), and accordingly wait for the scheduling grant (SG) from the serving BS over the DL control channel. Thus, Ψ_{ul} is given by

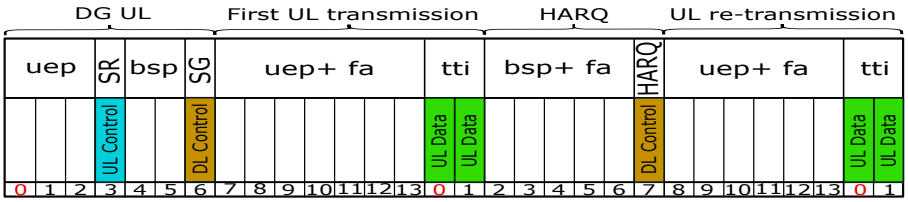
$$\Psi_{\text{ul}} = \varphi_{\text{dg}} + \varphi_{\text{td}} + \varphi_{\text{fa}} + \varphi_{\text{tti}} + \alpha\varphi_{\text{harq}} + \Lambda_{\text{bsp}}, \tag{G.4}$$

where φ_{dg}, φ_{td}, φ_{fa}, φ_{tti} and φ_{harq} are the UL DG delay, UL total buffering delay, UL frame alignment delay, UL payload transmission delay, and UL HARQ delay, respectively.

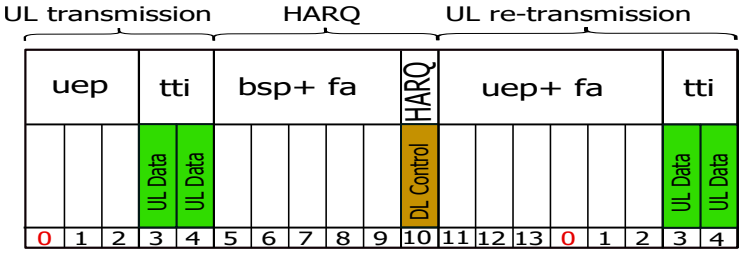
On another side, the grant-free (GF) UL scheduling [12] is considered as vital for URLLC UL transmissions. With UL grant-free, sporadic UL packets



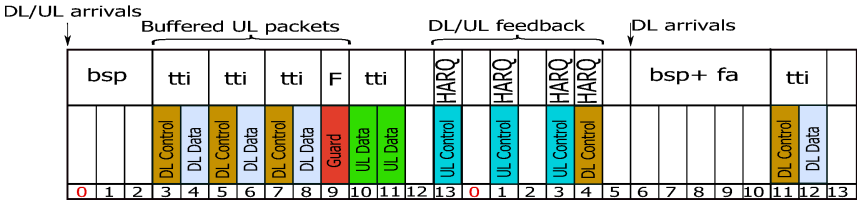
(a) Latency of a single DL packet arrival.



(b) Latency of a single UL packet arrival, with DG UL scheduling.



(c) Latency of a single UL packet arrival, with GF UL scheduling.



(d) Latency of multiple UL and DL packet arrivals.

Fig. G.2. URLLC one-way latency components with DG, and GF UL, for a TTI size = 2-OFDM symbols, and SCS = 30 kHz.

become immediately eligible for scheduling and transmission, i.e., no SR and SG delays, $\varphi_{\mathrm{dg}} = \varphi_{\mathrm{sr}} = \psi_{\mathrm{sg}} = 0$ ms, with φ_{sr} and ψ_{sg} as the transmission delays of the SR and SG, respectively; although, with the DG, φ_{dg} is then calculated as

$$\varphi_{\mathrm{dg}} = \Lambda_{\mathrm{uep}} + \varphi_{\mathrm{fa}} + \varphi_{\mathrm{sr}} + \Lambda_{\mathrm{bsp}} + \psi_{\mathrm{fa}} + \psi_{\mathrm{sg}} + \Lambda^{'}_{\mathrm{uep}}, \tag{G.5}$$

with $\Lambda^{'}_{\mathrm{uep}}$ as the UE processing delay to decode the SG, i.e., Λ_{uep}, as well as preparing the UL transport block, where $\Lambda^{'}_{\mathrm{uep}} > \Lambda_{\mathrm{uep}}$.

Equivalently to (G.2) and (G.3), the UL HARQ φ_{harq} and total queuing φ_{td} delays are given by

$$\varphi_{\mathrm{harq}} = \Lambda_{\mathrm{bsp}} + \psi_{\mathrm{fa}} + \psi_{\mathrm{nack}} + \Lambda_{\mathrm{uep}} + \varphi_{\mathrm{td}} + \varphi_{\mathrm{fa}} + \varphi_{\mathrm{tti}}. \tag{G.6}$$

$$\varphi_{\mathrm{td}} = \varphi_{\mathrm{q}} + \varphi_{\mathrm{tdd}}, \tag{G.7}$$

where φ_{q} and φ_{tdd} denote the UL packet queuing delay and the delay towards the first available UL transmission opportunity, where $\varphi_{\mathrm{td}} \neq \psi_{\mathrm{td}}$ due to the different UL and DL offered load, leading to varying UL and DL buffering performance, respectively. Fig. G.2.b and G.2.c depict the radio latency components which affect the average one-way URLLC UL latency Ψ_{ul}, for the DG and GF UL scheduling cases, respectively, and under the assumption of a single UL packet arrival without further multi-UE queuing delays. With the UL DG and one UL HARQ re-transmission, the URLLC UL packet gets delivered after 30-OFDM symbol duration, i.e., 1 ms, which does not allow for any further packet buffering due to the dynamic user scheduling; otherwise, the URLLC UL 1-ms latency target shall be violated.

Finally, Fig. G.2.d presents an example of multiple concurrent DL and UL packet arrivals, unlike Fig. G.2.a, G.2.b, and G.2.c, respectively. Herein, the BS decides multiple DL TTIs first to transmit the early-arriving DL packets. Accordingly, the UL packets are buffered over those DL TTIs as well as several guard symbols towards the first available UL TTI opportunity, i.e., $\varphi_{\mathrm{tdd}} \ggg 0$ ms, exceeding the UL latency budget. Next, the BS adopts alternating DL and UL TTI instances for the subsequent DL/UL HARQ feedback.

4 Traffic adaptation in dynamic-TDD systems

For a dynamic TDD deployment, BSs dynamically match their transmission link directions to the sporadic traffic arrivals. Hence, at each pattern update periodicity, which could be either per a slot or aggregated several slots, BSs select the slot formats, i.e., number of DL and UL symbols during the next slot(s), which best satisfy their individual link direction selection criteria.

We consider the amount of buffered DL and UL traffic to select the link directions. Thus, we define the buffered traffic ratio ϖ_c as

$$\varpi_c = \frac{Z_c^{\mathrm{dl}}}{Z_c^{\mathrm{dl}} + Z_c^{\mathrm{ul}}}, \tag{G.8}$$

where Z_c^{dl} and Z_c^{ul} are the aggregated buffered traffic size in the DL and UL directions, respectively. Herein, we assume perfect knowledge of Z_c^{ul} at the BSs from the UEs buffer status reports and pending SRs, respectively. The lower ϖ_c ratio, the larger the buffered UL traffic volume, and thus, BSs select slot formats with a majority of UL symbols. For instance, at an arbitrary BS with $\varpi_c = 0.2$, the buffered UL traffic volume is 4x the buffered DL traffic, thus, BS consequently selects a slot format of DL:UL symbol ratio as $\sim 1:4$. In case there are neither new packet arrivals nor buffered traffic at an arbitrary time instant, BSs fall back to a default slot format with equal DL and UL symbol share; however, BSs do not schedule any UEs though. This way, BSs tend to rapidly adapt to the accumulating buffered traffic, equalizing both the DL and UL TDD queuing performance, i.e., ψ_{tdd} and φ_{tdd}.

In this work, the order of the DL and UL OFDM symbols during the adopted slot format(s) is evenly distributed with a block size of 4 symbols, e.g., a selected slot pattern of $\sim 2:1$ DL:UL symbol ratio is configured as: [DDDDFUUUUDDDDF]. Such configuration allows for alternating DL and UL transmission opportunities during each slot duration for urgent packet arrivals; however, it comes at the expense of inflicting more guard symbols, i.e., [F] symbols, among each DL and UL symbol pair.

5 URLLC Outage Latency Assessment

We evaluate the URLLC radio performance using inclusive system level simulations [7], where the major functionalities of the physical and media access control layers, respectively, are implemented according to the latest 5G-NR specifications. The default simulation assumptions are listed in Table G.1, unless otherwise mentioned. We consider asynchronous Chase-combining HARQ, where the HARQ re-transmissions are dynamically scheduled and always prioritized over new transmissions. Finally, the URLLC outage latency, i.e., radio latency at the 10^{-5} outage probability, is assessed under various 5G system configurations.

URLLC outage latency with pattern update periodicity γ:

The pattern update periodicity implies how frequent the BSs update their corresponding slot formats, hence, how fast they adapt the network capacity towards the sporadic DL/UL packet arrivals. For instance, $\gamma = 1$ slot denotes that BSs update their adopted DL and UL symbol patterns per every slot duration, i.e., 14 OFDM symbols. Fig. G.3 holds a comparison of the

Table G.1: Default simulation parameters.

Parameter	Value
Environment	3GPP-UMA, one cluster, 21 cells
UL/DL channel bandwidth	20 MHz, SCS = 30 KHz, TDD
Antenna setup	$N_t = 4$, $M_r = 4$
UL power control	$\alpha = 1, P0 = -103$ dBm
Link adaptation	Adaptive modulation and coding
UE processing time	DL : 4.5/9-OFDM symbols UL : 5.5/11-OFDM symbols
Average user load per cell	$K^{\text{dl}} = K^{\text{ul}} = 1$, 10, 50, 100 and 200
TTI configuration	4-OFDM symbols
Traffic model	FTP3 $f^{dl} = f^{ul}$ = 400 bits $\lambda^{\text{ul}} = \lambda^{\text{dl}}$ = 100 pkts/sec
Interference conditions	Interference-free
DL/UL scheduling	Proportional fair; UL GF [12]
DL/UL receiver	LMMSE-IRC
Pattern update periodicity	1 radio frame (10 ms)

DL/UL combined URLLC outage latency under FDD, TDD with $\gamma = 1$ slot and a single frame duration, respectively, and for 20 MHz bandwidth. An equivalent FDD bandwidth allocation is also adopted, i.e., 10 MHz for UL transmissions and 10 MHz for DL transmissions. As can be clearly seen, at the lower load, i.e., $\Omega = 0.5$ Mbps, both duplexing schemes under evaluation achieve the 1-ms URLLC latency target. Although, by increasing the offered load up to $\Omega = 2.5$ Mbps, the FDD significantly outperforms the respective TDD, in terms of the URLLC outage latency due to the immediate availability of the DL and UL capacity, i.e., no TDD delays inflicted, and hence, $\psi_{\text{tdd}} = \varphi_{\text{tdd}} = 0$ ms. Accordingly, the TDD with a $\gamma = 1$ slot achieves a greatly improved URLLC outage latency, i.e., -218.7% outage latency reduction compared to the case of $\gamma = 1$ radio-frame, because of the faster link adaptation to the random DL/UL packet arrivals, leading to less traffic buffering delays. However, this comes with a significantly increased control overhead size, due to the guard time duration between each consecutive DL and UL symbol pair.

URLLC outage latency with dynamic and grant-free UL:

Based on the latency analysis in Section 3, grant-free UL has been demonstrated to significantly reduce the URLLC UL outage latency, compared to DG. Accordingly, Fig. G.4 depicts the complementary cumulative distribution function (CCDF) of the URLLC DL/UL combined latency when UL grant-free and DG are adopted. Herein, with DG UL, UEs transmit the

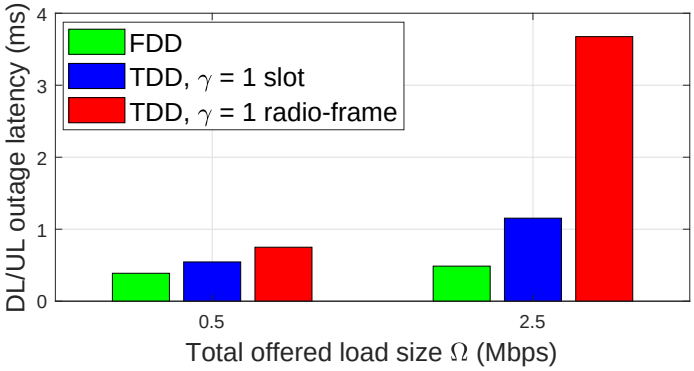


Fig. G.3. URLLC outage latency: with γ.

scheduling request on a periodicity of 16 TTIs, and hence, receive the corresponding scheduling grant 4 TTIs later. As noticed, the DG UL exhibits a linear offset in the UL outage latency by the additional latency component φ_{dg}, leading to $\sim +400\%$ increase in the URLLC outage latency, compared to the GF UL case, with $\varphi_{\text{dg}} = 0$ ms.

URLLC outage latency with the SCS size ρ:

The size of the channel SCS has a critical impact on the URLLC outage latency. Unlike the 4G standards, the 5G-NR specs adopt different SCSs for its diverse service classes, i.e., $\rho = 15$, 30, and 60 kHz, respectively, for the carrier frequencies below 6 GHz. However, it was recently agreed within the 3GPP community that $\rho = 15$ kHz is no longer appropriate for URLLC transmissions. Accordingly, the achievable URLLC UL outage latency with the SCS size is presented in Fig. G.5, for different offered loads Ω. The larger SCS size, i.e., $\rho = 60$ kHz, offers: (a) reduced BS and UE processing delays, i.e., Λ_{bsp} and Λ_{uep}, due to the shorter OFDM symbols in time, and (2) a higher probability of non-segmented URLLC transmissions, i.e., URLLC payload is transmitted in a single-shot without segmentation, reducing the DL ψ_{q} and UL φ_{q} buffering delays, respectively. Accordingly, a larger ρ allows for faster URLLC transmissions to compensate for the additional DL and UL switching delay of the dynamic-TDD systems, satisfying the stringent URLLC 1-ms outage latency.

URLLC outage latency with the TTI size μ:

The TTI length determines the packet transmission periodicity. Hence, it has a key impact on the maximum alignment delay that an arbitrary packet may inflict until the first available DL/UL TTI instance. As shown in Fig. G.6, the empirical CDF (ECDF) of the average scheduling delay of the combined DL ψ_{td} and UL φ_{td} transmissions, is introduced for $\mu = 4$, 7, and 14 OFDM symbols, respectively. Hence, the scheduling delay defines the delay

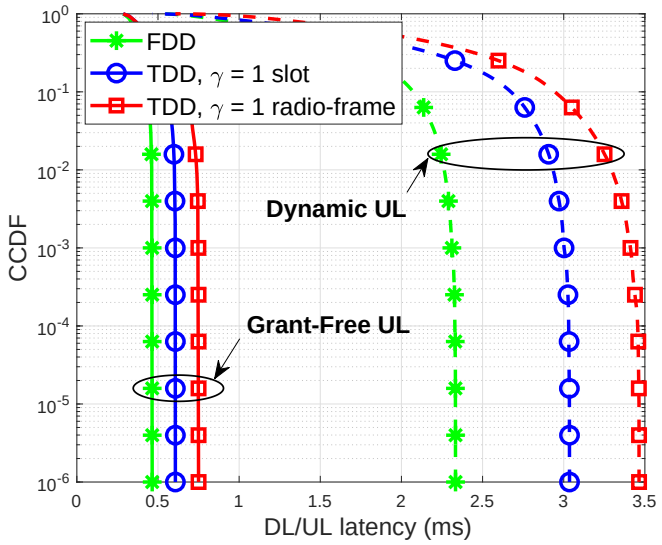


Fig. G.4. URLLC outage latency: with DG, GF, $\Omega = 0.5$ Mbps.

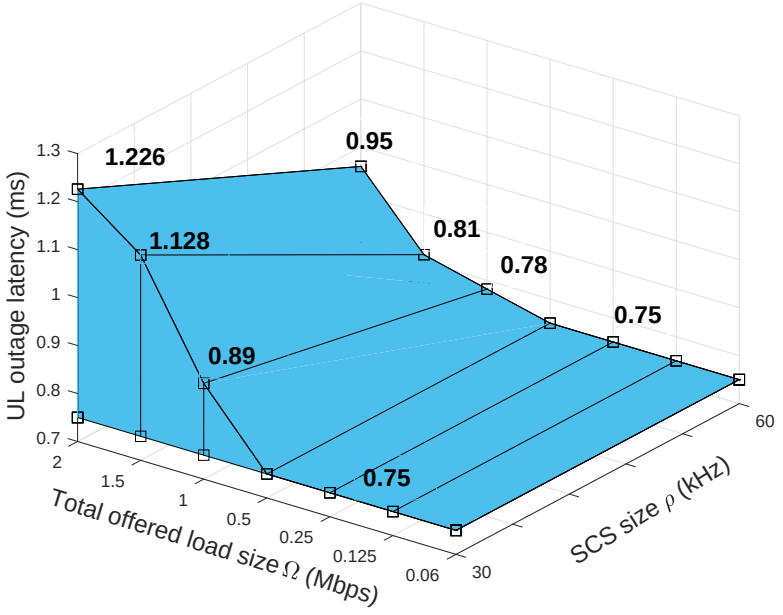


Fig. G.5. URLLC UL outage latency: with ρ.

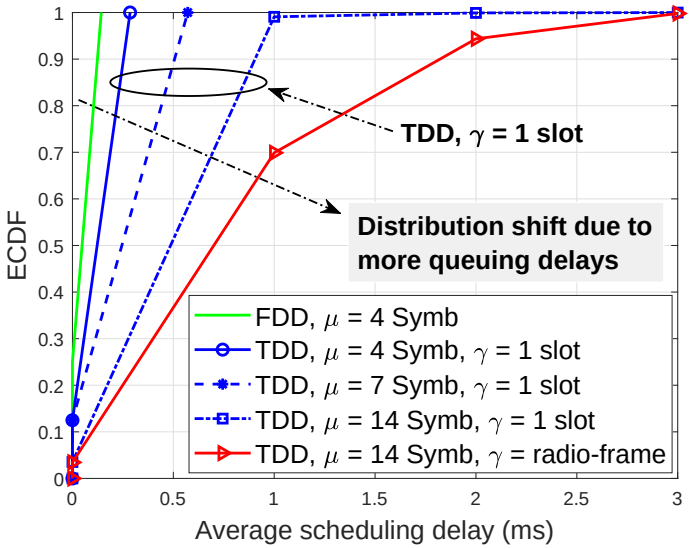


Fig. G.6. URLLC outage latency: with μ, $\Omega = 1$ Mbps.

between the time instant a packet arrives at the scheduling buffers until it is being transmitted, excluding the processing times. Obviously, the larger μ, the larger the time delay of which the incoming packets shall exhibit in the scheduling buffers. The TDD case with $\gamma = 1$ radio-frame and $\mu = 14$ OFDM symbols clearly provides the worst scheduling delay performance because of the slower traffic adaptation periodicity γ and the large TTI alignment delay, respectively. However, the FDD mode inflicts a lower scheduling delay due to the absence of the TDD switching delay, i.e., $\psi_{\text{tdd}} = \varphi_{\text{tdd}}, = 0$ ms.

URLLC outage latency with the inter-BS CLI:

The inter-BS CLI is considered as the most critical challenge against the 5G-NR dynamic-TDD systems. In this regard, Fig. G.7 depicts the CCDF of the URLLC UL latency with the FDD, and TDD duplexing, under CLI-non-free and CLI-free conditions, respectively. The latter case denotes a theoretical baseline, i.e., optimal inter-BS CLI cancellation is assumed, to which we compare the actual performance of the dynamic-TDD systems with CLI coexistence. The URLLC outage latency with unhandled CLI exhibits $+162.19\%$ increase compared to the CLI-free case. This is mainly because of the UL packets getting re-transmitted several times prior to a successful decoding, due to the severe BS-BS CLI, leading to significantly large φ_{td} and $\alpha\varphi_{\text{harq}}$ delays. On another side, the FDD case provides the best UL outage latency, mainly due to the absolute absence of the inter-BS CLI.

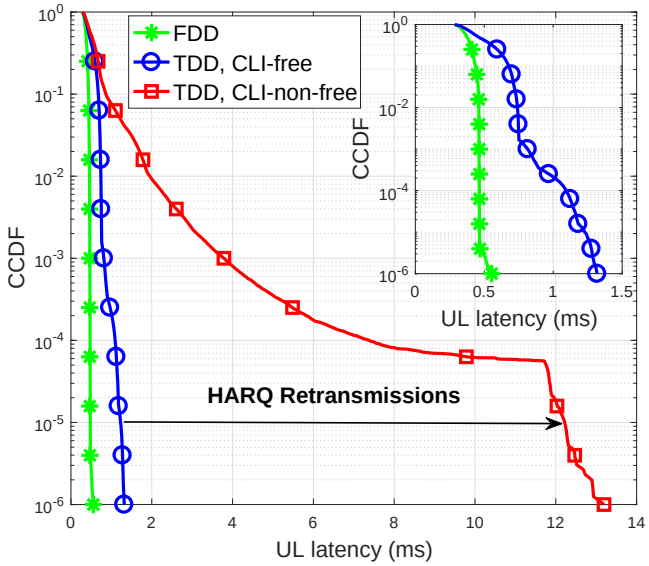


Fig. G.7. URLLC outage latency: with CLI for $\Omega = 1.5$ Mbps.

6 Discussions on state-of-the-art flexible-FDD

5G-NR dynamic-TDD systems offer a flexible link direction adaptation to the sporadic URLLC packet arrivals. However, throughout the paper, it has been demonstrated an extremely challenging task to achieve the URLLC outage latency and reliability targets in such systems.

In order to overcome the latency challenges of the dynamic-TDD operation, we next briefly consider the option of flexible-FDD operation for unpaired carriers. With the flexible-FDD, a single unpaired carrier frequency is utilized such that some PRBs are used for DL transmissions, while others are concurrently adopted for UL transmissions, as depicted by Fig. G.8. Herein, unlike the dynamic-TDD mode, simultaneous DL and UL transmissions are allowed, while still dynamically adjusting the amount of DL and UL frequency resources in line with the BS-specific link selection criterion. For instance, a BS with a buffered traffic ratio ϖ_c of 3:1, adopts 60% : 20% DL-to-UL PRB ratio, while the remaining frequency resources are flexibly configured as guard bands. The main advantages of the flexible-FDD over dynamic-TDD mode are as follows: (a) absence of the DL and UL switching delays, i.e., $\psi_{\text{tdd}} = \varphi_{\text{tdd}}, = 0$ ms, and (b) absence of the inter-BS CLI by simpler frequency coordination techniques.

However, flexible-FDD requires efficient self-interference mitigation techniques in practice, in order to cope with the power leakage problem, resulting

from the concurrent DL transmissions and UL receptions over the same PRB set. Accordingly, the self-interference mitigation operation is typically implemented as a hybrid process of analog interference suppression and digital interference cancellation. In that sense, a possible variant of a flexible-FDD deployment would therefore be to have BSs operating in the flexible-FDD mode, while connected UEs operate in half-duplex mode, either having an uplink or downlink link activation at time. Thereby, each BS shall simultaneously serve different UEs in opposite/same link directions over partially or fully shared frequency resources; though, without the need of self-interference mitigation capabilities at the UE-side.

7 Concluding Remarks

In this work, we studied the feasibility of the URLLC outage latency within the 5G new radio dynamic-TDD deployments. The URLLC radio performance is first evaluated under optimal interference-free conditions, with the various system design aspects of the 5G new radio, i.e., offered sporadic packet arrivals, channel sub-carrier spacing, transmission time interval duration, configured and grant-free uplink scheduling. Then, the impact of the inter-cell cross link interference on the achievable URLLC outage latency is identified. Finally, the state-of-the-art flexible-FDD duplexing mode is being introduced towards the upcoming 3GPP standards.

The main insights brought by this paper are summarized as follows: (1) with inter-BS interference-controlled conditions, the 30 kHz sub-carrier spacing (SCS) is proven suitable to satisfy the URLLC 1-ms radio latency target for offered loads up to 1 Mbps/BS, (2) with higher offered load levels, the SCS of 60 kHz and bandwidth allocation of 20 MHz should be adopted to further reduce the packet segmentation delay, user scheduling delay, TTI duration, UE and BS processing delays, (3) dynamic UL scheduling, the BS and UE processing delays, respectively, introduce a constant delay offset in the URLLC outage latency regardless of the other system design variants, and hence, they should be particularly optimized and (4) adding the BS-BS cross-link interference, the URLLC latency targets are almost not feasible due to the UL capacity blockage.

8 Acknowledgments

This work is partly funded by the Innovation Fund Denmark – File: 7038-00009B.

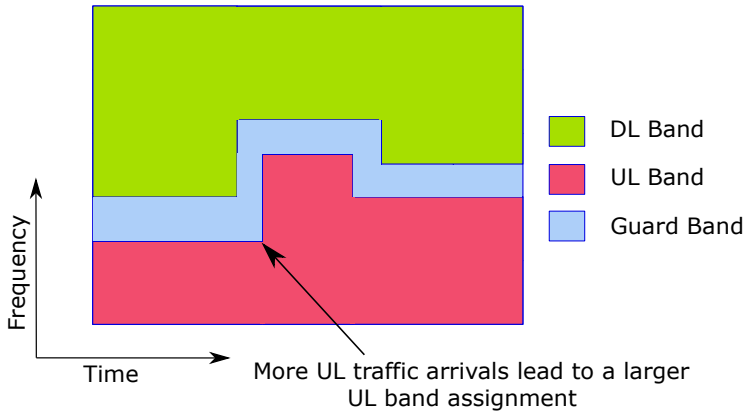


Fig. G.8. Flexible-FDD towards upcoming 3GPP 5G standards.

References

[1] Service requirements for the 5G system; Stage-1 (Release 16), 3GPP, TS 22.261, V16.6.0, Dec. 2018.

[2] Service requirements for cyber-physical control applications in vertical domains; Stage-1 (Release 16), 3GPP, TS 22.104, V16.0.0 , Dec. 2018.

[3] 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. 38451-38463, July 2018.

[4] J. Lee et al., "Spectrum for 5G: global status, challenges, and enabling techs," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 12-18, March 2018.

[5] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen and A. Szufarska, "A flexible 5G frame structure design for FDD cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53-59, March 2016.

[6] *5G; NR; Physical layer procedures for control*; (Release 15), 3GPP, TS 38.213, V15.3.0, Oct. 2018.

[7] Ali A. Esswie, K.I. Pedersen, and P. Mogensen, “Quasi-dynamic frame coordination for ultra- reliability and low-latency in 5G TDD systems,” *in Proc. IEEE ICC*, Shanghai, May 2019.

[8] A. Łukowa and V. Venkatasubramanian, "Performance of strong interference cancellation in flexible UL/DL TDD systems using coordinated muting, scheduling and rate allocation," *in Proc. IEEE WCNC*, Doha, 2016, pp. 1-7.

[9] Ali A. Esswie, and K.I. Pedersen, "Inter-cell radio frame coordination scheme based on sliding codebook for 5G TDD systems," *in Proc. IEEE VTC-spring,* Kuala Lumpur, 2019.

[10] R1-1701146, *Dynamic TDD interference mitigation concepts in NR*, Nokia, Alcatel-Lucent Shanghai Bell, 3GPP RAN1 #88, Feb. 2017.

[11] Ali A. Esswie, and K.I. Pedersen, "Cross link interference suppression by orthogonal projector in 5G dynamic-TDD URLLC systems," *Submitted to IEEE Commun. Lett*, 2019.

[12] R. Abreu, T. Jacobsen, K. Pedersen, G. Berardinelli, and P. Mogensen, "System level analysis of eMBB and grant-free URLLC multiplexing in uplink," *in Proc. IEEE VTC-spring*, Kuala Lumpur, 2019.

Paper H

Semi-Static Radio Frame Configuration for URLLC Deployments in 5G Macro TDD Networks

A. A. Esswie, K. I. Pedersen, and Preben E. Mogensen

The paper has been published in the
2020 IEEE Wireless Communications and Networking Conference (WCNC)

Abstract

Dynamic time division duplexing (TDD) is one of the major novelties of the 5G new radio standard. It notably improves the network resource utilization with sporadic directional packet arrivals. Although, the feasibility of the ultra-reliable and low-latency communications (URLLC) within such deployments is critically challenged, mainly due to the cross-link interference (CLI). In this work, we propose a semi-static and computationally-efficient TDD radio frame adaptation algorithm for 5G macro deployments. Particularly, we first identify the quasi-static variance of the cross-cell traffic buffering performance, with various CLI co-existence conditions. Accordingly, a common radio frame pattern is dynamically estimated based on the filtered multi-cell traffic statistics. Our system-level simulation results show that the proposed solution achieves a highly improved URLLC outage performance, i.e., offering $\sim$ 40% reduction gain of the achievable URLLC outage latency compared to perfect static-TDD, and approaching the optimal interference-free flexible-TDD case; though, with a significantly lower control overhead size.

Index Terms— Dynamic TDD; 5G new radio; URLLC; Traffic; Cross link interference (CLI).

1 Introduction

Ultra-reliable and low latency communication (URLLC) is the major service class of the upcoming fifth generation new radio (5G-NR) standards [1], where it enables a new set of cutting-edge and real-time applications over wireless mediums, e.g., interactive tactile-internet. URLLC entails sporadic radio transmissions of a small payload size, with stringent radio latency and reliability targets of one-way radio latency of 1 millisecond with a 99.999% success probability [2]. Most of the 5G-NR deployments are envisioned to be with the time division duplexing (TDD) due to its large spectrum availability [3]. Achieving the URLLC targets are particularly challenging for TDD systems [4] because of: (a) the non-concurrent downlink (DL) and uplink (UL) transmission opportunities, and (b) additional cross-link interference (CLI) among neighboring base-stations (BSs) adopting opposite transmission directions. Those challenges are particularly non-trivial for wide-area macro deployments and are the focus of this paper.

The 5G-NR has defined a flexible slot format design [5], where the TDD adaptation periodicity can be slot-based, i.e., in principal, per every 14 orthogonal frequency division multiplexing (OFDM) symbols. Thus, the DL/UL link switching delay is minimized down to less than a millisecond. However, the CLI still remains a critical capacity limitation of the flexible TDD deployments. In particular, for a macro setting, the DL-to-UL CLI, i.e., BS-BS CLI, is most problematic due to the power imbalance between the DL interfering

transmissions and the UL victim receptions.

In this study, we focus on achieving the URLLC-alike requirements for macro deployments at frequency range one (FR1), i.e. radio frequency (RF) operation below 7 GHz. For FR1, neighboring spectrum chunks are expected to be allocated for different operators. Accordingly, inter-operator co-existence must be considered, especially to handle the TDD inter-frequency interference. A study of the 5G-NR TDD RF co-existence was recently completed by 3GPP, concluding that fully-flexible and uncoordinated TDD deployments are not possible for FR1 macro deployments due to the severe BS-BS CLI [6], hence, recommending that operators must adopt fully-aligned TDD radio frame configurations (RFCs) to avoid the harmful inter-frequency CLI.

Furthermore, even for single-operator cases, co-channel CLI has been identified as a severe problem for macro deployments, leading the use of fully-dynamic TDD to be further challenging. Various methods to partially handle the co-channel CLI problem have therefore been proposed in the open literature. Those include CLI cancellation techniques [7-10] through inter-cell coordinated user scheduling, joint transceiver design, power control, and beam-forming. Simpler quasi-dynamic and opportunistic CLI avoidance schemes are also introduced based on hybrid RFC design [11-13]. However, although those techniques offer performance gain, the CLI problem remains non-negligible, and particularly harmful for URLLC use cases due to the strict requirements of the achievable latency and reliability. Needles to say, a simpler, overhead-limited and CLI-free adaptive RFC selection algorithm is still vital for 5G macro TDD deployments.

In this paper, a semi-static and fully-aligned RFC selection algorithm is proposed for 5G-NR TDD networks. Proposed solution offers CLI-free TDD transmissions while semi-statically adjusting the RFCs to manage the tail of the latency-reliability distribution of the experienced user performance, as the primary performance indicator for URLLC use cases. The cell-specific traffic load metrics are exchanged across coordinating cells, and are filtered to either adapt the upcoming RFCs to the average or individual cell outage performance. Hence, the proposed solution adaptively controls the tail distribution of the cluster capacity and latency, which contributes towards achieving a decent URLLC outage latency. As the RFC selection is NP-hard problem for multi-user URLLC deployments, we evaluate the performance of the proposed solution by means extensive dynamic system-level simulations to achieve results with high degree of realism. That is, we consider a dynamic multi-cell, multi-user environment in line with the 5G-NR specifications, and following the 3GPP simulation modeling guidelines, e.g. relying on advanced stochastic models for radio propagation, traffic generation, etc.). Care is taken to achieve trustworthy and statistical-reliable performance results as a basis for drawing conclusions.

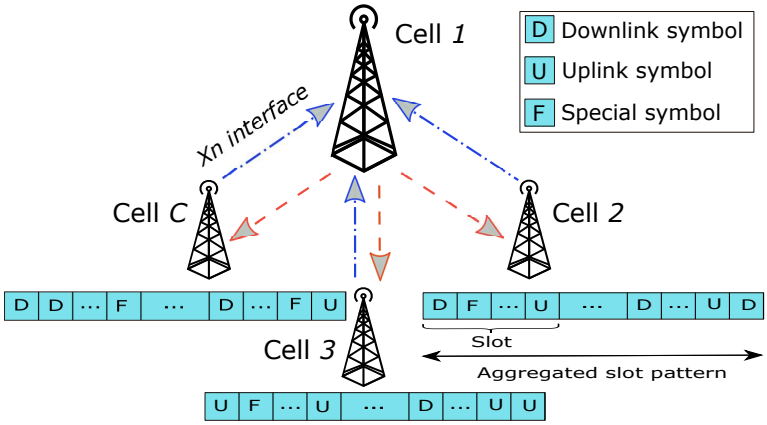


Fig. H.1. Flexible-TDD network deployment.

This paper is organized as follows. Section 2 presents our system setting. Section 3 discusses the problem formulation addressed by this work, while Section 4 introduces the proposed scheme. The performance evaluation appears in Section 5. Finally, conclusions are drawn in Section 6.

2 System Model

We assume a macro 5G-NR TDD network deployment, with C cells, each with N_t antennas. There exits an average number of K^{dl} and K^{ul} uniformly-distributed DL and UL active user-equipment's (UEs) per cell, each equipped with M_r antennas. Herein, we consider the FTP3 traffic model with payload sizes f^{dl} and f^{ul} bits, and Poisson Point Arrival Process, with mean packet arrivals λ^{dl} and λ^{ul}, for DL/UL links. Hence, the total offered average load per cell is given as: $\Omega = \Omega^{\text{dl}} + \Omega^{\text{ul}}$, with $\Omega^{\text{dl}} = K^{\text{dl}} \times f^{dl} \times \lambda^{\text{dl}}$, and $\Omega^{\text{ul}} = K^{\text{ul}} \times f^{ul} \times \lambda^{\text{ul}}$ as the average DL and UL offered load sizes, respectively.

We follow the latest 3GPP specifications for the 5G-NR TDD system design. Particularly, the 5G-NR flexible TDD slot format structure [5] is considered, as depicted by Fig. H.1. A slot format implies a certain placement of DL [D], UL [U] and flexible [F] symbols each 14-OFDM slot duration. In this work, we assume an even distribution of the DL and UL symbols over the slot in terms of a 4-symbol block size. For instance, a selected slot format with a DL:UL symbol ratio of 2 : 1 would be: [**DDDDFUUUUDDDDF**]. This configuration allows for sparse DL and UL transmission opportunities during a slot; though, at the expense of increased guard overhead, i.e., [**F**] symbols. During each slot, UEs are dynamically multiplexed using the OFDM access (OFDMA), with 30 kHz sub-carrier-spacing (SCS) and a physical resource block (PRB) of 12 consecutive SCs. The dynamic user scheduling is

performed based on the proportional fair (PF) criterion, and with a transmit time interval (TTI) duration of 4-OFDM symbols, for rapid URLLC radio transmissions. The achievable one-way outage URLLC latency is the main performance indicator of this work. It encompasses the delay from the moment the URLLC packet becomes available at the packet data convergence protocol (PDCP) layer until it has been successfully decoded, including the BS and UE processing delay, hybrid automatic repeat request (HARQ) retransmission delay, and scheduling buffering delay, respectively, in line with [4]. For UL transmissions, we assume a fast dynamic grant (DG) [4], where the UL packets become immediately available for scheduling upon availability at the UE PDCP layer. That is, the time from transmitting the UL scheduling request until receiving the DL scheduling grant is assumed negligible.

Lets define $\mathfrak{B}_{\text{dl}}$, $\mathfrak{B}_{\text{ul}}$, $\mathcal{K}_{\text{dl}}$ and $\mathcal{K}_{\text{ul}}$ as the sets of cells and UEs with active DL and UL transmissions, respectively. Hence, the DL received signal at the k^{th} UE, where $k \in \mathcal{K}_{\text{dl}}$, $c_k \in \mathfrak{B}_{\text{dl}}$, is given by

$$\mathrm{y}^{\text{dl}}_{k,c_k} = \underbrace{\mathbf{H}^{\text{dl}}_{k,c_k}\mathbf{v}_k s_k}_{\text{Useful signal}} + \mathfrak{T}^{\text{dl}}_k + \mathbf{n}^{\text{dl}}_k, \tag{H.1}$$

where $\mathbf{H}^{\text{dl}}_{k,c_k} \in \mathcal{C}^{M_r \times N_t}$ is the DL spatial channel from the cell serving the k^{th} UE, to the k^{th} UE, $\mathbf{v}_k \in \mathcal{C}^{N_t \times 1}$, and s_k are the single-stream pre-coding vector at the c_k^{th} cell, and data symbol of the k^{th} UE, respectively. $\mathbf{n}^{\text{dl}}_k$ is the additive white Gaussian noise, while $\mathfrak{T}^{\text{dl}}_k$ denotes the total interference seen at the k^{th} UE receiver end. Then, $\mathfrak{T}^{\text{dl}}_k$ is expressed by

$$\mathfrak{T}^{\text{dl}}_k = \begin{cases} \underbrace{\sum_{i \in \mathcal{K}_{\text{dl}} \setminus k} \mathbf{H}^{\text{dl}}_{k,c_i}\mathbf{v}_i s_i}_{\text{DL-to-DL interference}}, & \text{Aligned-TDD} \\ \underbrace{\sum_{i \in \mathcal{K}_{\text{dl}} \setminus k} \mathbf{H}^{\text{dl}}_{k,c_i}\mathbf{v}_i s_i}_{\text{DL-to-DL interference}} + \underbrace{\sum_{j \in \mathcal{K}_{\text{ul}}} \mathbf{G}_{k,j}\mathbf{w}_j s_j}_{\text{UL-to-DL interference}}, & \text{Flexible-TDD} \end{cases}, \tag{H.2}$$

where $\mathbf{w}_j \in \mathcal{C}^{M_r \times 1}$ is the pre-coding vector at the j^{th} UE, and $\mathbf{G}_{k,j} \in \mathcal{C}^{M_r \times M_r}$ is the the cross-link channel between the k^{th} and j^{th} UEs. Similarly, the received UL signal at the c_k^{th} cell, where $c_k \in \mathfrak{B}_{\text{ul}}$ from $k \in \mathcal{K}_{\text{ul}}$, is given as

$$\mathrm{y}^{\text{ul}}_{c_k,k} = \underbrace{\mathbf{H}^{\text{ul}}_{c_k,k}\mathbf{w}_k s_k}_{\text{Useful signal}} + \mathfrak{T}^{\text{ul}}_{c_k} + \mathbf{n}^{\text{ul}}_{c_k}, \tag{H.3}$$

with the total UL interference $\mathfrak{T}^{\text{ul}}_{c_k}$ calculated by

$$\mathfrak{T}_{c_k}^{\mathrm{ul}} = \begin{cases} \underbrace{\sum_{j \in \mathcal{K}_{\mathrm{ul}} \setminus k} \mathbf{H}_{c_k,j}^{\mathrm{ul}} \mathbf{w}_j s_j}_{\text{UL-to-UL interference}}, & \text{Aligned-TDD} \\ \underbrace{\sum_{j \in \mathcal{K}_{\mathrm{ul}} \setminus k} \mathbf{H}_{c_k,j}^{\mathrm{ul}} \mathbf{w}_j s_j}_{\text{UL-to-UL interference}} + \underbrace{\sum_{i \in \mathcal{K}_{\mathrm{dl}}} \mathbf{Q}_{c_k,c_i} \mathbf{v}_i s_i}_{\text{DL-to-UL interference}}, & \text{Flexible-TDD} \end{cases}, \tag{H.4}$$

where $\mathbf{Q}_{c_k,c_i} \in \mathcal{C}^{N_t \times N_t}$ is the cross-link channel between the cells serving the k^{th} and i^{th} UEs, $k \in \mathcal{K}_{\mathrm{ul}}$ and $i \in \mathcal{K}_{\mathrm{dl}}$, and it is measured by orchestrating inter-BS coordinated sounding measurements [10]. Accordingly, the achievable post-processing signal-to-interference (SIR) ratio in the DL direction γ_k^{dl} and UL direction $\gamma_{c_k}^{\mathrm{ul}}$ are given by

$$\gamma_k^{\mathrm{dl}} = \frac{\left\| \left(\mathbf{u}_k^{\mathrm{dl}}\right)^{\mathrm{H}} \mathbf{H}_{k,c_k}^{\mathrm{dl}} \mathbf{v}_k \right\|^2}{\sum\limits_{i \in \mathcal{K}_{\mathrm{dl}} \setminus k} \left\| \left(\mathbf{u}_k^{\mathrm{dl}}\right)^{\mathrm{H}} \mathbf{H}_{k,c_i}^{\mathrm{dl}} \mathbf{v}_i \right\|^2 + \sum\limits_{j \in \mathcal{K}_{\mathrm{ul}}} \left\| \left(\mathbf{u}_k^{\mathrm{dl}}\right)^{\mathrm{H}} \mathbf{G}_{k,j} \mathbf{w}_j \right\|^2}, \tag{H.5}$$

$$\gamma_{c_k}^{\mathrm{ul}} = \frac{\left\| \left(\mathbf{u}_k^{\mathrm{ul}}\right)^{\mathrm{H}} \mathbf{H}_{c_k,k}^{\mathrm{ul}} \mathbf{w}_k \right\|^2}{\sum\limits_{j \in \mathcal{K}_{\mathrm{ul}} \setminus k} \left\| \left(\mathbf{u}_k^{\mathrm{ul}}\right)^{\mathrm{H}} \mathbf{H}_{c_k,j}^{\mathrm{ul}} \mathbf{w}_j \right\|^2 + \sum\limits_{i \in \mathcal{K}_{\mathrm{dl}}} \left\| \left(\mathbf{u}_k^{\mathrm{ul}}\right)^{\mathrm{H}} \mathbf{Q}_{c_k,c_i} \mathbf{v}_i \right\|^2}, \tag{H.6}$$

with $\|\bullet\|^2$ as the second-norm, and $\mathbf{u}_k^{\kappa} \in \mathcal{C}^{N_t/M_r \times 1}$, $\mathcal{X}^{\kappa}, \kappa \in \{\mathrm{ul}, \mathrm{dl}\}$, is the linear minimum mean square error interference rejection combining (LMMSE-IRC) receiver [14], and $(\bullet)^{\mathrm{H}}$ denotes the Hermitian operation.

3 Problem Formulation

The URLLC outage performance is dominated by the achievable radio latency at the lower 10^{-5} outage probability. This implies a stringent latency bound with a rare violation occurrence. Thus, in TDD deployments, due to the non-concurrent DL and UL transmission, the URLLC latency and reliability targets become highly susceptible to the number and placement of the DL d_c and UL u_c symbols during an RFC. Accordingly, our objective is to optimize the RFC selection in order to minimize the URLLC outage radio latency as

$$\begin{gathered} \left(\frac{d_c}{u_c}\right)^* \triangleq \left\{ \frac{d^i}{u^i} \ : \ \frac{d^i}{u^i} \in \mathfrak{T} \right\} \\ \text{s.t: } \arg\min_k \left(\varphi_{c,k}\right), \forall k \in \mathcal{K}_{\mathrm{ul/dl}}, \end{gathered} \tag{H.7}$$

where $\mathfrak{T}$ is the set of all pre-defined possible RFC structures, $\varphi_{c,k}$ is the one-way URLLC radio latency [4]. Accordingly, there is no feasible optimal solution of the DL $d_c^{\text{opt.}}$ and UL $u_c^{\text{opt.}}$ symbol structure to satisfy the UE-specific latency and reliability requirements. For instance, in multi-UE URLLC deployments, and due to the time-variant sporadic traffic arrivals, multiple UEs may request simultaneous opposite link directions. Thus, BSs instead adapt the RFC structure, on a best effort basis, to offer faster transmissions of the UEs with the worst latency performance while buffering other UEs. Adding the severe BS-BS CLI on top, victim UL packets most likely inflict several HARQ re-transmissions before a successful decoding, violating the UE-specific latency budget as well as dictating the RFC adaptation by pending packets rather than the new packet arrivals. Thus, to tackle this issue, we propose a semi-static coordinated RFC selection algorithm, which offers fully CLI-free transmissions while adapting the RFC selection to the varying URLLC latency statistics.

4 Proposed Coordination Scheme

We propose a computationally-efficient RFC selection algorithm to offer a decent URLLC outage latency performance. First, cells estimate their average directional traffic size on a pre-defined periodicity. Then, a relative load metric is shared among the coordinating cells over the back-haul Xn-interface, i.e., multiple bits of feedback. Subsequently, a filtering window is applied on the reported traffic data-set, either to match the average traffic volume per cluster, i.e., equal-priority windowing, or biasing the RFC adaptation towards individual cells, e.g., typically those with the worst traffic buffering performance. Then, a common RFC is estimated to match the filtered traffic volume, and accordingly, cells within the cluster adopt the same RFC until the next RFC update instant.

4.1 Cell-specific directional traffic tracking

Cells seek to select the RFCs which minimize the achievable average URLLC outage latency, according to (H.7). In multi-user URLLC networks, there may exist several active UEs with simultaneous UL and DL transmission requests, respectively, and hence, cross-directional target latency conflict is exhibited.

For the considered URLLC use cases, the incoming traffic is only allowed to be buffered for a short time-duration before being transmitted. Otherwise, the URLLC latency constraint is violated. We can therefore observe the strong correlation between the amount of buffered data, i.e., queuing of data payloads, and the experienced latency. Selecting the RFC that offers the best URLLC outage performance is therefore translated to selecting the RFC

which minimizes the DL and UL buffering. As the offered traffic increases, buffering is obviously unavoidable, thus, the best feasible solution from the RFC selection point of view is to ensure that the traffic buffering of the two link directions is balanced.

Accordingly, at the ϱ^{th} slot of the radio frame, $\varrho = 1, 2, \ldots, \rho$, with ρ as the number of slots per radio frame, the c^{th} cell calculates the aggregated DL $Z_c^{\mathrm{dl}}(\varrho)$ and UL $Z_c^{\mathrm{ul}}(\varrho)$ buffered traffic size, respectively. Specifically, the UL traffic volume is identified at the cell side from the UE scheduling requests, and the associated buffer status reports. Thus, the normalized traffic ratio $\mu_c(\varrho)$ is defined as

$$\mu_c(\varrho) = \frac{Z_c^{\mathrm{dl}}(\varrho)}{Z_c^{\mathrm{dl}}(\varrho) + Z_c^{\mathrm{ul}}(\varrho)}. \tag{H.8}$$

Then, the instantaneous traffic ratios $\mu_c(\varrho)$ are linearly averaged across each frame duration as expressed by

$$\overline{\mu}_c = \frac{1}{\rho} \sum_{\varrho=1}^{\rho} \mu_c(\varrho), \tag{H.9}$$

where $\overline{\mu}_c$ implies the averaged traffic ratio of the c^{th} cell. In case there are neither DL and UL new packet arrivals or buffered traffic, BSs fall-back to a default RFC structure until the next update instant. The larger the $\overline{\mu}_c$, e.g., ~ 1, the larger the buffered DL traffic compared to the corresponding UL traffic volume. For instance, with $\overline{\mu}_c = 0.9$, the DL traffic volume is 9x the respective UL volume. Finally, per every RFC adaptation period, neighboring cells exchange the measured $\overline{\mu}_c$ over the Xn-interface.

4.2 Traffic filtering and common RFC selection

As the URLLC outage performance is mainly dominated by the cells of the worst latency and reliability performance, we apply a window filtering to dynamically control the URLLC latency tail distribution, i.e., outage latency. Thus, based on the exchange of $\overline{\mu}_c$, the most problematic cells are first identified as those inflicting the largest or smallest $\overline{\mu}_c$, i.e., having too large DL or UL buffered traffic volume, which implies the RFC adaptation is not properly configured for those victim cells. In this regard, cells calculate the absolute linear distance of the reported $\overline{\mu}_c$ data-set towards its mean value d, i.e., $d = 0.5$, and then, sort them in an descending order in terms of their respective absolute linear distance, as given by

$$\boldsymbol{\Psi} = \underset{x^{(a)}, y^{(b)}; a>b}{\mathbf{Sort}} \left[\overline{\mu}_1^{(|\overline{\mu}_1 - d|)}, \overline{\mu}_2^{(|\overline{\mu}_2 - d|)}, \ldots, \overline{\mu}_C^{(|\overline{\mu}_C - d|)} \right]. \tag{H.10}$$

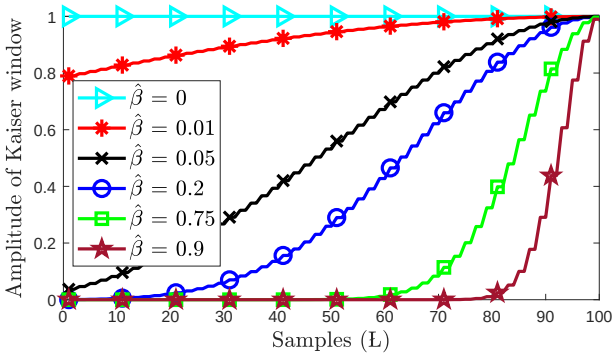


Fig. H.2. Mirrored Kaiser window with $\hat{\beta}$.

The ordered traffic data-set $\Psi = [\psi_1, \psi_2, \ldots, \psi_C]$ is then filtered using a spatial window. In this work, we consider the Kaiser window $w[l]$, due to its flexible response tunability, and is given in the discrete domain as

$$w[l] = \frac{I_0\left[\beta\sqrt{1-\left(\frac{2l}{L}-1\right)^2}\right]}{I_0[\beta]}, \; 0 \leq l \leq L \tag{H.11}$$

with I_0 as the zeroth-order modified Bessel function of the first kind, β is the window shaping factor, and $L+1$ denotes the window length. As depicted by Fig. H.2, the mirrored Kaiser window amplitude is shown for various normalized shaping factors $\hat{\beta} = \frac{\beta}{\beta_{\max}}$, with $\beta_{\max} = 100$ and $L = 100$. The larger the $\hat{\beta}$ factor, the more selective the Kaiser window. For instance, with $\hat{\beta} = 0$, the mirrored Kaiser window approaches a conventional band pass filter, i.e., all cells are equally prioritized; although, with a larger $\hat{\beta} = 0.9$, the window becomes highly selective over a subset of the sample space, i.e., certain cells are highly prioritized.

Accordingly, the Kaiser window coefficients are applied in a descending order on the sorted data-set Ψ, as

$$\Theta = \frac{\psi_1 w[0] + \psi_2 w[1] + \cdots + \psi_C w[L]}{w[0] + w[1] + \cdots + w[L]}, \; L = C - 1, \tag{H.12}$$

where Θ is the filtered traffic ratio per the entire cluster, with $w[0] > w[1] > \cdots > w[L]$. Based on the calculated Θ, a common RFC is selected and adopted by all cells within the cluster until the next RFC update instant. For example, with an estimated $\Theta = 0.2$, a common RFC of $d_c/u_c \simeq \frac{1}{4}$ is adopted across all cells, with the DL/UL symbol placement configured according to the strategy presented in Section 2.

4.3 Comparison to the state-of-the-art TDD studies

We compare the performance of the proposed solution against the state-of-the-art TDD solutions in the recent literature as follows:

Static-TDD (sTDD): a pre-defined RFC is globally configured for all cells across the entire network, where it matches the average network traffic demand. Herein, we define α as the normalized RFC mismatch error, where $\alpha = 0$ implies the global RFC is selected to perfectly match the average network traffic statistics and $\alpha = 0.35$ denotes 35% symbol mismatch of the configured RFC against the actual average offered traffic load. sTDD deployments offer CLI-free conditions; though, with a limited cross-cell traffic adaptation flexibility.

Dynamic TDD (dTDD): a fully flexible TDD operation is assumed, where at each RFC update period, each cell independently adopts the RFC which best meets its individual traffic demand. We consider two scenarios of the dTDD deployments as: (a) a dTDD setting with an optimal CLI cancellation (dTDD-CLI-free) [9], where the BS-BS and UE-UE are perfectly suppressed using full packet exchange over both the back-haul and radio interfaces, and (b) a dTDD deployment with CLI coexistence (dTDD-CLI).

5 Performance Evaluation

We assess the performance of the proposed solution using extensive system-level simulations, with a high degree of realism. The major simulation settings are listed in Table H.1, where the main assumptions of the 3GPP release-15 for TDD deployments are adopted. During every RFC update periodicity, cells estimate their buffered traffic ratio, according to (H.8), and hence, share it among the cluster in order to estimate a common RFC. Thus, during each TTI, DL and UL UEs are dynamically multiplexed using OFDMA based on the PF metric. The signal-to-interference-noise-ratio (SINR) points of the individual SCs are calculated by the LMMSE-IRC receiver, and combined into an effective SINR level using the exponential SNR mapping [15]. Finally, we adopt a dynamic link adaptation, i.e., adaptive modulation and coding selection, and asynchronous HARQ Chase combining, where the HARQ retransmissions are dynamically scheduled, and are always prioritized over new transmissions.

Fig. H.3 depicts the complementary cumulative distributive function (CCDF) of the DL/UL combined URLLC one-way radio latency in ms, of the proposed scheme, sTDD, and dTDD, respectively, for $\Omega = 1$ Mbps. Looking at the achievable latency at the 10^{-5} probability level, i.e., URLLC outage latency, the proposed solution clearly offers a decent URLLC outage performance, approaching the optimal dTDD-CLI-free. It achieves $\sim$ 20% and

Table H.1: Simulation parameters.

Parameter	Value
Environment	3GPP-UMA, one cluster, 21 cells
UL/DL channel bandwidth	10 MHz, SCS = 30 KHz, TDD
Carrier frequency	3.5 GHz
TDD mode	Synchronized
Antenna setup	$N_t = 8$, $M_r = 2$
Average user load per cell	$K^{\text{dl}} = K^{\text{ul}} = 10$
TTI duration	4-OFDM symbols
Traffic model	FTP3, $f^{dl} = f^{ul}$ = 400 bits λ^{dl} =125, and 375 pkts/sec λ^{ul} =125, and 375 pkts/sec
Offered load ratio	DL:UL = 1:1
Processing time	Preperation delay: 3-OFDM symbols PDSCH decoding: 4.5-OFDM symbols PUSCH decoding : 5.5-OFDM symbols
RFC update periodicity	10 ms (radio frame)
UL/DL receiver	LMMSE-IRC
Link adaptation	Adaptive modulation and coding
HARQ configuration	asynchronous with Chase Combining

$\sim$ 40% reduction of the URLLC outage latency, compared to the sTDD with a perfect RFC match, i.e., $\alpha = 0$ and non-perfect RFC match, i.e., $\alpha = 0.35$, respectively. However, the proposed solution inflicts $\sim$ 10% URLLC outage latency degradation against the ideal dTDD-CLI-free; although, this comes with a significantly lower control signaling overhead size. In that sense, the proposed solution relaxes one of the most challenging requirements of the conventional sTDD schemes, as it does not require a pre-configured RFC, while still preserving fully CLI-free transmissions with a semi-static traffic adaptation.

Furthermore, the dTDD-CLI-free achieves the best URLLC outage latency, i.e., $\sim$ 1.78 ms, due to the fully flexible, i.e., cell-wise, RFC adaptation to the individual sporadic traffic arrivals. Though, it comes with the assumption of optimal CLI-free conditions. The dTDD-CLI scheme exhibits an outage latency saturation, since the URLLC radio performance becomes dictated by the aggressive BS-BS CLI, instead of the RFC adaptation, leading to the consumption of the maximum HARQ attempts before UL packets are either dropped or successfully received after the Chase combining HARQ process.

Fig. H.4 shows the CCDF of the URLLC latency of the proposed solution, with different $\hat{\beta}$ settings, and $\Omega = 3$ Mbps. As can be noticed, with $\hat{\beta} = 0.9$, the tail of the URLLC latency distribution becomes more narrower, since the

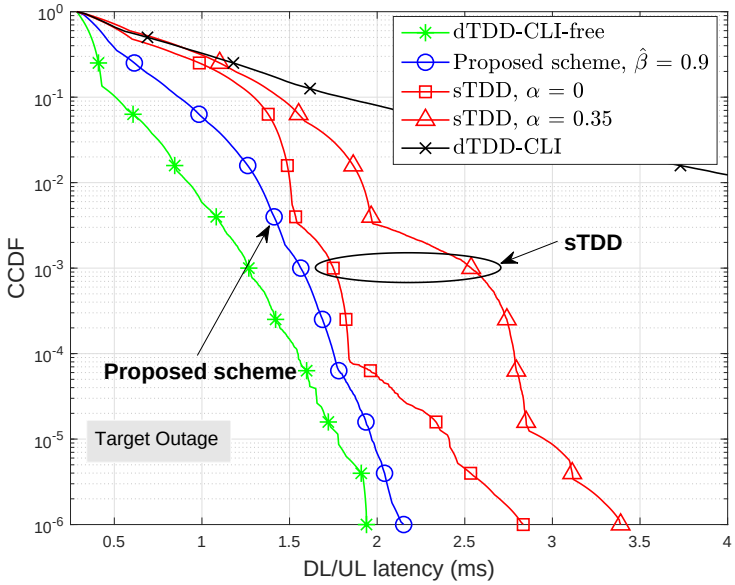


Fig. H.3. Achievable URLLC outage latency of proposed scheme, sTDD, and dTDD, with $\Omega = 1$ Mbps.

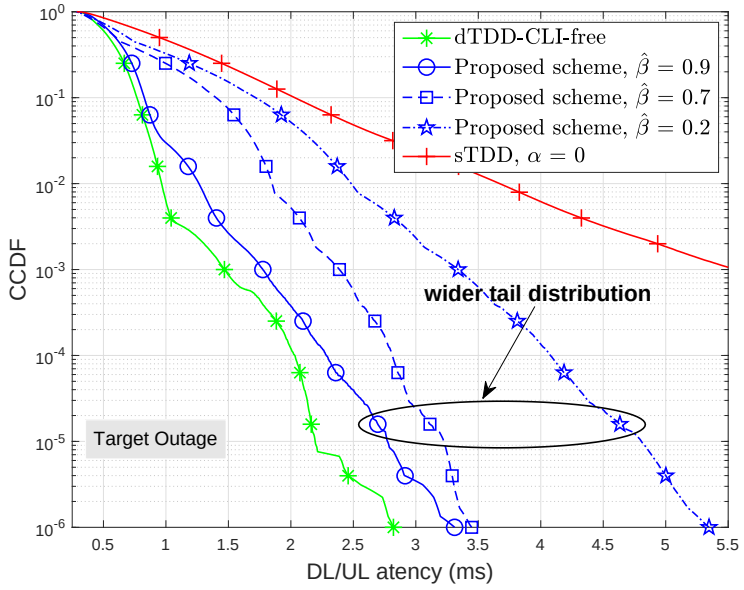
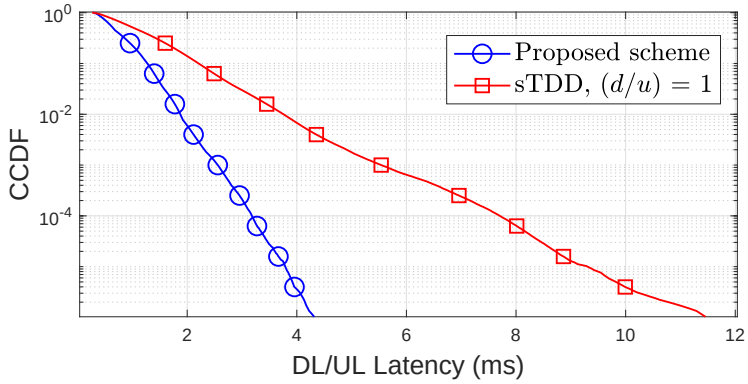


Fig. H.4. Achievable URLLC outage latency of proposed scheme, with $\hat{\beta}$ and $\Omega = 3$ Mbps.

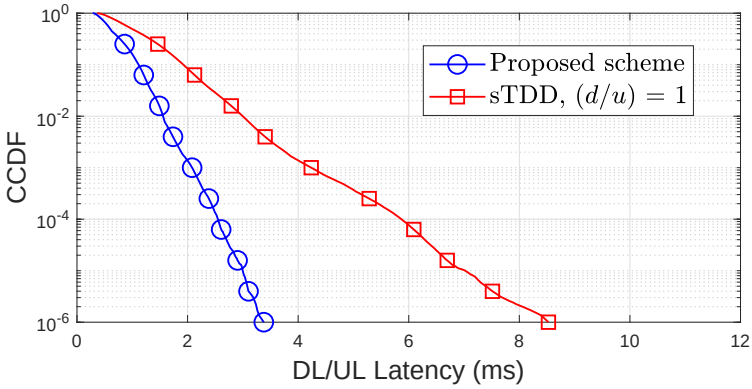
cells with the worst traffic buffering performance are highly prioritized in selecting the upcoming RFCs. With a smaller $\hat{\beta}$ factor, the reported traffic statistics of the coordinating cells are equally prioritized, leading to a wider latency tail distribution. That is $\sim +54\%$ increase in the URLLC outage latency with $\hat{\beta} = 0.2$, compared to the case with $\hat{\beta} = 0.9$. This consolidates the fact that the URLLC outage latency is dictated by the cells of the worst buffering imbalance. Hence, those should be given a higher priority when deciding the upcoming RFCs, in order to rapidly recover their respective outage targets.

Looking at the URLLC outage performance with different $\Omega^{\text{dl}}/\Omega^{\text{ul}}$ ratios, Fig. H.5 depicts the CCDF of the URLLC radio latency, under the proposed and the sTDD schemes, respectively, where the latter is configured with $d_c/u_c = 1$, for $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 3:1$ and $1:3$. The proposed solution offers a sufficient frame adaptation against the variable offered traffic ratio $\Omega^{\text{dl}}/\Omega^{\text{ul}}$, resulting in a decent URLLC outage latency, i.e., 3.8 ms for $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 3:1$ and 2.9 ms for $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 1:3$, respectively. The sTDD scheme clearly exhibits a significant outage latency increase due to the mismatch between the predefined $d_c/u_c = 1$ and the offered traffic ratio $\Omega^{\text{dl}}/\Omega^{\text{ul}}$, e.g., proposed solution offers $\sim 82\%$ outage latency reduction compared to the sTDD scheme. Accordingly, the proposed solution eliminates the rigid requirement of preconfiguring a global RFC while offering a semi-static RFC adaptation to the varying offered traffic.

Finally, Fig. H.6 depicts a comparison of the achievable combined DL/UL outage latency with various offered load levels, in reference to the dTDD-CLI-free scheme. At the very low offered region $\Omega = 0.25$ Mbps, with an average of a single active UE per cell, the traffic demand becomes highly variant among neighboring cells. Accordingly, the URLLC outage performance is dominated by how fast the cells adapt their individual RFCs to the sporadic traffic arrivals. Thus, both the proposed solution and sTDD schemes inflict a considerable outage latency degradation, compared to the optimal dTDD-CLI-free, i.e., $+29\%$ and $+63\%$ latency increase. However, the proposed scheme outperforms the corresponding sTDD by 35% outage latency reduction gain. This is mainly attributed to the semi-static cross-cell RFC adaptation of the proposed solution. Thus, unlike sTDD, cells with accumulating traffic size are given higher priority in selecting the RFCs. Over the high load region $\Omega = 5$ Mbps, similar conclusions are observed; although, with less outage latency relative degradation compared to the optimal dTDD-CLI-free, since in this case, the URLLC outage performance is mainly dictated by the scheduling queuing delay rather than the flexibility of the RFC adaptation operation.



(a) DL-heavy ($\Omega^{\text{dl}} : \Omega^{\text{ul}} = 3 : 1$)



(b) UL-heavy ($\Omega^{\text{dl}} : \Omega^{\text{ul}} = 1 : 3$)

Fig. H.5. Comparison of the URLLC latency performance with various DL and UL traffic ratios, $\Omega = 3$ Mbps.

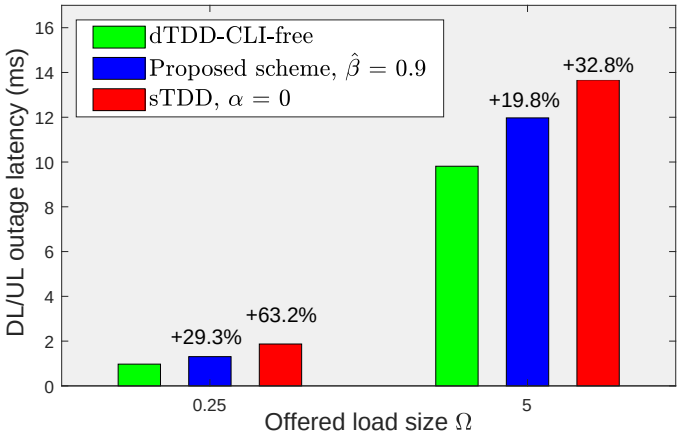


Fig. H.6. Comparison of the URLLC latency performance with various offered load levels Ω.

6 Concluding Remarks

A semi-static radio frame configuration (RFC) selection algorithm has been proposed for 5G TDD macro deployments. Proposed solution incorporates a simple inter-cell signaling exchange procedure of the relative traffic statistics, in order to estimate a common RFC of each cluster, which matches the time-variant and cell-specific traffic demand. Compared to the state-of-the-art TDD literature, the proposed solution demonstrates an attractive trade-off between the achievable URLLC outage performance and the signaling overhead size. It achieves $\sim$ 40% reduction of the URLLC outage latency, compared to the ideal static-TDD deployment, while approaching the optimal dynamic-TDD bound; though, with a significantly lower signaling overhead size.

The main insights brought by this paper are as follows: (a) within macro 5G new radio deployments, the cross-cell traffic statistics are of a low time-variance due to the sufficiently large number of active connected users, (b) accordingly, relaxing the requirements of fast traffic adaptation for the sake of controlling the critical cross-link interference (CLI) is of a more significance, and (c) the proposed solution offers a flexible and semi-static RFC adaptation to the sporadic cross-cell traffic demand, with fully CLI-free conditions and limited signaling overhead.

7 Acknowledgments

This work is partly funded by the Innovation Fund Denmark – File: 7038-00009B. 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] IMT vision – "Framework and overall objectives of the future development of IMT for 2020 and beyond", international telecommunication union (ITU), ITU-R M.2083-0, Feb. 2015.

[2] Service requirements for the 5G system; Stage-1 (Release 16), 3GPP, TS 22.261, V16.6.0, Dec. 2018.

[3] J. Lee et al., "Spectrum for 5G: global status, challenges, and enabling techs," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 12-18, March 2018.

[4] Ali A. Esswie, and K.I. Pedersen, "On the ultra-reliable and low-latency communications in flexible TDD/FDD 5G networks," *in Proc. IEEE CCNC,* Las Vegas, 2020.

[5] *5G; NR; Physical layer procedures for control*; (Release 15), 3GPP, TS 38.213, V15.3.0, Oct. 2018.

[6] *Cross link interference handling and remote interference management (RIM) for NR*; (Release 16); 3GPP, TR 38.828, V16.0.0, June 2019.

[7] A. Łukowa and V. Venkatasubramanian, "Performance of interference cancellation in flexible TDD systems using coordinated muting, scheduling and rate allocation," *in Proc. IEEE WCNC*, Doha, 2016, pp. 1-7.

[8] Z. Huo, N. Ma and B. Liu, "Joint user scheduling and transceiver design for cross-link interference suppression in MU-MIMO dynamic TDD systems," *in Proc. IEEE ICCC,* Chengdu, 2017, pp. 962-967.

[9] R1-1701146, *Dynamic TDD interference mitigation concepts in NR*, Nokia, Alcatel-Lucent Shanghai Bell, 3GPP RAN1 #88, Feb. 2017.

[10] Ali A. Esswie, and K.I. Pedersen, "Cross link interference suppression by orthogonal projector in 5G dynamic-TDD URLLC systems," *Submitted to IEEE Commun. Lett,* 2019.

[11] Ali A. Esswie, K.I. Pedersen, and P. Mogensen, “Quasi-dynamic frame coordination for ultra- reliability and low-latency in 5G TDD systems,” *in Proc. IEEE ICC*, Shanghai, China, 2019, pp. 1-6.

[12] J. W. Lee, C. G. Kang and M. J. Rim, "SINR-ordered cross link interference control scheme for dynamic TDD in 5G system," *in Proc. IEEE ICOIN*, Chiang Mai, 2018, pp. 359-361.

[13] A. A. Esswie and K. I. Pedersen, "Inter-cell radio frame coordination scheme based on sliding codebook for 5G TDD systems," *in Proc. IEEE VTC*, Kuala Lumpur, Malaysia, 2019, pp. 1-6.

[14] Tavares, F.M.L.; Berardinelli, G.; Mahmood, N.H.; Sorensen, T.B.; Mogensen, P., "On the potential of interference rejection combining in B4G networks," *in Proc. IEEE VTC*, Las Vegas, NV, 2013, pp. 1-5.

[15] 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 Trans. Wireless Commun.*, vol. 10, no. 10, pp. 3436–3448, Oct. 2011.

Paper I

Inter-Cell Radio Frame Coordination Scheme Based on Sliding Codebook for 5G TDD Systems

A. A. Esswie, and K. I. Pedersen

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

Abstract

The fifth generation (5G) of the wireless communication networks supports wide diversity of service classes, leading to a highly dynamic uplink (UL) and downlink (DL) traffic asymmetry. Thus, dynamic time division duplexing (TDD) technology has become of a significant importance, due to its radio frame flexibility. However, fully dynamic TDD systems suffer from potentially severe inter-cell cross link interference (CLI). In this paper, we propose a novel inter-cell radio frame coordination (RFC) scheme based on sliding codebook for fully dynamic TDD 5G networks. Proposed coordination scheme simultaneously addresses two optimization objectives of minimizing the average CLI while reliably maximizing the achievable DL/UL capacity, by virtually extending the RFC degrees of freedom through a sliding phase-offset RFC codebook design. Compared to the state-of-the-art TDD studies, the proposed scheme shows significantly improved ergodic capacity, i.e., at least $\sim$ 140% *gain under both the TCP and UDP protocols, and with much less signaling overhead, limited to* B-bit. *The paper offers valuable insights about how to most efficiently pre-mitigate potential CLI in Macro TDD systems.*

Index Terms— Dynamic TDD; 5G new radio; Cross link interference (CLI); Traffic; TCP; UDP.

1 Introduction

Time division duplexing (TDD) technology has drawn a major research attention since day-one of the long term evolution (LTE) development. The 3rd generation partnership project (3GPP) LTE-Advanced Rel-12 introduces an enhanced interference mitigation and traffic adaptation (eIMTA) [1, 2] to offer a more flexible TDD adaptation. eIMTA supports seven different TDD radio frame configuration (RFC) patterns with different downlink (DL) to uplink (UL) traffic ratios, where each cell dynamically in time adapts its radio frame based on its own link direction selection criteria, e.g., aggregated traffic demand [3]. Though, the fastest possible RFC adaptation periodicity of eIMTA is the LTE radio frame, i.e., 10 ms.

With the agile frame structure of the 5G new radio (5G-NR) [4], e.g., the flexible TDD slot formats and the variable transmission time interval (TTI) duration, a fully dynamic TDD with much faster and flexible adaptation becomes feasible. Accordingly, the link direction switching periodicity can be slot-based, i.e., $\leq$ 1 ms, instead of being RFC-based. Thus, 5G-NR TDD systems significantly improve the spectrum utilization and the ergodic capacity for services with fast-varying and asymmetric DL and uplink UL traffic [5]. However, the coexistence of different link directions over same frequency resources in adjacent cells results in potential cross link interference (CLI), i.e., user-to-user (UE-UE), and base-station to base-station (BS-BS) interference

[6]. In Macro deployments, the CLI, especially the BS-BS interference, is a critical problem due to the UL and DL power imbalance. Consequently, the gains of the adaptive RFCs in TDD may completely vanish due to severe CLI [7].

The state-of-the-art CLI suppression proposals from the open literature consider either CLI avoidance or post-cancellation. In [8], the combination of cell muting, liquid clustering and enhanced UL power control is suggested to minimize the average UE-UE and BS-BS CLI. Additionally, joint UE scheduling and advanced beam-forming techniques [9, 10] are envisioned as beneficial to counteract the CLI. Furthermore, a performance case study on the interaction of the transmission control protocol (TCP) with the 5G-NR TDD systems is presented [11]. In [12], a recent proposal introduces perfect CLI cancellation using full packet exchange, where DL-heavy cells signal neighboring UL-heavy cells with their respective DL UE transmission information for the UL-heavy cells to optimally cancel the critical BS-BS CLI.

Compared to the state-of-the-art coordinated TDD studies, significant inter-cell control signaling overhead and/or ideal periodic UE CLI measurements are usually assumed, which are infeasible in practice. Consequently, the overall capacity gains from the TDD RFC flexibility can be greatly limited due to cell muting or the abrupt changes in the joint scheduling decisions. Needless to say, an efficient and flexible coordination scheme is vital for macro TDD systems, to simultaneously improve the overall ergodic capacity in both UL and DL directions and with limited signaling overhead.

In this work, we propose an RFC based sliding codebook (RFCbCB) coordination scheme for 5G TDD systems. The proposed scheme effectively boosts the TDD system degrees of freedom, coming from its frame flexibility, with the size of a specially pre-designed RFC codebook. Consequently, the maximum possible ergodic capacity is achieved while simultaneously guaranteeing acceptable CLI levels and with a significantly reduced inter-cell control signaling overhead, limited to B-bit. Extensive system level simulations show that the proposed RFCbCB scheme significantly improves the ergodic capacity by the efficient CLI avoidance in both DL and UL directions simultaneously. Moreover, as various applications require different link reliability levels, e.g., TCP is commonly used with the 5G-NR enhanced mobile broadband service class and user data-gram protocol (UDP) with latency critical traffic, we evaluate the proposed scheme performance over both transport protocols to study the effect of the CLI on the TCP flow and congestion controls, respectively.

Due to the complexity of the 5G-NR [4] and addressed problem herein, we evaluate the performance by extensive system simulations, where the main TDD functionalities are calibrated against the 3GPP 5G-NR assumptions. This includes UL and DL channel modeling, dynamic modulation and coding schemes (MCS), dynamic hybrid automatic repeat request (HARQ),

and dynamic UE scheduling.

This paper is organized as follows. Section 2 presents the system model of this work. Section 3 introduces the problem formulation while Section 4 details the proposed RFCbCB coordination scheme and the reference studies to compare against. Section 5 discusses the performance evaluation results and paper is finally concluded in Section 6.

2 System Model

We consider a 5G-NR TDD system with a single cluster of C cells, each with N_t antennas. Each cell serves an average of K^{dl} and K^{ul} uniformly-distributed DL and UL UEs, each with M_r antennas. Without loss of generality, we assume the FTP3 traffic modeling with finite payload sizes f^{dl} and f^{ul} bits, and Poisson point arrival processes λ^{dl} and λ^{ul}, in the DL and UL directions, respectively. Thus, the total offered traffic load per cell in DL direction is: $K^{\text{dl}} \times f^{dl} \times \lambda^{\text{dl}}$ and in UL direction: $K^{\text{ul}} \times f^{ul} \times \lambda^{\text{ul}}$, respectively.

In the time domain, we assume an RFC of 10 sub-frames, each is 1-ms and can be either a DL, UL or special sub-frame. In the frequency domain, UEs are dynamically multiplexed by the orthogonal frequency division multiple access (OFDMA), with the smallest schedulable unit as the physical resource block (PRB) of 12-subcarriers, each is 15 kHz. Thus, a sub-frame is one slot of 14-OFDM symbols. Nonetheless, the proposed solution is also valid with different numerologies of the 5G-NR sub-carrier spacing, TTI duration, and number of TDD slots per sub-frame, respectively.

Within each cluster, an arbitrary master cell is declared where other cells act as slaves. Such master cell can be manually pre-configured since it is independent from time and the coordination technology. All cells within each cluster are assumed to be bi-bidirectionally inter-connected to the master cell through the *Xn interface*, as shown in Fig. I.1.

Let $\mathfrak{B}_{\text{dl}}$, $\mathfrak{B}_{\text{ul}}$, $\mathcal{K}_{\text{dl}}$ and $\mathcal{K}_{\text{ul}}$ indicate the sets of cells and UEs in the DL and UL transmission modes, respectively. Then, the received signal at the k^{th} UE, where $k \in \mathcal{K}_{\text{dl}}$, $c_k \in \mathfrak{B}_{\text{dl}}$, is

$$\mathrm{y}^{\text{dl}}_{k,c_k} = \underbrace{\mathbf{H}^{\text{dl}}_{k,c_k}\mathbf{v}_k s_k}_{\text{Useful signal}} + \underbrace{\sum_{i\in\mathcal{K}_{\text{dl}}\setminus k} \mathbf{H}^{\text{dl}}_{k,c_i}\mathbf{v}_i s_i}_{\text{BS to UE interference}} + \underbrace{\sum_{j\in\mathcal{K}_{\text{ul}}} \mathbf{G}_{k,j}\mathbf{w}_j s_j}_{\text{UE to UE interference}} + \mathbf{n}^{\text{dl}}_k, \tag{I.1}$$

where $\mathbf{H}^{\text{dl}}_{k,c_i} \in \mathcal{C}^{M_r \times N_t}$ denotes the DL channel from the cell, serving the i^{th} UE, to the k^{th} UE, $\mathbf{v}_k \in \mathcal{C}^{N_t \times 1}$ and s_k are the single-stream precoding vector at the c_k^{th} cell and data symbol of the k^{th} UE, respectively, $\mathbf{G}_{k,j} \in \mathcal{C}^{M_r \times M_r}$ is the channel between the k^{th} and j^{th} UEs. $\mathbf{w}_j \in \mathcal{C}^{M_r \times 1}$ is the single-stream

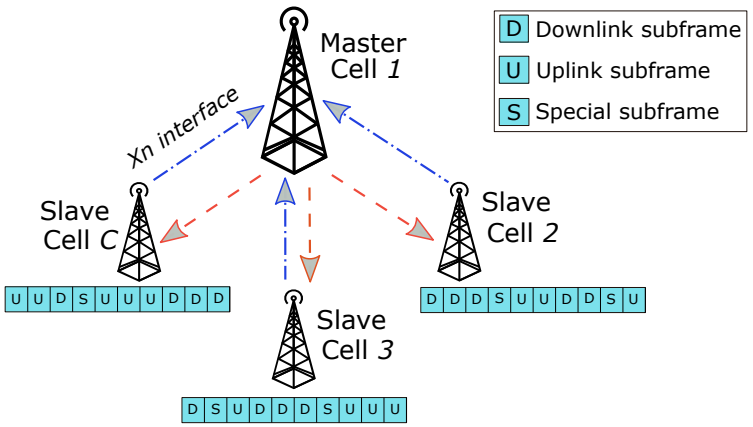


Fig. I.1. Proposed RFCbCB: system model.

precoding vector at the j^{th} UE, and $\mathbf{n}_k^{\mathrm{dl}}$ is the additive white Gaussian noise at the k^{th} UE. The first summation denotes the DL-to-DL inter-cell interference while the second summation represents the inter-cell UE-UE CLI. Similarly, the received signal at the c_k^{th} cell, where $c_k \in \mathfrak{B}_{\mathrm{ul}}$ from $k \in \mathcal{K}_{\mathrm{ul}}$, is

$$\mathrm{y}_{c_k,k}^{\mathrm{ul}} = \underbrace{\mathbf{H}_{c_k,k}^{\mathrm{ul}}\mathbf{w}_k s_k}_{\text{Useful signal}} + \underbrace{\sum_{j\in\mathcal{K}_{\mathrm{ul}}\setminus k} \mathbf{H}_{c_k,j}^{\mathrm{dl}}\mathbf{w}_j s_j}_{\text{UE to BS interference}} + \underbrace{\sum_{i\in\mathcal{K}_{\mathrm{dl}}} \mathbf{Q}_{c_k,c_i}\mathbf{v}_i s_i}_{\text{BS to BS interference}} + \mathbf{n}_{c_k}^{\mathrm{ul}}, \tag{I.2}$$

where $\mathbf{Q}_{c_k,c_i} \in \mathcal{C}^{N_t \times N_t}$ is the channel between the cells that serve the k^{th} and i^{th} UEs, respectively, where $k \in \mathcal{K}_{\mathrm{ul}}$ and $i \in \mathcal{K}_{\mathrm{dl}}$. The first summation implies the UL-to-UL inter-cell interference while the second summation denotes the inter-cell BS-BS CLI. Then, the received signal is decoded using the linear minimum mean square error interference rejection combining receiver (LMMSE-IRC) [4] matrix $\mathbf{a}$ as

$$\hat{s}_k^{\kappa} = (\mathbf{a}_k^{\kappa})^{\mathrm{H}}\,\mathrm{y}_k^{\kappa}, \tag{I.3}$$

where $\mathcal{X}^{\kappa}, \kappa \in \{\mathrm{ul}, \mathrm{dl}\}$, and $(\bullet)^{\mathrm{H}}$ denotes the Hermitian operation. Finally, the received signal-to-interference-noise-ratio (SINR) levels in the DL direction at the k^{th} UE and in the UL direction at the c_k^{th} cell, respectively, are expressed by

$$\gamma_k^{\mathrm{dl}} = \frac{p_{c_k}^{\mathrm{dl}} \left\| \mathbf{H}_{k,c_k}^{\mathrm{dl}}\mathbf{v}_k \right\|^2}{\sigma^2 + \sum\limits_{i\in\mathcal{K}_{\mathrm{dl}}\setminus k} p_{c_i}^{\mathrm{dl}} \left\| \mathbf{H}_{k,c_i}^{\mathrm{dl}}\mathbf{v}_i \right\|^2 + \sum\limits_{j\in\mathcal{K}_{\mathrm{ul}}} p_j^{\mathrm{ul}} \left\| \mathbf{G}_{k,j}\mathbf{w}_j \right\|^2}, \tag{I.4}$$

$$\gamma_{c_k}^{\mathrm{ul}} = \frac{p_k^{\mathrm{ul}} \left\| \mathbf{H}_{c_k,k}^{\mathrm{ul}} \mathbf{w}_k \right\|^2}{\sigma^2 + \sum\limits_{j \in \mathcal{K}_{\mathrm{ul}} \setminus k} p_j^{\mathrm{ul}} \left\| \mathbf{H}_{c_k,j}^{\mathrm{dl}} \mathbf{w}_j \right\|^2 + \sum\limits_{i \in \mathcal{K}_{\mathrm{dl}}} p_{c_i}^{\mathrm{dl}} \left\| \mathbf{Q}_{c_k,c_i} \mathbf{v}_i \right\|^2}, \tag{I.5}$$

where $p_{c_k}^{\mathrm{dl}}$ and p_k^{ul} are the transmission power of the c_k^{th} cell in the DL direction and the k^{th} UE in the UL direction, respectively. As can be observed from (I.5), the BS-BS CLI can significantly degrade the perceived UL SINR level due to the DL and UL power imbalance, i.e., $p_{c_i}^{\mathrm{dl}} \gg p_k^{\mathrm{ul}}$.

3 Problem Formulation - CLI Mitigation

In fully TDD systems, cells may not adopt exactly the same RFC. Thus, neighboring cells experience different transmission directions over several sub-frames, causing severe BS-BS and UE-UE CLI. Accordingly, the lower-power UL transmissions are severely degraded due to the strong CLI resulting from adjacent larger-power DL transmissions. As a result, the achievable UL capacity exhibits a significant loss, leading to more buffered UL traffic in those victim cells. Hence, these cells will be dictated by new and buffered UL traffic leading to a limited DL capacity and a highly degraded overall spectral efficiency as a consequence.

To address this problem, the proposed RFCbCB seeks to maximize the long-term ergodic capacity while simultaneously preserving limited inter-cell sub-frame misalignment, thus, an acceptable average CLI. Let u_c and d_c denote the estimated numbers of UL and DL sub-frames in an arbitrary RFC while $u_c^{\mathrm{opt.}}$ and $d_c^{\mathrm{opt.}}$ indicate the corresponding optimal numbers. Thus, we define a heuristic optimization problem as

$$R \triangleq \arg\max_{c} \sum_{c=1}^{C} \min\left(u_c, u_c^{\mathrm{opt.}}\right) F_c^u + \min\left(d_c, d_c^{\mathrm{opt.}}\right) F_c^d, \tag{I.6}$$

$$\phi_c\left(\eta_c\right) = \arg\min_{x} \frac{1}{C} \sum_{x=1, x \neq c}^{C} \varphi_{c,x}\left(\eta_c, \eta_x\right), \tag{I.7}$$

where R is the aggregate capacity of the cluster, F_c^u and F_c^d are the rate utility functions of the UL and DL transmissions in the c^{th} cell, i.e., the achievable capacity gain from having either UL or DL transmission. $\phi_c\left(\eta_c\right)$ and $\varphi_{c,x}\left(\eta_c, \eta_x\right)$ are the average and actual sub-frame misalignment of the RFC requested by the c^{th} cell η_c and between the RFCs of the c^{th} and x^{th} cells, i.e., η_c and η_x, respectively, $\forall x \neq c$. To maximize (I.6), $u_c = u_c^{\mathrm{opt.}}$ and $d_c = d_c^{\mathrm{opt.}}$ should be preserved; however, $u_c^{\mathrm{opt.}}$ and $d_c^{\mathrm{opt.}}$ may result in a large sub-frame misalignment, leading to severe CLI in the cluster and the overall capacity R shall be significantly degraded accordingly.

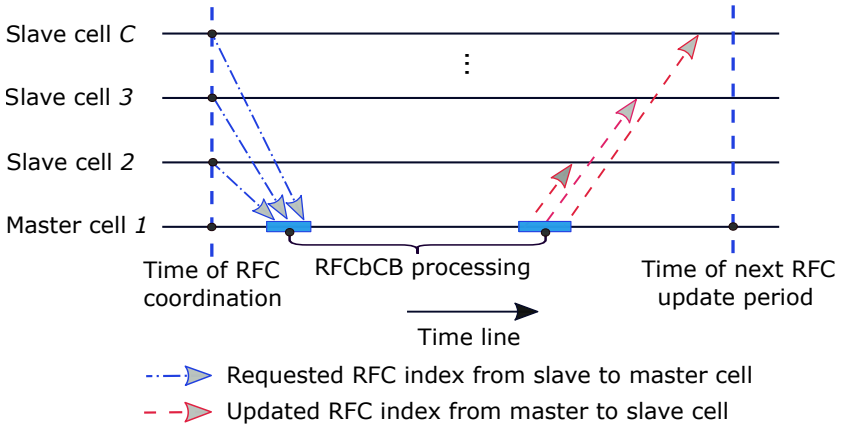


Fig. I.2. Timing diagram of the inter-cell RFC coordination.

4 Proposed RFCbCB Coordination

A specially designed RFC CB is pre-defined and assumed pre-known to all cells in each cluster. Slightly before each RFC update periodicity, each slave cell identifies its desired upcoming RFC from the CB that satisfies its link direction selection objectives. Next, it signals the master cell with the index of its desired RFC from the CB over *Xn interface*. The master BS then seeks to simultaneously satisfy both (I.6) and (I.7). Hence, the master cell may slightly change the RFC indices, which were requested by slave cells. Finally, it signals the updated RFC indices back to the slave cells, which should be used during the next RFC update period. Fig. I.2 depicts the generic timing diagram of the proposed solution.

4.1 Proposed Inter-Cell Coordination Scheme

The design of the RFC sliding codebook:

A pre-defined RFC CB of size $\mathcal{N}$ unique RFCs is constructed, where it is divided into L different sub-CBs. Each sub-CB contains RFCs with the same DL:UL sub-frame ratio in a radio frame, i.e., $d_c : u_c$; however, with a different DL and UL sub-frame placement as depicted in Fig. I.3, where each RFC is cyclic-shift of the other RFCs in the same sub-CB. The total number of RFCs, sub-CBs, and cyclic-shift in the CB are arbitrary design parameters.

At the slave cells within a cluster:

At each RFC update period, each slave cell selects an RFC from the CB that most meets its own link direction selection criteria. Without loss of generality, we consider the buffered traffic including pending re-transmissions as the main criterion with which each cell determines its required $d_c : u_c$ ratio. Let

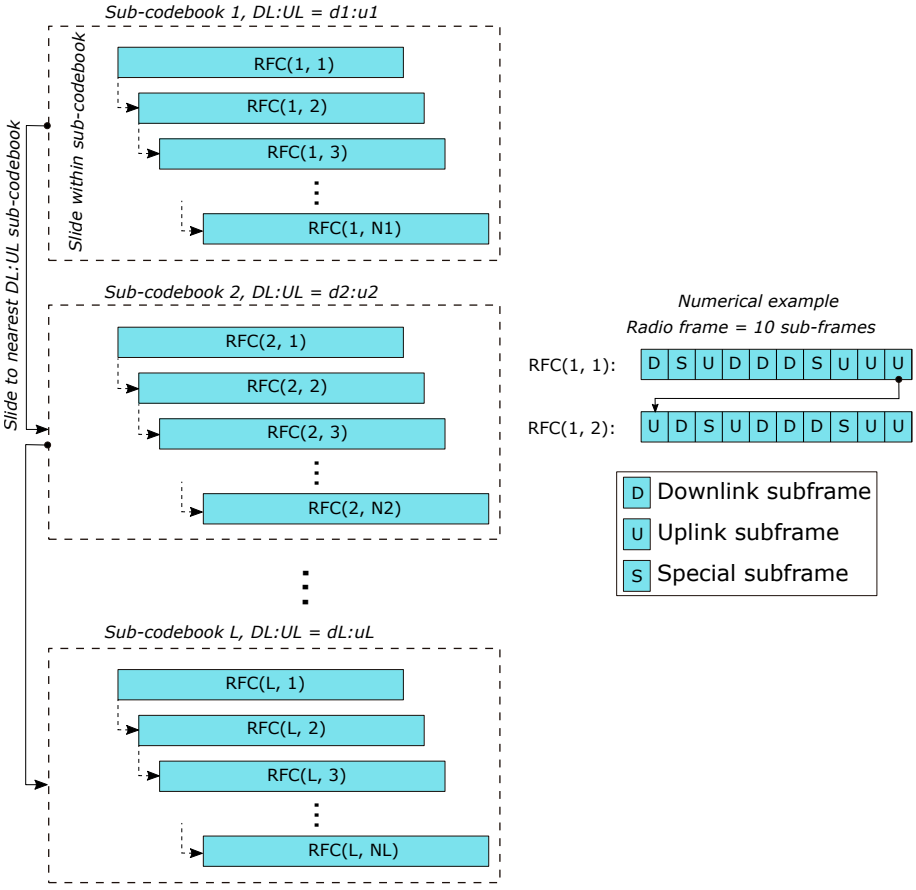


Fig. I.3. Proposed RFCbCB: RFC sliding CB design.

β_c implies the traffic load threshold of the c_k^{th} cell, as

$$\beta_c \leq \frac{\sum Z_c^{\mathrm{dl}}}{\sum Z_c^{\mathrm{dl}} + \sum Z_c^{\mathrm{ul}}}, \tag{I.8}$$

where $\sum Z_c^{\mathrm{dl}}$ and $\sum Z_c^{\mathrm{ul}}$ are the total buffered traffic in the DL and UL directions, respectively. The traffic load threshold β_c is used to bias the link direction selection to either DL or UL. For an instance, with fair $\beta_c = 0.5$, if $\sum Z_c^{\mathrm{dl}} \geq \sum Z_c^{\mathrm{ul}}$, a cell decides a DL-heavy RFC. Finally, slave cells feedback the master cell with $\mathrm{B} = \log_2(\mathcal{N})$ bits on the *Xn interface* to indicate the index of the desired RFC from the CB.

At the master cell within a cluster:

The master cell first identifies the *common RFC*, as the most requested RFC by the majority of the slave cells in the cluster. If not possible, the master cell considers any requested RFC as the common RFC of this cluster, and to which other requested RFCs from other slave cells must satisfy the minimum possible sub-frame misalignment with.

For each requested RFC η_c of the c_k^{th} cell, the master cell calculates the sub-frame misalignment to common RFC $\delta_x, \forall x \neq c$ as: $\varphi_{c,x}\left(\eta_c\left(u_c, d_c\right), \delta_x\left(u_x, d_x\right)\right)$. If $\varphi_{c,x} \leq \psi$, with ψ as a pre-defined sub-frame misalignment threshold, the master cell skips updating such requested RFC, i.e., does not change it, since it originally drives a limited CLI. Otherwise, the master cell slides over all RFCs within the same sub-CB of the requested RFC η_c. It calculates the corresponding misalignment values $\varphi_{c,x}$ and selects the RFC with the minimum sub-frame misalignment. Finally, it performs two variations:

- **Option-1**: if the sub-frame misalignment of the selected RFC is below ψ, the master cell considers such RFC as the updated RFC of this cell. Hence, an acceptable average CLI level is guaranteed across the upcoming RFC update period while still preserving the same required traffic service ratio $d_c : u_c$, improving both the slave cells and overall cluster capacity, respectively.

- **Option-2**: if $\varphi_{c,x} \leq \psi$ is not feasible with all RFCs in the same sub-CB, the master cell slides to a different sub-CB in the CB, with the nearest $d_c : u_c$ ratio to the requested ratio, e.g., $d_c : u_c = 2 : 6 \underset{\text{slide to}}{\rightarrow} d_c^{'} : u_c^{'} = 3 : 5$, and repeats the same operation. Herein, the master cell slightly sacrifices part of the full TDD RFC flexibility, due to the change in the requested $d_c : u_c$. However, such capacity loss is bounded over only a limited number of sub-frames and is reversibly proportional to the size of the CB, since the master cell slides only to the nearest $d_c : u_c$ sub-CB. As will be discussed in Section 5, this capacity loss is fully recovered on the long-term statistics due to the significantly reduced CLI.

As a last resort, if the sub-frame misalignment threshold can not be further satisfied, the master cell considers the RFC with the minimum misalignment $\varphi_{c,x}$, from either the same or different sub-CB as the requested one, as the updated RFC of this slave cell even it does not satisfy $\varphi_{c,x} \leq \psi$. Finally, the master cell feeds-back all slave cells back with B-bit indices over the *Xn interface* to indicate their respective updated RFCs to be used over the next RFC update period.

4.2 Comparison to the state-of-the-art TDD studies

We compare the performance of the proposed solution against the following state-of-the-art TDD proposals as:

Fully-uncoordinated TDD (FUC): all cells in the cluster independently select their respective RFCs from the CB based on the traffic criterion in (I.8). No inter-cell RFC coordination is assumed. Thus, a large inter-cell sub-frame misalignment and hence, severe CLI levels can be exhibited.

Ideal-UL interference coordination TDD (IUIC) [12]: within a cluster, cells independently select respective RFCs based on (I.8). Then, the DL-heavy cells feedback their respective DL payload, PRB mapping, UE MCS and precoding information to UL-heavy cells over the *Xn interface*. Accordingly, the UL-heavy cells are perfectly able to fully suppress the BS-BS CLI, i.e., BS-BS CLI = 0. Therefore, the IUIC is an UL-optimal TDD coordination scheme; however, with a significant signaling overhead over the back-haul links.

5 Performance Evaluation

The performance assessment of the proposed coordination scheme is based on highly dynamic system level simulations, where the main 3GPP assumptions are followed [4]. The major simulation setup parameters are listed in Table I.1. At each TTI, each cell dynamically and independently schedules UEs over system PRBs according to the proportional fair criterion. Herein, we assume fully dynamic link adaptation and Chase combining HARQ, respectively, where the DL/UL HARQ feedback is sent with a higher priority during the first available transmission opportunity of the adopted RFC. The sub-carrier SINR level is calculated using the LMMSE-IRC receiver. For MCS selection, sub-carrier SINR levels are combined using the effective exponential SNR mapping algorithm to obtain an effective wide-band SINR. Finally, we evaluate the performance of the proposed RFCbCB scheme under both TCP and UDP, with different offered traffic loads per cell and for the two proposed options.

Table I.1: Simulation parameters.

Parameter	Value
Environment	3GPP-UMA, one cluster, 21 cells
UL/DL channel bandwidth	10 MHz, TDD
TDD mode	Synchronized
Antenna setup	$N_t = 8$ Tx, $M_r = 2$ Rx
Average user load	$K^{\mathrm{dl}} = K^{\mathrm{ul}} = 10$ users per cell
UL/DL receiver	LMMSE-IRC
TTI configuration	1 ms (14-OFDM symbols)
HARQ	Chase combining
Link adaptation	Dynamic MCS
Traffic model	FTP3, $f^{dl} = f^{ul}$ = 4000 bits λ^{dl} = 500, 375, and 250 pkts/sec λ^{ul} = 500, 375, and 250 pkts/sec
User scheduler	Proportional fair
Offered average load per cell DL:UL	DL:UL = 2:1 (20:10) Mbps DL:UL = 1:1 (15:15) Mbps DL:UL = 1:2 (10:20) Mbps
Proposed RFCbCB setup	ψ = 3 sub-frames $\mathcal{N}$ = 55 RFCs L = 7 sub-CBs B = 6 bits
Transport layer setup	TCP/UDP max PDU: 1500 Bytes Congestion control: CUBIC Slow start threshold: 35 MSS

Table I.2: Proposed RFCbCB (option-1): achievable DL and UL throughput (Mbps) per cluster, with TCP.

Traffic Ratio	Offered load per cell (Mbps)	FUC		Proposed RFCbCB (option-1)		IUIC	
		DL	UL	DL	UL	DL	UL
DL:UL = 2:1	DL:UL = 20:10	131.52 0.0%	17.63 0.0%	167.87 +24.2%	109.23 +144.41%	228.38 +53.82%	158.61 +159.98%
DL:UL = 1:1	DL:UL = 15:15	117.73 0.0%	26.12 0.0%	149.53 +23.79%	154.39 +142.12%	191.09 +47.50%	195.16 +152.78%
DL:UL = 1:2	DL:UL = 10:20	106.07 0.0%	138.24 0.0%	113.10 +6.41%	198.65 +35.86%	139.79 +27.43%	264.6 +62.73%

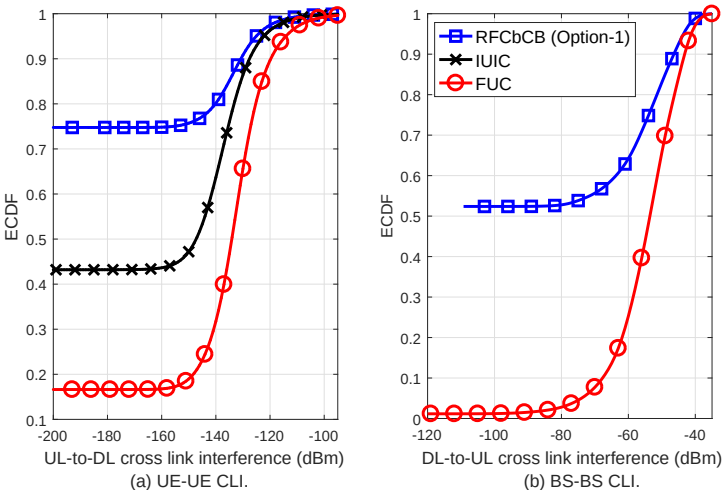


Fig. I.4. Cross link interference performance (dBm), with TCP.

Table I.2 shows the achievable DL and UL throughput per cluster under TCP of the FUC, proposed RFCbCB (option-1), and IUIC, for different DL:UL traffic ratios. As can be clearly observed, with all traffic load variations, the proposed RFCbCB (option-1) provides a significant capacity improvement in both the DL and UL directions, compared to the FUC. It also approaches the optimal IUIC, due to the significantly reduced average CLI. For instance, with a BS-BS CLI extreme case, i.e., DL:UL = 2:1, proposed RFCbCB (option-1) achieves $\sim +144.41\%$ gain in the UL capacity than the FUC. The optimal IUIC offers the best DL and UL throughput since the BS-BS CLI is assumed perfectly suppressed. Thus, UL traffic gets transmitted faster with zero CLI, i.e., UL PRBs become of higher capacity, leaving more time and resources for DL traffic. Accordingly, both UL and DL capacity are improved. The FUC exploits the full TDD RFC flexibility; however, the aggregated capacity is severely degraded due to the exhibited strong CLI levels.

The RFCbCB performance gain is mainly due to the significant reduction of the average CLI. Hence, Figs. I.4.a and I.4.b show the empirical cumulative distribution function (ECDF) of the BS-BS and UE-UE CLI, respectively, averaged over all system PRBs with DL:UL = 2:1. The proposed RFCbCB offers a highly improved CLI performance, i.e., more than 70% and 50% of the simulation time are UE-UE and BS-BS CLI-free, respectively. Compared to the optimal IUIC, the RFCbCB shows a further reduced UE-UE CLI since IUIC is only UL-optimal with no BS-BS CLI (no ECDF of the BS-BS CLI with IUIC in Fig. I.4.b). However, the proposed RFCbCB non-biasedly seeks for minimizing both BS-BS and UE-UE CLI, respectively.

Looking at the TCP performance, Fig. I.5 depicts the ECDF of the average

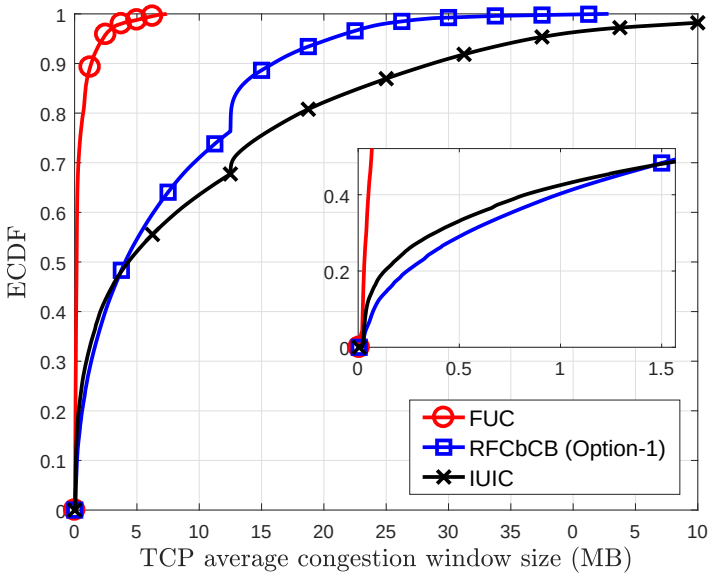


Fig. I.5. TCP average congestion control performance.

UL and DL TCP congestion window (CWND) in MB. The TCP CWND is a congestion control measure, applied at the transmitter side, to counteract network congestion. It defines the maximum rate bound that a transmitter can use towards a receiver such that it is exponentially increased when a successful TCP acknowledgment (ACK) is received, otherwise, it is decreased. Hence, the TCP transmission rate is restricted by either the transmitter CWND or the receiver advertised maximum window. Accordingly, the TCP CWND performance is highly correlated to the exhibited CLI in TDD systems. As shown in Fig. I.5, the FUC inflicts an extremely small CWND size, due to the exhibited severe CLI. The proposed RFCbCB achieves $\sim +168.25$ gain in the 90 percentile CWND size compared to the FUC. However, the optimal IUIC offers the best average CWND performance, despite that it exhibits a larger average UE-UE CLI than the proposed RFCbCB, as shown in Fig. I.4.a. This consolidates the fact that the BS-BS CLI has a stronger impact on overall capacity than the UE-UE CLI due to the power imbalance between UL and DL transmissions. Though, the proposed RFCbCB achieves an average $\sim +58.1\%$ gain in the CWND size than IUIC for the percentiles below 40%, due to the achievable $\sim -52.1\%$ reduction in the UE-UE CLI as in Fig. I.4.a. Thus, with the proposed RFCbCB, cell-edge DL UEs inflict much less CLI from adjacent inter-cell-edge UL UEs.

Furthermore, the proposed RFCbCB scheme is demonstrated as best effort since the minimum sub-frame misalignment threshold may not be satisfied

over all RFCs within the same requested $d_c : u_c$ sub-CB. Thus, we investigate cases where the master cell slides to the nearest $d_c : u_c$ sub-CB to the requested one, i.e., RFCbCB (option-2), sacrificing part of the TDD RFC flexibility due to the $d_c : u_c$ change. Fig. I.6 introduces the post-detection UL SINR, after the IRC decoding as in eq. (I.3), of the RFCbCB (option-1) (slide only within requested sub-CB), RFCbCB (option-2) (if applicable, slide to nearest sub-CB), FUC, and IUIC, with UDP on top for DL:UL = 2:1. The RFCbCB (option-1) provides substantial improvements in the UL SINR compared to the FUC, i.e., an average of $+7.9$ dB increase. Moreover, RFCbCB (option-2) further improves the perceived UL SINR level by an average of $+2.9$ dBs than RFCbCB (option-1), and with a bounded average loss of -2.7 dB to the optimal IUIC.

Interestingly, the proposed RFCbCB (option-2) further improves the overall DL and UL capacity per cluster, as depicted in Fig. I.7, closely approaching the UL-optimal IUIC. The RFCbCB (option-2) further significantly reduces the probability of the CLI occurrence than RFCbCB (option-1) by sliding to other RFC sub-CBs, at the expense of slightly changing the traffic service ratio $d_c : u_c$, requested by slave cells. Thus, the instantaneous UE rates may inflict a capacity loss, though, being limited due to the conservative $d_c : u_c$ change. However, on the average traffic statistics, the UL and DL traffic gets scheduled and successfully decoded faster due to limited CLI, and thus, an enhanced decoding ability, leaving more time and resources for incoming traffic. As a result, the total UL and DL capacity per cell is further improved.

6 Concluding Remarks

In this work, a radio frame configuration based on sliding codebook coordination scheme has been proposed for dynamic TDD 5G macro systems. The proposed solution, with its introduced variations, offer a significant capacity improvement, i.e., more than $\sim 140.0\%$ gain, under both TCP and UDP, and with a highly reduced inter-cell signaling overhead size, limited to B-bit. Compared to the state-of-the-art TDD solutions from industry and academia, the proposed scheme has been demonstrated as a flexible, high-performance and low-complexity way to control the critical cross link interference (CLI) in dynamic TDD networks.

The main insights brought by this work are summarized as: (1) the achievable capacity gains from the frame direction flexibility in dynamic TDD macro systems can fully vanish or revert to a capacity loss due to severe CLI, (2) the majority of the state-of-the-art dynamic TDD coordination schemes assumes sophisticated inter-cell communications to share the scheduling decisions, and transmission information. This leads to a significant amount of control overhead, which is infeasible in practice, and (3) proposed solu-

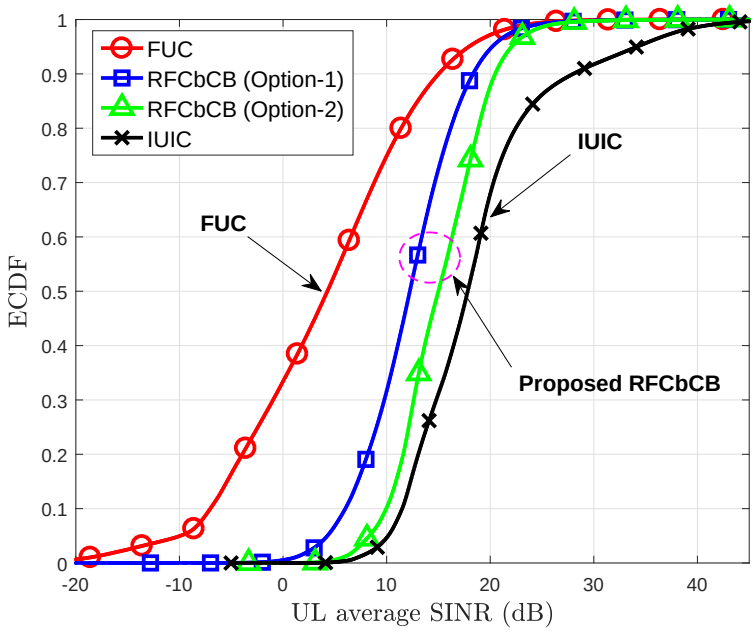


Fig. I.6. UL average SINR performance (dB), with UDP.

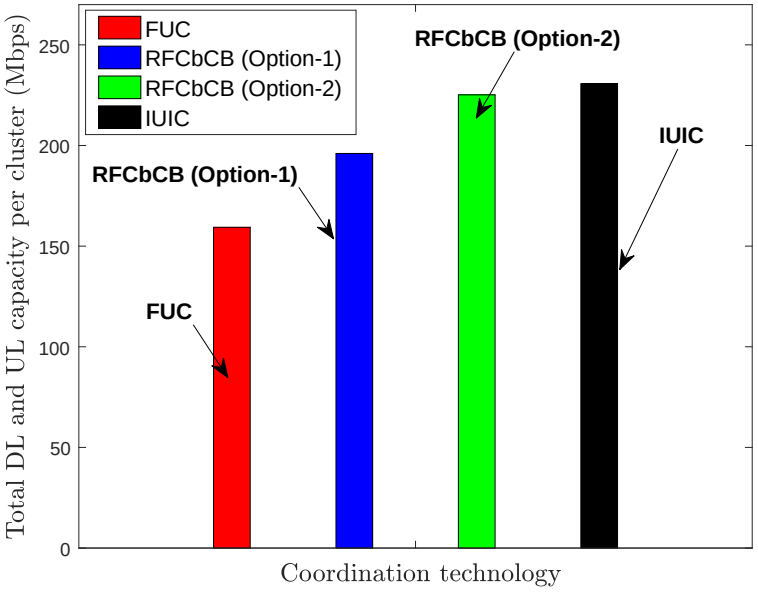


Fig. I.7. Total DL and UL capacity per cluster (Mbps), with UDP.

tion demonstrates a flexible coordination scheme that dynamically exploits the fully dynamic TDD frame flexibility when moderate levels of CLI are accepted. Otherwise, it slightly relaxes the requirement of the fully flexible frame configuration, trading-off an intended small capacity loss in the UE instantaneous rates for the sake of a significant improvement in the overall capacity. A further study with an analytical demonstration on the radio latency optimization of the proposed solution will be conducted in a future work.

7 Acknowledgments

This work is partly funded by the Innovation Fund Denmark, Grant: 7038-00009B. Also, part of this work is performed in the framework of the Horizon 2020 project ONE5G (ICT-760809) receiving funds from the European Union.

References

[1] P. Chang, Y. Chang, Y. Han, C. Zhang and D. Yang, "Interference analysis and performance evaluation for LTE TDD systems," *in Proc. IEEE ICACC*, Shenyang, 2010, pp. 410-414.

[2] D. Yun and W. Lee, "LTE-TDD interference analysis in spatial, time and frequency domain," *in Proc. IEEE ICUFN*, Milan, 2017, pp. 785-787.

[3] A. Roessler, J. Schlienz, S. Merkel, and M. Kottkamp, "LTE- advanced (3GPP Rel.12) technology introduction," Rohde & Schwarz, 1MA252_2E, USA, Aug. 2015.

[4] Ali A. Esswie, and K.I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Netw.*, vol. 6, pp. 38451-38463, June 2018.

[5] K. I. Pedersen, G. Berardinelli, F. Frederiksen and P. Mogensen, "A flexible 5G wide area solution for TDD with asymmetric link operation," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 122-128, April 2017.

[6] K. Lee, Y. Park, M. Na, H. Wang and D. Hong, "Aligned reverse frame structure for interference mitigation in dynamic TDD systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6967-6978, Oct. 2017.

[7] Y. Long and Z. Chen, "Interference-canceled asymmetric traffic cellular networks: dynamic TDD meets massive MIMO," *IEEE Trans. Veh. Technol.*, early access.

[8] A. Łukowa and V. Venkatasubramanian, "Performance of strong interference cancellation in flexible UL/DL TDD systems using coordinated muting, scheduling and rate allocation," *in Proc. IEEE WCNC*, Doha, 2016, pp. 1-7.

[9] Z. Huo, N. Ma and B. Liu, "Joint user scheduling and transceiver design for cross-link interference suppression in MU-MIMO dynamic TDD systems," *in Proc. IEEE ICCC,* Chengdu, 2017, pp. 962-967.

[10] E. d. O. Cavalcante, G. Fodor, Y. C. B. Silva and W. C. Freitas, "Distributed beamforming in dynamic TDD MIMO networks with cell to cell interference constraints," *IEEE Wireless Commun. Lett,* early access.

[11] D. Catania, M. Sarret, A. Cattoni, F. Frederiksen, G. Berardinelli and P. Mogensen, "Flexible UL/DL in small cell TDD systems: a Performance study with TCP traffic," *in Proc. IEEE VTC*, Glasgow, 2015, pp. 1-6.

[12] R1-1701146, *Dynamic TDD interference mitigation concepts in NR*, Nokia, Alcatel-Lucent Shanghai Bell, 3GPP RAN1 #88, Feb. 2017.

Paper J

Quasi-Dynamic Frame Coordination For Ultra-Reliability and Low-Latency in 5G TDD Systems

A. A. Esswie, K. I. Pedersen, and Preben E. Mogensen

The paper has been published in the
2019 IEEE International Conference on Communications (ICC)

Abstract

The fifth generation (5G) mobile technology features the ultra-reliable and low-latency communications (URLLC) as a major service class. URLLC applications demand a tight radio latency with extreme link reliability. In 5G dynamic time division duplexing (TDD) systems, URLLC requirements become further challenging to achieve due to the severe and fast-varying cross link interference (CLI) and the switching time of the radio frame configurations (RFCs). In this work, we propose a quasi-dynamic inter-cell frame coordination algorithm using hybrid frame design and a cyclic-offset-based RFC code-book. The proposed solution adaptively updates the RFCs in time such that both the average CLI and the user-centric radio latency are minimized. Compared to state-of-the-art dynamic TDD studies, the proposed scheme shows a significant improvement in the URLLC outage latency, i.e., $\sim$ 92% reduction gain, while boosting the cell-edge capacity by $\sim$ 189% and with a greatly reduced coordination overhead space, limited to B-bit.

Index Terms— Dynamic TDD; 5G new radio; URLLC; Cross link interference (CLI); Traffic; UDP.

1 Introduction

Ultra-reliable low-latency communication (URLLC) is a key driver of the fifth generation (5G) mobile networks [1]. Various URLLC use cases require one-way radio latency of one or several milliseconds with an outage probability below 10^{-5} [2]. As most of the 5G URLLC deployments are envisioned over the 3.5 GHz band, the time division duplexing (TDD) becomes a vital candidate transmission mode due to its frame adaptation, in order to dynamically match the sporadic URLLC capacity in both downlink (DL) and uplink (UL) directions [3].

With the 5G new radio (NR), the agile frame structure with variable transmission time interval (TTI) duration is introduced [3, 4]. Thus, 5G-NR TDD offers more adaptation flexibility with much faster link-direction update periodicity, that is slot-dependent instead of being frame-based, i.e., $\leq$ 1 ms. However, the coexistence of different transmission directions in adjacent cells results in cross link interference (CLI) [5], i.e., base-station to base-station (BS-BS) and user-equipment to user-equipment (UE-UE) CLI, respectively. Hence, URLLC performance is highly impacted by the degraded decoding ability, due to the fast-varying CLI, and the waiting interval to the first DL/UL transmission opportunity.

To the best of our knowledge, no prior work has assessed the performance of the URLLC outage with the 5G-NR dynamic TDD technology. The state-of-the-art TDD proposals consider joint multi-cell scheduling, cell muting, and enhanced power control [6, 7] to minimize the average network CLI.

Furthermore, advanced massive multi-antenna processing and beam-forming [8] are envisioned as vital to counteract the CLI by utilizing the channel hardening phenomenon. Opportunistic inter-cell coordination algorithms [9] are also proven attractive to boost the cell capacity of the dynamic TDD systems; however, at the expense of a sub-optimal URLLC outage performance.

In this work, we propose a hybrid-frame based coordination scheme (HFCS) for 5G-NR dynamic TDD systems. The proposed HFCS introduces a multi-objective and slot-dependent dynamic user scheduling. A hybrid radio frame structure and sliding radio frame configuration (RFC) code-book are designed to virtually extend the degrees of freedom of the TDD dynamicity. Thus, the URLLC users with the worst radio conditions always guarantee semi-preemptive, i.e., immediate scheduling over pre-set time slots, and CLI-free transmissions, leading to a significant reduction of the URLLC tail latency. The proposed coordination scheme shows a significant enhancement in the URLLC outage performance as well as maximizing the ergodic capacity, and with a confined coordination overhead span.

The performance of the proposed scheme is assessed by realistic system level simulations, due to the complexity of the 5G-NR and addressed problem herein. The major functionalities of the physical and media access control layers of the 5G-NR are incorporated and calibrated against latest 3GPP assumptions, including UL and DL channel modeling, hybrid automatic repeat request (HARQ), adaptive modulation and coding selection (MCS) and dynamic user scheduling.

This paper is organized as follows. Section 2 introduces the system modeling of this work while Section 3 presents the problem formulation. Section 4 details the proposed solution and Section 5 discusses the numerical results of the proposed scheme. Finally, conclusions are drawn in Section 6.

2 System Modeling

A macro 5G-NR TDD system is considered, with a single cluster of C cells, each with N_t antennas. Each cell has an average of K^{dl} and K^{ul} uniformly-distributed DL and UL active UEs, respectively, each with M_r antennas. We assume a URLLC dedicated network where the sporadic FTP3 traffic is adopted with finite packet sizes of f^{dl} and f^{ul} bits, and Poisson arrival processes λ^{dl} and λ^{ul}, in the DL and UL directions. Accordingly, the average offered load per cell in DL direction is: $K^{\mathrm{dl}} \times f^{dl} \times \lambda^{\mathrm{dl}}$ and in UL direction as: $K^{\mathrm{ul}} \times f^{ul} \times \lambda^{\mathrm{ul}}$.

We assume an RFC of 10 sub-frames, each can be DL, UL or a special sub-frame. UEs are dynamically multiplexed by the orthogonal frequency division multiple access with 15 kHz sub-carrier spacing. The smallest scheduling unit is the physical resource block (PRB) of 12 consecutive sub-carriers.

Furthermore, we adopt a user scheduling per a mini-slot duration of 7-OFDM symbols for faster URLLC transmissions.

Furthermore, an arbitrary master cell is initially identified in each cluster, where other cells are considered as slaves. All slave cells within the cluster are bidirectionally inter-connected to the master cell through the *Xn interface*.

We define $\mathfrak{B}_{\mathrm{dl}}$, $\mathfrak{B}_{\mathrm{ul}}$, $\mathcal{K}_{\mathrm{dl}}$ and $\mathcal{K}_{\mathrm{ul}}$ as the inclusive sets of cells and UE with DL and UL transmission directions, respectively. Hence, the pre-decoding received signal at the k^{th} UE, where $k \in \mathcal{K}_{\mathrm{dl}}$, $c_k \in \mathfrak{B}_{\mathrm{dl}}$, is expressed by

$$\mathrm{y}^{\mathrm{dl}}_{k,c_k} = \underbrace{\mathbf{H}^{\mathrm{dl}}_{k,c_k}\mathbf{v}_k s_k}_{\text{Useful signal}} + \underbrace{\sum_{i\in\mathcal{K}_{\mathrm{dl}}\setminus k}\mathbf{H}^{\mathrm{dl}}_{k,c_i}\mathbf{v}_i s_i}_{\text{BS to UE interference}} + \underbrace{\sum_{j\in\mathcal{K}_{\mathrm{ul}}}\mathbf{G}_{k,j}\mathbf{w}_j s_j}_{\text{UE to UE interference}} + \mathbf{n}^{\mathrm{dl}}_k, \tag{J.1}$$

where $\mathbf{H}^{\mathrm{dl}}_{k,c_i} \in \mathcal{C}^{M_r\times N_t}$ is the DL fading channel from the cell serving the i^{th} UE, to the k^{th} UE, $\mathbf{v}_k \in \mathcal{C}^{N_t\times 1}$, $\mathbf{w}_j \in \mathcal{C}^{M_r\times 1}$ and s_k denote the single-stream zero-forcing precoding vector at the c^{th}_k cell, precoding vector at the j^{th} UE, and transmitted data symbol of the k^{th} UE, respectively, $\mathbf{G}_{k,j} \in \mathcal{C}^{M_r\times M_r}$ represents the the cross-link channel between the k^{th} and j^{th} UEs. $\mathbf{n}^{\mathrm{dl}}_k$ denotes the additive white Gaussian noise at the k^{th} UE. In the UL direction, the received signal at the c^{th}_k cell, where $c_k \in \mathfrak{B}_{\mathrm{ul}}$ from $k \in \mathcal{K}_{\mathrm{ul}}$, is modeled by

$$\mathrm{y}^{\mathrm{ul}}_{c_k,k} = \underbrace{\mathbf{H}^{\mathrm{ul}}_{c_k,k}\mathbf{w}_k s_k}_{\text{Useful signal}} + \underbrace{\sum_{j\in\mathcal{K}_{\mathrm{ul}}\setminus k}\mathbf{H}^{\mathrm{ul}}_{c_k,j}\mathbf{w}_j s_j}_{\text{UE to BS interference}} + \underbrace{\sum_{i\in\mathcal{K}_{\mathrm{dl}}}\mathbf{Q}_{c_k,c_i}\mathbf{v}_i s_i}_{\text{BS to BS interference}} + \mathbf{n}^{\mathrm{ul}}_{c_k}, \tag{J.2}$$

where $\mathbf{Q}_{c_k,c_i} \in \mathcal{C}^{N_t\times N_t}$ denotes the cross-link fading channel between the cells that serve the k^{th} and i^{th} UEs, respectively, $k \in \mathcal{K}_{\mathrm{ul}}$ and $i \in \mathcal{K}_{\mathrm{dl}}$. The pre-detection signal-to-interference-noise-ratio (SINR) in the DL direction at the k^{th} UE γ^{dl}_k and in the UL direction at the c^{th}_k cell $\gamma^{\mathrm{ul}}_{c_k}$, are given by

$$\gamma^{\mathrm{dl}}_k = \frac{p^{\mathrm{dl}}_{c_k}\left\|\mathbf{H}^{\mathrm{dl}}_{k,c_k}\mathbf{v}_k\right\|^2}{\sigma^2 + \sum_{i\in\mathcal{K}_{\mathrm{dl}}\setminus k} p^{\mathrm{dl}}_{c_i}\left\|\mathbf{H}^{\mathrm{dl}}_{k,c_i}\mathbf{v}_i\right\|^2 + \sum_{j\in\mathcal{K}_{\mathrm{ul}}} p^{\mathrm{ul}}_j\left\|\mathbf{G}_{k,j}\mathbf{w}_j\right\|^2}, \tag{J.3}$$

$$\gamma^{\mathrm{ul}}_{c_k} = s\frac{p^{\mathrm{ul}}_k\left\|\mathbf{H}^{\mathrm{ul}}_{c_k,k}\mathbf{w}_k\right\|^2}{\sigma^2 + \sum_{j\in\mathcal{K}_{\mathrm{ul}}\setminus k} p^{\mathrm{ul}}_j\left\|\mathbf{H}^{\mathrm{ul}}_{c_k,j}\mathbf{w}_j\right\|^2 + \sum_{i\in\mathcal{K}_{\mathrm{dl}}} p^{\mathrm{dl}}_{c_i}\left\|\mathbf{Q}_{c_k,c_i}\mathbf{v}_i\right\|^2}, \tag{J.4}$$

where $p^{\mathrm{dl}}_{c_k}$ and p^{ul}_k denote the transmission powers of the c^{th}_k cell and the k^{th} UE, respectively. Finally, the received UL/DL signals are decoded by the linear minimum mean square error interference rejection combining receiver (LMMSE-IRC) [4] vector $\mathbf{a}$, expressed as: $\hat{s}^{\kappa}_k = \left(\mathbf{a}^{\kappa}_k\right)^{\mathrm{H}}\mathrm{y}^{\kappa}_k$, $\mathcal{X}^{\kappa}, \kappa \in \{\mathrm{ul}, \mathrm{dl}\}$, with $(\bullet)^{\mathrm{H}}$ as the Hermitian operation.

3 Problem Formulation

The URLLC latency and reliability requirements are further challenging to achieve in 5G-NR dynamic TDD systems, mainly due to the link-direction switching time and the degraded URLLC decoding performance. The former is significantly minimized by the flexible 5G frame structure; however, the latter still remains an open issue.

In fully dynamic TDD macro networks, neighboring cells may have simultaneous cross-directional transmissions, leading to a strong CLI which varies per the link-direction update periodicity. With the 5G-NR, such periodicity is slot-based, i.e., ≤ 1 ms, leading to highly varying CLI fluctuations. As a result, URLLC UEs inflict significantly degraded decoding performance. In particular, lower-power URLLC UL transmissions suffer from a strong CLI from adjacent higher-power DL transmissions, leading to several HARQ retransmissions prior to a successful decoding, not satisfying the URLLC targets.

Let u_c and d_c present the estimated numbers of UL and DL slots during a given RFC while $u_c^{\text{opt.}}$ and $d_c^{\text{opt.}}$ are the respective optimal numbers. Hence, the proposed HFCS defines a programming optimization problem as:

$$R \triangleq \arg\max_{c} \sum_{c=1}^{C} \min\left(u_c, u_c^{\text{opt.}}\right) \digamma_c^{u} + \min\left(d_c, d_c^{\text{opt.}}\right) \digamma_c^{d},$$

subject to:

$$\begin{cases} \arg\min_{c} \phi_c\left(\eta_c\right) = \frac{1}{C} \sum_{x=1, x\neq c}^{C} \varphi_{c,x}\left(\eta_c, \eta_x\right), \\ \forall k \in \mathcal{K}_{\text{ul/dl}} : \arg\min_{k} \left(\Psi_{c,k}\right), \Psi_{c,k} \leq \epsilon \text{ ms}, \end{cases} \tag{J.5}$$

where R is total capacity of each cluster, $\digamma_c^{u}$ and $\digamma_c^{d}$ denote rate utility functions of the UL and DL transmissions, i.e., capacity gain due to an UL or DL transmission. $\phi_c\left(\eta_c\right)$ and $\varphi_{c,x}\left(\eta_c, \eta_x\right)$ represent the average and actual slot misalignment of the requested RFC by the c^{th} cell η_c and between the RFCs of the c^{th} and x^{th} cells, i.e., η_c and η_x, respectively, $\forall x \neq c$, and $\Psi_{c,k}$ is the one-way radio latency of the k^{th} UL or DL user which is confined by ϵ ms.

For best RFC adaptation and highest ergodic capacity, $u_c = u_c^{\text{opt.}}$ and $d_c = d_c^{\text{opt.}}$ should be arbitrarily set in (J.5). However, $u_c^{\text{opt.}}$ and $d_c^{\text{opt.}}$ may introduce a large inter-cell slot misalignment ϕ_c, resulting in severe CLI within the cluster, and thus, a significant degradation of the overall capacity R and URLLC latency performance. As such problem is non-convex, we propose a heuristic approach using complexity-efficient coordination with hybrid-frame design, multi-objective user scheduling and a sliding-based RFC code-book.

4 Proposed HFCS Coordination

The proposed HFCS combines a hybrid RFC design, multi-objective distributed user scheduling, and a cyclic-offset-based RFC code-book. A pre-defined RFC code-book is constructed and presumed pre-known to all cells within the cluster, where all RFCs have a set combination of static and dynamic slots. At each RFC update instant, each slave cell selects the one RFC from the code-book that most satisfies its individual link-direction selection criterion. The slave cells signal the index of the selected RFC to the master cell where it seeks to improve the joint capacity. Thus, it may slightly change the RFCs requested by slave cells. Accordingly, the master cell feeds-back the updated RFC indices to the slave cells, to be adopted until the next RFC update. During each RFC period, each cell considers a dual-objective dynamic user scheduling.

4.1 Proposed Inter-Cell Coordination Scheme

Hybrid RFC design and sliding RFC code-book

A hybrid RFC design is adopted, where each RFC is divided into arbitrary static and dynamic slot sets (SSS, DSS). A SSS denotes the radio slots which are fixed across all RFCs in the code-book, i.e., static TDD slots with CLI-free transmissions. However, a DSS implies fully dynamic radio slots.

Accordingly, a pre-defined RFC code-book of $\mathcal{N}$ unique RFCs is constructed such that it is divided into L groups. The RFCs within each group share the same DL:UL slot ratio, i.e., $d_c : u_c$; though, with a different placement during the DSS. For instance, the DSS of each RFC is a cyclic-shift of the other RFCs, as depicted in Fig. J.1. The structure of the SSS, DSS, and size of the RFC code-book are design parameters.

At slave cells – Traffic and latency adaptation

During each RFC update instant, each slave cell selects the one RFC from the code-book which best satisfies its link-direction selection criterion. Without loss of generality, we consider the DL/UL buffered traffic size including pending HARQ re-transmissions as the major criterion to select the RFC, and accordingly the best $d_c : u_c$ ratio. Then, the traffic load threshold β_c is defined as

$$\beta_c \leq \frac{\sum Z_c^{\mathrm{dl}}}{\sum Z_c^{\mathrm{dl}} + \sum Z_c^{\mathrm{ul}}}, \tag{J.6}$$

where $\sum Z_c^{\mathrm{dl}}$ and $\sum Z_c^{\mathrm{ul}}$ imply the aggregate traffic in the DL and UL directions, respectively. With $\beta_c = 0.5$, if $\sum Z_c^{\mathrm{dl}} \gg \sum Z_c^{\mathrm{ul}}$, a cell selects an RFC with a majority of DSS DL slots. Although, the free selection of the best RFCs requested by each slave cell may result in severe CLI, hence, several

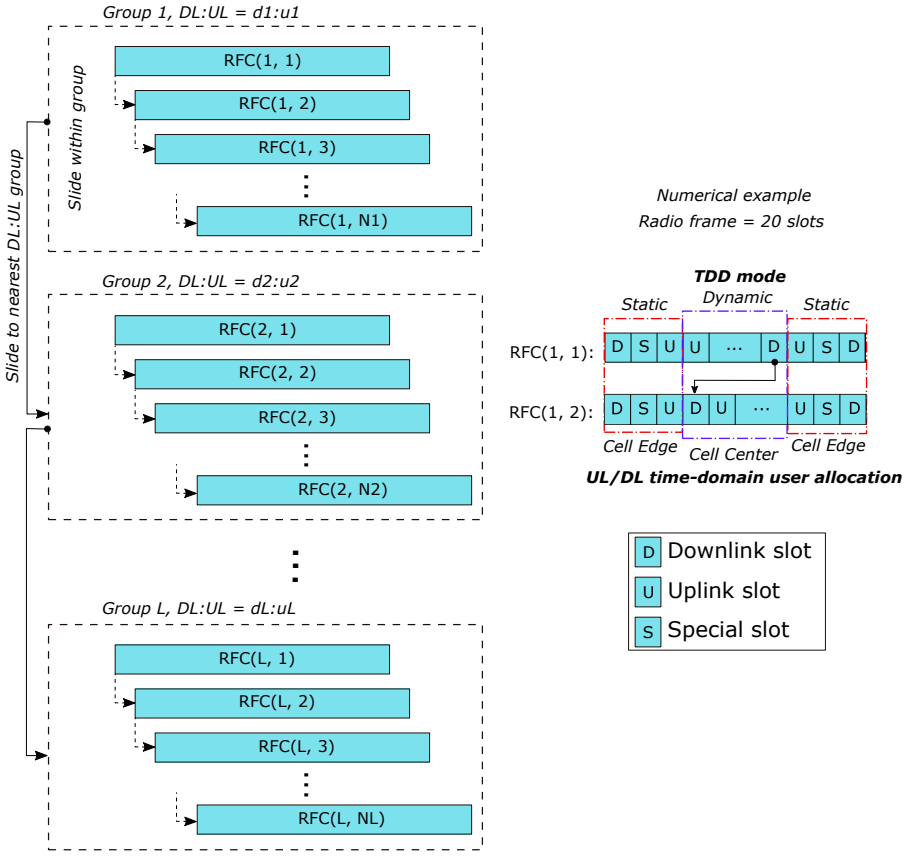


Fig. J.1. Hybrid RFC design and sliding code-book.

HARQ re-transmissions may be inflicted, leading to a significant radio latency and reliability. On the other side, an abrupt change of these RFCs to reduce the average CLI leads to significant queuing delays up to the first DL/UL transmission opportunities. Thus, to address the constraints in (J.5), each cell adaptively estimates a dynamic sliding threshold $\psi_c(t)$, where t is the link-direction update time, with which it instructs the master cell about the maximum allowable change of its desired RFC, in order to achieve an adequate joint URLLC and ergodic capacity performance.

Let Θ^{BS} and Θ^{UE} denote the BS-BS and UE-UE CLI at the BS and UE, respectively. These CLI estimates can be obtained at the BS through radio feedback links from UEs; however, there is no a standardized mechanism of the CLI measurement reporting available yet. Then, each BS calculates the average experienced CLI using an arbitrary filter function. In this work, we assume a weighted average filter as

$$\Xi_c^{\mathrm{avg.}} = \frac{\tilde{\beta}_c \times \Theta_c^{\mathrm{BS}} + \tilde{\mu_c} \times \Theta_c^{\mathrm{UE}}}{\Theta_c^{\mathrm{BS}} + \Theta_c^{\mathrm{UE}}}, \tag{J.7}$$

$$\tilde{\beta}_c,\, \tilde{\mu_c} = \begin{cases} \frac{1}{\beta_c}, \frac{1}{\mu_c} & \forall\, \Theta_c^{\mathrm{BS}}, \Theta_c^{\mathrm{UE}} \leq \varrho\,\mathrm{dBm} \\ \beta_c, \mu_c & \forall\, \Theta_c^{\mathrm{BS}}, \Theta_c^{\mathrm{UE}} > \varrho\,\mathrm{dBm} \end{cases}, \tag{J.8}$$

$$\beta_c = \frac{\sum Z_c^{\mathrm{dl}}}{\sum Z_c^{\mathrm{ul}}} \times \mu_c, \tag{J.9}$$

where β_c and μ_c are the BS-BS and UE-UE CLI weights, and ϱ is a CLI threshold. The ratio of both weights is set to the ratio of the buffered traffic as in (J.9), such that a cell with $\sum Z_c^{\mathrm{dl}} \gg \sum Z_c^{\mathrm{ul}}$, and accordingly a DL-heavy RFC, shall impose severe BS-BS CLI to adjacent cells. Hence, under this condition, $\Xi_c^{\mathrm{avg.}}$ is biasedly maximized and the RFC adaptation is enforced towards the CLI minimization. In the delay domain, cells measure the head of line delay (HoLD) within their DL and UL transmission buffers. HoLD indicates an estimate of the maximum time required to transmit the last packet in the buffer, based on the expected UL/DL transmission constraints of the current RFC. Such metric is of a significant importance with URLLC since a packet can be considered of no use if its latency deadline is not fulfilled. Hence, to reduce the average HoLD, selected traffic-based RFCs should be used without a significant change in order to quickly transmit the data buffers.

Thus, we propose a simple and dynamic sliding threshold for a best-effort trade-off between CLI and radio latency. Fig. J.2 shows a numerical example of such approach. A SSS to DSS ratio of 8:12 is assumed. Thus, the slot misalignment threshold is bounded by the size of the DSS. Accordingly, the range of the CLI and HoLD values is quantized over the DSS size. For an arbitrary cell, if the average CLI, experienced over the previous measurement

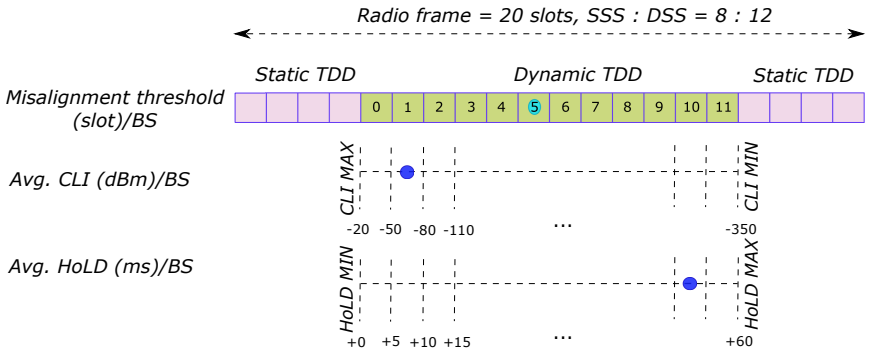


Fig. J.2. Dynamic misalignment threshold $\psi_c(t)$.

cycle, is at maximum, e.g., $\Xi_c^{\text{avg.}} = -60$ dBm, it implies a tight slot misalignment threshold should be enforced to promptly reduce such severe CLI over the upcoming RFC period, e.g., $\psi_c(t) = 1$ slot. However, if such cell simultaneously inflicts a large HoLD, e.g., HoLD = 52 ms, the slot misalignment constraint shall be relaxed, e.g., $\psi_c(t) = 10$ slots, in order not to allow the master cell to change the RFC of this cell, hence, having faster transmissions for the respective traffic. Without loss of generality, we apply a fair averaging of both misalignment thresholds, i.e., $\psi_c(t) = 5$ slots.

Finally, at each RFC update periodicity, slave cells signal the master cell with the requested RFC indices of $\mathrm{B} = \log_2(\mathcal{N})$ bits on the *Xn interface* as well as the maximum allowable slot misalignment thresholds $\psi_c(t)$.

At master cell – CLI minimization

When the master cell receives all RFC information from slave cells, it first identifies a *common RFC*, which is requested by the majority of the slave cells. If not feasible, the master cell randomly selects any reported RFC as the common one, to which all other RFCs shall maintain the respective slot misalignment thresholds. Thus, for each RFC η_c of the c_k^{th} cell, master cell calculates the slot misalignment to the common RFC $\delta_x, \forall x \neq c$ as in (J.5). Then, the master cell does not alter such requested RFC if the following condition is fulfilled:

$$\varphi_{c,x}(t) \leq \psi_c(t). \tag{J.10}$$

Hence, the respective slave cell utilizes its best matching RFC to its latency and capacity outage. Otherwise, the master cell slides over all RFCs within the same group as the desired RFC of the c^{th} cell η_c. Accordingly, it estimates the corresponding slot misalignment values and considers the one RFC with $\varphi_{c,x}(t)$ that has the closest linear distance to the requested $\psi_c(t)$. If the slot misalignment constraint in (J.10) is satisfied, master cell adopts such RFC as the updated RFC of the current cell. This way, an acceptable average CLI is

guaranteed at the slave cells while still preserving the same requested traffic service ratio $d_c : u_c$, leading to a significant improvement of the capacity and outage latency performance.

If the slot misalignment constraint is not yet feasible across all RFCs from the same group as the requested one, master cell progressively slides to the other RFC groups from the RFC code-book with the nearest possible $d_c : u_c$ ratio to the requested ratio, e.g., $d_c : u_c = 4 : 12 \underset{\text{slide to}}{\rightarrow} d_c^{'} : u_c^{'} = 3 : 13$, and repeats the same process. Herein, the master cell partly relaxes the target outage requirements of the salve cells due to the abrupt change in the $d_c : u_c$ ratio. However, such outage degradation is bounded across a limited number of slots during the RFC and is reversely proportional to the size of the RFC code-book $\mathcal{N}$. As a last best-effort resort, if the constraint in (J.10) could not be satisfied across all RFCs, either from same or different group(s), the master cell considers the one RFC with the closest possible estimated slot misalignment to desired $\psi_c(t)$, and then, it signals all slave cells within the cluster over the *Xn interface* with the updated RFC indices that should be used over the upcoming RFC periodicity.

4.2 Distributed multi-objective user scheduling

During each RFC periodicity, each cell applies a slot-dependent dynamic user scheduling. During the DSS instances, cells may adapt an arbitrarily capacity maximizing user scheduling. Without loss of generality, and since we assume an equally-prioritized URLLC setup, we adopt the proportional fair (PF) criterion ω in both the time and frequency domains to maintain a global scheduling fairness as

$$\omega\left\{\mathrm{PF}_{k_{\mathrm{ul/dl}}}\right\} = \frac{r_{k_{\mathrm{ul/dl}},rb}}{\bar{r}_{k_{\mathrm{ul/dl}},rb}}, \tag{J.11}$$

$$k^{*}_{\mathrm{ul/dl}} = \underset{k_{\mathrm{ul/dl}} \in \mathcal{K}_{\mathrm{ul/dl}}}{\arg\max} \quad \omega\left\{\mathrm{PF}_{k_{\mathrm{ul/dl}}}\right\}, \tag{J.12}$$

where $r_{k_{\mathrm{ul/dl}},rb}$ and $\bar{r}_{k_{\mathrm{ul/dl}},rb}$ denote the instantaneous and average delivered rates of the k^{th} UL/DL user. However, during the SSS periods, each cell preemptively interrupts its individual time-domain scheduling metric by immediately allocating the users with the worst radio conditions, i.e., potentially cell-edge users. These users are identified based on the reported channel quality indication (CQI) reports. To avoid threshold-based user identification, the UL/DL time-domain scheduler sorts active users in an ascending-order list in terms of their reported CQI levels, i.e., users from the top of the list are of worst radio conditions, thus, scheduler grants them a higher priority for immediate scheduling during the CLI-free SSS. In the frequency domain, the PF metric is used to preserve fairness among cell-edge URLLC

users. Thus, cell-edge URLLC users achieve a better decoding ability with faster transmissions, avoiding the latency-costly HARQ re-transmissions.

4.3 Comparison to the state-of-the-art TDD studies

We evaluate the performance of the proposed scheme against the state-of-the-art coordinated TDD proposals as:

Non-coordinated TDD (**NC-TDD**): no RFC coordination is assumed. Cells independently and dynamically in time pick the RFCs from the code-book which most meet their individual traffic demand, as in (J.6). Hence, maximum TDD RFC flexibility is achieved with no coordination overhead; however, associated with potentially a large slot misalignment and severe average CLI levels accordingly.

Sliding code-book based coordinated TDD (**SCC-TDD**) [9]: in our prior work, we introduced a simple inter-cell coordination algorithm, mainly for broadband services, to significantly reduce the average slot misalignment, based on a preset global misalignment threshold Ω, and hence, the aggregate CLI, resulting in greatly improved ergodic capacity. Though, it has been demonstrated not suitable for URLLC transmissions due to the monotonic scheduling objective.

CLI-free coordinated TDD (**CFC-TDD**): cells dynamically select their respective RFC according to (J.6). A sophisticated BS-BS and UE-UE coordination is artificially assumed. That is, BSs and UEs exchange PRB mapping, UE MCS and precoding information, for them to perfectly suppress the BS-BS and UE-UE CLI. However, such coordination introduces a significant control overhead over both the back-haul and radio interfaces, respectively. In [10], a 3GPP technical study introduces a sub-optimal CFC-TDD approach with a lower overhead space. However, CFC-TDD holds an optimal theoretical baseline, where both maximum TDD RFC flexibility and CLI-free transmissions are always guaranteed.

5 Performance Evaluation

The major simulation assumptions are presented in Table J.1. During each TTI, each cell dynamically multiplexes users over system PRBs using the PF metric, if it is within the DSS of the current RFC or by preemptive cell-edge user allocations when it is within the SSS. We consider a fully dynamic MCS selection and adaptive Chase-combining HARQ re-transmissions, where the HARQ feedback is always prioritized over new transmissions. The post-detection SINR levels are estimated by the LMMSE-IRC receiver, where the average interference is identified by its mean covariance. Finally, we assess

Table J.1: Simulation parameters.

Parameter	Value
Environment	3GPP-UMA, one cluster, 21 cells
UL/DL channel bandwidth	10 MHz, SCS = 30 KHz, TDD
Antenna setup	$N_t = 8$ Tx, $M_r = 2$ Rx
UL power control	LTE-alike, $\alpha = 1$, $P0 = -103$ dBm
Average user load per cell	$K^{\mathrm{dl}} = K^{\mathrm{ul}} = 10$ and 20
TTI configuration	0.5 ms (7-OFDM symbols)
Traffic model	FTP3, $f^{dl} = f^{ul}$ = 400 bits λ^{dl} =167, and 620 pkts/sec λ^{ul} =334, and 620 pkts/sec
Offered average load per cell DL:UL	DL:UL = 1:2 (0.6:1.2) Mbps DL:UL = 1:1 (5:5) Mbps
Proposed HFCS setup	$\mathcal{N}$ = 55 RFCs L = 7 groups B = 6 bits

the proposed solution under the latency-efficient user data-gram protocol for several offered cell loading conditions.

Fig. J.3 depicts a comparison of the complementary cumulative distribution function (CCDF) of the URLLC outage latency in the UL direction for all TDD coordination schemes under assessment, for an average offered load of 2 Mbps/cell with a DL:UL traffic ratio of 1:2. Furthermore, we present the latency performance of the best static-TDD case where the static pattern is pre-selected to perfectly match the DL-to-UL average traffic ratio, i.e., 6 DL mini-slots, 12 UL mini-slots and 2 guard mini-slots. The optimal CFC-TDD achieves the best URLLC latency performance, i.e., 42 ms at 10^{-5} outage probability. However, it comes under the ideal assumption of perfect elimination of any experienced CLI, and with an infinite coordination overhead, which is infeasible in practice. The proposed HFCS clearly provides a significant improvement of the UL URLLC latency, approaching the optimal CFC-TDD; however, with greatly reduced overhead span, mainly limited to $\log_2(\mathcal{N})$ bits. That is, it achieves 92% and 67% reduction gain in the UL outage latency compared to SCC-TDD and NC-TDD. The best static-TDD case out-performs proposed HFCS scheme, i.e., 9% reduction in the outage latency, due to the absence of the CLI, approaching CFC-TDD; though, this comes with the assumption that the static RFC pattern is pre-defined to perfectly align with the traffic demands.

The significant latency improvements of the proposed HFCS are attributed to the guaranteed preemptive cell-edge user scheduling with CLI-free transmissions, where these users majorly control the latency tail, i.e., outage, per-

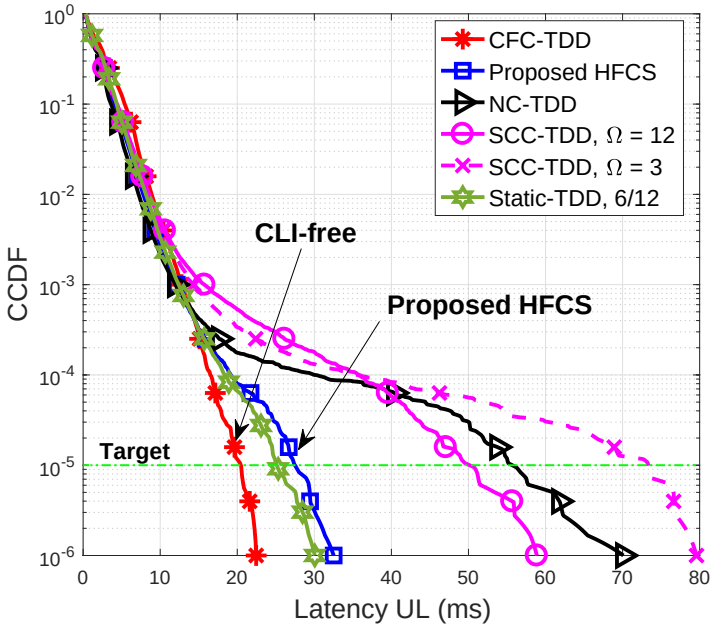


Fig. J.3. URLLC outage latency in UL direction (ms).

formance. Thus, less costly HARQ re-transmissions are experienced. The SCC-TDD latency performance depends on the preset misalignment threshold Ω. For instance, with a tight $\Omega = 3$, the master cell may aggressively change the requested RFC of a given slave cell, in order to only allow for an average misalignment of three slots. As a result, slave cells may adopt RFCs that do not best match their current traffic demands, leading to a more queuing delay to the first transmission opportunity. Finally, the NC-TDD offers a fair URLLC latency performance since the maximum possible TDD RFC flexibility is utilized; however, with severe CLI levels.

Similar observations are obtained from the URLLC outage latency in the DL direction, as shown in Fig. J.4. All considered TDD coordination schemes provide a decent DL latency , i.e., ≤ 8 ms. This is due to the larger desired DL transmission power, i.e., compared to the interfering UL power, hence, less impactful CLI. However, static-TDD case inflicts a longer queuing delay due to the fixed DL and UL slot placement.

Fig. J.5 shows the empirical CDF (ECDF) of the post-receiver UL interference performance in dBm, including both cross and same link inter-cell interference, respectively. Due to the absence of the CLI, CFC-TDD offers an attractive interference performance. However, due to the dual-scheduling metrics during the DSS and SSS periods, the proposed HFCS achieves the

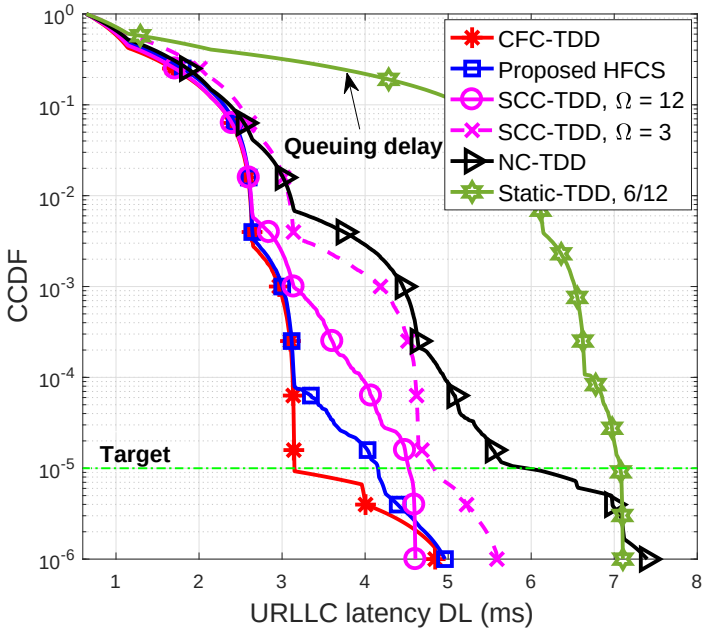


Fig. J.4. URLLC outage latency in DL direction (ms).

same interference suppression capability as the optimal CFC-TDD for the critical lower percentiles below 20%, i.e., cell-edge users. Furthermore, the proposed HFCS offers 39% and 45% reduction of the post-receiver interference at the 20^{th} percentile, compared to SCC-TDD and NC-TDD, respectively. The SCC-TDD exhibits a monotonic interference suppression performance where cell-edge users, get most impacted, while NC-TDD inflicts the worst interference performance due to the extreme slot misalignment, hence, the sever CLI levels.

Fig. J.6 presents the average cell throughput per TTI in the UL direction, with an average total offered load per cell of 10 Mbps. As can be noted, proposed solution boosts the cell-edge capacity, e.g., 189% capacity gain is achieved against SCC-TDD at the 30^{th} percentile. The change of the distribution slope of the proposed HFCS is due to the slot-based dual scheduling objectives, i.e., joint latency-capacity scheduling. However, the proposed HFCS still exhibits a capacity loss of 45% at the 95^{th} percentile compared to ideal SCC-TDD, due to the preemptive scheduling of cell-edge users during the SSS of each RFC, despite that they may not be the best capacity/fairness maximizing set of users. The fully dynamic NC-TDD fails to offer an acceptable cell-edge capacity due to the extreme CLI, i.e., $\sim$ 48% of the scheduling TTI instances have no sufficient capacity.

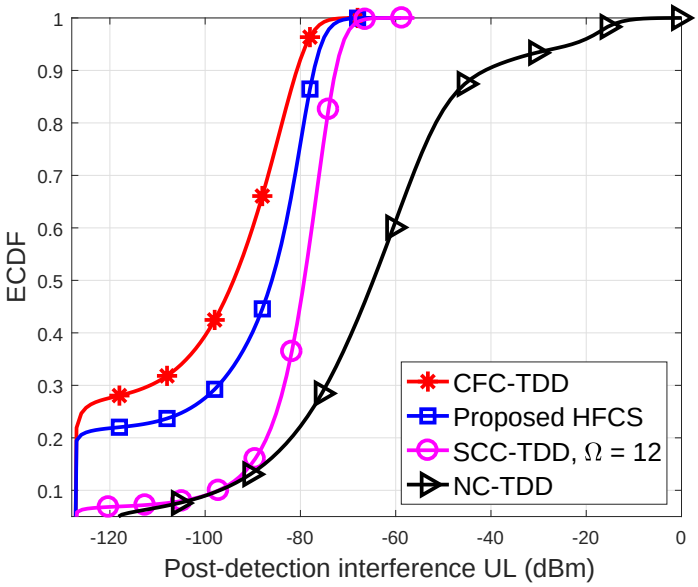


Fig. J.5. Post-receiver interference in UL direction (dBm).

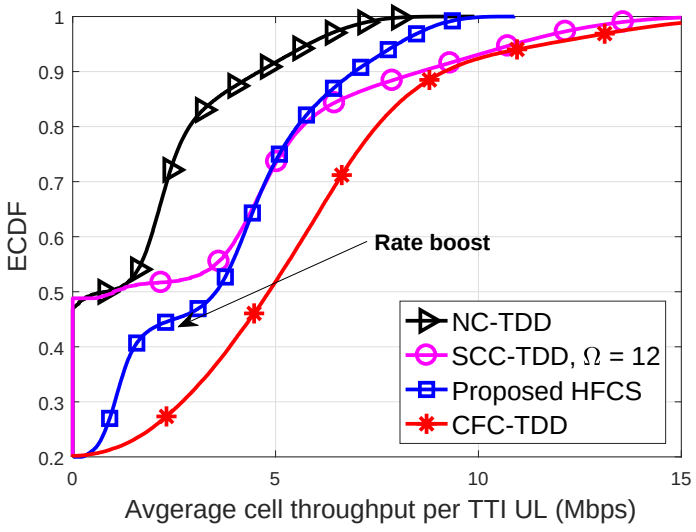


Fig. J.6. Average cell throughput in UL direction (Mbps).

6 Concluding Remarks

A quasi-dynamic coordination scheme has been introduced for ultra-reliable and low-latency communications (URLLC) in 5G TDD networks. The proposed solution combines hybrid radio frame design, distributed multi-objective user scheduling and a cyclic-offset-based radio frame code-book. Compared to the state-of-the-art coordinated TDD proposals from industry and academia, proposed scheme offers a significant improvement of the URLLC outage performance, e.g., 92% latency reduction gain, in addition to achieving aggregated cell capacity gain of 189%, and with a limited control overhead space, bounded to B-bit.

7 Acknowledgments

This work is partly funded by the Innovation Fund Denmark – File: 7038-00009B. Also, part of this work has been performed in the framework of the Horizon 2020 project ONE5G (ICT-760809) receiving funds from the European Union.

References

[1] IMT vision – "Framework and overall objectives of the future development of IMT for 2020 and beyond", international telecommunication union (ITU), ITU-R M.2083-0, Feb. 2015.

[2] Service requirements for the 5G system; Stage-1 (Release 16), 3GPP, TS 22.261, V16.6.0, Dec. 2018.

[3] K. I. Pedersen, G. Berardinelli, F. Frederiksen and P. Mogensen, "A flexible 5G wide area solution for TDD with asymmetric link operation," *IEEE Wireless Commun.* , vol. 24, no. 2, pp. 122-128, April 2017.

[4] Ali A. Esswie, and K.I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Netw.*, vol. 6, pp. 38451-38463, June 2018.

[5] K. Lee, Y. Park, M. Na, H. Wang and D. Hong, "Aligned reverse frame structure for interference mitigation in dynamic TDD systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6967-6978, Oct. 2017.

[6] A. Łukowa and V. Venkatasubramanian, "Performance of strong interference cancellation in flexible UL/DL TDD systems using coordinated

muting, scheduling and rate allocation," *in Proc. IEEE WCNC*, Doha, 2016, pp. 1-7.

[7] Z. Huo, N. Ma and B. Liu, "Joint user scheduling and transceiver design for cross-link interference suppression in MU-MIMO dynamic TDD systems," *in Proc. IEEE ICCC,* Chengdu, 2017, pp. 962-967.

[8] E. d. O. Cavalcante, G. Fodor, Y. C. B. Silva and W. C. Freitas, "Distributed beamforming in dynamic TDD MIMO networks with cell to cell interference constraints," *IEEE Wireless Commun. Lett,* early access.

[9] Ali A. Esswie, and K.I. Pedersen, "Inter-cell radio frame coordination scheme based on sliding codebook for 5G TDD systems," *in Proc. IEEE VTC-spring,* Kuala Lumpur, 2019.

[10] R1-1701146, *Dynamic TDD interference mitigation concepts in NR*, Nokia, Alcatel-Lucent Shanghai Bell, 3GPP RAN1 #88, Feb. 2017.

Paper K

Cross-Link Interference Suppression By Orthogonal Projector For 5G Dynamic TDD URLLC Systems

Ali A. Esswie and Klaus I. Pedersen

The paper has been published in the
2020 IEEE Wireless Communications and Networking Conference (WCNC)

Abstract

Dynamic time division duplexing (TDD) is envisioned as a vital transmission technology of the 5G new radio, due to its reciprocal propagation characteristics. However, the potential cross-link interference (CLI) imposes a fundamental limitation against the feasibility of the ultra-reliable and low latency communications (URLLC) in dynamic-TDD systems. In this work, we propose a near-optimal and complexity-efficient CLI suppression scheme using orthogonal spatial projection, while the signaling overhead is limited to B-bit*, over the back-haul links. Compared to the state-of-the-art dynamic-TDD studies, proposed solution offers a significant improvement of the URLLC outage latency, e.g., $\sim -199\%$ reduction, while boosting the achievable capacity per the URLLC packet by $\sim +156\%$.*

Index Terms— URLLC; Cross link interference; TDD; 5G.

1 Introduction

Ultra-reliable and low-latency communication (URLLC) is the major service class of the 5G new radio (NR) [1]. URLLC denotes short and stochastic packet transmissions with extreme reliability and radio latency bounds, i.e., couple of milli-seconds with a success probability of 99.999% [2]. Furthermore, the global regulatory bodies have envisioned early 5G deployments over the 3.5 GHz spectrum due to its abundant available unpaired bands. Accordingly, dynamic time division duplexing (TDD) has become of a great significance [3]. With dynamic TDD, base-stations (BSs) independently and dynamically in time select their respective link directions based on individual objective functions, leading to an improved transmission adaptation to the sporadic traffic arrivals.

However, the URLLC reliability and latency targets are further challenging to achieve in dynamic TDD systems [2] due to: (a) the switching time between the downlink (DL) and uplink (UL) sub-frames, and (b) the potential inter-cell cross-link interference (CLI) between neighboring BSs of different directional transmissions [4]. That is, the DL-to-UL CLI (BS-BS) and UL-to-DL CLI (user-equipment to user-equipment (UE-UE)). The former is tackled by the flexible frame design of the 5G-NR, where variable transmission time intervals (TTIs) and a scalable sub-carrier spacing (SCS) are supported [1]. Thus, the DL and UL switching delay can be slot-dependent, i.e., $\ll 1$ ms. Although, the latter issue, especially the BS-BS CLI due to the power imbalance between the DL and UL transmissions, remains a critical issue against practical implementation of the dynamic TDD macro systems.

As part of the long-term evolution, i.e., 4G, standards, advanced linear interference rejection combining (IRC) transceivers [5] are adopted to suppress the inter-cell interference sub-space from that is of the useful signal.

Although, within dense macro deployments, there exist multiple dominant and sparse BS-BS CLI interferers, degrading the IRC decoding performance due to the linear interference averaging. Accordingly, optimal BS-BS CLI cancellation [6] is discussed within 3GPP, where inter-cell full-packet exchange is assumed. Moreover, coordinated dynamic scheduling and beam-forming [7, 8] are proposed to counteract the CLI by globalizing the BS scheduling decisions. Furthermore, joint beam-forming schemes are suggested [9, 10] in order to control the inflicted inter-cell CLI in the spatial domain. On another side, opportunistic CLI pre-avoidance [4, 11, 12] schemes have been introduced based on ordered signal-to-interference-noise-ratio (SINR) lists and a sliding radio frame configuration (RFC) code-book design, respectively.

In this paper, we propose a high-performance and low-complexity BS-BS CLI suppression algorithm (CSA) for 5G-NR dynamic TDD macro systems. The proposed scheme utilizes a linear estimation of an orthonormal subspace projector to reliably suppress the BS-BS CLI on-the-fly, while it combines a hybrid radio frame design, cyclic-offset based frame code-book, and dual-objective dynamic user scheduling to opportunistically pre-avoid the UE-UE CLI occurrence. Compared to state-of-the-art dynamic-TDD studies, the proposed scheme offers a significant enhancement of the URLLC UL and DL outage latency, while improving the ergodic capacity, approaching the optimal CLI-free case. However, the proposed scheme neither requires periodic user CLI measurements nor significant signaling overhead. Particularly, the contribution aspects of this paper are as follows:

- Unlike the standard linear IRC receiver, we utilize a newly proposed inter-BS exchange of the user DL spatial signatures to manipulate the estimated interference covariance. Hence, we introduce an enhanced formulation of the standard IRC receiver, where the BS-BS CLI spatial span is regularized *on-the-fly*, leading the IRC receiver be further directive to the user effective channel.
- The proposed solution requires a modest inter-BS signaling overhead.
- The proposed enhanced IRC receiver provides $\sim$ 199% gain of the achievable URLLC outage latency, compared to state-of-the-art relevant IRC literature.

Due to the complexity of the addressed problem herein and the 5G-NR system dynamics, the performance of the proposed solution is assessed using a highly-detailed system level simulator, with a high degree of realism. Following the same simulation methodology in [4], these simulations are based on widely-accepted mathematical models and being validated against the latest 3GPP 5G-NR assumptions. The main functionalities of Layer 1 and 2 of the 5G-NR protocol stack are integrated including the hybrid automatic

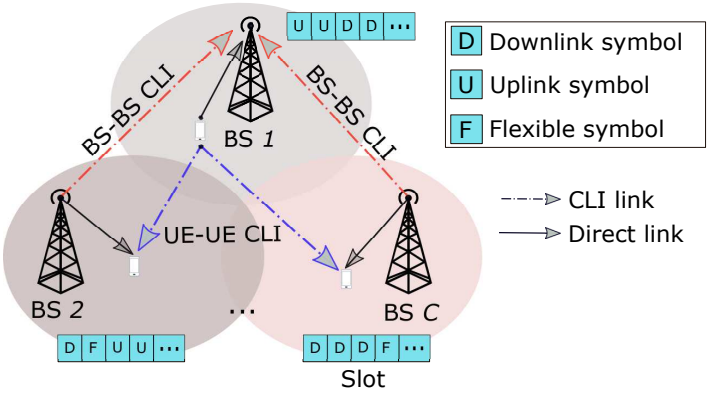


Fig. K.1. Flexible TDD system modeling.

repeat request (HARQ) re-transmissions, 3D spatial channel modeling, adaptive modulation and coding.

The paper is organized as follows. Section 2 introduces the system model of this work. Section 3 presents the proposed solution while Section 4 discusses the performance assessment metrics. Conclusions are drawn in Section 5.

2 System Model

We consider a synchronous dynamic-TDD 5G-NR network of a single cluster of C BSs, each equipped with N antennas. There are K^{dl} and K^{ul} uniformly-distributed DL and UL active UEs per BS, respectively, each with M antennas. The URLLC stochastic FTP3 traffic model is assumed, with finite payload sizes of f^{dl} and f^{ul} bits, and Poisson arrival processes λ^{dl} and λ^{ul}, in the DL and UL directions. Hence, the directional offered loads per BS are given by: $\Omega^{\{\text{dl,ul}\}} = K^{\{\text{dl, ul}\}} \times f^{\{\text{dl,ul}\}} \times \lambda^{\{\text{dl,ul}\}}$, with $\Omega = \Omega^{\text{dl}} + \Omega^{\text{ul}}$ as the total load per cell.

We adopt the latest system assumptions of the 3GPP specifications for URLLC [2]. Hence, a 10-ms RFC is composed of 10 sub-frames, each is constructed of a scalable number of slots. Accordingly, we consider the dynamic 3GPP release-15 slot format design [13], with a flexible structure of the DL, UL and special symbols, respectively, as shown in Fig. K.1. UEs are dynamically multiplexed by the orthogonal frequency division multiple access (OFDMA), with 30 KHz SCS and a physical resource block (PRB) of 12 consecutive SCs. Furthermore, a short TTI duration of 4-OFDM symbols is adopted.

Consider $\mathfrak{B}_{\text{dl}}$, $\mathfrak{B}_{\text{ul}}$, $\mathcal{K}_{\text{dl}}$ and $\mathcal{K}_{\text{ul}}$ as the sets of BSs and UEs with DL and UL transmissions, respectively. Thus, the DL signal at the k^{th} UE, where

$k \in \mathcal{K}_{\text{dl}}$, $c_k \in \mathfrak{B}_{\text{dl}}$, is expressed by

$$\mathrm{y}^{\text{dl}}_{k,c_k} = \underbrace{\mathbf{H}^{\text{dl}}_{k,c_k}\mathbf{v}_k s_k}_{\text{Useful signal}} + \underbrace{\sum_{i\in\mathcal{K}_{\text{dl}}\setminus k}\mathbf{H}^{\text{dl}}_{k,c_i}\mathbf{v}_i s_i}_{\text{BS to UE interference}} + \underbrace{\sum_{j\in\mathcal{K}_{\text{ul}}}\mathbf{G}_{k,j}\mathbf{w}_j s_j}_{\text{UE to UE interference}} + \mathbf{n}^{\text{dl}}_k, \tag{K.1}$$

where $\mathbf{H}^{\text{ul}}_{c_k,k} \in \mathcal{C}^{N\times M}$ denotes the 3GPP 3D-UMA spatial channel [4] from the k^{th} UE to its c^{th}_k BS serving BS, $\mathbf{v}_i \in \mathcal{C}^{N\times 1}$, $\mathbf{w}_k \in \mathcal{C}^{M\times 1}$ and s_k are the zero-forcing pre-coding vector at the c^{th}_i BS, pre-coding vector of the the k^{th} UE, and the transmitted data symbol of the k^{th} UE, respectively, while $\mathbf{n}^{\text{ul}}_{c_k}$ implies the additive white Gaussian noise. Similarly, the UL signal at the c^{th}_k cell, $c_k \in \mathfrak{B}_{\text{ul}}$ from $k \in \mathcal{K}_{\text{ul}}$, is expressed by

$$\mathrm{y}^{\text{ul}}_{c_k,k} = \underbrace{\mathbf{H}^{\text{ul}}_{c_k,k}\mathbf{w}_k s_k}_{\text{Useful signal}} + \underbrace{\sum_{j\in\mathcal{K}_{\text{ul}}\setminus k}\mathbf{H}^{\text{ul}}_{c_k,j}\mathbf{w}_j s_j}_{\text{UE to BS interference}} + \underbrace{\sum_{i\in\mathcal{K}_{\text{dl}}}\mathbf{Q}_{c_k,c_i}\mathbf{v}_i s_i}_{\text{BS to BS interference}} + \mathbf{n}^{\text{ul}}_{c_k}, \tag{K.2}$$

where $\mathbf{Q}_{c_k,c_i} \in \mathcal{C}^{N\times N}$ is the cross-link BS-BS channel between the serving BSs of the k^{th} and i^{th} UEs, $k \in \mathcal{K}_{\text{ul}}$ and $i \in \mathcal{K}_{\text{dl}}$. Then, the post-receiver signal-to-interference ratio (SIR) in the DL γ^{dl}_k and UL $\gamma^{\text{ul}}_{c_k}$ directions are given by,

$$\gamma^{\text{dl}}_k = \frac{\left\|\left(\mathbf{u}^{\text{dl}}_k\right)^{\text{H}}\mathbf{H}^{\text{dl}}_{k,c_k}\mathbf{v}_k\right\|^2}{\sum\limits_{i\in\mathcal{K}_{\text{dl}}\setminus k}\left\|\left(\mathbf{u}^{\text{dl}}_k\right)^{\text{H}}\mathbf{H}^{\text{dl}}_{k,c_i}\mathbf{v}_i\right\|^2 + \sum\limits_{j\in\mathcal{K}_{\text{ul}}}\left\|\left(\mathbf{u}^{\text{dl}}_k\right)^{\text{H}}\mathbf{G}_{k,j}\mathbf{w}_j\right\|^2}, \tag{K.3}$$

$$\gamma^{\text{ul}}_{c_k} = \frac{\left\|\left(\mathbf{u}^{\text{ul}}_k\right)^{\text{H}}\mathbf{H}^{\text{ul}}_{c_k,k}\mathbf{w}_k\right\|^2}{\sum\limits_{j\in\mathcal{K}_{\text{ul}}\setminus k}\left\|\left(\mathbf{u}^{\text{ul}}_k\right)^{\text{H}}\mathbf{H}^{\text{ul}}_{c_k,j}\mathbf{w}_j\right\|^2 + \sum\limits_{i\in\mathcal{K}_{\text{dl}}}\left\|\left(\mathbf{u}^{\text{ul}}_k\right)^{\text{H}}\mathbf{Q}_{c_k,c_i}\mathbf{v}_i\right\|^2}, \tag{K.4}$$

where $\|\bullet\|^2$ is the second-norm, $\mathbf{u}^{\kappa}_k \in \mathcal{C}^{N/M\times 1}$, $\mathcal{X}^{\kappa}, \kappa \in \{\text{ul}, \text{dl}\}$, is the linear minimum mean square error interference rejection combining (LMMSE-IRC) receiver vector [5], with $(\bullet)^{\text{H}}$ as the Hermitian operation.

3 Proposed BS-BS CLI suppression algorithm

The proposed CSA offers an efficient BS-BS CLI cancellation with a limited and 3GPP-compliant overhead space. First, based on [4], the UE-UE CLI is reliably pre-avoided. Then, during the BS-BS CLI slots, victim UL BSs identify the basis of the principal BS-BS CLI interfering sub-space using a *DL precoder map* signaling over the *Xn-interface*. Then, UL BSs estimate the corresponding orthonormal projector sub-space. Finally, for every impacted UL transmission, UL BSs spatially project the estimated IRC interference covariance onto the projector sub-space, prior to decoding, as shown in Fig. K.2.

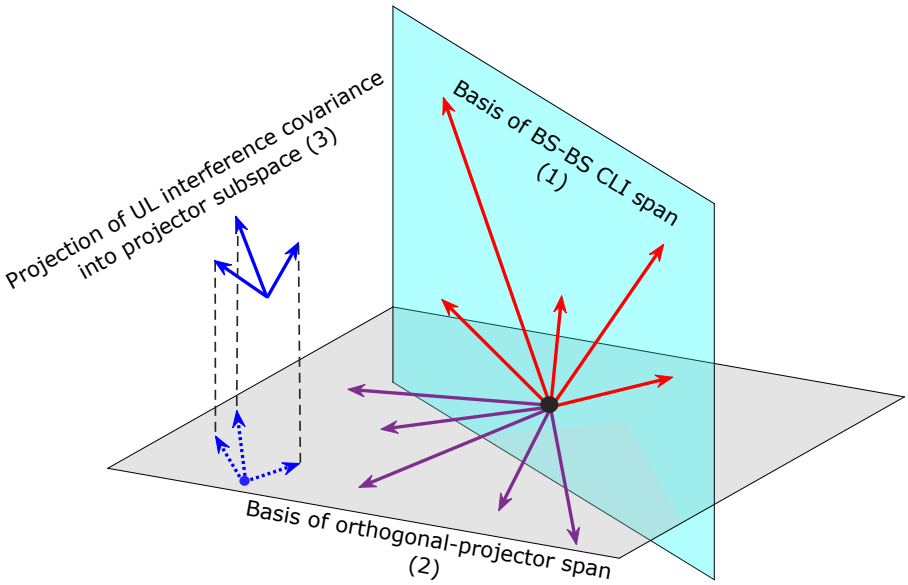


Fig. K.2. BS-BS CSA: CLI projection onto projector sub-space.

3.1 Link-direction adaptation

During each RFC update instance, each BS independently selects an RFC from the RFC code-book which best satisfies its individual link-direction selection criterion, with a respective DL-to-UL symbol ratio, i.e., $d_c : u_c$. We adopt the DL and UL buffered traffic size as the main criterion to select an RFC. The buffered traffic ratio $\mu_c(t)$ is defined as

$$\mu_c(t) = \frac{Z_c^{\mathrm{dl}}(t)}{Z_c^{\mathrm{dl}}(t) + Z_c^{\mathrm{ul}}(t)}, \tag{K.5}$$

where $Z_c^{\mathrm{dl}}(t)$ and $Z_c^{\mathrm{ul}}(t)$ are the total buffered DL and UL traffic of the c_k^{th} BS at the RFC update time t. For example, at the c^{th} BS with $\mu_c(t) = 0.3$, the buffered UL traffic volume is 2.3x the buffered DL traffic, thus, BS consequently selects a slot format of DL:UL symbol ratio as $\sim 1 : 2.3$. The placement of the DL and UL symbols during a slot duration is set evenly to allow for multiple scattered DL and UL transmission opportunities. Accordingly, the achievable capacity T of each cluster is given by

$$T = \sum_{c=1}^{C} \min\left(u_c, u_c^{\mathrm{opt.}}\right) F_c^u + \min\left(d_c, d_c^{\mathrm{opt.}}\right) F_c^d, \tag{K.6}$$

where F_c^u and F_c^d represent the rate utility functions of the UL and DL directions, respectively. $u_c^{\mathrm{opt.}}$ and $d_c^{\mathrm{opt.}}$ are the optimal numbers of the UL and

DL slots that should be adopted during the current RFC to perfectly match the current traffic variations. Thus, UL χ^{ul} and DL χ^{dl} symbol mismatch are inflicted due to the insufficient RFC quantization as

$$\chi^{\text{ul}} = \left| u_c - u_c^{\text{opt.}} \right|. \tag{K.7}$$

$$\chi^{\text{dl}} = \left| d_c - d_c^{\text{opt.}} \right|. \tag{K.8}$$

To maximize capacity T, $u_c = u_c^{\text{opt.}}$ and $d_c = d_c^{\text{opt.}}$ should be always satisfied. Although, $u_c^{\text{opt.}}$ and $d_c^{\text{opt.}}$ may introduce severe BS-BS CLI which severely degrades the UL capacity.

3.2 Proposed BS-BS CSA

During the inter-BS CLI slots within an RFC, the DL-aggressor BSs signal adjacent victim UL BSs with a *DL precoder map* over the *Xn-interface*. Such on-demand signaling denotes a vector of the DL sub-band pre-coding matrix indices (PMIs), which will be used during the next slot by the scheduled DL users. For instance, with 10 MHz bandwidth, i.e., 50 PRBs, 4 antenna port setup, i.e., 4-bit PMI, 3 BS-BS CLI slots, 8-PRB sub-bands, the size of the *DL precoder map* $\mathbb{O}$ can be calculated as:

$$\mathbb{O} = 3 \times \left(\frac{50}{8} \times \left(\log_2 \left(\frac{50}{8} \right) + 4 \right) \right) \simeq 124 \text{ bits per 10 ms.} \tag{K.9}$$

Accordingly, the victim UL BSs seek to identify the strongest $N-1$ sub-band BS-BS interferers as

$$\Lambda^{l}_{b_{\text{ul}},b_{\text{dl}}} = \left\| \mathbf{Q}^{l}_{b_{\text{ul}},b_{\text{dl}}} \mathbf{v}^{l}_{b_{\text{dl}}} \right\|^2, \;\; b_{\text{dl}} \in \mathfrak{B}_{\text{dl}}, b_{\text{ul}} \in \mathfrak{B}_{\text{ul}}, l \in L \tag{K.10}$$

$$\left(\mathfrak{J}_{\;1}^{\;b_{\text{ul}},l}, \ldots, \mathfrak{J}_{\;N-1}^{\;b_{\text{ul}},l} \right) = \left\{ \mathbf{Q}^{l}_{b_{\text{ul}},b_{\text{dl}}} \mathbf{v}^{l}_{b_{\text{dl}}} \rightarrow \underset{b_{\text{dl}},l}{\arg\max} \left(\Lambda^{l}_{b_{\text{ul}},b_{\text{dl}}} \right) \right\}, \tag{K.11}$$

where $\mathbf{Q}^{l}_{b_{\text{ul}},b_{\text{dl}}}$ is the BS-BS channel between the b^{th}_{ul} and b^{th}_{dl} BSs over the l^{th} sub-band, with L as the number of DL aggressor sub-bands. $\mathbf{v}^{l}_{b_{\text{dl}}}$ implies the DL precoder of the scheduled user over the l^{th} sub-band at the b^{th}_{dl} BS, and $\mathfrak{J}_{\;i}^{\;b_{\text{ul}},l} \in \mathcal{C}^{N \times 1}$, with $i = 1, 2, \ldots N-1$, are the identified strongest BS-BS interfering vectors at the b^{th}_{ul} BS. Since the strongest BS-BS interferers, i.e., $\mathbf{Q}^{l}_{b_{\text{ul}},b_{\text{dl}}} \mathbf{v}^{l}_{b_{\text{dl}}}$, are linearly independent due to the independent inter-cell user

scheduling, we can utilize the Gram Schmidt orthogonalization [14] for victim UL BSs to estimate the basis vectors $\beta_i^{b_{\mathrm{ul}},l} \in \mathcal{C}^{N\times 1}$ of a spatial sub-space that spans all $N-1$ BS-BS interferers, as

$$\beta_i^{b_{\mathrm{ul}},l} = \begin{cases} \mathfrak{J}_{\ 1}^{\ b_{\mathrm{ul}},l}, & i = 1 \\ \mathfrak{J}_{\ i}^{\ b_{\mathrm{ul}},l} - \sum\limits_{\tau=1}^{i-1} \mathrm{proj}_{\beta_\tau}\left(\mathfrak{J}_{\ i}^{\ b_{\mathrm{ul}},l}\right), & 2 \geq i \leq N-1, \end{cases} \tag{K.12}$$

$$\mathrm{proj}_{\beta_\tau}\left(\mathfrak{J}_{\ i}^{\ b_{\mathrm{ul}},l}\right) = \left(\frac{\mathfrak{J}_i^{b_{\mathrm{ul}},l} \cdot \beta_\tau}{\|\beta_\tau\|^2}\right)\beta_\tau, \tag{K.13}$$

where $\mathrm{proj}_{\mathsf{X}}(\mathsf{Y})$ implies the spatial line-projection of vector Y on vector X, while $(\mathsf{X}.\mathsf{Y})$ is the dot product. Then, the BS-BS CLI basis matrix $\mathcal{A} \in \mathcal{C}^{N\times N-1}$ is constructed as

$$\mathcal{A} = \left[\beta_1^{b_{\mathrm{ul}},l}, \beta_2^{b_{\mathrm{ul}},l}, \ldots, \beta_{N-1}^{b_{\mathrm{ul}},l}\right]. \tag{K.14}$$

The UL BSs accordingly estimate an orthonormal projector subspace $\mathcal{A}^{\perp} \in \mathcal{C}^{N\times N}$ by the orthogonal projection, as

$$\mathcal{A}^{\perp} = \mathcal{A}\left(\mathcal{A}^{\mathrm{T}}\mathcal{A}\right)^{-1}\mathcal{A}^{\mathrm{T}}, \tag{K.15}$$

where $(\bullet)^{-1}$ and $(\bullet)^{\mathrm{T}}$ are the inverse and transpose operations. Finally, for each UL transmission during the current BS-BS CLI slot, UL BSs calculate the average UL interference covariance matrix $\mathbf{R}_k^{\mathrm{ul}} \in \mathcal{C}^{N\times N}$, in order to construct the LMMSE-IRC receiver matrix for decoding, expressed as

$$\Xi_k^{\mathrm{ul}} = \underbrace{\sum_{j\in\mathcal{K}_{\mathrm{ul}}\setminus k} \mathbf{H}_{c_k,j}^{\mathrm{ul}}\mathbf{w}_j}_{\text{Same-link}} + \underbrace{\sum_{i\in\mathcal{K}_{\mathrm{dl}}} \mathbf{Q}_{c_k,c_i}\mathbf{v}_i}_{\text{Cross-link}}. \tag{K.16}$$

$$\mathbf{R}_k^{\mathrm{ul}} = \Xi_k^{\mathrm{ul}} \times \left(\Xi_k^{\mathrm{ul}}\right)^{\mathrm{H}}. \tag{K.17}$$

Such interference estimate is highly sparse in the spatial domain due to the BS-BS CLI summation, leading to a degraded linear-IRC decoding performance. Thus, prior to decoding, the UL BSs spatially project the interference column vectors of $\mathbf{R}_k^{\mathrm{ul}}$, i.e., $\mathbf{r}_\rho^{\mathrm{ul}}$, onto the projector sub-space basis as

$$\breve{\mathbf{r}}_\rho^{\mathrm{ul}} = \mathrm{proj}_{\ \mathbf{a}_\rho^{\perp}}\left(\mathbf{r}_\rho^{\mathrm{ul}}\right) = \frac{\mathbf{r}_\rho^{\mathrm{ul}} . \mathbf{a}_\rho^{\perp}}{\left\|\mathbf{a}_\rho^{\perp}\right\|^2} \times \mathbf{a}_\rho^{\perp}, \quad \forall \rho = 1, 2, \ldots, N. \tag{K.18}$$

with $\overset{\perp}{\mathbf{a}}_{\rho}$ and $\check{\mathbf{r}}_{\rho}^{\mathrm{ul}}$ are the column vectors of the projector sub-space $\overset{\perp}{\mathcal{A}}$ and the updated interference covariance matrix $\check{\mathbf{R}}_{k}^{\mathrm{ul}}$. Hence, the spatial span of $\check{\mathbf{R}}_{k}^{\mathrm{ul}}$ is regularized by suppressing the sparse $N-1$ BS-BS CLI strongest aggressors, i.e., $\sim$ removing the second summation of eq. (K.16). Finally, the UL LMMSE-IRC receiver matrix is then designed as

$$\mathbf{u}_{k}^{\mathrm{ul}} = \left(\mathbf{H}_{c_k,k}^{\mathrm{ul}}\mathbf{w}_{k}\left(\mathbf{H}_{c_k,k}^{\mathrm{ul}}\mathbf{w}_{k}\right)^{\mathrm{H}} + \check{\mathbf{R}}_{k}^{\mathrm{ul}}\right)^{-1}\mathbf{H}_{c_k,k}^{\mathrm{ul}}\mathbf{w}_{k}. \tag{K.19}$$

Therefore, the UL decoder becomes highly directive towards the span of the direct effective channel, and outside the subspace spanned by the principal BS-BS CLI basis, leading to a significant improvement of the URLLC UL performance.

4 Simulation Results

We adopt extensive system-level simulations to evaluate the performance of the proposed BS-BS CSA, where the major 3GPP 5G-NR assumptions for URLLC [4] are followed, and as listed in Table K.1. A 8×2 antenna setup along with 10 MHz bandwidth of 30 KHz SCS are configured, while the DL transmission power is set to 40 dBm. The offered DL traffic is set to 2x times the UL traffic. During each TTI, BSs dynamically schedule active UEs using the proportional fair (PF) criterion. The achievable SC SINRs are combined using the exponential SNR mapping [15] in order to estimate an effective SINR level. Accordingly, fully dynamic modulation and coding selection (MCS) and Chase combining HARQ re-transmissions are utilized. Pending HARQ re-transmissions are always prioritized over new transmissions during the first available DL/UL slot transmission opportunity. We assess the performance of the proposed solution against the state-of-the-art dynamic-TDD studies as follows:

CLI-free TDD (CF-TDD) [6]: a fully dynamic TDD setup, where BSs independently select the RFCs that best meet their individual traffic demands; however, with the assumption of a perfect UE-UE and BS-BS CLI cancellation. We consider such optimal; although, theoretical baseline, as the reference case.

Non-coordinated TDD (NC-TDD): a fully dynamic TDD is assumed; however, neither inter-BS coordination nor UE-UE and BS-BS CLI cancellation are supported. Herein, BSs achieve the maximum dynamic-TDD adaptation; though, with potentially severe BS-BS and UE-UE CLI, respectively.

Coordinated-RFC TDD (CRFC-TDD) [4]: a hybrid frame design along with a cyclic-offset-based RFC code-book are constructed to reliably pre-avoid the UE-UE CLI. That is, UEs with the worst radio conditions, are pre-

Table K.1: Default simulation parameters.

Parameter	Value
Environment	3GPP-UMA, one cluster, 21 cells
UL/DL channel bandwidth	10 MHz, SCS = 30 KHz, TDD
TDD mode	Synchronized
Antenna setup	$N = 4$, $M = 4$
UL power control	$\alpha = 1, P0 = -103$ dBm
Link adaptation	Adaptive modulation and coding
HARQ configuration	Asynchronous, Chase Combining
Processing times	PDSCH : 4.5-OFDM symbols PUSCH : 5.5-OFDM symbols
TTI configuration	4-OFDM symbols
Traffic model	FTP3 $f^{dl} = f^{ul} = 400$ bits
Offered traffic ratio	DL:UL = 2 : 1
DL/UL scheduling	Proportional fair
DL/UL receiver	LMMSE-IRC
Pattern update periodicity	Slot duration
Transport layer setup	UDP, MTU = 1500 Bytes
User scheduler	Proportional fair

emptively scheduled during certain CLI-free slots, i.e., static slots within all RFCs. Hence, CRFC-TDD boosts the cell-edge capacity; though, performance is highly limited by the more critical BS-BS CLI.

We first evaluate the performance of the proposed scheme in terms of the URLLC outage latency. That is, the achievable URLLC radio latency at 10^{-5} outage probability. It implies the one-way radio latency from the moment a packet arrives at transmitter until it has been successfully decoded at the receiver end, including the standard BS and UE processing delays, dynamic user scheduling delay, and the HARQ re-transmission buffering delay, respectively. Thus, Fig. K.3 and K.4 depict the complementary cumulative distribution function (CCDF) of the UL and DL URLLC latency, respectively, under various offered load levels for the proposed CSA, NC-TDD, and the hypothetical; though, optimal, interference-free (I-free) case, where we assume a perfect inter-cell interference cancellation, including the same-link and cross-link interference. As clearly shown, the proposed CSA scheme offers a decent URLLC outage latency due to the enhanced suppression of the principal BS-BS CLI interferers. The degraded outage latency under the high offered load region is attributed to the inflicted queuing delay due to the dynamic user scheduling, and the increasing same-link inter-cell interference. The NC-TDD with the standard IRC receiver design clearly inflicts a signif-

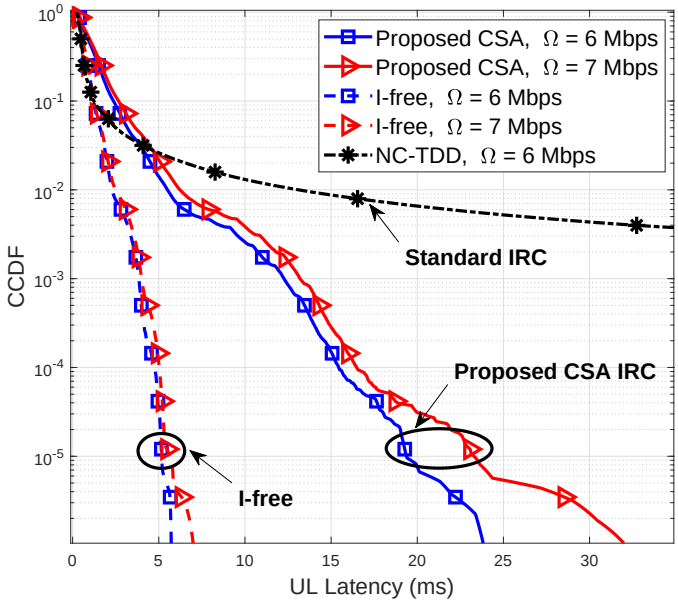


Fig. K.3. BS-BS CSA: UL latency performance.

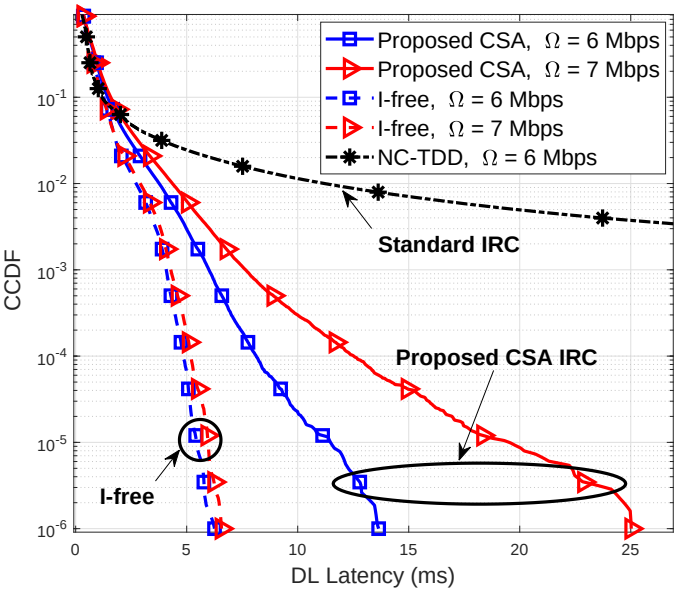


Fig. K.4. BS-BS CSA: DL latency performance.

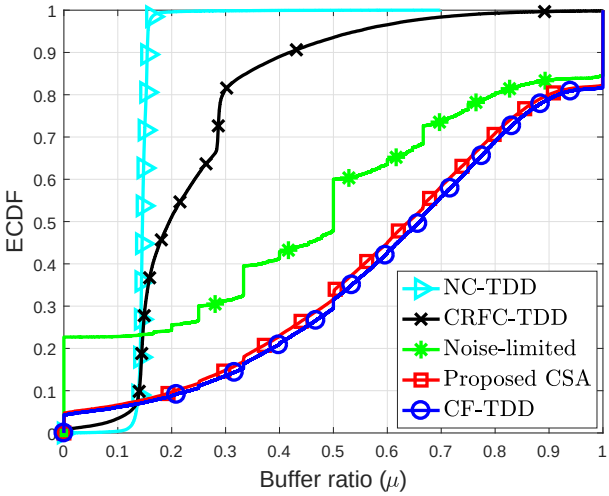


Fig. K.5. BS-BS CSA: traffic buffering performance.

icant degradation of the achievable URLLC latency due to the severe BS-BS CLI.

Table K.2 holds a comparison of the URLLC radio latency in ms, for all schemes under evaluation at different offered traffic loads per BS. To reflect the URLLC reliability targets, the URLLC outage latency at the 10^{-5} outage probability level is evaluated. The CF-TDD clearly provides the best URLLC outage latency performance due to the absolute absence of the UE-UE and BS-BS CLI. The NC-TDD and CRFC-TDD schemes fail to offer a decent URLLC DL and UL outage latency, mainly due to the severe and unhandled BS-BS CLI. Under high offered loads, their respective outage latency increases dramatically due to the inflicted UL re-transmissions.

The proposed BS-BS CSA offers a significant improvement of the URLLC DL and UL outage latency, clearly approaching the optimal CF-TDD under all offered loads; however, with a significantly reduced control overhead size. Due to the sufficient BS-BS CLI suppression, the proposed solution guarantees faster UL transmissions without several HARQ re-transmissions, leaving more time and resources for DL traffic.

These conclusions are confirmed by examining the empirical CDF (ECDF) of the buffered traffic ratio μ as in eq. (K.5), and shown by Fig. K.5. The lower μ, the higher the buffered UL traffic in the scheduling queues. Herein, we introduce a hypothetical case, where the system is only noise-limited, i.e., inter-cell same-link and cross-link interference is assumed to be perfectly suppressed (I-free case as depicted by Fig. K.3 and K.4). This case provides a fairer buffer ratio, i.e., $\mu = 0.5$ at the 50 percentile since all DL and UL

Table K.2: Comparison of the URLLC outage latency, with offered load per BS, and DL:UL = 2 : 1.

Offered load	CF-TDD		NC-TDD		CRFC-TDD		Proposed BS-BS CSA	
	DL	UL	DL	UL	DL	UL	DL	UL
4 Mbps	7.15 0.0%	14.76 0.0%	8.47 +16.9%	105.34 +150.8%	7.75 +8.0%	24.12 +48.1%	7.36 +2.89%	17.4 +16.4%
5 Mbps	8.04 0.0%	15.17 0.0%	1663 +198%	6063 +199%	14.24 +55.6%	201.6 +172%	8.43 +4.7%	18.0 +17%
6 Mbps	11.04 0.0%	16.29 0.0%	7394 +199.4%	18390 +199.6%	3150 +198.6%	12540 +199.4%	11.47 +3.82%	19.32 +17%
7 Mbps	17.28 0.0%	18.23 0.0%	12480 +199.4%	25610 +199.7%	6575 +198.9%	19470 +199.6%	19.8 +13.5%	23.07 +23.4%

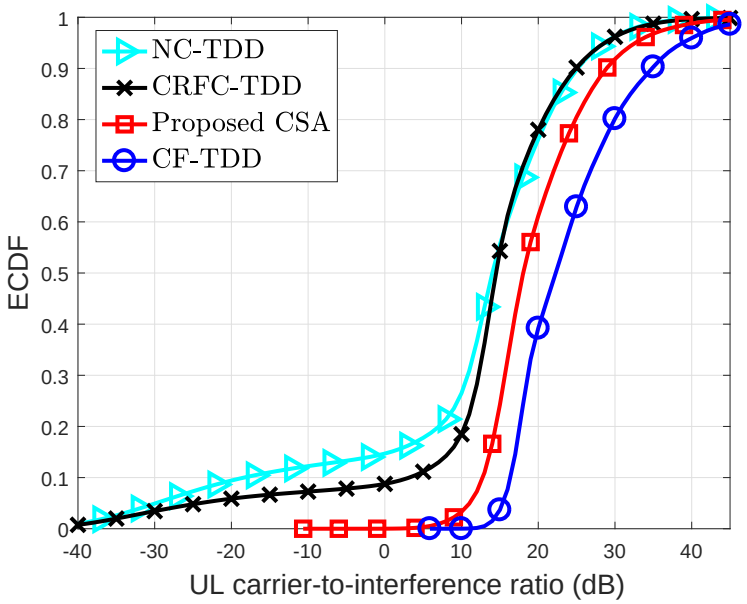


Fig. K.6. BS-BS CSA: UL interference performance.

payloads get successfully decoded from the first time. The NC-TDD and CRFC-TDD offer an extremely low μ, i.e., $\mu = 0.15$ and 0.2 at the 50 percentile. That is, the buffered UL traffic is 5.6x and 4x times the buffered DL traffic, respectively, despite that the offered DL traffic is twice the offered UL traffic. This is due to the UL traffic excessive buffering, due to the consistent consumption of the maximum UL HARQ attempts before failure, and caused by the severe BS-BS CLI. This denotes the link direction adaptation of the dynamic TDD becomes dictated by the HARQ performance, rather than by the new packet arrivals. However, the proposed BS-BS CSA and optimal CF-TDD offer a smooth buffering performance, i.e., $\mu = 0.66$, which implies that buffered UL traffic is 0.525x times the buffered DL traffic, that perfectly aligns with the configured offered traffic ratio.

Fig. K.6 presents the ECDF of the UL carrier-to-interference ratio (CIR) in dB. For a proper presentation, the artificial noise-limited case is excluded. The NC-TDD obviously exhibits the worst CIR performance. The CRFC-TDD only outperforms the NC-TDD over the lower percentiles (cell edge UEs), i.e., +22 dB increase at the 10 percentile, due to the reliable UE-UE CLI preavoidance. Proposed solution offers +31 dB and +9 dB CIR improvements at the 10 percentile, compared to the NC-TDD, and CRFC-TDD, respectively. Unlike the CRFC-TDD, the CIR gain of the proposed solution does not vanish over the higher percentiles, due to the sufficient BS-BS CLI suppression.

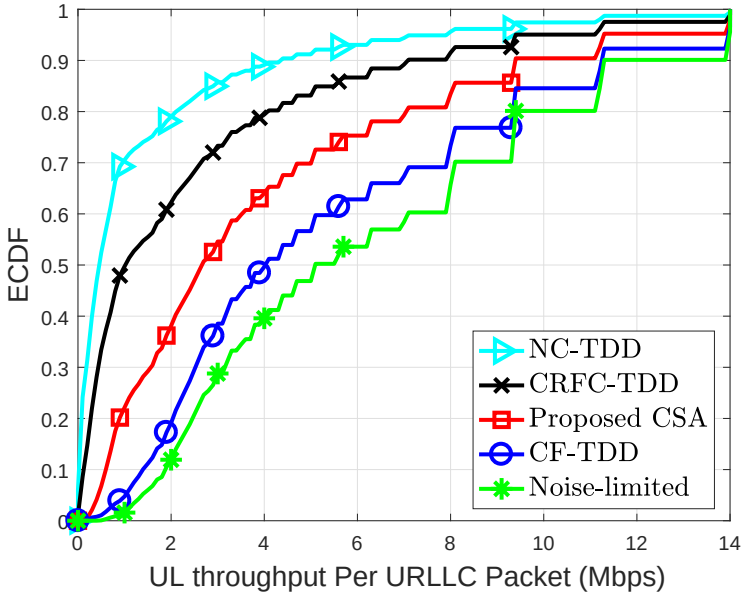


Fig. K.7. BS-BS CSA: UL packet throughput performance.

Proposed scheme approaches the optimal CF-TDD with an average loss of -4 dB.

Similar conclusions are also drawn from Fig. K.7, where the ECDF of the UL throughput per packet is depicted. At the 10 percentile, the proposed BS-BS CSA offers $\sim +156\%$ increase in the achievable URLLC packet throughput, compared to the NC-TDD scheme. This is mainly attributed to the achievable CIR gain of the proposed CSA solution.

5 Concluding Remarks

A high-performance and computation-efficient cross-link interference (CLI) suppression algorithm has been proposed in this work, for 5G dynamic-TDD macro systems. The proposed solution utilizes a BS-BS CLI orthonormal projector sub-space to near-optimally suppress the critical BS-BS CLI *on-the-fly*. Compared to the state-of-the-art dynamic-TDD proposals from industry and academia, the proposed algorithm offers a significant improvement of the URLLC outage latency performance and the ergodic capacity accordingly, while greatly minimizing the control signaling overhead space to B-bit.

The main insights brought by this paper are as follows: (a) achieving the URLLC outage targets in dynamic TDD systems are highly challenged

because of the switching delay among the DL and UL transmission opportunities, and the resultant CLI, (b) the 5G new radio introduces a flexible slot format design, which in turn minimizes the DL/UL switching delay to less than a single millisecond, (c) however, within macro deployments, the BS-BS CLI dominates the URLLC outage performance due to the higher power DL interfering transmissions, (d) thus, inter-cell CLI coordination techniques become vital in order to reap the benefits the flexible TDD systems, and (e) proposed solution demonstrates a near-optimal BS-BS CLI suppression capability while preserving the transmission flexibility of the dynamic TDD technology, and with a limited signaling overhead size.

6 Acknowledgments

This work is partly funded by the Innovation Fund Denmark – File: 7038-00009B.

References

[1] IMT vision – "Framework and overall objectives of the future development of IMT for 2020 and beyond", international telecommunication union (ITU), ITU-R M.2083-0, Feb. 2015.

[2] Ali A. Esswie, and K.I. Pedersen, "On the ultra-reliable and low-latency communications in flexible TDD/FDD 5G networks," *in Proc. IEEE CCNC,* Las Vegas, 2020.

[3] K. I. Pedersen, G. Berardinelli, F. Frederiksen and P. Mogensen, "A flexible 5G wide area solution for TDD with asymmetric link operation," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 122-128, April 2017.

[4] Ali A. Esswie, K.I. Pedersen, and P. Mogensen, "Quasi-dynamic frame coordination for ultra- reliability and low-latency in 5G TDD systems," *in Proc. IEEE ICC*, Shanghai, China, 2019, pp. 1-6.

[5] Tavares, F.M.L.; Berardinelli, G.; Mahmood, N.H.; Sorensen, T.B.; Mogensen, P., "On the potential of interference rejection combining in B4G networks," *in Proc. IEEE VTC*, Las Vegas, NV, 2013, pp. 1-5.

[6] R1-1701146, *Dynamic TDD interference mitigation concepts in NR*, Nokia, Alcatel-Lucent Shanghai Bell, 3GPP RAN1 #88, Feb. 2017.

[7] A. Łukowa and V. Venkatasubramanian, "Coordinated user scheduling in 5G dynamic TDD systems with beamforming," *in Proc. IEEE PIMRC*, Bologna, 2018, pp. 596-597.

[8] A. Łukowa and V. Venkatasubramanian, "Performance of strong interference cancellation in flexible UL/DL TDD systems using coordinated muting, scheduling and rate allocation," *in Proc. IEEE WCNC*, Doha, 2016, pp. 1-7.

[9] E. d. O. Cavalcante, G. Fodor, Y. C. B. Silva and W. C. Freitas, "Distributed beamforming in dynamic TDD MIMO networks with BS to BS interference constraints," *IEEE Wireless Commun. Lett. , vol. 7, no. 5, pp. 788-791, Oct. 2018.*

[10] Z. Huo, N. Ma and B. Liu, "Joint user scheduling and transceiver design for cross-link interference suppression in MU-MIMO dynamic TDD systems," *in Proc. IEEE ICCC*, Chengdu, 2017, pp. 962-967.

[11] Ali A. Esswie, and K.I. Pedersen, "Inter-cell radio frame coordination scheme based on sliding codebook for 5G TDD systems," *in Proc. IEEE VTC-spring*, Kuala Lumpur, Malaysia, 2019, pp. 1-6.

[12] J. W. Lee, C. G. Kang and M. J. Rim, "SINR-ordered cross link interference control scheme for dynamic TDD in 5G system," *in Proc. IEEE ICOIN*, Chiang Mai, 2018, pp. 359-361.

[13] 5G; NR; *Physical layer procedures for control*; (Release 15), 3GPP, TS 38.213, V15.3.0, Oct. 2018.

[14] S. Haykin, *Digital commun. systems*. Hoboken, Wiley & Sons, 2014.

[15] 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 Trans. Wireless Commun.*, vol. 10, no. 10, pp. 3436–3448, Oct. 2011.

Part IV

Machine Learning Potential Towards Improved Dynamic-TDD Operation

Machine Learning Potential Towards Improved Dynamic-TDD Operation

This part of the thesis presents the potential of the classical machine learning (ML) algorithms for the cellular TDD 5G-NR deployments. In particular, a reinforcement learning (RL) based framework is developed for BSs to autonomously design the TDD radio frames that offer the best possible URLLC outage latency performance. The proposed solution is implemented and evaluated through extensive system level simulations with a high degree of realism, and finally compared to the TDD frame coordination schemes introduced in Part III of the thesis.

1 Problem Formulation

The design and selection of the TDD radio frames have been demonstrated to a have a vital impact on the achievable radio latency [1-3]. For multi-cell multi-UE URLLC dynamic TDD deployments, with downlink and uplink sporadic packet arrivals, the TDD radio frame selection problem is further challenging, i.e., an NP-hard problem [3], due to the simultaneously opposite inter-user link requirements. The developed TDD schemes in addition to the quoted state-of-the-art relevant literature in Part III offer considerable URLLC performance merits [4-7]. However, each of those schemes is best functioning within a certain offered load region, i.e., achievable performance gain is mainly dependent on the offered load region. Furthermore, those schemes do not consider the actual and time-varying latency performance in designing the BS-specific TDD radio frames.

Therefore, in this part of the thesis, we take one step further where our main hypothesis is that an ML based solution is viable for BSs to dynamically design the TDD radio frame structures. The time-varying latency statistics

are considered as an input for the ML-based TDD framework to learn and select the TDD radio frame structure which provides the best possible URLLC outage performance without the requirement of the inter-BS coordination. That is, by monitoring the past and current traffic and latency statistics, an ML based solution can learn and predict the upcoming radio frame switching pattern that ensures a balanced downlink and uplink traffic handling and accordingly, a minimal URLLC outage latency. However, it is a challenging task to ensure a robust convergence and learning performance, particularly when the URLLC target outage probability of 10^{-5} is considered. Therefore, a significant amount of work in this Part is devoted to ensuring the proper convergence of the developed latency-aware ML-based TDD scheme. Finally, for joint URLLC-eMBB coexistence deployments, QoS-awareness is vital to satisfy the diverse performance requirements of the URLLC and eMBB services, respectively. Hence, for such deployments, we develop QoS-aware TDD link selection and dynamic UE scheduling framework.

2 Objectives

The objective of this part of the PhD thesis are as follows:

- Study the available ML models in recent literature and the corresponding overall operation, performance merits, learning potential, and processing complexity, respectively. The purpose is to select the ML solution that offers robust convergence and learning performance, implementable and requires lower processing complexity.
- Develop an ML TDD frame design scheme that is adaptable to the various offered loads and network deployments. In this part, we consider the Q-reinforcement-learning (QRL) [8] due its simple implementation and performance potential.
- Develop an RL based distributed solution for BSs to dynamically in time design the radio frame switching pattern in macro networks. The high-level objective is to guarantee a balanced downlink and uplink traffic handling of the future packet arrivals across the next radio frames while minimizing the tail latency distribution.
- Explore the developed ML based TDD framework for the indoor industrial factory deployments where multi-QoS transmissions are adopted, i.e., eMBB-URLLC coexistence.

3 Included Articles

The main relevant papers of this PhD part are listed as follows:

Paper L: Online Radio Pattern Optimization Based On Dual Reinforcement Learning Approach For URLLC 5G Networks

In this paper, we propose and develop a dual RL approach to dynamically determine the TDD radio pattern in multi-cell multi-UE dynamic TDD macro URLLC deployments. A primary RL network is defined such that it seeks to periodically select the sufficient number of downlink and uplink transmission opportunities across each upcoming radio frame. The objective is to ensure a balanced downlink and uplink traffic service such that traffic accumulation in either direction is not likely to occur. The secondary RL layer seeks to learn the corresponding best possible downlink and uplink link switching pattern, which offers the minimum URLLC outage latency. The reward function of the primary RL network is defined as the combined downlink and uplink buffered traffic ratio for which the the primary learning algorithm seeks to preserve around its mean. That is to achieve a balanced downlink and uplink traffic buffering performance. For the secondary learning network, the downlink and uplink buffer latency statistics are considered as the learning input. Non-uniform Kaiser window filter weights are applied on each of the downlink and uplink UE latency samples to calculate an overall single cell-specific latency indication for both directions. The filtering is vital in order to smooth out any potential abrupt latency changes, and hence, to avoid disturbing the learning convergence. The reward function of the secondary network is defined as the combined downlink and uplink latency ratio for which the secondary learning instance aims to maintain around its mean.

As verifying the convergence performance of the RL solutions is a challenging task, a significant amount of work of this Paper has been devoted to ensuring a robust convergence performance. We adopt the real-time temporal difference (TD) of the defined reward functions for both the primary and secondary networks as a good indication of the convergence of the proposed RL solution. The TD implies the transition rate of the relative learning outcomes to the state of the learning agent. After 1 second (in real time) of the convergence delay, the proposed RL framework has shown a stable transition rate of the learning outcomes, i.e., an average TD rate of only 8%. This is achieved by tuning the simulation warm-up time, where the system consumes it until it gets loaded, in order to allow for the proposed RL solution to converge, over which action exploration is prioritized over greedy exploitation. During the actual simulation time (RL inference time), where the performance statistics are captured, the proposed RL solution always prioritizes the best actions (TDD patterns) that contribute to achieving the minimum possible radio latency performance.

The performance of the proposed RL framework has been assessed using highly-detailed system level simulations. Accordingly, performance comparisons of the achievable URLLC outage latency are performed with the state-of-the-art TDD schemes in addition to the developed proposals in Part III of the thesis. For instance, the proposed learning framework offers 70% and 53% URLLC outage latency reduction compared to the fully dynamic and static TDD schemes, respectively, for an offered load of 1 Mbps with even downlink and uplink distribution. In general, the achievable latency performance of the proposed learning solution is dependent on the offered load region. Specifically, at the high offered load region, the proposed learning algorithm displays a similar URLLC outage latency performance as the developed semi-static TDD scheme in Part III. This is resulting from the fact that over such high load region, the CLI dominates the URLLC outage performance. Hence, avoiding the CLI occurrence by explicit inter-BS signaling exchange (i.e., semi-static TDD scheme) is as attractive as the developed learning scheme without the requirement of the inter-BS coordination.

Paper M: Analysis of Outage Latency and Throughput Performance in Industrial Factory 5G TDD Deployments

This paper analyzes the achievable URLLC outage latency performance in the emerging industrial factory (InF) automation deployments. We consider the state-of-the-art InF channel modeling alongside an optimized set of the system configurations which are specific to such deployments. We start by analyzing the impact of the uplink transmit power control settings as well as the network CLI on the achievable URLLC performance. Furthermore, we consider the multi-QoS coexistence scenarios, where a mixture of the latency-critical URLLC and the capacity-demanding enhanced mobile broadband (eMBB) services are incorporated. For the latter case, we develop a quality of service (QoS)-aware TDD link selection framework and dynamic UE scheduling in order to balance achieving a decent URLLC outage latency while satisfying the eMBB capacity demands. Finally, we apply the RL based TDD scheme, developed in Paper L in such InF deployments. The main conclusions of this paper are as follows: (1) the accurate setting of the UL transmit power control configurations is vital to achieve a decent URLLC outage latency. In particular, setting those to be too high leads to an increased inter-cell interference; however, setting the UL power control configurations to be too low results in a highly degraded uplink spectral efficiency. Based on our extensive sensitivity analysis, we recommend configuring the UL power control P_0 in range of -65 to -60 dBm in such InF networks, (2) for multi-service coexistence scenarios, the QoS-awareness of the dynamic UE scheduler as well as the TDD adaptation process is vital to achieve a decent URLLC outage latency, and (3) the RL-based TDD solution has been shown as a suitable solution for such InF networks, where either a similar or clear performance gain is achieved over the former schemes in Part III, depending the offered

load region.

4 Main Findings and Recommendations

Main Findings

As depicted by Fig. IV.1 [9, Paper L], the proposed RL solution incorporates a dual RL layers. The primary network considers the past and current traffic buffering performance in order to estimate the sufficient number of the downlink and uplink transmission opportunities across the next TDD radio frame. The objective is to achieve a balanced downlink and uplink transmission performance regardless of the time-varying downlink and uplink packet arrivals. The corresponding secondary RL sub-network is therefore activated to select the specific TDD radio switching pattern that provides the best possible inter-UE URLLC outage latency.

It is a challenging task to achieve a robust RL convergence performance, due to its sparse value function. Therefore, we have performed an extensive sensitivity analysis in order to achieve the best possible RL convergence performance under the presumed system dynamics and settings. That is, by running a large set of system level simulations with different warm-up periods. The warm-up duration denotes the time needed for the simulations to get loaded, and over which no performance statistics are extracted. We also utilize such period for the proposed RL solution to converge to the near-optimal policy.

During the warm-up time (considered as the RL convergence delay, over which the RL algorithm visits the majority of possible actions and the corresponding reward functions stably converge), we prioritize the action exploration rather than action exploitation. This denotes that the adopted RL algorithm tends to select random TDD radio frames to rapidly explore all possible actions during the warm-up. After the warm-up time, when the performance statistics start to be extracted, the proposed RL algorithm always prioritizes the actions (TDD radio frames) that minimize the defined cost function (radio latency), i.e., inference time. Based on our sensitivity analysis of a large system level simulation of different convergence delays and action exploration probabilities, we therefore adopt 25% probability of action exploration during warm-up and 0% exploration probability during the actual simulation time, i.e., inference time. The warm-up time is set to approximately 2100 slots (1 second in real time).

The proposed solution is shown to offer a considerable improvement of the URLLC latency compared to the dynamic TDD (dTDD) and static TDD (sTDD) schemes, respectively. Fig. IV.2 [9, Paper L] shows the complementary cumulative distribution function (CCDF) of the combined URLLC down-

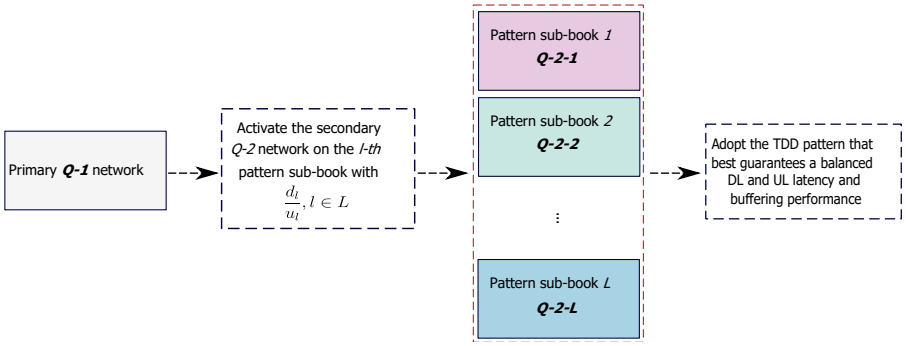


Fig. IV.1: The dual reinforcement learning approach for online optimization of the TDD radio frame [9, Paper L].

link/uplink latency for the proposed solution, dTDD, sTDD, and FDD cases, respectively. This is for a total offered load of 1 Mbps/cell with an equal downlink and uplink traffic split. As obviously shown, the proposed RL framework offers a clear URLLC outage latency enhancements compared to dTDD and sTDD schemes, approaching the FDD case. In particular, the proposed RL framework offers 70% and 53% outage latency reduction compared to the dTDD and sTDD schemes, respectively. The performance gain of the proposed RL solution is due to the achievable learning potential of the TDD radio frame structures, in terms of the number and placement of the downlink and uplink transmission opportunities across the radio frame, in order to continuously reduce the packet buffering delay.

Furthermore, we investigate the achievable URLLC outage performance in the emerging industrial factory (InF) deployments [10]. Fig. IV.3 shows the CCDF of the achievable DL/UL URLLC radio latency alongside with dynamic TDD frame adaptation for different offered loads. Obviously, at the very low load region Ω =0.5 Mbps, the URLLC latency target of 1 ms is fulfilled due to the lower resource utilization. At the higher offered loads, the packet queuing delay starts to be more visible. Moreover, the CLI is shown to be less of a problem for the InF networks due to the indoor propagation conditions and the smaller difference between the downlink and uplink transmission power, compared to the macro case.

We also apply the developed learning solution of Paper L for such networks, where Fig. IV.4 [11, Paper M] depicts the URLLC latency comparison between the TDD and proposed RL-based TDD schemes, respectively. At the low load region, and due to the very low resource utilization, there is no visible queuing delay neither CLI. Therefore, the RL-based TDD and conventional TDD offer a similar URLLC latency performance. Over the higher load region Ω =3 Mbps, the RL-based solution clearly provides 45% latency reduction compared to the TDD scheme. Herein, the main performance

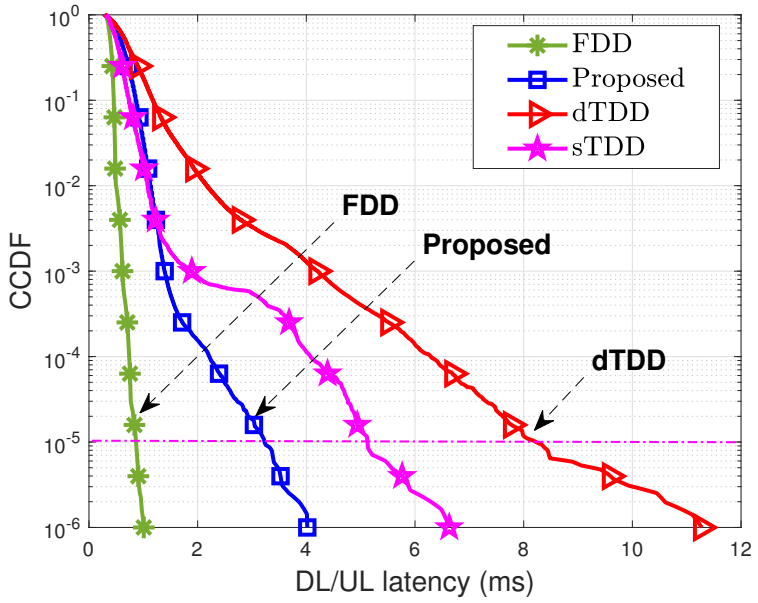


Fig. IV.2: The achievable URLLC latency performance of the proposed dual reinforcement learning approach [9, Paper L].

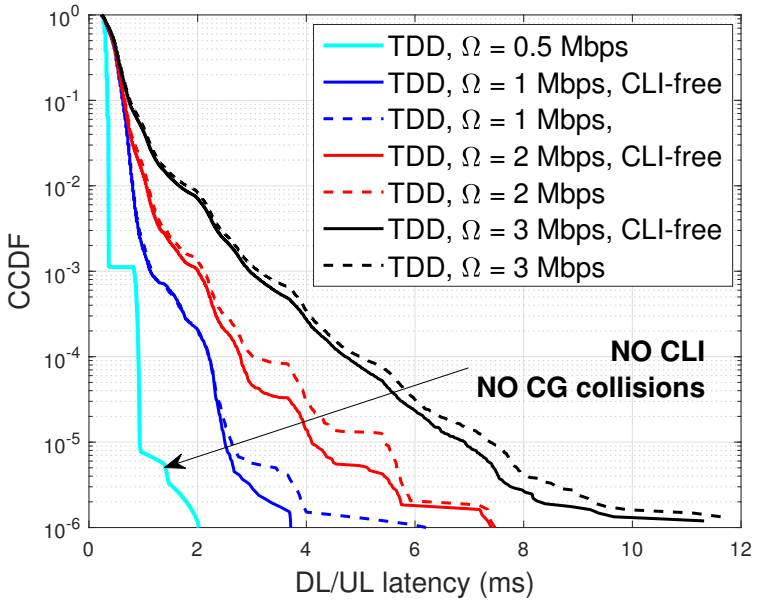


Fig. IV.3: The achievable URLLC latency performance of within the InF networks [11, Paper M].

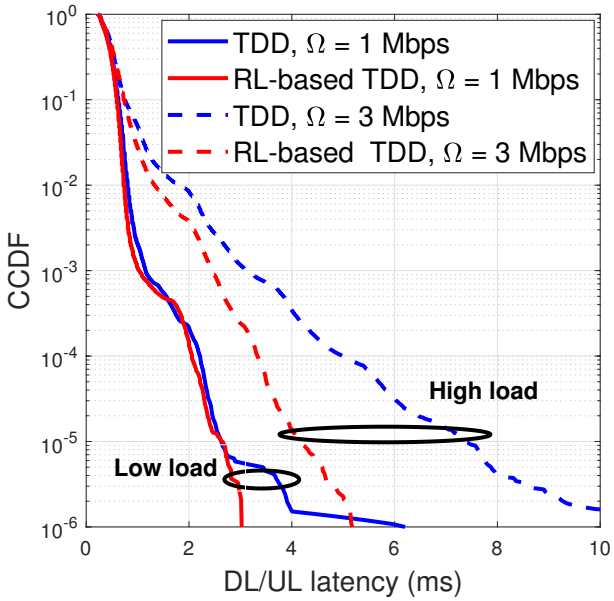


Fig. IV.4: The achievable URLLC latency performance of the proposed dual reinforcement learning approach for InF networks [11, Paper M].

merit results from the latency-awareness of RL-based TDD scheme, where the structure of each TDD radio frame is continuously optimized such that the combined DL/UL packet buffering delay is minimized.

Finally, Fig. IV.5 [11, Paper M] shows the achievable URLLC latency in InF networks for an eMBB-URLLC service coexistence scenario. For such deployments, the QoS-awareness of the dynamic user scheduler and the TDD link selection criterion becomes vital to achieve a decent URLLC outage latency. We compare the obtained performance of the URLLC when latency-aware [12] and latency-unaware dynamic user schedulers are adopted. The former considers the packet queuing delay into the scheduling criterion in order to reduce the packet segmentation probability. In case the packet segmentation is unavoidable, the scheduler seeks to select the packet segments that lead to the lowest control overhead, where the traffic of a maximum single UE is segmented per TTI. However, the latter resembles the well-known proportional fair (PF) scheduling criterion which is latency-unaware, i.e., only considers the achievable user capacity. Furthermore, the TDD link selection criterion is defined as the combined downlink and uplink total (eMBB+URLLC) buffered traffic ratio. For eMBB-URLLC deployments, such QoS-unaware criterion can be dictated by the offered eMBB traffic. This leads to a degraded URLLC latency performance due to the additional URLLC packet queuing delays.

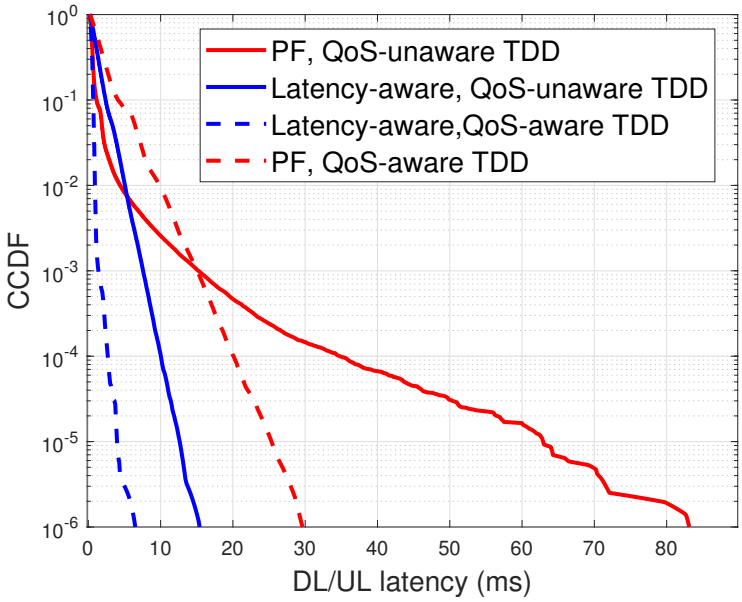


Fig. IV.5: The achievable URLLC latency performance for eMBB-URLLC coexistence in InF networks [11, Paper M].

Hence, we adopt a QoS-aware TDD link selection criterion where only the URLLC buffered traffic size is considered in the TDD link selection operation.

As clearly observed from Fig. IV.5 [11, Paper M], the latency and QoS awareness is key to achieving a decent URLLC outage latency performance in joint eMBB-URLLC deployments. For instance, using the the QoS-aware TDD link selection, the achievable URLLC outage latency of the PF scheduler is reduced by more than 38 ms compared to the case with the QoS-unaware TDD link selection.

Main recommendations

In the following, we summarize the major research recommendations of this part of the thesis as follows:

1. The proposed RL-based TDD solution is demonstrated as an effective ML solution to dynamically learn and predict the TDD radio pattern in dynamic TDD URLLC deployments. The proposed solution is applicable to a diversity of offered load regions and various network deployments such as the InF and UmA.

2. The proposed RL-based TDD framework offers a significant or a similar URLLC performance improvement compared to the TDD solutions which are tailored to specific offered load regions.

3. For joint URLLC-eMBB coexistence deployments, the QoS-awareness is vital to achieve a decent URLLC outage performance. This includes the dynamic user scheduling and the selection criterion of the TDD radio frame.

4. For InF network, the network CLI is shown to be less of a problem. This is attributed to the indoor propagation conditions and the smaller difference between the downlink and uplink transmission power, compared to the macro case. The optimization of the uplink power control settings (for InF case, $P_o = -61$ dBm) is key to achieve a decent URLLC outage latency.

5. Finally, the proposed RL framework is fully compliant with the latest 3GPP 5G-NR system specifications. It requires a light processing complexity without inter-BS coordination signaling.

References

[1] Ali A. Esswie, and K.I. Pedersen, “Cross link interference suppression by orthogonal projector in 5G dynamic-TDD URLLC systems,” *in Proc. IEEE WCNC*, May 2020.

[2] *Cross link interference handling and remote interference management (RIM) for NR*; (Release 16); 3GPP, TR 38.828, V16.0.0, June 2019.

[3] Ali A. Esswie, and K.I. Pedersen, “On the ultra-reliable and low-latency communications in flexible TDD/FDD 5G networks,” *in Proc. IEEE CCNC*, , Las Vegas, NV, USA, Jan. 2020, pp. 1-6.

[4] A. Łukowa and V. Venkatasubramanian, "Performance of strong interference cancellation in flexible UL/DL TDD systems using coordinated muting, scheduling and rate allocation," *in Proc. IEEE WCNC*, Doha, April 2016, pp. 1-7.

[5] E. d. O. Cavalcante, G. Fodor, Y. C. B. Silva and W. C. Freitas, "Distributed beamforming in dynamic TDD MIMO networks with cell to cell interference constraints," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 788-791, Oct. 2018.

[6] Ali A. Esswie, K.I. Pedersen, and P. Mogensen, “Quasi-dynamic frame coordination for ultra- reliability and low-latency in 5G TDD systems,” *in Proc. IEEE ICC*, Shanghai, China, May 2019, pp. 1-6.

[7] A. A. Esswie and K. I. Pedersen, "Inter-cell radio frame coordination scheme based on sliding codebook for 5G TDD systems," *in Proc. VTC*, Kuala Lumpur, Malaysia, April 2019, pp. 1-6.

[8] Richard S. Sutton and Andrew G. Barto,. *Reinforcement Learning: An Introduction*. A MIT Press, Cambridge, MA, 2018

[9] A. A. Esswie, K. I. Pedersen, and P. Mogensen, "Online radio pattern optimization based on dual reinforcement-learning approach for 5G URLLC networks," *in IEEE Access*, 2020.

[10] *Field measurement results from an operational factory floor at 3.5 GHz and 28 GHz*, 3GPP, RAN WG1, R1-1810659, Chengdu, Oct. 2018.

[11] A. A. Esswie, and K. I. Pedersen, "Analysis of URLLC outage Performance in ondustrial factory 5G TDD deployments," Submitted to VTC-spring, 2021.

[12] 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. IEEE VTC-Spring*, Kuala Lumpur, Malaysia, 2019, pp. 1-6

Paper L

Online Radio Pattern Optimization Based on Dual Reinforcement-Learning Approach for 5G URLLC Networks

Ali A. Esswie, Klaus I. Pedersen Preben E. Mogensen

The paper has been published in the
2020 IEEE Access

Abstract

— The fifth generation (5G) radio access technology is designed to support highly delay-sensitive applications, i.e., ultra-reliable and low-latency communications (URLLC). For dynamic time division duplex (TDD) systems, the real-time optimization of the radio pattern selection becomes of a vital significance in achieving decent URLLC outage latency. In this study, a dual reinforcement machine learning (RML) approach is developed for online pattern optimization in 5G new radio TDD deployments. The proposed solution seeks to minimizing the maximum URLLC tail latency, i.e., min-max problem, by introducing nested RML instances. The directional and real-time traffic statistics are monitored and given to the primary RML layer to estimate the sufficient number of downlink (DL) and uplink (UL) symbols across the upcoming radio pattern. The secondary RML sub-networks determine the DL and UL symbol structure which best minimizes the URLLC outage latency. The proposed solution is evaluated by extensive and highly-detailed system level simulations, where our results demonstrate a considerable URLLC outage latency improvement with the proposed scheme, compared to the state-of-the-art dynamic-TDD proposals.

Index Terms— Dynamic-TDD; URLLC; 5G new radio; Machine learning; Reinforcement learning; Q-learning; Cross link interference (CLI).

1 Introduction

One of the main drivers of the fifth generation (5G) radio standardization is the ultra-reliable and low-latency communications (URLLC) service class [1]. URLLC entail the transmission of sporadically-arriving small-payload packets with one-way radio latency of 1 ms and 99.999% success probability [2]. As the early 5G commercial enrollments are foreseen over the 3.5 GHz unpaired spectrum, due to its wide spectrum availability [3], time-division duplexing (TDD) technology is vital for the success of the 5G. With dynamic TDD, base-stations (BSs) independently utilize either a downlink (DL) or uplink (UL) transmission opportunity at a time in order to meet their capacity and latency demands, respectively [4].

Achieving the URLLC targets for dynamic TDD deployments is highly challenging [5] due to: (i) the non-concurrent availability of the DL and UL transmission opportunities, and (ii) the potentially strong cross-link interference (CLI) between neighboring BSs and user-equipment's (UEs), adopting opposite transmission directions. The fine selection of the DL and UL symbol structure during a TDD radio pattern has been demonstrated to immensely impact the achievable URLLC outage latency, even under hypothetically CLI-free conditions [5]. Moreover, the TDD radio pattern selection is an NP-hard problem for multi-cell multi-UE deployments, due to the simultaneous requests of conflicting link directions, and thus, this is the problem addressed

in this work.

1.1 State of The Art Dynamic TDD Studies

The third generation partnership project (3GPP) has recently standardized a flexible frame structure for dynamic TDD 5G systems [6]. That is, BSs configure a 10-ms radio frame, consisting of multiple slot formats, each is composed of DL [D], UL [U], and flexible [F] symbols, respectively. The latter indicates the symbol set that can be dynamically configured, through a dedicated radio signaling from BS to UEs, either as DL or UL or act as a guard time among successive DL and UL symbols, respectively. Such design offers a highly resilient framework for adapting the radio patterns to the time-variant offered traffic needs. One simple way to approach such frame flexibility is to semi-statically adapt the radio pattern configuration to the current average traffic conditions [7]. In particular, a common radio pattern is periodically updated and adopted by all neighboring BSs in order to meet the average network capacity demands, with minimal inter-BS signaling overhead.

Recent prior-art proposals seek to utilize the standardized pattern update flexibility. In [8, 9], a predefined set of radio frame configurations is adopted, with different possible DL and UL symbol ratios and pre-determined structures (aka - a frame-book). Thus, BSs dynamically select those patterns from the frame-book which best satisfy their individual link selection criteria, e.g., the currently buffered traffic.

However, as a consequence to the BS-specific pattern adaptation, neighboring BSs may simultaneously adopt opposite link directions, resulting in a severe CLI. For instance, the BS-BS CLI is demonstrated as a fundamental limitation of the achievable UL capacity [5], mainly due to the larger DL transmit power compared to the victim UL power. CLI mitigation and coordination schemes have therefore been widely investigated over recent prior art. In [10-12], coordinated cross-cell beam-forming, UL power control and cell muting are proposed to limit the residual network CLI, especially towards the more CLI-sensitive cell-edge UEs. Joint UL transceiver design [13-15], based on inter-cell signaling of the UEs' spatial signatures, is also introduced in order to isolate the BS-BS CLI spatial subspace from that is of the desired UL transmission. The drawback of those proposals is mainly the requirement of a large inter-cell signaling overhead space. Therefore, simpler and less-coordination-overhead demanding opportunistic CLI avoidance schemes [16, 17] have been suggested to offer attractive capacity and latency merits, where the BS-BS and UE-UE CLI is pre-averted on a best effort basis. This encompasses the design of a hybrid TDD pattern with a slot-aware dynamic UE scheduling. Although those proposals require simpler implementation complexity, they optimize the URLLC performance on a heuristic basis, which may jeopardize the achievable URLLC reliability and latency performance.

1.2 Machine Learning Potential in Dynamic TDD Systems

Although the quoted TDD studies present clear advancements and valuable findings, the radio pattern selection procedure is yet deemed as a challenging problem towards the success of the 5G TDD deployments. This is particularly relevant for dynamic URLLC multi-cell multi-user TDD deployments, where the DL and UL traffic arrivals are highly sporadic in time, and with strict latency and reliability constraints. As stated, the problem of selecting the optimum TDD switching pattern is NP-hard and has so far been addressed by means of rather simple heuristic solutions. In this study, we go one step further where our hypothesis is that machine learning (ML) is a viable solution to be utilized at the BS nodes to dynamically select the best possible TDD switching pattern. That is, based on monitoring the past and current traffic and latency performance per BS, an ML capability shall learn and predict the best TDD switching pattern for the next radio frame.

ML techniques have been notably studied with the 5G wireless radio communications [18] for various radio design aspects such as interference management [19] and radio resource management [20 - 24]. Generally, ML can be divided into three categories [25] as: (1) supervised-ML (SML), where the input data is a priori known and well-labeled for model training. The SML model is continuously trained with the right *question-answer* pairs until it approaches the optimal model, (2) unsupervised-ML (UML), where the input data is neither a priori known nor labeled. Accordingly, data clustering and dimensionality reduction become necessary to extract the meaningful and independent feature vectors, and (3) reinforcement-ML (RML), where unlike SML and UML, it does not require offline model training. Thus, RML has been widely employed towards the real-time decision-making applications. RML algorithms are goal-oriented which consistently in time learn how to achieve a complex objective, through an iterative; however, simple, process of action exploration and environment observation, respectively. The model-free RML algorithms are mainly categorized to on-policy and off-policy techniques [26], respectively. The former directly learns the optimal policy while the latter approaches the near-optimal policy through more conservative exploration. On-policy ML algorithms, such as state–action–reward–state–action (SARSA) [27], have been demonstrated particularly attractive for the critical use cases where the learning agent is critically challenged with a tight training duration, and over which it cannot employ a sub-optimal policy, e.g., walking robots over a cliff.

For the latency-critical URLLC traffic, SML and UML are substantially challenging for practical deployments due to the required large size of dedicated training samples to reach a sufficient learning of the target URLLC 10^{-5} outage probability. Therefore, SML and UML methods are not adopted in this study as deemed too demanding for achieving the required level of

model training. We prioritize RML as being more suitable for the type of system and objectives addressed in this paper, and hence, this is the focus of this study.

1.3 Paper Contribution

In this paper, a dual-RML based pattern optimization scheme is proposed for dynamic TDD 5G systems. The proposed solution targets minimizing the inflicted URLLC radio latency on a real-time basis, and accordingly, improving the achievable URLLC outage performance. The proposed scheme utilizes nested RML layers, where the primary layer estimates the number of the DL and UL symbols of the upcoming radio pattern to satisfy the foreseen offered traffic. Subsequently, the secondary RML sub-layers determine the DL and UL symbol structure that achieves the minimum possible URLLC radio latency. The proposed algorithm neither requires inter-cell signaling exchange overhead nor offline dedicated training, i.e., online and distributed pattern optimization. Performance results show a significant improvement of the URLLC outage latency with the proposed solution, compared to state-of-the-art dynamic TDD proposals. The major contributions of this paper are listed as follows:

- We propose a novel dual reinforcement machine learning (RML) approach for online URLLC outage optimization for 5G-NR TDD networks.
- Unlike the state-of-the-art relevant TDD solutions [7-19], the proposed solution considers the joint capacity and latency statistics to optimize the URLLC outage latency performance. It is fully compliant with the current 3GPP 5G-NR standard specifications for dynamic-TDD deployments. The proposed framework neither requires inter-cell signaling exchange nor high processing complexity.
- Compared to the state-of-the-art TDD literature, the proposed scheme offers a considerable URLLC latency and reliability enhancement, under various DL and UL offered loads. It achieves 70% outage latency reduction compared to the standard dynamic TDD scheme.

Due to the complexity of the 5G new radio system design and the addressed problems herein, the proposed solution has been evaluated by extensive and highly-detailed system level simulations. Those simulations incorporate the major functionalities of the 5G new radio protocol stack, e.g., dynamic resource allocation and user scheduling, adaptive modulation and coding schemes (MCS), hybrid automatic repeat request (HARQ) re-transmissions, and the 3GPP 3D spatial channel modeling, respectively. Special care is given to ensure statistically-reliable results.

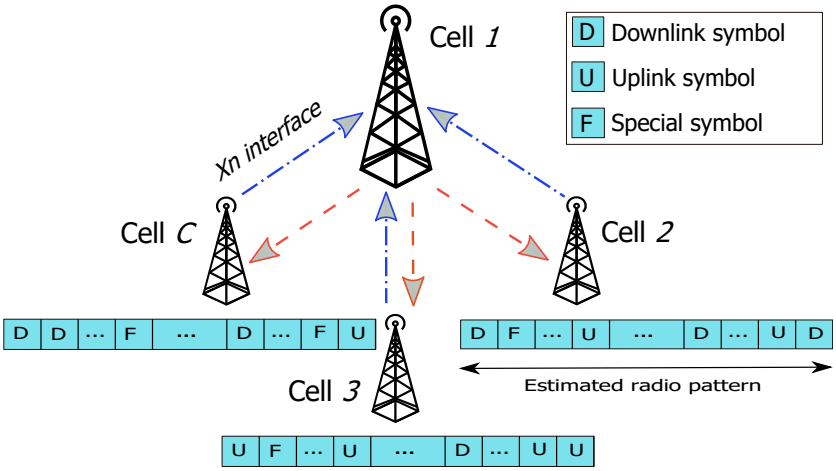


Fig. L.1. System model: dynamic-TDD macro deployment.

The paper is organized as follows. Section 2 presents the system modeling, while Section 3 formulates the problem addressed in this work. Section 4 introduces a brief overview of the Q-reinforcement learning and Section 5 presents the detailed description of the proposed solution. Section 6 introduces the state-of-the-art dynamic-TDD schemes, against which we evaluate the performance of the proposed solution. The performance evaluation results appear in Section 7, while conclusions are drawn in Section 8.

2 Setting the Scene

2.1 System Model

We consider a macro 5G dynamic TDD deployment, where base-stations (BSs) are configured with 3-sector cell setting. Thus, there are a total of C cells, each is equipped with N antennas. Each cell serves an average of $K = K^{\text{dl}} + K^{\text{ul}}$ uniformly-distributed UEs, each equipped with M antennas, K^{dl} and K^{ul} denote the average numbers of the DL and UL UEs per cell. In this study, we assume that UEs are requesting DL and UL traffic with different DL and UL packet arrival rates, respectively. We adopt the URLLC-alike FTP3 traffic modeling with packet sizes of f^{dl} and f^{ul} bits, and a Poisson Arrival Process, with mean packet arrival rates of λ^{dl} and λ^{ul}, in the DL and UL directions, respectively [28]. The average offered load per cell in the DL direction is: $\Omega^{\text{dl}} = K^{\text{dl}} \times f^{dl} \times \lambda^{\text{dl}}$, and UL direction: $\Omega^{\text{ul}} = K^{\text{ul}} \times f^{ul} \times \lambda^{\text{ul}}$. The total offered load per cell is given as: $\Omega = \Omega^{\text{dl}} + \Omega^{\text{ul}}$.

We follow the 3GPP guidelines for the 5G TDD system modeling, as

shown by Fig. L.1. UEs are dynamically multiplexed using the orthogonal frequency division multiple access (OFDMA). In line with the 3GPP URLLC studies [28], the SCS is selected to equal 30 kHz as it offers sufficiently short symbol durations to fulfill the considered latency requirements, while still having enough cyclic prefix duration to cope with time-dispersion for the considered macro scenario, with the physical resource block (PRB) of twelve consecutive SCSs. Furthermore, we assume a short transmission time interval (TTI) duration of 4 OFDM symbols towards faster URLLC transmissions. Prior to the start of each radio frame [29], i.e., every 10-ms, the BS decides the next radio frame pattern based on the proposed RML solution. In this work, we assume a single guard OFDM symbol between every DL and UL symbols in the radio frame, in order to account for the DL channel delay spread before the UL transmissions are triggered.

Accordingly, when a DL packet arrives at the cell, it is first processed by the serving cell, and thereafter, is buffered towards the first available DL transmission opportunity of the current TDD radio pattern, i.e., TDD pattern switching delay. The time to prepare a DL transmission block is taken explicitly into account in line with 3GPP 5G-NR specifications [30]. Then, the cell scheduler dynamically multiplexes all pending DL packets using the proportional fair criterion, where some DL packets can be further queued to the next DL transmission instant, i.e., scheduling queuing delay. The HARQ re-transmissions are always prioritized over new transmissions. Herein, dynamic link adaptation is also adopted, where the DL transmission MCS is adaptively selected such that it corresponds to a first-transmission block error rate (BLER) of 1%. The MCS selection is typically based on the most recently received channel quality indication (CQI) report from the UE. The scheduled users are notified with a scheduling grant (aka DL control information – DCI), and the overhead from the corresponding physical-layer control signaling is taken explicitly into account in line with [31]. At the UE-side, DL reception is subject to processing time for decoding of the DL transmission. In case the transmitted DL packet is not successfully decoded by the intended UE, the UE triggers the transmission of a HARQ negative acknowledgment (NACK) during the next available UL transmission opportunity of the radio pattern, where the appropriate radio resources are allocated. Correspondingly, serving cell re-transmits the respective DL packet to be soft-combined at the UE.

For UL packet transmissions, we assume configured grant (CG) transmission (aka grant-free) with fixed MCS per UE [32], The use of CG means that as soon as a packet arrives at the UE, it is immediately prepared for UL transmission, and transmitted at the first coming UL TTI opportunity. Each CG transmission includes a robust preamble, so the receiving BS is able to detect from which UE the transmission is coming. The CG parameterization is such that UEs with high path-loss are transmitting on the full bandwidth with

a conservative MCS corresponding to QPSK rate 1/8, such that one URLLC payload of 32 bytes can be transmitted. In line with [32], UEs with better path-loss conditions are configured to transmit on one quarter of the carrier bandwidth with MCS QPSK rate 1/2. Such UE classification, of high or low path-loss conditions, is based on a predefined coupling gain threshold $\hat{c}$. The UL transmit power $\Sigma\,[dBm]$ is configured to equal

$$\Sigma\,[dBm] = \min\left\{\Sigma_{\max},\ P0 + 10\log_{10}\left(\wp\right) + \alpha\eth + \nabla_{\text{MCS}}\right\}, \qquad \text{(L.1)}$$

where $\Sigma_{\max}$ is the max UE transmit power, $P0$ is the target power spectral density, $\wp$ is the number of granted UL PRBs, α and $\eth$ denote the path-loss compensation factor and path-loss, respectively. ∇_{MCS} is an UL power boost factor where $\nabla_{\text{MCS}} = 10$ dB for QPSK1/2 and $\nabla_{\text{MCS}} = 0$ dB for QPSK1/8 in line with [32]. As CG transmissions from multiple users may occur at the same time on overlapping resources, uplink transmissions from UEs are subject to potential intra-cell interference, which only to a certain extent can be combated by the a linear BS multi-antenna receiver. If the BS fails to correctly decode a CG transmission from an UL UE, it immediately sends an uplink scheduling grant for the UE in the next coming DL TTI, issuing an UL HARQ re-transmission from the UE in the next UL TTI. The UL HARQ re-transmission is sent using the same configuration (bandwidth and MCS configurations) as the original transmission, but with a +3 dB transmission power boost to enhance the probability of decoding the HARQ re-transmission at the BS [32].

As an input to the proposed RML algorithm to dynamically select the radio frame configuration, cells should be aware of the directional traffic and latency statistics. Hence, in this work, we assume a realistic knowledge of those statistics at the cell side. Particularly, the DL traffic size, including buffered and new packets, is spontaneously known at the cell stack. However, in the UL direction, new UL packet transmissions are not a priori known at the cell. Those are only identified at the cell side when the first UL transmission attempt is either failed or correctly received. Thus, we only assume the UL HARQ-buffered traffic size is known at the cell side.

For capturing the latency statistics of the corresponding DL/UL buffers, we define the head of line delay (HoLD) per packet per UE as the time from the moment a DL/UL packet arrives at the transmitter packet data convergence protocol (PDCP) layer until it is successfully received at the receiver end, and forwarded to the PDCP layer. The exact DL HoLD is known at the cell.

For the UL direction, it is not known at the BS-side when a packet arrives at the UE-side. The BS only becomes aware of pending UL transmissions from the UE when it first tries to transmit those to the BS. The UL HoDL is therefore only monitored at the BS-side as the time from the first UL trans-

mission attempt until successful decoding, i.e., essentially corresponding to the effective HARQ retransmission round trip time. Due to the adaptation of the TDD switching pattern and presence of both inter-cell and intra-cell interference from other UEs, as well as the potential BS-BS CLI, the UL HARQ round trip time is time-variant, and often dominant for the tail of the UL packet distribution.

Finally, the achievable one-way radio URLLC latency at the 10^{-5} outage probability is the main performance metric [5] of this work. It implies the delay from the moment when a URLLC packet arrives at the packet data convergence protocol layer of the transmitter until it is successfully received at the intended receiver, summing the BS and UE processing delays, buffering delay due to dynamic UE scheduling, delay to the first DL/UL transmission opportunity, and HARQ re-transmission delay.

2.2 Signal Model

Assume $\mathfrak{B}_{\text{dl}}$, $\mathfrak{B}_{\text{ul}}$, $\mathcal{K}_{\text{dl}}$ and $\mathcal{K}_{\text{ul}}$ as the BS and UE sets with DL and UL transmissions, respectively. Thus, the DL signal at the k^{th} UE, where $k \in \mathcal{K}_{\text{dl}}$, $c_k \in \mathfrak{B}_{\text{dl}}$, is given as

$$\mathrm{y}^{\text{dl}}_{k,c_k} = \underbrace{\mathbf{H}^{\text{dl}}_{k,c_k}\mathbf{h}_k x_k}_{\text{Useful signal}} + \underbrace{\sum_{i\in\mathcal{K}_{\text{dl}}\setminus k}\mathbf{H}^{\text{dl}}_{k,c_i}\mathbf{h}_i x_i}_{\text{BS to UE interference}} + \underbrace{\sum_{j\in\mathcal{K}_{\text{ul}}}\mathbf{G}_{k,j}\mathbf{o}_j x_j}_{\text{UE to UE interference}} + \mathbf{n}^{\text{dl}}_k, \tag{L.2}$$

where $\mathbf{H}^{\text{dl}}_{k,c_i} \in \mathcal{C}^{M\times N}$ is the DL 3D-UMA fading channel [33] from the cell serving the i^{th} UE, to the k^{th} UE, $\mathbf{h}_i \in \mathcal{C}^{N\times 1}$, $\mathbf{o}_k \in \mathcal{C}^{M\times 1}$ and x_k are the zero-forcing precoding vector at the c_i^{th} BS, precoding vector of the k^{th} UE, and the transmitted data symbol of the k^{th} UE, respectively, while $\mathbf{G}_{k,j} \in \mathcal{C}^{M\times M}$ implies the cross-link channel between the k^{th} and j^{th} UEs, and $\mathbf{n}^{\text{dl}}_k$ represents the additive white Gaussian noise. The UL signal at the c_k^{th} cell, $c_k \in \mathfrak{B}_{\text{ul}}$ from $k \in \mathcal{K}_{\text{ul}}$, is expressed by

$$\mathrm{y}^{\text{ul}}_{c_k,k} = \underbrace{\mathbf{H}^{\text{ul}}_{c_k,k}\mathbf{o}_k x_k}_{\text{Useful signal}} + \underbrace{\sum_{j\in\mathcal{K}_{\text{ul}}\setminus k}\mathbf{H}^{\text{ul}}_{c_k,j}\mathbf{o}_j x_j}_{\text{UE to BS interference}} + \underbrace{\sum_{i\in\mathcal{K}_{\text{dl}}}\mathbf{P}_{c_k,c_i}\mathbf{h}_i x_i}_{\text{BS to BS interference}} + \mathbf{n}^{\text{ul}}_{c_k}, \tag{L.3}$$

where $\mathbf{P}_{c_k,c_i} \in \mathcal{C}^{N\times N}$ is the BS-BS channel between the serving BSs of the k^{th} and i^{th} UEs, $k \in \mathcal{K}_{\text{ul}}$ and $i \in \mathcal{K}_{\text{dl}}$. Then, the post-receiver signal-to-interference ratio in the DL γ^{dl}_k and UL $\gamma^{\text{ul}}_{c_k}$ directions are expressed by,

$$\gamma^{\text{dl}}_k = \frac{\left\|\left(\mathbf{u}^{\text{dl}}_k\right)^{\text{H}}\mathbf{H}^{\text{dl}}_{k,c_k}\mathbf{h}_k\right\|^2}{\sum\limits_{i\in\mathcal{K}_{\text{dl}}\setminus k}\left\|\left(\mathbf{u}^{\text{dl}}_k\right)^{\text{H}}\mathbf{H}^{\text{dl}}_{k,c_i}\mathbf{h}_i\right\|^2 + \sum\limits_{j\in\mathcal{K}_{\text{ul}}}\left\|\left(\mathbf{u}^{\text{dl}}_k\right)^{\text{H}}\mathbf{G}_{k,j}\mathbf{o}_j\right\|^2}, \tag{L.4}$$

$$\gamma_{c_k}^{\mathrm{ul}} = \frac{\left\| \left(\mathbf{u}_k^{\mathrm{ul}}\right)^{\mathrm{H}} \mathbf{H}_{c_k,k}^{\mathrm{ul}} \mathbf{o}_k \right\|^2}{\sum\limits_{j \in \mathcal{K}_{\mathrm{ul}} \setminus k} \left\| \left(\mathbf{u}_k^{\mathrm{ul}}\right)^{\mathrm{H}} \mathbf{H}_{c_k,j}^{\mathrm{ul}} \mathbf{o}_j \right\|^2 + \sum\limits_{i \in \mathcal{K}_{\mathrm{dl}}} \left\| \left(\mathbf{u}_k^{\mathrm{ul}}\right)^{\mathrm{H}} \mathbf{P}_{c_k,c_i} \mathbf{h}_i \right\|^2}, \tag{L.5}$$

where $\|\cdot\|^2$ is the second-norm, $\mathbf{u}_k^{\kappa} \in \mathcal{C}^{N/M \times 1}$, $\mathcal{X}^{\kappa}, \kappa \in \{\mathrm{ul}, \mathrm{dl}\}$, is the linear minimum mean square error interference rejection combining (LMMSE-IRC) receiver vector [34], with $(\bullet)^{\mathrm{H}}$ as the Hermitian operation.

3 Problem Formulation

The URLLC applications require a stringent radio latency bound and with a rare per-packet violation probability. In dynamic TDD systems, the URLLC outage performance is dominated by the number and the structure of the DL d_c and UL u_c symbols across the configured radio pattern. In this study, our objective is to optimize the radio pattern configuration, i.e., to determine the number and structure of d_c and u_c, for a faster and DL-UL balanced traffic transmission, and thus, an improved URLLC outage latency, as

$$\left(\frac{d_c}{u_c}\right)^* \triangleq \left\{ \frac{d^i}{u^i} \ : \ \frac{d^i}{u^i} \in \mathfrak{T} \right\}, \tag{6.a}$$

$$(\hat{w}_c)^* \triangleq \left\{ \hat{w}^j \ : \ \hat{w}^j \in \hat{\mathrm{W}} \right\}, \tag{6.b}$$

(L.6)

Subject to:

$$\begin{cases} \arg\min\limits_{c,t} \left(\Upsilon_{c,t}\right) \\ \arg\min\limits_{k} \left(\varphi_{c,k}\right), \forall k \in \mathcal{K}_{\mathrm{ul/dl}} \end{cases}$$

where $\mathfrak{T}$ and $\hat{\mathrm{W}}$ are the inclusive sets of all possible d_c/u_c ratios and structures, respectively. $\Upsilon_{c,t}$ denotes the buffered traffic difference of the c^{th} cell at time t between the amount of buffered DL and UL traffic volume, and $|\cdot|$ denotes the absolute value. $\varphi_{c,k}$ indicates the achievable one-way radio latency of the k^{th} UE.

The first constraint (L.6.a) implies that the selected TDD pattern at an arbitrary time should contribute to closing the gap among the buffered DL and UL traffic size over the pattern duration, regardless of the variant DL and UL PRB capacity and the offered traffic ratio $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}}$. The second constraint (L.6.b) ensures that the UE-specific latency performance is monotonically optimized.

4 Overview of the Q Reinforcement Machine Learning (Q-RML)

The RML [35, 36] is a vital branch of the machine learning. It has been widely applied in real-time decision-making problems such as autonomous driving and robot control. RML follows the mathematical framework of the Markov decision process [37], where the learning outcomes are partially random and tightly related to the environment. Accordingly, the goal of an RML agent is to obtain an optimal policy $\pi^* : S \rightarrow A$, which determines an action $a \in A$ under state $s \in S$, thus, to optimally maximize or minimize a pre-defined value function V^{π}. The value function is typically expressed in terms of the expected discounted cumulative reward or penalty at time epoch t, as

$$V^{\pi}(s_t) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma \; r_t(s_t, a_t) \,|s_0 = s\right], \tag{L.7}$$

$$V^{\pi}(s_t) = \mathbb{E}_{\pi}\left[r_t(s_t, a_t) + \gamma \; V^{\pi}(s_{t+1}) \,|s_0 = s\right], \tag{L.8}$$

where $\mathbb{E}(\cdot)$ implies the statistical mean, $r_t(s_t, a_t)$ is the immediate reward or penalty, observed from the environment after taking an action a under state s at time epoch t, and $\gamma \in [0, 1]$ is the discount factor on future rewards or penalties. Simple dynamic programming schemes can be utilized to solve eq. (7), when the state transition probabilities are a priori known. The RML aims to finding the optimal policy π^* when the system dynamics are not known through an iterative process of continuously adjusting its policy. In that sense, Q-RML is one of the most effective RML techniques. In this study, we adopt the baseline off-policy Q-learning approach to rapidly learn the optimal policy during the warm-up time. Thus, unlike the case with the on-policy RML techniques, we preserve a sufficiently enough pre-training time in order for the Q-RML approach to converge to the optimal greedy policy before impacting the actual inference performance.

A Q-RML agent applies the actions which closes the gap between the current policy π and the optimal target policy π^*, i.e., $\pi \overset{t}{\rightarrow} \pi^*$, such that the observed reward or penalty from the environment is monotonically optimized as

$$V(s_t) = F\left[r_t(s_t, a_t) + \gamma \; V^{\pi}(s_{t+1})\right], \tag{L.9}$$

where the optimization function F is the optimization function, which defines the Q-RML learning goal, in terms of the corresponding value function, as given by

$$F \cong \begin{cases} \arg\max\limits_{a \in A}(\mathscr{F}), V(s) \rightarrow \text{reward} \\ \arg\min\limits_{a \in A}(\mathscr{F}), V(s) \rightarrow \text{penality} \end{cases}, \tag{L.10}$$

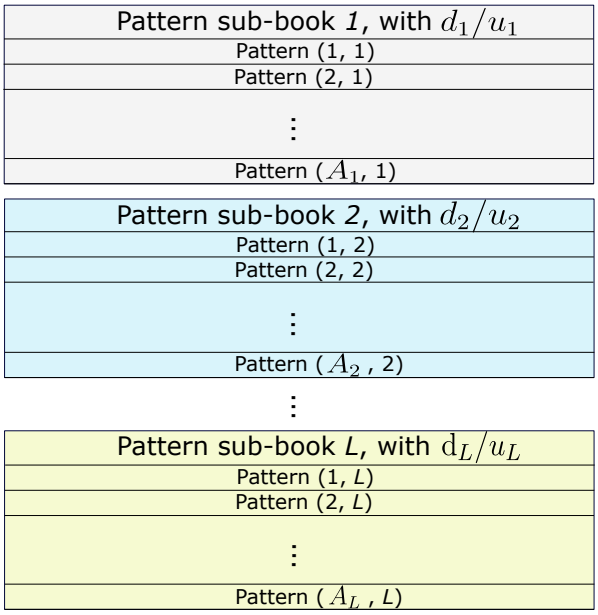


Fig. L.2. Proposed algorithm: pattern-book design.

where $\mathscr{F}$ is the actual environment value function of the Q-RML instance, defined as a reward or penalty.

5 Proposed RML based pattern optimization

The proposed solution incorporates a dual RML approach for joint capacity and latency online optimization. We consider the model-free Q-reinforcement-machine-learning (Q-RML) algorithm [35] for its performance merits and low implementation complexity under a moderate state-action space size. As depicted by Fig. L.2, a pattern-book is constructed, where there are L pattern sub-books, each is of a size **Card** $(\mathcal{A}_l)$ radio patterns, where **Card** $(\cdot)$ denotes the cardinality of a set, and $\mathcal{A}_l$ is the set of radio patterns in the l^{th} sub-book, $\forall l \in L$. All radio patterns within a single sub-book share the same d^l/u^l symbol ratio; although, with different symbol structures. The nested pattern book design allows for utilizing independent Q-RML instances to estimate the DL and UL symbol ratio as well as the respective symbol structure.

As depicted by Fig. L.3, the primary Q-RML network, i.e., $Q-1$, estimates the number of the DL d_c and UL u_c symbols of the upcoming radio pattern. The $Q-1$ target is to select the symbol ratio which contributes into a faster; though, balanced DL and UL, traffic service over the pattern duration; however, adopting a default symbol structure. Then, the secondary Q-RML

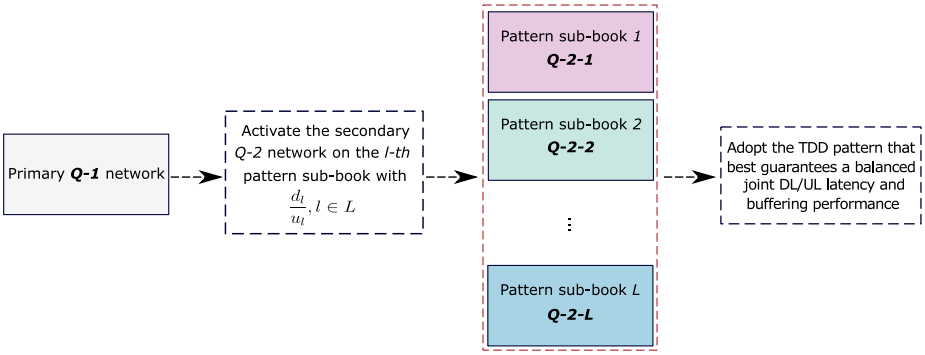


Fig. L.3. Proposed algorithm: nested Q-network design for TDD pattern selection.

sub-networks, i.e., $Q-2-l$, determine the best possible DL and UL symbol structure, following the calculated $^{d^l}/_{u^l}$ ratio from $Q-1$, in order to minimize the filtered HoLD statistics, leading to a significantly improved and DL/UL fair URLLC outage performance. In the following, we represent the sole operation of $Q-1$ layer as Algorithm-1, and as Algorithm-2 when $Q-1$ and $Q-2-l$ layers are simultaneously incorporated.

5.1 Primary Q-RML Network For Balanced DL/UL Buffering

In dynamic TDD macro systems, the achievable UL capacity is highly variant from the corresponding DL capacity, mainly due to the severe BS-BS CLI. For instance, a linear mapping from $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}} = {}^1/_1$ to $^{d^c}/_{u^c} = {}^1/_1$ may not be sufficient. Accordingly, the TDD pattern adaptation process becomes fully dictated by the residual UL traffic, i.e., including buffered, and retransmitted traffic, leading to a highly degraded DL capacity accordingly due to the subsequent starvation of the DL transmission opportunities. Thus, the Algorithm-1 RML instance seeks a rapid; but; balanced DL and UL traffic transmission, by estimating the sufficient $^{d^c}/_{u^c}$ ratio for a given DL and UL traffic statistic every radio pattern duration.

In that regard, at the ς^{th} slot of the radio pattern, $\varsigma = 1, 2, \ldots, \xi$, with ξ as the number of slots per the configured radio pattern, the relative traffic ratio $\mu_{[t,c]}(\varrho)$ of the c^{th} BS at time epoch t is defined as

$$\mu_{[t,c]}(\varsigma) = \frac{\mathrm{Z}^{\mathrm{dl}}_{[t,c]}(\varsigma)}{\mathrm{Z}^{\mathrm{dl}}_{[t,c]}(\varsigma) + ({}^1/_\iota)\ \mathrm{Z}^{\mathrm{ul}}_{[t,c]}(\varsigma)}, \tag{L.11}$$

where $\mathrm{Z}^{\mathrm{dl}}_{[t,c]}(\varsigma)$ and $\mathrm{Z}^{\mathrm{ul}}_{[t,c]}(\varsigma)$ are the aggregated DL and UL buffered traffic size of the ς^{th} slot during the current pattern, and ι is the first-transmission average UL BLER, experienced at the BS side. As discussed in Section 2.A, $\mathrm{Z}^{\mathrm{ul}}_{c}(\varsigma)$ implies only the UL HARQ-buffered packets. Accordingly, to ensure

fairness against $Z^{\mathrm{dl}}_{[t,c]}(\varsigma)$, the average ι is incorporated in eq. (L.11) such that the term $(1/\iota)\ Z^{\mathrm{ul}}_{[t,c]}(\varsigma)$ reflects the average total UL offered traffic size at the BS. The instantaneous traffic ratios $\mu_{[t,c]}(\varsigma)$ are linearly averaged over the duration of the TDD pattern as

$$\overline{\mu}_{[t,c]} = \frac{1}{\xi} \sum_{\varsigma=1}^{\xi} \mu_{[t,c]}(\varsigma), \tag{L.12}$$

with $\overline{\mu}_{[t,c]}$ as the relative traffic ratio at time epoch t. The traffic ratio $\overline{\mu}_{[t,c]} \rightarrow [0,1]$ reflects the combined buffering performance of the DL and UL traffic. For instance, $\overline{\mu}_{[t,c]} = 0.1$ denotes that the buffered UL traffic is 9x times the DL traffic. Accordingly, a state space $S^{(1)}$ is defined to represent the DL and UL traffic buffering conditions at an arbitrary time epoch t, as

$$S^{(1)}_t = \left\{ s^{(1)}_{1,t},\ s^{(1)}_{2,t},\ \ldots,\ s^{(1)}_{\mathfrak{J}_1,t} \right\}, \tag{L.13}$$

with $\mathfrak{J}_1$ as the size of the $Q-1$ state space. In principal, the state of the learning agent is determined as a function of the input performance metric, by an arbitrary mapping structure. In this work, we adopt a linear mapping of the quantized traffic volume to determine the BS state. Accordingly, the traffic-to-state mapping is designed as

$$s^{(1)}_t = \begin{cases} s^{(1)}_{1,t}, & \overline{\mu}_{[t,c]} < \mu_{\min} \\ s^{(1)}_{2,t}, & \mu_{\min} \leq \overline{\mu}_{[t,c]} < \mu_{\min} + \sigma \\ s^{(1)}_{3,t}, & \mu_{\min} + \sigma \leq \overline{\mu}_{[t,c]} < \mu_{\min} + 2\sigma \\ \vdots & \vdots \\ s^{(1)}_{\mathfrak{J}_1,t}, & \overline{\mu}_{[t,c]} \geq \mu_{\max} \end{cases}, \tag{L.14}$$

where the traffic ratio quantization step σ is given as:

$$\sigma = \frac{\mu_{\max} - \mu_{\min}}{\mathfrak{J}_1 - 2}, \tag{L.15}$$

where $\mu_{\max}$ and $\mu_{\min}$ indicate the pre-defined minimum and maximum allowable levels of the traffic ratio $\overline{\mu}_{[t,c]}$. In that sense, $s^{(1)}_{1,t}$ indicates a traffic state where the buffered UL traffic is much larger than of the DL direction. Thus, an intermediate state is the system favorable target state to offer a balanced DL and UL buffering performance.

The action space $A^{(1)}$ is constructed to represent the set of all possible Algorithm-1 outcomes as

$$A^{(1)}_t = \left\{ a^{(1)}_{1,t},\ a^{(1)}_{2,t},\ \ldots,\ a^{(1)}_{L,t} \right\}, \tag{L.16}$$

where $a^{(1)}_{l,t} \equiv {}^{d^l}/_{u^l}, \forall l \in L$. Particularly, the Algorithm-1 instance determines the pattern sub-book, and hence, the corresponding ${}^{d^l}/_{u^l}$ ratio, to be adopted

over the upcoming radio pattern. Herein, we assume the immediate environment return $\Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right)$ of Algorithm-1 represents a performance penalty, as

$$\Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right) = \left|\overline{\mu}_{[t,c]} - \mathfrak{m}^{(1)}\right|, \forall i \in \mathfrak{J}_1,\, l \in L, \tag{L.17}$$

where $\mathfrak{m}^{(1)}$ denotes the mean value of the traffic ratio distribution $\overline{\mu}_{[t,c]}$. The mean value of the buffered traffic ratio $\mathfrak{m}^{(1)}$ is selected as the target of the primary Q-RML learning, since it allows for selecting the TDD pattern, with a certain DL-to-UL symbol ratio that is likely to preserve a balanced downlink and uplink buffered traffic performance. Specifically, $\Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right)$ indicates the immediate cost, observed from the environment upon taking an action $a^{(1)}_{l,t}$ under state $s^{(1)}_{i,t}$, and is calculated in terms of how much deviant the traffic ratio $\overline{\mu}_{[t,c]}$ is from its balanced mean $\mathfrak{m}^{(1)}$. That is, a large Θ implies either unfavorable much buffered DL or UL traffic. At an arbitrary time epoch, the Algorithm-1 instance selects the action $a^{(1)}_{l,t} \equiv {}^{d^l}/_{u^l}$ which best minimizes the immediate cost as

$$\left(a^{(1)}_{l,t}\right)^* = \underset{a_{l,t} \in A^{(1)}}{\arg\min}\ \Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right). \tag{L.18}$$

Furthermore, the $\epsilon-$greedy policy is adopted to trade-off action exploration versus exploitation. Thus, at each step, a random number is drawn from a uniform distribution $\varrho^{(1)} \in \mathcal{U}\ (0,1)$, and is compared against the predefined exploration probability $\epsilon^{(1)}$. If $\varrho^{(1)} \leq \epsilon^{(1)}$ is satisfied, a random action is selected; otherwise, a greedy action according to eq. (L.18) is adopted. Finally, the value function entries $Q^{(1)}_{[t,c]}$ are iteratively updated to reflect the learning experiences as follows:

$$Q^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right) \leftarrow \left(1-\alpha^{(1)}\right)\ Q^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right) + \alpha^{(1)} \left[\Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right) + \gamma^{(1)}\ \underset{a_l \in A^{(1)}}{\arg\min}\ Q^{(1)}_{[t+1,c]}\left(s^{(1)}_{i,t+1}, a^{(1)}_{l}\right)\right], \tag{L.19}$$

where $\alpha^{(1)} \rightarrow [0,1]$ is the learning rate, which specifies how fast the learning occurs. For instance, if $\alpha^{(1)}$ is small, the learning rate of Algorithm-1 network shall exhibit a longer convergence time. $\gamma^{(1)} \rightarrow [0,1]$ implies the discounted factor, which determines how much significance is considered on future penalties. If $\gamma^{(1)}$ is large, the Algorithm-1 RML instance is biased towards adopting actions at time epoch t, which are highly probable to result in a further favorable state at $t+1$. The detailed primary RML network is summarized in Algorithm-1.

Algorithm 1 : Algorithm-1 for balanced DL/UL capacity

Initialize:
  for each $s^{(1)} \in S^{(1)}$ and $a^{(1)} \in A^{(1)}$ **do**
    Initialize the Q-value $Q^{(1)}_{[t_0,c]}\left(s^{(1)}_{i,t_0}, a^{(1)}_{l,t_0}\right)$
  end for
top:
  At the next pattern update time epoch t:
  Generate a random number $\varrho^{(1)} \in \mathcal{U}\ (0,1)$
  if $\left(\varrho^{(1)} \leq \epsilon^{(1)}\right)$, **then**
    Apply a random action $a^{(1)}_t \in A^{(1)}$
  else
    Apply the action $a^{(1)}_t \in A^{(1)}$, *accord. to eq. (L.18)*
  end if
  Observe DL and UL traffic statistics $\overline{\mu}_{[t,c]}$
  Get current cost $\Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right)$, *accord. to eq. (L.17)*
  Determine system next state $s^{(1)}_{t+1}$, *accord. to eq. (L.14)*
  Update Q-value $Q^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right)$, *accord. to eq. (L.19)*
  Move time indexer: $t = t+1$, $s^{(1)}_t = s^{(1)}_{t+1}$, ***goto*** top.

5.2 Secondary Q-RML Sub-Networks For URLLC Latency Minimization

After the DL-to-UL symbol ratio $^{d^l}/_{u^l}$ is estimated from Algorithm-1 (layer 1), the corresponding Algorithm-2 sub-network is activated to estimate the best DL and UL symbol structure $\hat{w}_c$. For that, the DL and UL buffer latency samples per UE are monitored. Although, having monitored the latency for all the DL and UL packets for all active UEs in each cell represents a significant amount of statistics. Those samples are therefore further compressed into a more manageable metric that is meaningful for Algorithm-2 to learn and predict the best DL and UL symbol structure to minimize the overall cell latency outage performance. In this paper, the adopted method is to separately filter the DL and UL latency samples using a Kaiser filter. The motivation for using such a filter is its flexibility to provide higher filter weights (priorities) to the most critical input latency samples. This fits in achieving the stringent URLLC performance, where the achievable overall URLLC outage latency target is typically dictated by the worst latency samples. Therefore, a single scalar latency indication for the cell is calculated to reflect the overall DL and UL latency performance of each cell, guiding the learning process of Algorithm-2.

The inter-UE DL/UL HoLD samples are filtered using a non-uniform spatial window. Precisely, the filtering is applied on the HoLD statistics in order

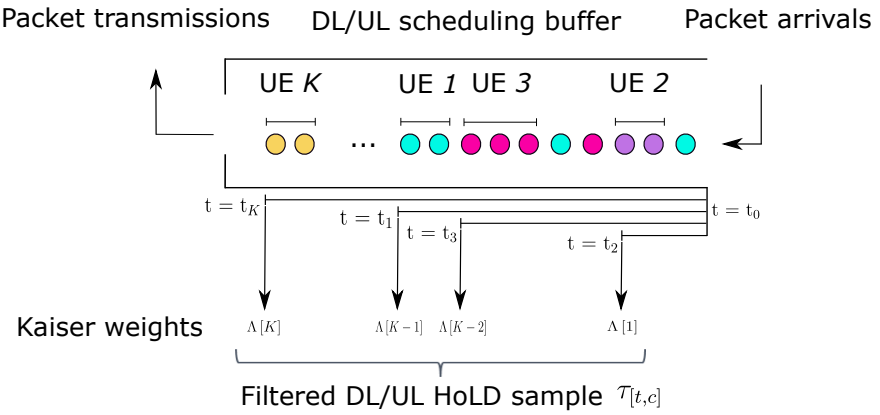


Fig. L.4. URLLC outage latency in DL/UL direction (ms).

to: (1) prioritize the delay samples of the UEs with the largest HoLD by assigning delay-proportional weights, and (2) safeguard Algorithm-2 learning convergence against the sudden changes of the per-packet HoLD samples. As shown by Fig. L.4, we apply a mirrored Kaiser window $\Lambda[\vartheta]$ over the inter-UE HoLD statistics [7], where $\Lambda[\vartheta]$ is expressed in the digital domain by

$$\Lambda[\vartheta] = \frac{I_0\left[\beta\sqrt{1-\left(\frac{2\vartheta}{\theta}-1\right)^2}\right]}{I_0[\beta]}, 1 \leq \vartheta \leq \theta+1, \tag{L.20}$$

where I_0 implies the zero-order modified Bessel function, β is a shaping factor, and $\theta+1$ denotes the window length, where $\Lambda[K] > \Lambda[K-1] > \ldots > \Lambda[1]$. Accordingly, the HoLD ratio $\tau_{[t,c]}(\varsigma)$ of the ς^{th} slot is defined as

$$\tau_{[t,c]}(\varsigma) = \frac{\tau^{\mathrm{dl}}_{[t,c]}(\varsigma)}{\tau^{\mathrm{dl}}_{[t,c]}(\varsigma) + \tau^{\mathrm{ul}}_{[t,c]}(\varsigma)}, \tag{L.21}$$

where $\tau^{\mathrm{dl}}_{[t,c]}(\varsigma)$ and $\tau^{\mathrm{ul}}_{[t,c]}(\varsigma)$ are the Kaiser-filtered cell-specific HoLD samples in the DL and UL directions, respectively. Then, the average HoLD $\overline{\tau}_{[t,c]}$ across the radio pattern is then calculated by

$$\overline{\tau}_{[t,c]} = \frac{1}{\tilde{\xi}} \sum_{\varsigma=1}^{\tilde{\xi}} \tau_{[t,c]}(\varsigma). \tag{L.22}$$

Equivalently to $\overline{\mu}_{[t,c]}$ of Algorithm-1, $\overline{\tau}_{[t,c]} \rightarrow [0,1]$ captures the directional HoLD performance. For instance, $\overline{\tau}_{[t,c]} = 0.8$ denotes that the DL HoLD is 4x times the corresponding UL HoLD. The state space of Algorithm-2 sub-networks is accordingly defined as

$$S_t^{(2,l)} = \left\{ s_{1,t}^{(2,l)},\ s_{2,t}^{(2,l)},\ \ldots,\ s_{\mathfrak{J}_{2,l},t}^{(2,l)} \right\}, \tag{L.23}$$

where $\mathfrak{J}_{2,l}$ is the state space size of $Q-2-l$. Then, the corresponding HoLD-to-state mapping is defined as

$$s_t^{(2,l)} = \begin{cases} s_{1,t}^{(2,l)}, & \overline{\tau}_{[t,c]} < \tau_{\min} \\ s_{2,t}^{(2,l)}, & \tau_{\min} \leq \overline{\tau}_{[t,c]} < \tau_{\min} + \mathrm{Y}_l \\ s_{3,t}^{(2,l)}, & \tau_{\min} + \mathrm{Y}_l \leq \overline{\tau}_{[t,c]} < \tau_{\min} + 2\mathrm{Y}_l \\ \vdots & \vdots \\ s_{\mathfrak{J}_{2,l},t}^{(2,l)}, & \overline{\tau}_{[t,c]} \geq \tau_{\max} \end{cases}, \tag{L.24}$$

with the HoLD quantization step Y given by

$$\mathrm{Y}_l = \frac{\tau_{\max} - \tau_{\min}}{\mathfrak{J}_{2,l} - 2}, \tag{L.25}$$

where $\tau_{\max}$ and $\tau_{\min}$ are the pre-defined maximum and minimum allowable bounds of the HoLD ratio $\overline{\tau}_{[t,c]}$. The intermediate $s_{i,t}^{(2,l)}, \forall i \in \mathfrak{J}_{2,l}$ states are the favorable state set of Algorithm-2 RML sub-networks, in order to preserve the minimum possible; though, balanced DL and UL HoLD performance.

The action space $A^{(2,l)}$ is built to present all the possible DL and UL symbol structures of the l^{th} pattern sub-book as

$$A_t^{(2,l)} = \left\{ a_{1,t}^{(2,l)},\ a_{2,t}^{(2,l)},\ \ldots,\ a_{\mathbf{Card}\left(A^{(2,l)}\right),t}^{(2,l)} \right\}, \tag{L.26}$$

where $a_{j,t}^{(2,l)} \equiv \hat{w}_j, \forall j \in \mathbf{Card}\left(A^{(2,l)}\right)$, with $A^{(2,l)}$ as the set of all radio structures in the l^{th} sub-book. Accordingly, the immediate environment return $\Theta_{[t,c]}^{(2,l)}\left(s_{i,t}^{(2,l)}, a_{j,t}^{(2,l)}\right), \forall i \in \mathfrak{J}_{2,l},\ j \in \mathbf{Card}\left(A^{(2,l)}\right)$ is defined by how much average HoLD $\overline{\tau}_{[t,c]}$ deviation is observed from its balanced mean $\mathfrak{m}^{(2,l)}$ as follows:

$$\Theta_{[t,c]}^{(2,l)}\left(s_{i,t}^{(2,l)}, a_{j,t}^{(2,l)}\right) = \left| \overline{\tau}_{[t,c]} - \mathfrak{m}^{(2,l)} \right|, \tag{L.27}$$

where the mean value of the HoLD ratio $\mathfrak{m}^{(2,l)}$ is adopted as the optimization target of the secondary Q-RML sub-networks, as it ensures a balanced DL and UL HoLD performance. Then, the secondary RML instances adopt the action, i.e, symbol structure $\hat{w}_j$, which offers the minimum variance of the relative HoLD performance as given by

$$\left(a_{j,t}^{(2,l)}\right)^* = \underset{a_j \in A^{(2,l)}}{\arg\min}\ \Theta_{[t,c]}^{(2,l)}\left(s_{i,t}^{(2,l)}, a_{j,t}^{(2,l)}\right). \tag{L.28}$$

Similarly to eq. (L.19), the value function entries $Q_{[t,c]}^{(2,l)}$ of Algorithm-2 are iteratively updated to reflect the learning experiences, as expressed by

$$Q_{[t,c]}^{(2,l)}\left(s_{i,t}^{(2,l)}, a_{j,t}^{(2,l)}\right) \leftarrow \left(1-\alpha^{(2)}\right)\ Q_{[t,c]}^{(2,l)}\left(s_{i,t}^{(2,l)}, a_{j,t}^{(2,l)}\right) + \alpha^{(2)}$$

$$\left[\Theta^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right) + \gamma^{(2)} \underset{a_j \in A^{(2,l)}}{\arg\min}\, Q^{(2,l)}_{[t+1,c]}\left(s^{(2,l)}_{i,t+1}, a^{(2,l)}_{j}\right)\right], \tag{L.29}$$

where $\alpha^{(2)}$ and $\gamma^{(2)}$ are the learning rate and discount factor of the secondary sub-network, respectively. The detailed steps of secondary RML instance is described by Algorithm 2.

Algorithm 2 : Algorithm-2 for outage latency minimization

Initialize:
  for each $s^{(2,l)} \in S^{(2,l)}$ and $a^{(2,l)} \in A^{(2,l)}$ **do**
    Initialize the Q-value $Q^{(2,l)}_{[t_0,c]}\left(s^{(2,l)}_{i,t_0}, a^{(2,l)}_{j,t_0}\right)$
  end for
top:
  At the next pattern update time epoch t:
    Activate the $Q-2-l$, to selected $\frac{d^l}{u^l}$ from $Q-1$
    Generate a random number $\varrho^{(2,l)} \in \mathcal{U}\ (0,1)$
    if $\left(\varrho^{(2,l)} \leq \epsilon^{(2,l)}\right)$, **then**
      Apply a random action $a^{(2,l)}_{j,t} \in A^{(2,l)}$
    else
      Apply the action $a^{(2,l)}_{j,t} \in A^{(2,l)}$, *accord. to eq. (L.28)*
    end if
    Observe DL and UL HoLD statistics $\overline{\tau}_{[t,c]}$
    Get the cost $\Theta^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right)$, *accord. to eq. (L.27)*
    Determine system next state $s^{(2,l)}_{t+1}$, *accord. to eq. (L.24)*
    Update Q-value $Q^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right)$, *accord. to eq. (L.29)*
    Move time indexer: $t = t+1$, $s^{(2,l)}_{t} = s^{(2,l)}_{t+1}$, ***goto*** top.

6 State-of-The-Art Duplexing Schemes

We compare the performance of the proposed solution against the most widely adopted duplexing schemes, for different directional traffic offered loads. The proposed solution is evaluated under two main variants, i.e., when either Algorithm-1 learning is solely adopted or both Algorithm-1 and Algorithm-2 are activated. For the former case, a default DL/UL evenly-distributed pattern structure is employed, following the estimated d/u from Algorithm-1. The duplexing deployments under investigation are as follows:

Frequency division duplexing (**FDD**): for a comprehensive URLLC latency analysis, FDD is considered as the reference case. The FDD DL and UL bandwidth allocations are configured equivalently to the TDD cases, such

that the total bandwidth is fixed. That is, the bandwidth allocation for each of the UL and DL direction is half of the TDD bandwidth.

Dynamic TDD (**dTDD**) [5]: neighboring BSs independently and dynamically in time select the radio patterns which better satisfy their link selection criteria. Herein, for the sake of cross-scheme fairness, we adopt the same buffered traffic criterion of Algorithm-1 as per eq. (L.12). The structure of the selected radio pattern, in terms of the placement of the DL and UL symbols, is presumed to be always evenly distributed, and with a symbol block size of 4 symbols. For example, a 14-symbol slot with $^{d}/_{u} = {}^{2}/_{1}$ is configured as [DDDDFUUUUDDDDF]. Such strategy allows for distributed DL and UL transmission opportunities across the pattern duration. Herein, no inter-BS coordination is assumed, hence, BS-BS and UE-UE CLI can be inflicted.

Static TDD (**sTDD**): a pre-defined global radio pattern is configured for all neighboring BSs, that meets the average traffic demands of the cluster. We assume a perfect knowledge of the average offered traffic ratio $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}}$, thus, configuring the global radio pattern with a perfect-matching $^{d}/_{u}$. Although sTDD requires the simplest implementation complexity, without CLI infliction, it offers no pattern adaptation to the BS-specific varying traffic and latency demands.

Semi-static TDD (**Semi-sTDD**) [7]: it is built on top of the sTDD scheme in order to offer an extended TDD adaptation flexibility. Basically, Semi-sTDD follows the same setup as the sTDD scheme; however, the common radio pattern is periodically updated to meet the varying cross-BS traffic demands, and accordingly re-used by all coordinated BSs. In that regard, neighboring BSs continuously exchange indications to their respective traffic needs over the *Xn-interface*.

7 Performance Evaluation

7.1 Simulation Methodology

We evaluate the performance of the proposed solution using extensive dynamic system level simulations, where the main modeling assumptions are listed in Table I. The simulations follow the system model described in Section 2, and are in line with agreed 3GPP system level simulation methodology. The simulated scenario is the Urban Macro (UMa) with three sector base station sites placed in a regular hexagonal grid and UEs randomly positioned, following a spatial uniform distribution. Time-variant dynamic traffic is simulated for each UE as per the description in Section 2.A. Each UE is served by the cell corresponding to the highest received reference signal received power. The advanced three-dimensional 3GPP UMa radio propagation model is assumed [38]. The simulator includes explicit modeling of all

the major MAC and PHY layer functionalities, and related RRM functionalities. For each transmission, the per subcarrier symbol SINR is calculated. Such SINR calculations assume LMMSE-IRC and include both the effect of the co-channel and potential CLI into account in line with the SINR calculations in (L.4) and (L.5). Based on all the subcarrier symbols SINR for the transmission, the combined mean mutual information per coded bit (MMIB) mapping [39] is applied for calculation of the effective SINR level. The respective transmission packet error probability (PEP) is calculated based on look-up tables, obtained from extensive link level simulations. Based on the calculated PEP, the corresponding packet is determined as either successful or failed. During the DL TTIs, DL UEs are dynamically scheduled based on the proportional fair criterion, assuming also dynamic link adaptation with adaptive selection of the MCS based on the most recent received CQI reports, including also outer loop link adaptation. UL UEs are served using the CG baseline as outlined in Section 2.A. HARQ re-transmissions are always prioritized over new packet transmissions. For each frame periodicity (10 ms), the proposed learning framework in Section 5 runs in a distributed manner for each cell to determine the next radio pattern configuration.

The simulator is validated via so-called calibration exercises, where baseline statistics are reported and compared between 3GPP partners [40]. Simulations are run for a sufficiently long-time period to ensure statistical reliable results, and thereby a solid basis for drawing mature conclusions. In line with [41], the default simulation length is 5 million successfully decoded URLLC payloads. Thus, assuming that the URLLC packets are fully uncorrelated, the target 99.999% percentile of the URLLC latency distribution is calculated with a maximum error margin of $\pm 5\%$, and therefore, with a 95% statistical confidence level [42].

Due to the nature of the simulations where the UEs are created at the start, traffic is dynamic (i.e. payloads are generated according to Poisson point processes), and the various control loops (e.g. for link adaptation, TDD frame adaptation, etc.), we apply a so-called warm-up time. Only after the warm-up time, the performance statistics are collected from the simulations. By default, the warm-up time is configured to equal 1 second as this is found to be enough time for the network performance to stabilize.

7.2 Baseline Performance Comparison

Fig. L.5 depicts the complementary cumulative distribution function (CCDF) of the combined DL and UL achievable latency in ms, under the proposed scheme, FDD, dTDD, and the sTDD deployments. Clearly, the FDD scheme always outperforms the dTDD scheme. This is mainly attributed to the absence of the CLI as well as the concurrent availability of the DL and UL transmission opportunities. The sTDD is configured with the assumption of

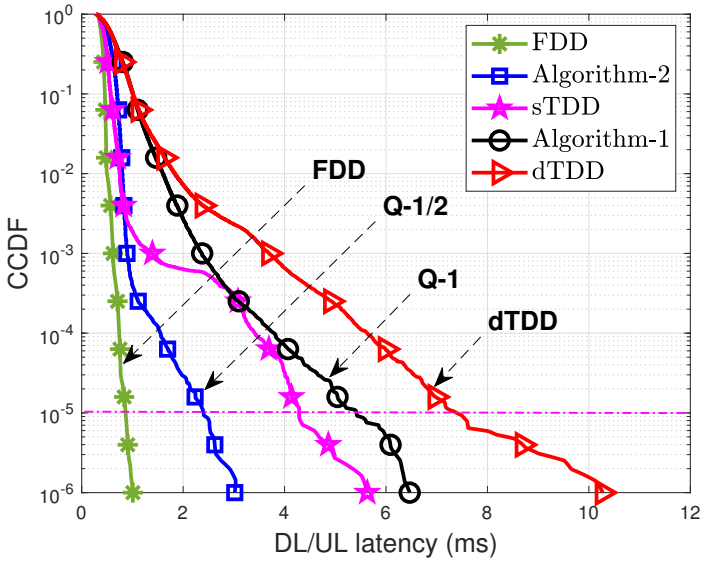


Fig. L.5. Achievable latency, with $\Omega = 1$ Mbps, and $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 1$.

the optimal knowledge of the directional offered load. Hence, it is configured with a perfect-matching pattern configuration, i.e., $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 1 \rightarrow d/u = 1$. Looking particularly at the outage URLLC latency at the 10^{-5} probability, the proposed Algorithm-2 clearly offers a significant outage latency improvement. That is, 70% and 53% outage latency reduction compared to dTDD and sTDD, respectively. Although, it inflicts $\sim$ 51% outage latency increase compared to the FDD case. The performance merits of the proposed solution are mainly due to the sufficient learning gain to compensate for the directional HoLD in designing the radio pattern configuration. The sTDD, with the optimal knowledge of $\Omega^{\text{dl}}/\Omega^{\text{ul}}$, offers a slight latency enhancement than the proposed Algorithm-1, due to the non-existent CLI. Though, it exhibits a clear performance loss compared to the proposed Algorithm-2, as the latter introduces an additional latency-aware RML layer.

Fig. L.6 shows the empirical CDF (ECDF) of the traffic ratio $\overline{\mu}$ for all schemes under evaluation. Clearly, the larger the $\overline{\mu}$, the larger the buffered DL traffic compared to that is of the UL direction. The dTDD scheme obviously inflicts the lowest $\overline{\mu}$, with $\overline{\mu} = 0.15$ at the 50%ile, indicating that the UL traffic is consistently blocked by the BS-BS CLI, i.e., the buffered UL traffic is 5.6x times the corresponding DL traffic, despite the configured $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 1$. The sTDD provides a marginally improved UL buffering performance, compared to dTDD, due to the CLI-free UL. However, it does not account for the DL and UL traffic variations. The proposed Algorithm-1 and

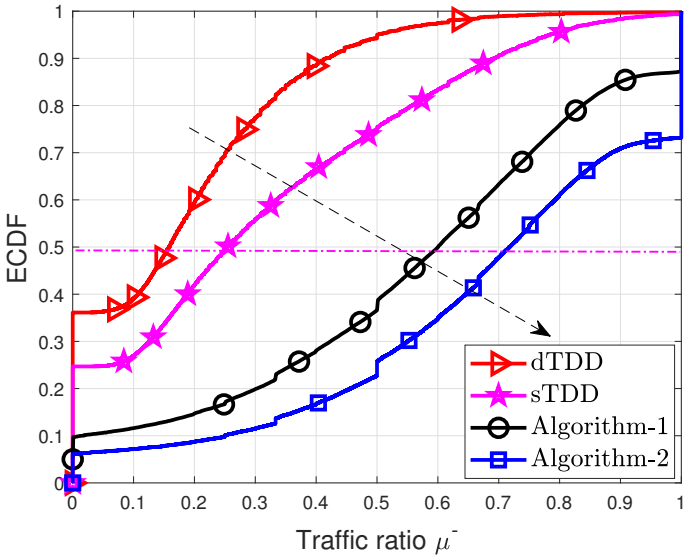


Fig. L.6. Buffering performance, with $\Omega = 1$ Mbps, and $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 1$.

Algorithm-2 solutions offer a smooth traffic buffering performance, clearly without the UL traffic accumulation problem, i.e., $\overline{\mu} = 0.57$ and 0.71, respectively. This denotes the buffered UL traffic size is 0.75x and 0.48x times the buffered DL traffic, respectively. Accordingly, the proposed learning solution dynamically compensates for the degraded UL capacity by assigning more UL transmission opportunities across the radio pattern, leading to a faster UL traffic recovery. Though, this comes at the expense of an additional DL traffic buffering, i.e., 25% more buffered DL traffic with Algorithm-2.

Fig. L.7 shows the CCDF of the achievable URLLC latency under the proposed algorithm and the Semi-sTDD scheme, respectively, for both light and high offered load cases. With $\Omega = 0.25$ Mbps, the proposed learning algorithm clearly achieves a significant enhancement of the DL/UL URLLC outage latency, offering 1.06 ms at the 10^{-5} probability, with 60% outage latency reduction, compared to Semi-sTDD. For such a lightly-loaded case, the URLLC outage latency is dominated by the structure of the DL and UL symbols across the radio pattern, rather than the CLI intensity. The proposed learning solution autonomously optimizes the pattern structure to provide a faster and BS-specific DL and UL link switching design. Though, the Semi-sTDD inflicts a clear URLLC outage degradation due to the high DL and UL traffic fluctuations across neighboring BSs, thus, adopting a common radio pattern offers limited TDD adaptability. For the high-load region $\Omega = 1$ Mbps, the CLI becomes vital to control because of the increased DL traf-

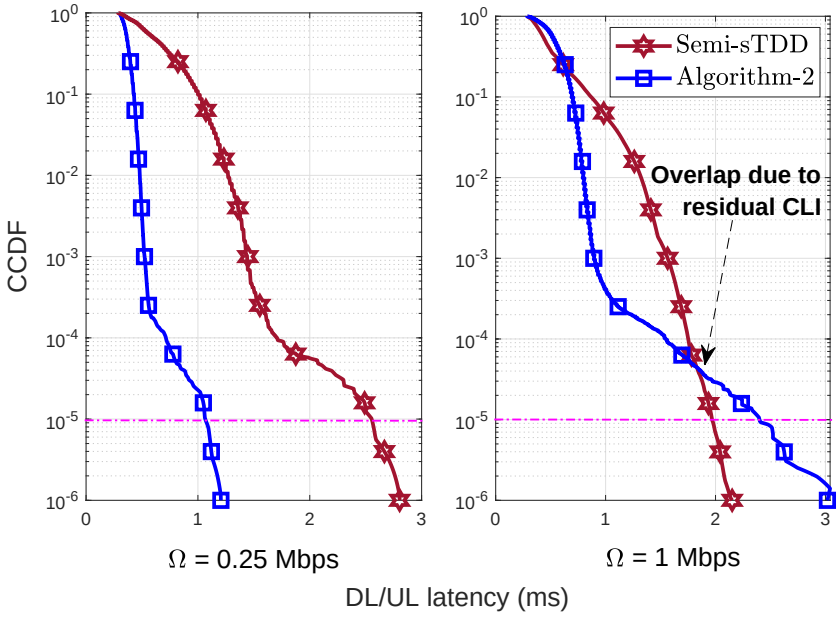


Fig. L.7. Latency comparison to Semi-sTDD, with $\Omega^{\mathrm{dl}}/\Omega^{\mathrm{ul}} = 1$.

fic size, and thus, the critical BS-BS CLI. The proposed solution therefore exhibits limited degrees of freedom in designing the sufficient DL and UL switching structure, in order to control the severe CLI accordingly. The Semi-sTDD scheme offers 21% latency reduction, compared the proposed solution, mainly due to the absence of the CLI. This case, unlike the lightly-loaded setup, the cross-BS traffic statistics converge to the same average, hence, the Semi-sTDD with a global radio pattern becomes more efficient to achieve a decent URLLC outage latency.

7.3 Q-RML Convergence Performance

Achieving a robust convergence performance of the RL-based solutions is demonstrated to be a challenging task [26] mainly due to the sparse reward function observed from surrounding environment. Furthermore, since the system-model adopted in this work incorporates time-variant channel conditions with sporadic and UE-specific packet arrivals, analyzing the convergence performance of the proposed learning approach becomes vital. We performed a large set of the system level simulations with various warm-up periods in order to obtain the best possible RL settings which offer the best achievable URLLC outage latency. As described in Section 7.A, the warm-up

duration implies the starting period of the simulation until the system gets loaded. We also utilize such time as the convergence delay of the proposed QRL framework where the action exploration is prioritized to stabilize all corresponding Q-value functions during the warm-up. That is, we adopted warm-up periods from 0.25 to 1.5 second alongside with adopting different action exploration-exploitation probabilities from 0 to 0.7 for both Algorithm 1 and 2, respectively. Therefore, based on our extensive sensitivity analysis, we adopt ~1 second of warm-up time over which the action exploration probability for both the primary and secondary learning instances is set to $\epsilon^{(1)} = \epsilon^{(2)} = 0.25$. During the actual simulation time, i.e,. QRL inference time, the actions which offer the lowest possible cost functions are always utilized, i.e., $\epsilon^{(1)} = \epsilon^{(2)} = 0.0$ during inference (no action exploration). This setting offers the shortest convergence delay and accordingly, the best achievable URLLC outage performance for the considered system configurations.

To monitor the actual convergence performance of the proposed QRL framework, we calculate the learning temporal difference (TD). The TD reflects how well the Q-learning is converging towards the optimal policy in time. In particular, it captures the difference among the current learning samples and the former learning experiences as

$$\mathbf{TD}_{Q_1} = \Theta^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right) + \gamma^{(1)} \underset{a_l \in A^{(1)}}{\arg\min}\, Q^{(1)}_{[t+1,c]}\left(s^{(1)}_{i,t+1}, a^{(1)}_{l}\right) - Q^{(1)}_{[t,c]}\left(s^{(1)}_{i,t}, a^{(1)}_{l,t}\right). \tag{L.30}$$

$$\mathbf{TD}_{Q_{2,l}} = \Theta^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right) + \gamma^{(2)} \underset{a_j \in A^{(2,l)}}{\arg\min}\, Q^{(2,l)}_{[t+1,c]}\left(s^{(2,l)}_{i,t+1}, a^{(2,l)}_{j}\right) - Q^{(2,l)}_{[t,c]}\left(s^{(2,l)}_{i,t}, a^{(2,l)}_{j,t}\right). \tag{L.31}$$

As depicted by Fig. L.8, the TD distribution of both Algorithm-1 and Algorithm-2 is quite compressed, where Algorithm-2 tends to experience a faster learning convergence than Algorithm-1, due to the already refined learning of the symbol ratio d/u. Upon convergence, the new learning observations do not significantly change the applied actions, leading to a slower transition rate over the *state-action* pairs. That is, at the 50%-ile of the TD distribution, the secondary learning exhibits a normalized TD of 0.08. This denotes that, upon convergence, the cost values of the proposed Algorithm-2 are fluctuating in time by only $\pm 8\%$, due to the sufficient learning of the pattern structure. Such convergence performance is obtained with the baseline system setting as indicated by Table I. That is, an offered traffic load of 1 Mbps/cell and equal DL and UL traffic load split, where the action exploration probabilities are set as: $\epsilon^{(1)} = \epsilon^{(2)} = 0.25$ during the warm-up time.

In particular, the modeling of the learning objectives, i.e., learning targets, learning inputs and outputs are shown to significantly impact the achievable convergence performance. As the main learning objective of the primary Q-RML is the aggregated buffered traffic, it imposes partial stationarity due to

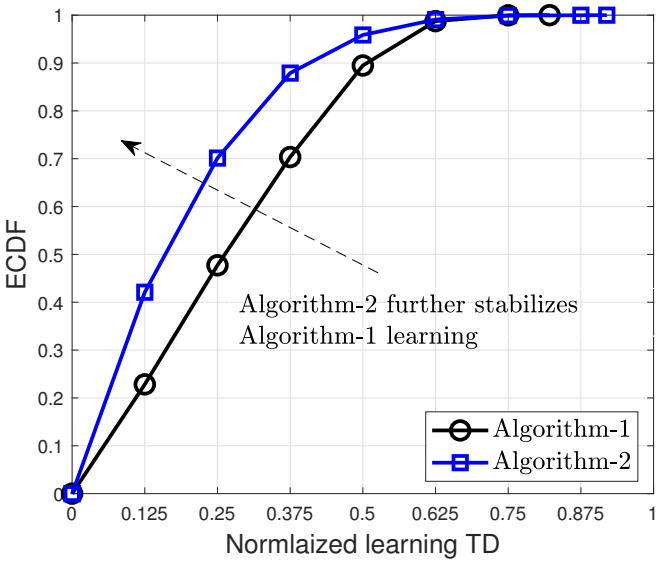


Fig. L.8. TD performance, with $\Omega = 1$ Mbps, and $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 1$.

the several active users at the same time. That is, an abrupt change of the aggregate buffered traffic is not highly likely. For the secondary Q-RML networks, the Kaiser-window filtered delay statistics of the buffered users are considered instead of the actual latency values, as the latter could potentially rapidly change, disturbing the learning convergence. Thus, the convergence of the proposed approach has a quick time cycle. Furthermore, as the learnable action set are the set of all possible TDD radio frame configurations, the complexity of the proposed solution scales mainly with the number of possible TDD radio patterns. That is typically limited by couple of hundreds, allowing for a further quicker convergence delay, i.e., the complexity for calculating and updating the Q-values of each possible action (TDD pattern).

7.4 Cross Link Interference Performance

As a consequence to the achievable radio frame learning potential, the proposed solution tends to realize an autonomous trade-off between the DL and UL symbol switching periodicity versus the subsequent CLI performance. In particular, a faster DL and UL switching periodicity during the radio pattern is favored; though, it is likely to result in frequent CLI occurrences, due to the higher probability of adjacent BSs adopting opposite link directions. Accordingly, the latency merits, obtained from the fast link switching, are completely wiped out, and reverted into an outage latency loss due to the severe CLI. Therefore, as shown by Fig. L.9, the proposed solution clearly offers

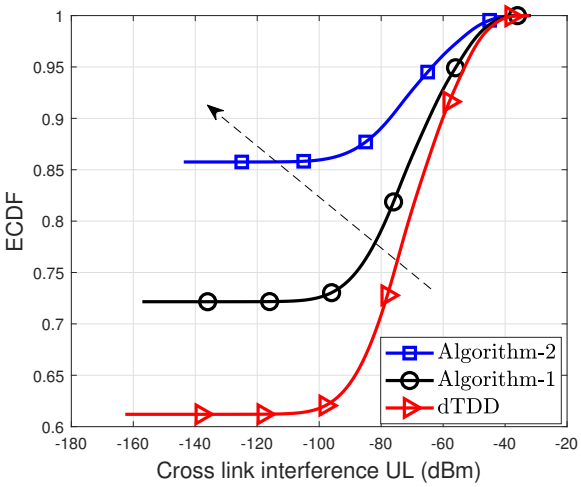


Fig. L.9. BS-BS CLI performance, with $\Omega = 1$ Mbps, and $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 1$.

a substantial reduction of the BS-BS CLI, compared to the dTDD scheme.

7.5 Performance Evaluation With Different Offered Traffic Ratios

Examining the proposed solution under different offered load ratios, Fig. L.10 presents the achievable latency performance with $\Omega = 3$ Mbps, and $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 1/2$ and $2/1$, respectively. Particularly, with $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 2/1$, the URLLC outage latency becomes dictated by the severe CLI, and especially the BS-BS CLI, due to the larger DL traffic portion. The proposed solution dynamically compensates for the highly-degraded UL PRB capacity by allocating more UL transmission opportunities, leading to 67% outage UL latency reduction, compared to dTDD. However, it comes at the expense of further increased DL traffic buffering, i.e., 49% outage DL latency increase. With $\Omega^{\text{dl}}/\Omega^{\text{ul}} = 1/2$, where the BS-BS CLI is negligible, proposed solution achieves a reliable outage latency improvement for both link directions.

To explore how the schemes under evaluation re-act to the directional traffic variations, we define the symbol ratio $\eta^c \rightarrow [0, 1]$ as given by

$$\eta^c = \frac{d^c}{d^c + u^c}. \tag{L.32}$$

Accordingly, Fig. L.11 shows the average symbol ratio η^c of the proposed solution, sTDD, dTDD, and Semi-sTDD schemes, respectively, for different $\Omega^{\text{dl}}/\Omega^{\text{ul}}$ ratios. Clearly, the sTDD scheme always adopts a linear mapping from $\Omega^{\text{dl}}/\Omega^{\text{ul}}$ to d^c/u^c due to the fixed pattern configuration. That is, $\eta^c = 0.33$,

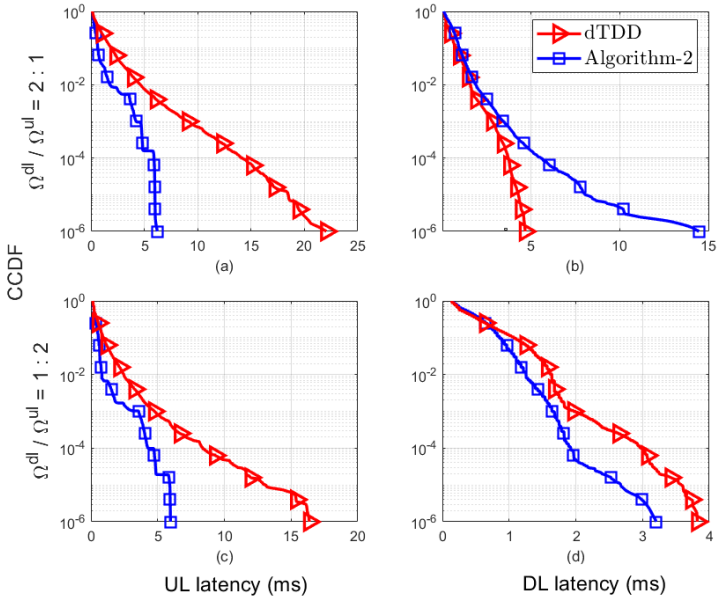


Fig. L.10. Latency performance (ms), with $\Omega = 3$ Mbps, and Ω^{dl}/Ω^{ul}.

0.5, and 0.66 for $\Omega^{dl}/\Omega^{ul} = d^c/u^c = 1/2$, $1/1$, and $2/1$, respectively. The Semi-sTDD scheme follows the sTDD in terms of the dynamically configured average symbol ratio η^c; however, with moderate variations due to the additional TDD pattern adaptation gain, e.g., adopting +12% UL symbols on average than the sTDD scheme with $\Omega^{dl}/\Omega^{ul} = 2/1$. The dTDD scheme performs quite efficiently under light CLI intensity. That is, with $\Omega^{dl}/\Omega^{ul} = 1/2$, an almost-balanced DL and UL adaptation is achieved, where an average $\eta^c = 0.29$ is observed. It implies that the $u^c = 2.4\, d^c$ symbol configuration is favored by the dTDD pattern adaptation process, to allow for the degraded UL capacity due to the residual CLI. However, the dTDD scheme obviously inflicts an UL capacity blocking under high CLI intensity conditions, i.e., $\Omega^{dl}/\Omega^{ul} = 2/1$, where $\eta^c = 0.34$ is exhibited. That denotes the $u^c = 1.9\, d^c$ configuration is adopted on average, and subsequently, the DL capacity inflicts a starvation of the transmission opportunities across the configured radio patterns.

Moreover, Fig. L.11 shows that the proposed solution preserves a balanced symbol configuration performance under all considered directional load cases. Unlike the sTDD and Semi-sTDD schemes, proposed learning solution tends to bias the pattern configuration towards even more UL transmission opportunities to compensate for the severe BS-BS CLI. Although, unlike the dTDD solution, proposed solution does not exhibit the UL capacity blocking issue, even under severe CLI conditions, i.e., $\eta^c = 0.49$ with

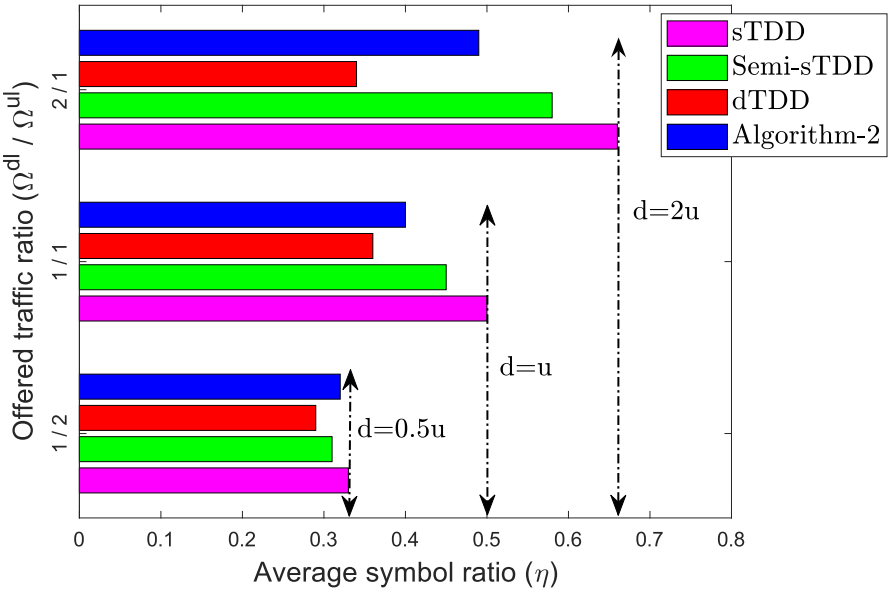


Fig. L.11. Symbol configuration, with $\Omega = 1$ Mbps, and Ω^{dl}/Ω^{ul}.

$\Omega^{dl}/\Omega^{ul} = 2/1$. This is mainly attributed to the well-learned trade-off among the residual CLI and the link switching periodicity.

Finally, Fig. L.12 depicts the achievable per-TTI UL throughput performance in Mbps of the proposed solution and dTDD case, respectively. The proposed solution achieves a considerable capacity improvement due to the faster traffic transmissions. Obviously, the major capacity gain of the proposed is realized at the lower percentiles, i.e., BS-edge UL UEs, since those are the most impacted by the obtained CLI enhancement and the faster UL transmissions accordingly.

8 Concluding remarks

In this paper, a radio pattern optimization scheme has been proposed for 5G new radio TDD systems. The proposed solution encompasses dual reinforcement Q-reinforcement-learning (QRL) instances for online optimization of the achievable URLLC outage latency, tackling a *min-max* URLLC problem. The primary QRL-network seeks to estimate the number of the DL and UL symbols across the next radio pattern, which best satisfies a faster; but, balanced downlink and uplink traffic handling. The secondary QRL-sub-networks select the corresponding pattern structure to achieve a decent

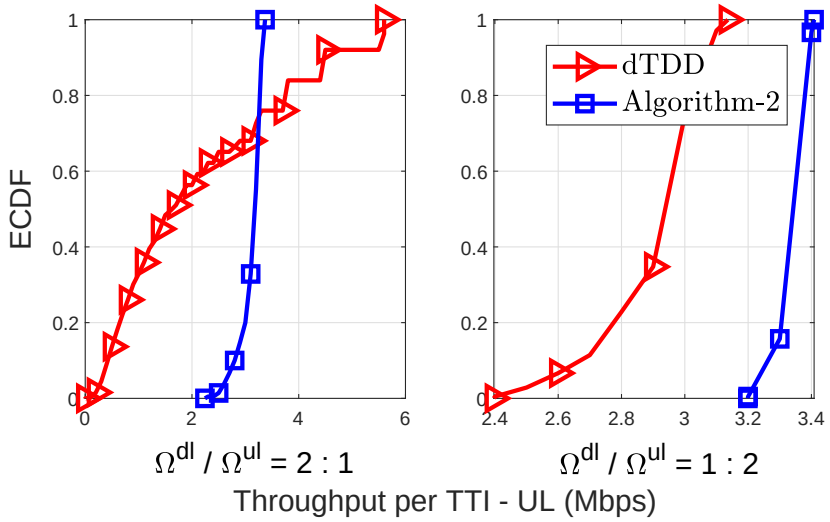


Fig. L.12. Throughput performance, with $\Omega = 3$ Mbps.

URLLC outage latency accordingly.

Through extensive system-level simulations, the proposed solution demonstrates a significant URLLC outage latency improvement compared to state-of-the-art dynamic TDD proposals. As an example, the URLLC outage latency is reduced by 70% and 53% compared to the fully dynamic and static TDD solutions, respectively, when assessed at high offered loads. The proposed solution achieves URLLC outage latency of 1 ms at the modest offered load of 250 kbps, while the semi-static TDD solution with inter-cell coordination achieves 2.7 ms latency, i.e. a latency reduction of 60%. Such impressive gain is achieved while the proposed ML solution runs independently for each cell. The semi-static TDD solution utilizes explicit inter-cell coordination. However, at high offered load, where the outage latency is in orders of magnitude higher than the 1 ms URLLC target, the semi-static TDD with explicit inter-cell coordination to avoid any CLI displays as good performance as the proposed solution.

The main insights brought by this paper are as follows: (1) URLLC latency and reliability performance is highly challenged in dynamic TDD deployments, due to the non-concurrent downlink and uplink transmission opportunities, and the additional cross-link interference (CLI), (2) thus, the real-time optimization of the radio pattern structure becomes vital towards a decent URLLC outage performance, (3) accordingly, machine learning techniques can be efficiently utilized to offer a proactive pattern estimation learning gain, (4) in this regard, reinforcement Q-learning has been adopted due to its online (real-time) learning capabilities, and simple implementation com-

plexity under the adopted system model, and (5) proposed solution demonstrates a flexible and dynamic radio pattern selection strategy to autonomously trade-off the CLI intensity with the URLLC outage performance; however, the achievable gain is shown to be load-dependent. As a future extension of this study, various learning approaches such as the state-action-reward-state-action (SARSA) shall be considered in order to learn and further optimize the selection of TDD radio patterns. Furthermore, extending the ML-driven solution for TDD pattern optimization to include explicit inter-cell coordination may offer further performance benefits; including also faster learning convergence and robustness.

9 Acknowledgments

This work is partly funded by the Innovation Fund Denmark (IFD) – case number: 7038-00009B.

References

[1] M. Bennis, M. Debbah and H. V. Poor, "Ultrareliable and low-latency wireless communication: tail, risk, and scale," *Proc. of the IEEE*, vol. 106, no. 10, pp. 1834-1853, Oct. 2018.

[2] *Service requirements for the 5G system*; Stage-1 (Release 16), 3GPP, TS 22.261, V16.6.0, Dec. 2018.

[3] J. Lee et al., "Spectrum for 5G: global status, challenges, and enabling techs," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 12-18, Mar. 2018.

[4] K. I. Pedersen, G. Berardinelli, F. Frederiksen and P. Mogensen, "A flexible 5G wide area solution for TDD with asymmetric link operation," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 122-128, April 2017.

[5] Ali A. Esswie, and K.I. Pedersen, "On the ultra-reliable and low-latency communications in flexible TDD/FDD 5G networks," *in Proc. IEEE CCNC,* Las Vegas, Jan. 2020.

[6] *NR; Physical layer procedures for control*; Release 16, V16.0.0, TS 38.213, 3GPP, Dec. 2019.

[7] Ali A. Esswie, K.I. Pedersen, and P. Mogensen, "Semi-static radio frame configuration for URLLC deployments in 5G macro TDD networks," *in Proc. IEEE WCNC,* April 2020.

[8] Ali A. Esswie, and K.I. Pedersen, “Inter-cell radio frame coordination scheme based on sliding codebook for 5G TDD systems,” *in Proc. IEEE VTC-spring,* Kuala Lumpur, Malaysia, May 2019, pp. 1-6.

[9] L. Binyong, and C. Gang, "Dynamic TDD DL/UL reconfiguration based on shift," *in Proc. IEEE ICCC*, Chengdu, Oct. 2016, pp. 1561-1568.

[10] Z. Huo, N. Ma and B. Liu, "Joint user scheduling and transceiver design for cross-link interference suppression in MU-MIMO dynamic TDD systems," *in Proc. IEEE ICCC,* Chengdu, Dec. 2017, pp. 962-967.

[11] A. Łukowa and V. Venkatasubramanian, "Performance of strong interference cancellation in flexible UL/DL TDD systems using coordinated muting, scheduling and rate allocation," *in Proc. IEEE WCNC*, Doha, April 2016, pp. 1-7.

[12] A. Łukowa and V. Venkatasubramanian, "Coordinated user scheduling in 5G dynamic TDD systems with beamforming," *in Proc. IEEE PIMRC*, Bologna, Sep. 2018, pp. 596-597.

[13] Ali A. Esswie, and K.I. Pedersen, “Cross link interference suppression by orthogonal projector in 5G dynamic-TDD URLLC systems,” *in Proc. IEEE WCNC,* April 2020.

[14] E. d. O. Cavalcante, G. Fodor, Y. C. B. Silva and W. C. Freitas, "Distributed beamforming in dynamic TDD MIMO networks with BS to BS interference constraints," *IEEE Wireless Commun. Lett. , vol. 7, no. 5, pp. 788-791, Oct. 2018.*

[15] Z. Huo, N. Ma and B. Liu, "Joint user scheduling and transceiver design for cross-link interference suppression in MU-MIMO dynamic TDD systems," *in Proc. IEEE ICCC*, Chengdu, Dec. 2017, pp. 962-967.

[16] Ali A. Esswie, K.I. Pedersen, and P. Mogensen, “Quasi-dynamic frame coordination for ultra- reliability and low-latency in 5G TDD systems,” *in Proc. IEEE ICC*, Shanghai, China, May 2019, pp. 1-6.

[17] J. W. Lee, C. G. Kang and M. J. Rim, "SINR-ordered cross link interference control scheme for dynamic TDD in 5G system," *in Proc. IEEE ICOIN*, Chiang Mai, Jan. 2018, pp. 359-361.

[18] M. E. Morocho-Cayamcela, H. Lee and W. Lim, "Machine learning for 5G/B5G mobile and wireless communications: potential, limitations, and future directions," *IEEE Access*, vol. 7, pp. 137184-137206, Sep. 2019.

[19] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu and N. D. Sidiropoulos, "Learning to optimize: training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438-5453, Oct. 2018.

[20] I. Comşa et al., "Towards 5G: a reinforcement learning-based scheduling solution for data traffic management," *IEEE Trans. Netw. Service Manag.*, vol. 15, no. 4, pp. 1661-1675, Dec. 2018.

[21] A. Azari, M. Ozger and C. Cavdar, "Risk-aware resource allocation for URLLC: challenges and strategies with machine learning," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 42-48, Mar. 2019.

[22] F. D. Calabrese, L. Wang, E. Ghadimi, G. Peters, L. Hanzo and P. Soldati, "Learning radio resource management in RANs: framework, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 138-145, Sept. 2018.

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

[24] C. She, R. Dong, Z. Gu, Z. Hou, Y. Li, W. Hardjawana, C. Yang, L. Song, and B. Vucetic, "Deep learning for ultra-reliable and low-latency communications in 6G networks ," *in IEEE Netw.*, Feb. 2020.

[25] P. Louridas and C. Ebert, "Machine learning," *IEEE Software*, vol. 33, no. 5, pp. 110-115, Oct. 2016.

[26] Richard S. Sutton and Andrew G. Barto,. *Reinforcement Learning: An Introduction*. A MIT Press, Cambridge, MA, 2018.

[27] B. M. Assaouy, O. Zytoune and D. Aboutajdine, "Policy iteration vs Q-Sarsa approach optimization for embedded system communications with energy harvesting," *in Proc. IEEE ATSIP*, Fez, Oct. 2017, pp. 1-6.

[28] *Study on new radio access technology physical layer aspects*; Release 14, V14.2.0, TR 38.802, 3GPP, Sep. 2017.

[29] *NR; Physical channels and modulation* ; Release 16, V16.1.0, TS 38.211, 3GPP, Mar. 2020.

[30] *NR; Physical layer procedures for data*; Release 16, V16.0.0, TS 38.214, 3GPP, Dec. 2019.

[31] 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. 28912-28922, May 2018.

[32] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, I. Z. Kovcs and P. Mogensen, "Joint resource configuration and MCS selection scheme for uplink grant-free URLLC," *in Proc. Globecom*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1-6.

[33] *Study on 3D Channel Model for LTE*; Release 12, V12.7.0, TR 36.873, 3GPP, Dec. 2014.

[34] Y. Ohwatari, N. Miki, Y. Sagae, and Y. Okumura, "Investigation on interference rejection combining receiver for space–frequency block code transmit diversity in LTE-advanced downlink," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 191–203, Jan. 2014.

[35] B. Jang, M. Kim, G. Harerimana and J. W. Kim, "Q-learning algorithms: a comprehensive classification and applications," *IEEE Access*, vol. 7, pp. 133653-133667, Sep. 2019.

[36] S. alışır and M. K. Pehlivanoğlu, "Model-free reinforcement learning algorithms: a survey," *in Proc. IEEE SIU*, Sivas, Turkey, Aug. 2019, pp. 1-4.

[37] M. Abu Alsheikh, D. T. Hoang, D. Niyato, H. Tan and S. Lin, "Markov decision processes with applications in wireless sensor networks: a survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1239-1267, Oct. 2015.

[38] *Study on channel model for frequencies from 0.5 to 100 GHz*, Release 16, V16.1.0, TR 38.901, 3GPP, Dec. 2019.

[39] D. G. Popescu, M. Varga and V. Bota, "Comparison between measured and computed values of the mean mutual information per coded bits in OFDM based wireless transmissions," *in Proc. IEEE TSP*, Rome, July 2013, pp. 380-384.

[40] *Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC)*, Release 16, TR 38.824, 3GPP, Mar. 2019.

[41] G. Pocovi, K. I. Pedersen and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *in IEEE Access*, vol. 6, pp. 28912-28922, 2018.

[42] L. D. Brown, T. T. Cai, and A. Dasgupta, "Confidence intervals for a binomial proportion and asymptotic expansions," *Annu. Statist.*, vol. 30, no. 1, pp. 160–201, 2002.

Table L.1: Simulation setup and major parameters

Parameter	Value
Network environment	3GPP Urban Macro (UMa) network with 21 cells and 500 meter inter-site distance, 3D SCM channel [33]
Carrier configuration	10 MHz carrier bandwidth at 3.5 GHz; synchronous TDD
PHY numerology	30 kHz subcarrier spacing; 12 subcarriers per PRB; TTI of 4-OFDM symbol duration
Max transmit power	BS: 43 dBm; UE: 23 dBm
Control channel	Error-free control signalling with dynamic link adaptation
DL data channel	QPSK to 64QAM modulation
UL data channel	QPSK modulation
Channel state information	Channel quality indication and precoding matrix indication, reported every 5 ms; Sub-band size: 8 PRBs
Antenna configuration	$N = 8$, $M = 2$; LMMSE-IRC DL/UL receiver
UL power control	Open-loop, $\alpha = 1$, $P0 = -96$ dBm, variable power offset for low and high pathloss UEs, $\hat{c}$ = -110 dB
Packet scheduler	DL: proportional fair, UL: configured grant
BLER target	DL: 1 percent with dynamic MCS selection, UL: pre-defined MCS configuration [32]
HARQ	Asynchronous HARQ with Chase combining; Maximum number of re-transmissions: 6
Average UE load	$K^{\text{dl}} = K^{\text{ul}} = 40$, and 118; Uniformly distributed
Traffic composition	FTP3, $f^{dl} = f^{ul}$ = 256 bits; λ^{dl} = 50 packets/sec; λ^{ul} = 50 packets/sec
Offered load $\Omega^{\text{dl}}/\Omega^{\text{ul}}$	1 : 1 [0.5 : 0.5 Mbps]; 1 : 2 [1 : 2 Mbps]; 2 : 1 [2 : 1 Mbps]
Processing time	PDSCH preperation: 2.5-OFDM symbols PDSCH decoding: 4.5-OFDM symbols PUSCH preperation: 5.5-OFDM symbols PUSCH decoding: 3-OFDM symbols
TDD pattern periodicity	10 ms
Default symbol structure	DL/UL evenly distributed in blocks of four symbols with same link direction
Proposed solution setup	$L = 9$ $\mathfrak{J}_1 = \mathfrak{J}_2 = 3$ $\mu_{\max} = 0.75$ $\mu_{\min} = \tau_{\min} = 0.2$ $\mathfrak{y}^{(1)} = \mathfrak{y}^{(2)} = \alpha^{(1)} = \gamma^{(1)} = 0.5$ $\epsilon^{(1)} = \epsilon^{(2)} = 0.25$ $\tau_{\max} = 0.8$ **Card** $A_t^{(2)} = \{11, 19, 27, 34, 61, 48, 31, 10, 7\}$ $\alpha^{(2)} = \gamma^{(2)} = 0.7$

References

Paper M

Analysis of Outage Latency and Throughput Performance in Industrial Factory 5G TDD Deployments

Ali A. Esswie, and Klaus I. Pedersen

The paper has been submitted to the
2021 IEEE 93rd Vehicular Technology Conference: VTC2021-Spring

Abstract

The fifth generation (5G) new radio supports a diversity of network deployments. The industrial factory (InF) wireless automation use cases are emerging and drawing an increasing attention of the 5G new radio standardization groups. Therefore, in this paper, we propose a service-aware time division duplexing (TDD) frame selection framework for multi-traffic deployments. We evaluate the performance of the InF network deployments with the state-of-the-art 3GPP modeling assumptions. In particular, we consider the dynamic TDD mode along with optimized uplink power control settings. Multi-traffic coexistence scenarios are also incorporated such that quality of service (QoS) aware dynamic user scheduling and TDD link selection are introduced. Extensive system level simulations are performed in order to evaluate the performance of the proposed solutions, where the proposed QoS-aware scheme shows 68% URLLC outage latency reduction compared to the QoS-unaware solutions. Finally, the paper offers insightful conclusions and design recommendations on the TDD radio frame selection, uplink power control settings and the best QoS-coexistence practices, in order to achieve a decent URLLC outage latency performance in the state-of-the-art InF deployments.

Index Terms— Dynamic TDD; Indoor factory automation (inF); URLLC; eMBB; Cross link interference (CLI); 5G new radio.

1 Introduction

The 5G new radio (5G-NR) supports multiple service classes such as the ultra-reliable and low latency communications (URLLC), and the enhanced mobile broadband (eMBB) [1]. The URLLC services require stringent radio and reliability targets, i.e., one way radio latency of 1 ms with 99.999% success probability, where the eMBB applications demand extreme data rates [2, 3]. The indoor factory automation (InF) [4, 5] use cases are emerging where the 5G-NR cellular communications are envisioned to replace the Ethernet-based interconnections. The early 5G commercial roll-outs are expected over the unpaired spectrum due to the available large free bandwidth [6, 7]. Therefore, the time division duplexing (TDD) is vital for the 5G success. For TDD deployments, base-stations (BSs) are able to dynamically change their respective radio frame configurations in order to meet the time-varying traffic demands.

Although dynamic TDD systems offer greater flexibility of the network resources in line with the directional traffic demands, the stringent URLLC latency and reliability targets are highly challenging in those networks [8]. This is attributed to: (a) the non-concurrent availability of the downlink (DL) and uplink (UL) transmission opportunities, and (b) the additional cross link

interference (CLI) of BSs and user-equipment's (UEs) with concurrent opposite transmission links.

The achievable URLLC outage performance has been widely investigated for the indoor deployments [9-11], where the indoor office deployments are mainly considered. Although, to the best of our knowledge, there is a lack of prior art of the URLLC performance analysis in the InF dynamic TDD deployments and with the corresponding channel modeling and design assumptions. Furthermore, in [12], authors investigate the achievable radio outage latency in the time-sensitive communications, where a tighter synchronization and on-time delivery of packets are considered. In TDD deployments, The structure of the DL and UL link switching of the TDD radio frame and the BS-BS CLI have been proved to have a dominating impact on the URLLC outage radio latency [8]. Therefore, a diversity of inter-BS TDD radio frame coordination schemes are introduced in the open literature. In [13-15], coordinated DL beam-forming and receiver design are proposed in order to isolate the subspace of the BS-BS CLI in the spatial domain from the useful signal subspace. Moreover, smarter dynamic UE scheduling and optimized power control [16] are essential to control the network CLI. Those schemes typically require an inter-BS coordination signaling overhead, e.g., for exchanging the UE-specific allocation information.

Opportunistic TDD frame coordination schemes are also developed in order to partially or fully avoid the network CLI with simpler processing requirements and less coordination overhead. In [17], a set of TDD system optimizations, such as hybrid frame design and slot-aware dynamic UE scheduling, is combined in order to offer CLI-free channels for the UEs of the worst channel conditions. Furthermore, a semi-static TDD adaptation algorithm [18] is proposed to avoid the network CLI while offering a semi-static dynamicity of the network TDD radio frame to traffic demands. Finally, a reinforcement-learning (RL) based TDD frame optimization scheme [19] has been proposed to autonomously optimize the BS-specific TDD frame selection in a distributed manner, where the achievable learning gain offers a considerable URLLC performance improvement compared to reactive TDD adaptation schemes. Therein, two learning instances have been defined. The first learning instance estimates the best DL to UL symbol ratio to adopt during a radio frame where the second learning network seeks the best corresponding symbol placement across the frame such that the latency statistics are minimized.

In this paper, we propose a QoS-aware TDD system framework for emerging InF TDD deployments. This includes service-aware dynamic UE scheduling, TDD radio frame selection criterion. We comprehensively evaluate the achievable URLLC outage latency performance within such deployments, in combination with the eMBB services. First, we investigate the impact of the UL power control setting and CLI on the URLLC outage performance.

Secondly, joint URLLC and eMBB QoS coexistence scenarios are considered. QoS-aware TDD link selection and dynamic UE scheduling are incorporated to balance among the feasibility of a decent URLLC outage latency performance and the achievable eMBB capacity. Finally, we adopt an RL based solution to dynamically optimize the selection of the BS-specific TDD frame configuration for different load regions. The presented performance evaluations are obtained through extensive system level simulations where the latest 3GPP modeling guidelines are followed. The paper offers insightful recommendations of the optimized TDD system design aspects for the InF deployments to fulfill the URLLC stringent targets.

This paper is organized as follows. Section 2 introduces the system modeling. Section 3 presents the considered QoS-aware dynamic user scheduling and TDD link selection strategy. Section 4 discusses the simulation methodology and the major performance evaluation of the proposed solution. Finally, conclusions are drawn in Section 5.

2 System Model

We consider an InF TDD network with C cells, each is equipped with N antennas. As depicted by Fig. M.1, the network deployment follows the 3GPP modeling guidelines for InF networks [4, 5]. There are $K = K^{\mathrm{dl}} + K^{\mathrm{ul}}$ uniformly-distributed UEs per cell, where K^{dl} and K^{ul} imply the number of the DL and UL UEs per cell. Each UE is equipped with M antennas, and is assumed to request both DL and UL transmissions, respectively. The URLLC service is modeled with the FTP3 traffic model [20], where the DL and UL URLLC packets are of a finite size f^{dl} and f^{ul} bits, respectively. URLLC packets arrive at the transmitter according to a Poisson Arrival Process with mean packet arrival rates of λ^{dl} and λ^{ul}, in the DL and UL directions, respectively. Therefore, the offered URLLC load per cell in the DL direction is calculated by: $\Omega^{\mathrm{dl}} = K^{\mathrm{dl}} \times f^{dl} \times \lambda^{\mathrm{dl}}$, and in the corresponding UL direction as: $\Omega^{\mathrm{ul}} = K^{\mathrm{ul}} \times f^{ul} \times \lambda^{\mathrm{ul}}$. The total offered URLLC load is expressed as: $\Omega = \Omega^{\mathrm{dl}} + \Omega^{\mathrm{ul}}$. In this paper, we also assume the eMBB-URLLC coexistence scenarios solely in the DL direction, where $k^{\mathrm{dl}}_{\mathrm{eMBB}} \subset K^{\mathrm{dl}}$. The eMBB traffic is modeled by a constant bit rate (CBR) per each eMBB UE [21], i.e., emulates a broadband video streaming service. Specifically, it implies finite-size eMBB packets $\rho -$ bits which arrive at the transmitter with a constant arrival rate in time. For those scenarios, the total offered load in DL direction is calculated as: $\Omega^{\mathrm{dl}} = \left(\Omega^{\mathrm{dl}}\right)^{\mathrm{eMBB}} + \left(\Omega^{\mathrm{dl}}\right)^{\mathrm{urllc}}$, where $\left(\Omega^{\mathrm{dl}}\right)^{\mathrm{eMBB}}$ is the eMBB offered load.

The UEs are dynamically multiplexed using the orthogonal frequency division multiple access (OFDMA). In line with the 3GPP assumptions for URLLC, we adopt a sub-carrier spacing (SCS) of 30 kHz with a physical resource block (PRB) of twelve consecutive SCSs. We assume a short trans-

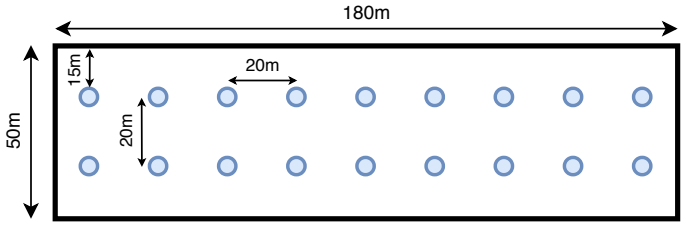


Fig. M.1. System model: InF network deployment.

mission time interval (TTI) of 4 OFDM symbol duration for both URLLC and eMBB transmissions. Before the start of each radio frame (10 ms), each BS decides the structure of the selected slot formats within the radio frame, where there is a guard symbol between each DL and UL TTI transition to compensate for the channel propagation delay.

In line with the system modeling assumptions in [19], we explicitly consider the major functionalists of the 5G NR PHY and MAC layers. In the DL direction, arriving packets are first processed, and are buffered towards the first available DL TTI of the current TDD radio frame. The DL UEs are dynamically scheduled using the adopted MAC scheduler. The DL packets are subject to a further processing delay at the UE-side. In case the DL packets are not successfully decoded, UEs trigger a HARQ negative ACK (NACK) during the first available UL TTI. Subsequently, the serving BSs re-transmit the failed DL packets during the next available DL transmission opportunity. We adopt dynamic link adaptation in the DL direction, based on periodic reporting of the channel quality indications (CQIs) to select the best corresponding modulation and coding scheme (MCS) to achieve the target block error rate (BLER).

In the UL direction, in line with [22], we consider the configured grant (CG) transmission with a fixed MCS per UE. With CG, the arriving UL packets at the UE-side side are immediately prepared for UL transmissions during the first available UL TTI. This removes the delay for transmitting the scheduling request until receiving the corresponding scheduling grant. All CG-based UL transmissions include a robust preamble such that the BS is able to distinguish from which UEs the UL transmission is initiated. The CG UL configurations are set such that all active UL UEs transmit over a randomly selected sub-band, of one quarter of the carrier bandwidth, with a predefined MCS level of QPSK rate $1/2$. This setting allows for transmitting one full URLLC packet in a single shot without segmentation.

The UL transmission power is configured as

$$\Sigma\left[dBm\right] = \min\left\{\Sigma_{\max},\ P0 + 10\log_{10}\left(\wp\right) + \alpha\eth\right\}, \tag{M.1}$$

where $\Sigma_{\max}$ is the maximum UE transmit power, $P0$ is the power spectral density, $\wp$ is the number of granted UL PRBs, α and $\eth$ are the path-loss compensation factor and path-loss. As CG transmissions from different UEs can be transmitted over overlapping resources, those are subject to intra-cell interference. In case the BS fails to decode UL transmissions from different UL UEs, the BS triggers a re-transmission request during the first available DL TTI with a dedicated scheduling grant for the UE. Correspondingly, the UL UE initiates a packet re-transmission using the same MCS and bandwidth configuration as the first UL transmission, with a 3 dB transmission power boost to improve the decoding probability of the HARQ re-transmission [22].

3 Key Factors Impacting The URLLC Performance in InF Deployments

In the following subsections, we show the critical 5G NR system design aspects impacting the achievable URLLC outage performance within the emerging InF TDD deployments. Those span the optimization of the UL power control settings, dynamic UE scheduling, and the TDD link selection framework, respectively.

3.1 QoS-aware TDD Radio Frame Selection

In dynamic TDD networks, BSs independently select the radio frame structures, in terms of the number of the DL and UL transmission opportunities across the frame duration, that best meet their respective traffic needs. Therefore, BSs continuously monitor their offered traffic demands in the DL and UL directions. We formulate the relative buffered traffic ratio $\mu_{[t,c]}(\varsigma)$ at the ς^{th} slot of the radio frame, $\varsigma = 1, 2, \ldots, \xi$, and ξ is the number of slots per the radio frame as

$$\mu_{[t,c]}(\varsigma) = \frac{Z^{\mathrm{dl}}_{[t,c]}(\varsigma)}{Z^{\mathrm{dl}}_{[t,c]}(\varsigma) + (1/\iota)\ Z^{\mathrm{ul}}_{[t,c]}(\varsigma)}, \tag{M.2}$$

where $Z^{\mathrm{dl}}_{[t,c]}(\varsigma)$ and $(1/\iota)\, Z^{\mathrm{ul}}_{[t,c]}(\varsigma)$ denote the total DL and UL buffered traffic size of the ς^{th} slot during the current frame, and ι implies the first-transmission average UL BLER at the BS side. The latter is linearly averaged across all UL transmissions and updated using a sliding window per UE. The intuition of such formulation is derived by the fact that the BS has different knowledge of the $Z^{\mathrm{dl}}_{[t,c]}(\varsigma)$ and $Z^{\mathrm{ul}}_{[t,c]}(\varsigma)$ information. In particular, the knowledge of the $Z^{\mathrm{dl}}_{[t,c]}(\varsigma)$ is available at the BS. However, in the UL direction, the buffered first UL transmission size per UE $Z^{\mathrm{ul}}_{[t,c]}(\varsigma)$ is not im-

mediately accessible at the BS until it is received at the BS side. Therefore, the term $(1/\iota)\, Z^{\mathrm{ul}}_{[t,c]}(\varsigma)$ is adopted to reflect the actual offered UL traffic size, i.e., equivalent to $Z^{\mathrm{dl}}_{[t,c]}(\varsigma)$ in the DL direction.

For multi-traffic deployments with joint URLLC-eMBB, the terms $Z^{\mathrm{dl}}_{[t,c]}(\varsigma)$ and $(1/\iota)\, Z^{\mathrm{ul}}_{[t,c]}(\varsigma)$ represent the aggregate URLLC-eMBB buffered traffic sizes in the DL and UL directions, respectively. As the eMBB traffic demand is typically much larger than of the corresponding URLLC, the buffer ratio in (M.2) and the selection of the TDD frame are both dominated by the eMBB traffic statistics instead. This could be problematic to achieve a decent URLLC outage latency due to the additional URLLC packet buffering, i.e., due to the selection of a radio frame configuration that does mainly satisfy with the buffered URLLC traffic. Therefore, we adopt a QoS-aware TDD link selection criterion such as the buffered traffic statistic in (M.2) is biased towards the URLLC QoS as follows

$$\begin{matrix} Z^{\mathrm{dl/ul}}_{[t,c]}(\varsigma) \rightarrow \left(Z^{\mathrm{dl/ul}}_{[t,c]}(\varsigma)\right)^{\mathrm{urllc}}, \text{URLLC-only} \\ Z^{\mathrm{dl/ul}}_{[t,c]}(\varsigma) \rightarrow \left(Z^{\mathrm{dl/ul}}_{[t,c]}(\varsigma)\right)^{\mathrm{eMBB}}, \text{eMBB-only} \end{matrix}, \tag{M.3}$$

where $\left(Z^{\mathrm{dl}}_{[t,c]}(\varsigma)\right)^{\mathrm{urllc}}$, $\left(Z^{\mathrm{dl}}_{[t,c]}(\varsigma)\right)^{\mathrm{eMBB}}$, $\left(1/\iota Z^{\mathrm{ul}}_{[t,c]}(\varsigma)\right)^{\mathrm{urllc}}$, and $\left(1/\iota Z^{\mathrm{ul}}_{[t,c]}(\varsigma)\right)^{\mathrm{eMBB}}$ are the aggregate DL and UL buffered traffic sizes for the URLLC and eMBB UEs, respectively. The instantaneous buffered traffic ratios $\mu_{[t,c]}(\varsigma)$ are averaged over the duration of the TDD radio frame given as

$$\overline{\mu}_{[t,c]} = \frac{1}{\xi} \sum_{\varsigma=1}^{\xi} \mu_{[t,c]}(\varsigma), \tag{M.4}$$

where $\overline{\mu}_{[t,c]}$ is the average traffic ratio of the current radio frame. The traffic ratio $\overline{\mu}_{[t,c]} \rightarrow [0,1]$ implies the combined buffering performance of the DL and UL traffic size. For example, $\overline{\mu}_{[t,c]} = 0.1$ implies that the buffered UL traffic is 9x times the corresponding DL traffic. Therefore, the corresponding BS shall select a TDD radio frame with 90% of time allocation to the UL transmission opportunities, assuming a similar UL and DL spectral efficiency. The DL and UL symbols of the selected radio frames are evenly distributed in terms of 4 OFDM symbol blocks, following the adopted DL-to-UL symbol ratio.

3.2 QoS-aware Dynamic UE Scheduling

To highlight the impact of the UE scheduler, we adopt two frameworks of the multi-QoS dynamic UE schedulers. First, we consider the well-known weighted proportional fair (PF) criterion [23] to dynamically schedule different URLLC and eMBB UEs in the time and frequency domains. UEs are

sorted in the time domain such as the URLLC UEs are always given a higher priority than the eMBB UEs, i.e., URLLC UEs are given a higher weight in the PF criterion. Therefore, the higher PF weight of the URLLC UEs aims to always schedule the active URLLC UEs before the respective eMBB UEs in the time domain. Thereafter, active URLLC and eMBB UEs are both scheduled based on the PF criterion in the frequency domain. That is, according to their achievable instantaneous throughput relative to the total received capacity. This way, the scheduling fairness is always guaranteed in the frequency domain among each set of the URLLC and eMBB UEs, respectively. The main drawback of such scheduling framework is that the URLLC latency statistics are not considered in the scheduling criterion, and therefore, it could lead to a degraded URLLC outage latency performance.

Secondly, we adopt the scheduling framework introduced in [24]. Instead of the throughput-based PF scheduling criterion, the head of line delay (HoLD) is the basic scheduling criterion. The HoLD per packet per UE is defined as the time from the DL packet arrives at the transmitter end until it is successfully decoded at the intended receiver end. The scheduler always prioritizes an immediate scheduling for the URLLC UEs with the largest HoLD statistics while requiring the least packet segmentation. The intuition is that the scheduler seeks to minimize the probability of the URLLC packet segmentation probability, therefore, reducing the URLLC outage latency. In case packet segmentation is not avoidable due to the resource shortage, the scheduler seeks to segment a single URLLC packet that leads to the minimum control overhead per TTI, hence, leaving more resource for data transmissions. In joint URLLC-eMBB deployments, such scheduler is proved to offer considerable eMBB capacity, due to the faster transmissions of the concurrent URLLC packets, therefore, leaving more resources for the corresponding eMBB traffic.

4 Performance Evaluation

4.1 Simulation Methodology

We adopt a highly-detailed system level simulations to evaluate the performance of the proposed solutions. The main set of the simulation parameters is listed in Table M.I. We adopt the dense clutter - high BS propagation model of the InF deployments [4, 5], where the BSs are elevated as compared to active UEs. The simulator used for the system level evaluations has a timing resolution of a single OFDM symbol and includes the main functionalities of the 5G NR protocol stack. The simulator is validated via calibration exercises, where baseline statistics for predefined simulation scenarios are reported and compared between the various 3GPP partners [25]. For each radio frame of 10

Table M.1: Simulation parameters.

Parameter	Value
Environment	3GPP-InF, one cluster, 18 cells
UL/DL channel bandwidth	20 MHz, SCS = 30 KHz, TDD
Channel model	InF-DH (dense clutter and high BS) [5]
BS and UE transmit power	BS: 30 dBm, UE: 23dBm
Carrier frequency	3.5 GHz
BS and UE heights	BS: 10m, UE: 1.5m
Antenna setup	$N = 4$, $M = 4$
Average UEs per cell	$K^{\mathrm{dl}} = K^{\mathrm{ul}}$ = 8-16
TTI configuration	4-OFDM symbols
URLLC Traffic model	FTP3, $f^{dl} = f^{ul}$ = 256 bits λ^{dl} = 50 pkts/sec λ^{ul} = 50 pkts/sec
eMBB Traffic model	CBR, ρ = 16k bits, rate/UE = 0.5 Mbps
DL scheduling	PF, min-HoLD [24]
UL scheduling	CG, QPSK1/2, $P0$= -61 dBm, α= 1
Processing time	PDSCH prep. delay: 2.5-OFDM symbols PUSCH prep. delay: 5.5-OFDM symbols PDSCH decoding : 4.5-OFDM symbols PUSCH decoding: 5.5-OFDM symbols
DL/UL receiver	L-MMSE-IRC
TDD frame	10 ms

ms, the BSs select the radio frame configurations which best suit their current DL and UL traffic demand. During the DL TTIs, UEs are dynamically scheduled using either the PF or min-HoLD [24] criterion. During the UL TTIs, UEs transmit their UL packets using the CG UL following the settings presented in Section 2. For DL/UL packets, the signal to interference noise ratios (SINR) of the granted sub-carriers are calculated using the linear minimum mean square error interference rejection and combining receiver (L-MMSE-IRC). Those are combined using the mean mutual information per coded bit (MMIB) mapping [26] in order to estimate the effective SINR point. Based on the effective SINR, the corresponding error probability is calculated using look-up tables, obtained from extensive link level simulations, considering the received effective SINR and the adopted MCS.

4.2 Performance Results

The UL power control settings have a vital impact on the overall URLLC performance. Fig. M.2 presents the complementary cumulative distribution

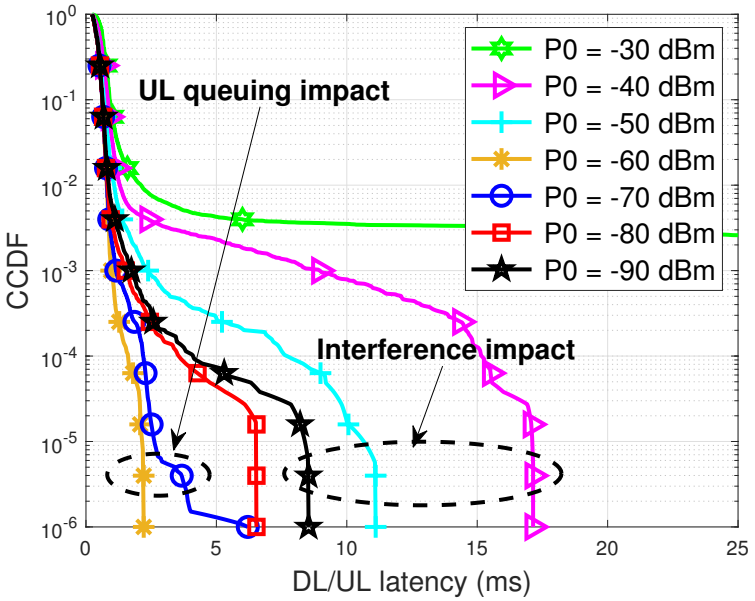


Fig. M.2. Achievable URLLC latency with dynamic TDD for different $P0$.

function (CCDF) of the achievable URLLC latency with different UL power control settings, i.e., for several $P0$ configurations. As can be clearly observed, with $P0 = -90$ to -60 dBm, a decent URLLC outage latency performance is obtained. Herein, the majority of the UL UEs transmit their UL packets with a lower transmission power. Therefore, the inter-cell interference is controlled while the UL packet queuing delay dominates the achievable URLLC outage latency. With very high $P0 = -40$ to -30 dBm, the majority of the UL UEs transmit their payload with the maximum permissible transmission power, resulting in a significant increase of the inter-cell interference. Hence, the interference starts to dominate the URLLC outage latency where the packets require multiple HARQ re-transmission combing attempts before a successful decode, leading to a highly degraded URLLC outage latency performance. Based on the obtained URLLC performance in Fig. M.2, we adopt $P0 = -61$ dBm for the rest of the results in order to achieve the best possible URLLC outage latency.

Fig. M.3 depicts the CCDF of the achievable combined DL/UL URLLC latency for different offered loads. For the low load region with $\Omega = 0.5$ Mbps, the URLLC target is achieved, i.e., a fully dynamic TDD satisfies the URLLC outage latency target of 1 ms. This is mainly attributed to the low CLI intensity and the smaller queuing delays under such very low offered load. For higher offered loads of $\Omega = 3$ Mbps, the achievable URLLC outage

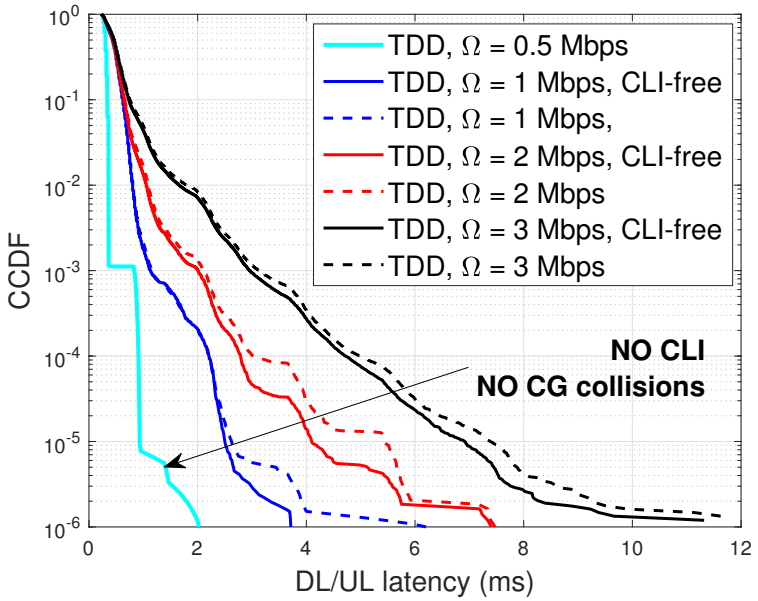


Fig. M.3. Achievable URLLC latency with dynamic TDD for different loads.

latency inflicts a clear increase due to the packet queuing delay. Moreover, the CLI is shown to have a minor effect on the achievable URLLC latency for the low and moderate offered load levels.

Next, we investigate the URLLC and eMBB coexistence performance under different scheduling policies. Fig. M.4 depicts the CCDF of the achievable URLLC latency, where $\Omega = 2$ Mbps, and the eMBB traffic is only incorporated in the DL directions with 3 eMBB UEs per cell, each has a CBR of 0.5 Mbps. The throughput-based PF dynamic UE scheduler fails to achieve a decent URLLC outage compared to the HoLD-aware scheduler [24]. This is mainly because the latter considers the latency statistics of pending URLLC UEs in the scheduling criterion. It seeks to schedule the URLLC UEs with the largest HoLD statistics while reducing the probability of the packet segmentation. Furthermore, adopting a QoS-aware TDD link selection criterion tends to significantly improve the achievable URLLC performance, i.e., 68% outage latency reduction compared to the URLLC QoS-unaware TDD selection criterion. This is attributed to the fact that with the QoS-aware TDD link selection, the selection of the TDD frame configuration is dictated by the URLLC offered traffic size, instead of the aggregate URLLC/eMBB traffic, reducing the TDD link switching delay of the urgent URLLC packets.

Fig. M.5 shows the empirical CDF (ECDF) of the achievable throughput per eMBB UE. The source CBR rate is pre-configured as 0.5 Mbps per UE. As

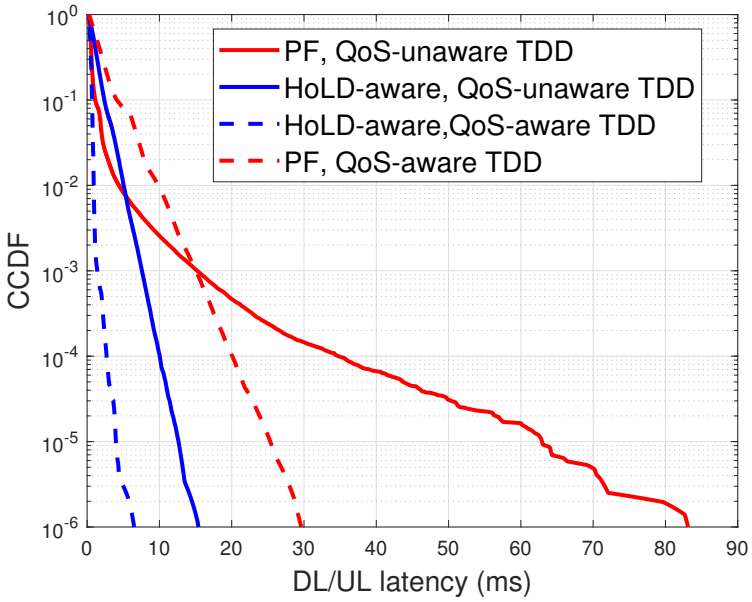


Fig. M.4. Achievable URLLC latency for different dynamic UE scheduling and TDD link selection.

depicted, the achievable eMBB throughput with the HoLD-aware scheduler approaches the source CBR rate, while significantly outperforming the case with the PF scheduler. This is mainly because: (1) the URLLC transmissions are scheduled in a faster basis while the URLLC packet segmentation is reduced, and (2) in the frequency domain, URLLC packets are scheduled based on the throughput-to-average criterion which further minimizes the required total number of PRBs to allocate the active URLLC UEs. The HoLD-aware scheduler attempts to avoid the URLLC packet segmentation, resulting from the insufficiently available free resources. In case this is not possible, the scheduler seeks to inflict segmentation of the URLLC packets that result in the lowest possible control overhead. Therefore, it leaves more resources for the respective eMBB traffic, and accordingly, achieving a highly optimized eMBB capacity compared to the case with the PF scheduler.

Finally, we investigate the potential of the RL-based TDD frame selection solution [19] compared to the non-RL based TDD frame selection schemes, i.e., reactive TDD, where only the URLLC traffic is considered. Fig. M.6 depicts the achievable URLLC latency performance when such learning approach is adopted for different load regions. As can be observed, at the low load region, both the RL-based TDD and reactive dynamic TDD schemes offer a similar URLLC outage performance. This is mainly due to the low

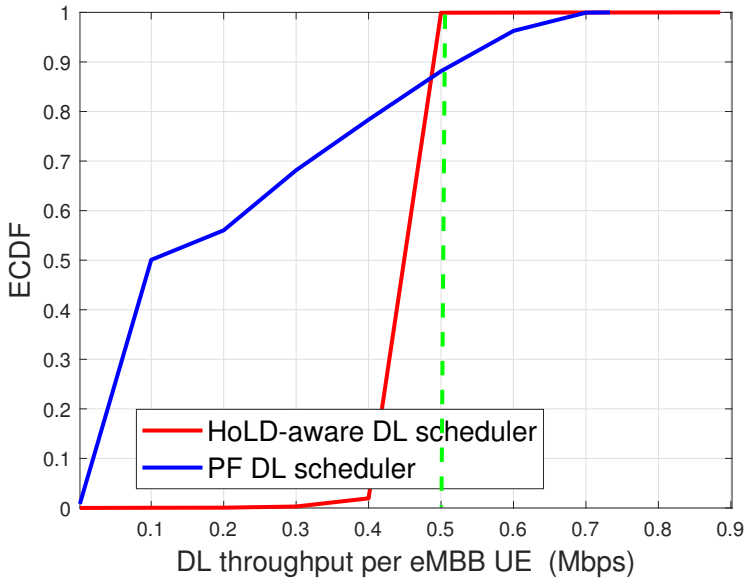


Fig. M.5. Achievable eMBB CBR rate for URLLC-eMBB coexistence.

resource utilization, thus, adopting predefined random UL/DL allocations during the selected TDD frame, in line with the buffered traffic ratio, is sufficient. At the high load region, the resource utilization increases, introducing additional queuing delays for urgent URLLC packets in both the DL and UL directions. Therefore, due to the smarter and latency-aware adaptation of the RL-based TDD solution, the TDD learning approach obviously outperforms the basic dynamic TDD scheme.

5 Concluding Remarks

We have evaluated the achievable URLLC performance for the emerging indoor factory automation 5G network deployments. We have analyzed the state-of-the-art dynamic TDD duplexing scheme with optimized uplink power control settings. For the URLLC-eMBB coexistence scenarios, we adopt quality of service (QoS)-aware dynamic user scheduling and TDD link selection strategy, respectively. The main recommendations offered by this paper are as follows: (a) for the indoor factory deployments, the CLI is not a major critical performance bottleneck as the case with the macro networks, due to the smaller difference between the uplink and downlink transmission power, (2) the optimization of the uplink control settings has a vital impact on the achievable URLLC outage performance. Unoptimized uplink power control

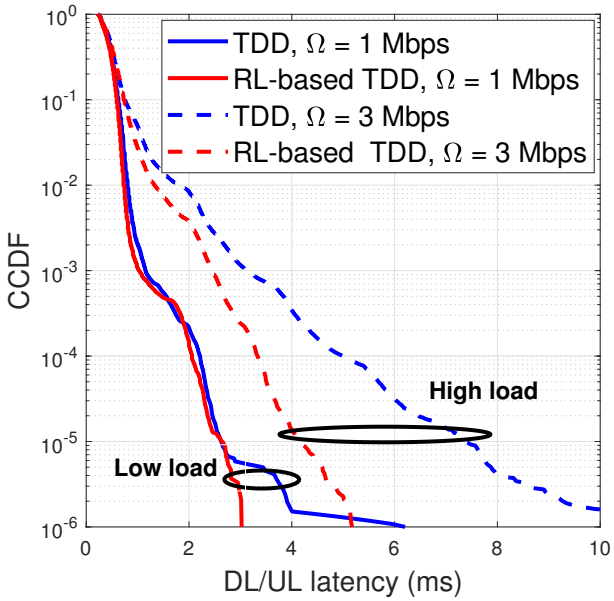


Fig. M.6. Achievable URLLC latency with the reinforcement learning approach.

configurations could either lead to a further uplink queuing delay or a significantly higher inter-cell same and cross-link interference. Therefore, we recommend setting $P0 = -61$ dBm within the indoor factory deployments to achieve the best possible URLLC outage latency, (3) within multi-QoS coexistence scenarios, latency-aware dynamic user scheduling and TDD frame selection strategies are vital to achieve a decent URLLC latency performance, and (4) reinforcement learning (RL) based TDD frame adaptation is effective in achieving a decent URLLC outage latency within InF deployments, through the dynamic selection of the number and placement of the downlink and uplink transmission opportunities across the TDD radio frame which best reduces the overall radio latency. However, it requires a careful modeling of the learning objectives, inputs, and outputs.

6 Acknowledgments

This work is partly funded by the Innovation Fund Denmark – File: 7038-00009B. The authors would like to acknowledge the contributions of their colleagues in the project.

References

[1] *Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, ITU-R M.2083-0, International Telecommunication Union (ITU), Feb. 2015.

[2] *Service requirements for the 5G system*; Stage-1 (Release 16), 3GPP, TS 22.261, V16.6.0, Dec. 2018.

[3] M. Bennis, M. Debbah and H. V. Poor, "Ultrareliable and low-latency wireless communication: tail, risk, and scale," *Proc. of the IEEE*, vol. 106, no. 10, pp. 1834-1853, Oct. 2018.

[4] *Service requirements for cyber-physical control applications in vertical domains*; Stage-1 (Release 17), 3GPP, TS 22.104, V17.3.0, July 2020.

[5] *Study on channel model for frequencies from 0.5 to 100 GHz*; Release 16, 3GPP, TR 38.901, V16.1.0, Dec. 2019.

[6] J. Lee et al., "Spectrum for 5G: global status, challenges, and enabling techs," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 12-18, Mar. 2018.

[7] K. I. Pedersen, G. Berardinelli, F. Frederiksen and P. Mogensen, "A flexible 5G wide area solution for TDD with asymmetric link operation," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 122-128, April 2017.

[8] Ali A. Esswie, and K.I. Pedersen, "On the ultra-reliable and low-latency communications in flexible TDD/FDD 5G networks," *in Proc. IEEE CCNC,* Las Vegas, Jan. 2020.

[9] L. Wang and H. Zhang, "Analysis of joint scheduling and power control for predictable URLLC in industrial wireless networks," *in Proc. IEEE ICII* , Orlando, FL, USA, 2019, pp. 160-169.

[10] M. Alonzo, P. Baracca, S. R. Khosravirad and S. Buzzi, "URLLC for factory automation: an extensive throughput-reliability analysis of D-MIMO," *in Proc. IEEE WSA,* Hamburg, Germany, 2020, pp. 1-6.

[11] V. Hytnen, Z. Li, B. Soret and V. Nurmela, "Coordinated multi-cell resource allocation for 5G ultra-reliable low latency communications," *in Proc. IEEE EuCNC*, Oulu, 2017, pp. 1-5.

[12] R. B. Abreu, G. Pocovi, T. H. Jacobsen, M. Centenaro, K. I. Pedersen and T. E. Kolding, "Scheduling enhancements and performance evaluation of downlink 5G time-sensitive communications," *IEEE Access*, vol. 8, pp. 128106-128115, 2020.

[13] Z. Huo, N. Ma and B. Liu, "Joint user scheduling and transceiver design for cross-link interference suppression in MU-MIMO dynamic TDD systems," *in Proc. IEEE ICCC,* Chengdu, Dec. 2017, pp. 962-967.

[14] A. Łukowa and V. Venkatasubramanian, "Performance of strong interference cancellation in flexible UL/DL TDD systems using coordinated muting, scheduling and rate allocation," *in Proc. IEEE WCNC*, Doha, April 2016, pp. 1-7.

[15] Ali A. Esswie, and K.I. Pedersen, "Cross link interference suppression by orthogonal projector in 5G dynamic-TDD URLLC systems," *in Proc. IEEE WCNC,* April 2020.

[16] A. Łukowa and V. Venkatasubramanian, "Coordinated user scheduling in 5G dynamic TDD systems with beamforming," *in Proc. IEEE PIMRC*, Bologna, Sep. 2018, pp. 596-597.

[17] Ali A. Esswie, K.I. Pedersen, and P. Mogensen, "Quasi-dynamic frame coordination for ultra- reliability and low-latency in 5G TDD systems," *in Proc. IEEE ICC*, Shanghai, China, May 2019, pp. 1-6.

[18] Ali A. Esswie, K.I. Pedersen, and P. Mogensen, "Semi-static radio frame configuration for URLLC deployments in 5G macro TDD networks," *in Proc. IEEE WCNC,* April 2020.

[19] A. A. Esswie, K. I. Pedersen, and P. Mogensen, "Online radio pattern optimization based on dual reinforcement-learning approach for 5G URLLC networks," *IEEE Access*, 2020.

[20] *Study on new radio access technology physical layer aspects*; Release 14, V14.2.0, TR 38.802, 3GPP, Sep. 2017.

[21] Tong Shan, O. Yang and Genzao Zhang, "Scheduling jittered CBR traffic in broadband wireless access systems," *in Proc. IEEE ICC*, Anchorage, AK, 2003, pp. 22-26.

[22] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, I. Z. Kovacs and P. Mogensen, "Joint resource configuration and MCS selection scheme for uplink grant-free URLLC," *in Proc. Globecom*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1-6.

[23] M. K. Mller, S. Schwarz and M. Rupp, "QoS investigation of proportional fair scheduling in LTE networks," *in Proc. IEEE IFIP (WD)*, Valencia, 2013, pp. 1-4.

[24] 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. IEEE VTC-Spring*, Kuala Lumpur, Malaysia, 2019, pp. 1-6.

[25] *Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC)*, Release 16, TR 38.824, 3GPP, Mar. 2019.

[26] D. G. Popescu, M. Varga and V. Bota, "Comparison between measured and computed values of the mean mutual information per coded bits in OFDM based wireless transmissions," *in Proc. IEEE Conference on Telecommunications and Signal Processing (TSP*), Rome, July 2013, pp. 380-384.

Part V

Conclusions

Conclusions

In this PhD dissertation, several broader system design enhancements are proposed and developed for multi-QoS 5G-NR coexistence scenarios. Particularly, multiple novel RRM schemes have been developed for multi-service 5G networks where a diverse set of the target radio latency, reliability and spectral efficiency are consistently optimized over the same spectrum. The developed schemes in this thesis mainly consider the URLLC and eMBB service classes, and span both the FDD and TDD duplexing modes.

1 Summary of the Main Findings

Throughout the thesis, we have demonstrated that the achievability of the URLLC latency and reliability targets is highly challenging, specially when coexisted with the eMBB service class over the same spectrum. Accordingly, several multi-service RRM schemes have been developed to flexibly satisfy the diverse URLLC and eMBB performance targets.

User-Centric Spatial Dynamic User Schedulers

The spatial schedulers are of a significant importance for those multi-QoS coexistence scenarios. They dynamically utilize the spatial degrees of freedom (sDoFs), offered by the BS antenna array, in order to achieve the diverse QoS targets. The developed QoS-aware spatial schedulers have demonstrated a further flexible multiplexing gain among the URLLC-eMBB services. The main idea is that, within the coexistence deployments, the spatial dynamic user scheduler is able to tune the available sDoFs to achieve the stringent URLLC targets while maximizing the eMBB achievable capacity. However, unlike the conventional spatial schedulers, in case the sufficient free sDoFs are not immediately available, the developed spatial schedulers enforce such sDoFs to immediately schedule the urgent URLLC packets while controllably compromising the eMBB ergodic capacity. Therefore, the introduced QoS-aware spatial schedulers offer a significant improvement of the achievable URLLC outage performance compared to network-specific spatial schedulers.

For instance, 50% reduction gain of the achievable URLLC outage latency is realized with the proposed QoS-aware spatial schedulers compared to the state-of-the-art spatial schedulers.

Flexible Inter-BS Coordination For Dynamic TDD Systems

The TDD duplexing mode is vital for the early 5G deployments, mainly due to the abundantly available free bandwidth over the unpaired TDD spectrum. Though, achieving the stringent URLLC performance targets within TDD roll-outs is highly challenged by the non-concurrent availability of the downlink and uplink transmission opportunities. Furthermore, for dynamic TDD systems, the additional cross-link-interference (CLI), from neighboring BSs and UEs transmitting and receiving at the same time, has been shown to be a critical performance bottleneck of achieving the stringent URLLC latency and reliability targets.

First, to overcome the former issue, flexible inter-BS dynamic coordination schemes are recommended to control the CLI, and particularly the severe BS-BS CLI. Therefore, we have developed several TDD radio frame coordination schemes throughout the thesis. The objective is to either avoid or suppress the network CLI while preserving the frame selection flexibility to match the time-varying and BS-specific traffic demands. The developed CLI suppression schemes, using cross-BS coordinated interference rejection and combining, have been demonstrated more effective in eliminating the BS-BS CLI compared to conventional CLI avoidance schemes. However, this comes at the expense of increased inter-BS coordination signaling overhead over the back-haul links. Using extensive system level simulations, the developed coordination schemes are shown to provide considerable CLI reduction improvements, resulting in achieving a decent URLLC outage performance compared to the static and fully uncoordinated dynamic TDD schemes, respectively. In particular, the frame-book based TDD solution (Paper I) offers a fully dynamic TDD framework that best work within the low and medium offered load region. It avoids the occurrence of the network CLI on a best effort basis. The quasi dynamic TDD solution (Paper J) provides a semi-dynamic TDD frame adaptation through hybrid TDD frame deign to allow for BS-BS CLI-free uplink transmission opportunities. Therefore, it best fits within the medium load region where the CLI intensity is moderate. For the highly loaded scenarios, the proposed semi-static TDD (Paper H) and the coordination IRC (Paper K) solutions offer the best URLLC performance through the full avoidance/suppression of the network CLI.

Finally, we have utilized a reinforcement learning (RL) based approach in predicting the TDD radio frame pattern of each BS. The main objective is for BSs to dynamically in time select the TDD radio link switching patterns that best match the incoming packet arrivals across the upcoming radio frames, hence, reducing the packet queuing delay. Unlike the former developed TDD

coordination schemes, the learning algorithm enables selecting not only the number of the downlink and uplink transmission opportunities across the TDD frame but also the best possible downlink and uplink symbol structure to adopt. The proposed learning approach shows a considerable URLLC outage improvement compared to reactive TDD schemes. The performance gain is shown to be varying with the size of the offered load. For the low load region, the RL-based TDD offers a similar outage performance as the conventional TDD scheme. For the higher load region, the packet queuing delay and the network CLI becomes more visible, and therefore, the QoS-aware RL-based TDD scheme provides a clear latency performance than the reactive TDD scheme. Finally, we utilize the developed learning framework for the emerging industrial factory (inF) deployments. For joint eMBB-URLLC coexistence inF scenarios, QoS-aware TDD link selection and dynamic user scheduling are demonstrated vital to achieve the diverse QoS targets of the eMBB and URLLC services, respectively.

2 Recommendations

Derived by the research findings of the PhD thesis, in the following, we present the list of recommendations that correspond to the research questions assumed in Part I as

Q1 How the spatial DoFs of the BS antenna array can be utilized for an enhanced URLLC-eMBB multiplexing performance?

R1 The developed QoS-aware spatial schedulers offer a flexible multiplexing gain for URLLC and eMBB service coexistence scenarios. They provide at least 50% reduction of the achievable URLLC outage latency compared to the traditional spatial schedulers. Developed schedulers require at least 8 antenna elements at the BS side and new minimal radio signaling over the downlink radio control channel. To reap the full benefits of those schedulers, we recommend defining the required new radio signaling in the upcoming 3GPP specifications.

Q2 How sensitive is the achievable URLLC outage latency performance to the TDD frame settings for various 5G-NR system configurations?

R2 The URLLC latency targets are proved to be highly sensitive to the TDD frame configurations. A faster TDD pattern periodicity enables more rapid adaptation of the network resources to the time-varying sporadic traffic arrivals. A smaller TTI duration of 4 OFDM symbols with 30 KHz sub-carrier spacing and a TDD pattern adaptation of multiple radio slots are recommended towards achieving the URLLC 1 ms latency target.

Q3 How to design flexible and computationally-efficient CLI control mechanisms?

R3 For macro dynamic TDD deployments, the BS-BS CLI is a major performance bottleneck, and leads to a poor URLLC outage latency performance. For such deployments, we recommend using the developed Semi-static TDD and RL-based solutions to control the network CLI, and hence, achieving a decent URLLC outage performance. For the InF deployments, the CLI is less of a problem than the case of the macro networks, and therefore, we recommend using the dynamic TDD and the developed quasi-dynamic TDD solutions.

Q4 What is the ML learning potential to offer a better TDD radio frame adaptation?

R4 The developed RL-based TDD solution is robust to the offered load region, where it offers a clear gain of the URLLC outage latency over the higher load region and a similar URLLC performance over the lower load region. It is shown to be an appropriate solution for various network deployments such as the InF and UmA networks, in a fully distributed manner, and with a limited convergence delay of 1 second (real-time). Therefore, we recommend adopting the developed simple RL-based TDD solution.

3 Future Work

Based on the main research findings and knowledge obtained throughout the course of this PhD project, we list the potential research extensions as follows:

- Further validating the proposed TDD solutions for frequency range 2 (FR2) and the potential needed extensions to be combined with the richer usage of gNB beam-forming and UE multi-panel antenna schemes.
- Exploring the centralized spatial schedulers option for joint URLLC-eMBB coexistence deployments. Particularly, incorporating the developed spatial schedulers in a centralized deployment option could be an interesting research topic to allow for inter-BS coordinated spatial dynamic user scheduling.
- Exploring the centralized spatial schedulers option for joint URLLC-eMBB coexistence deployments. Particularly, incorporating the developed spatial schedulers in a centralized deployment option could be an interesting research topic to allow for inter-BS coordinated spatial dynamic user scheduling.

- Adopting more agile and accurate learning models for predicting the TDD radio frame in dynamic TDD URLLC networks. This could include the on-policy RL algorithms which offer a faster convergence delay with more conservative inference performance. Examples are double Q-RL, state-action-reward-state-action (SARSA), and supervised learning based algorithms. Furthermore, centralized and multi-BS coordination learning is a potential extension to offer an inter-BS coordinated action exploration and inference.

The fifth generation (5G) of the cellular technology offers greater support for three main service classes; the ultra-reliable and low-latency communications (URLLC), enhanced mobile broadband (eMBB), and the massive machine type communication (mMTC). URLLC services require the transmission of sporadic and small-payload packets with stringent radio latency and reliability targets. The eMBB applications demand wide-band transmissions with extreme peak data rates. Finally, for mMTC, the network is required to simultaneously serve a large number of connected devices, each is associated with strict energy consumption constraints. However, there is a fundamental tradeoff between the achievable latency, reliability, and network spectral efficiency. Concurrently optimizing the quality of service (QoS) of those service classes is one of the major challenges of the 5G new radio and neither been addressed for the former wireless standards. Furthermore, the 5G new radio is designed to support both the frequency and time division duplexing (FDD, TDD) modes. And due to the abundantly available bandwidth at the 3.5 GHz unpaired spectrum, most of the early 5G deployments are envisioned with the TDD duplexing technology. However, achieving such an efficient multi-service-aware resource management is further challenging with TDD. The broader scope of this PhD. project is to research and develop novel and multi-service-aware radio resource management algorithms for multi-QoS 5G networks, spanning both FDD and TDD modes.

The first part addresses the multi-QoS (URLLC-eMBB) multiplexing problem. A QoS-aware multi-user multiple-input multiple-output (MU-MIMO) downlink scheduler is developed based subspace projections. The key idea is to eliminate the scheduling queuing delay of the newly-arriving URLLC packets in case the sufficient radio re-sources are not immediately available. The incoming URLLC transmissions are instantly paired with the active eMBB users which spatial signatures are closest possible to a pre-defined subspace. To control the inter-user interference at the critical URLLC users, the co-scheduled eMBB transmissions are spatially projected on-the-fly into an arbitrary spatial sub-space, to which the paired URLLC users align their respective transceivers into the orthonormal subspace, exhibiting substantially zero eMBB interference. Moreover, we have developed several variants of the proposed scheduler for eMBB capacity recovering and spectral efficiency optimization. We adopt highly-detailed system level simulations, with a high degree of realism in line with 3GPP NR assumptions, to evaluate the performance of the proposed schemes. Our simulation results demonstrate considerable improvements of the URLLC outage latency and the network capacity, e.g., minimizing the URLLC outage latency by 50 percent while enhancing the network capacity by 79 percent, compared to Rel-15 standard URLLC scheduler.

In the second part of the study, we target achieving the stringent URLLC outage targets in TDD 5G networks. We first demonstrate that the URLLC QoS is further harder to achieve in TDD deployments, mainly due to the TDD frame structure, i.e., no simultaneous downlink and uplink transmissions are possible, and the severe cross-link interference (CLI) when neighboring base-stations or users are adopting opposite transmission directions. A diversity of novel inter-cell coordination schemes are developed for mitigation of the critical CLI. Those schemes incorporate a new set of TDD system design improvements such as semi-static frame configuration, sliding frame-book design, joint hybrid frame design and slot-aware user scheduling, and coordinated transceiver design. Accordingly, developed coordination techniques offer a wide variety of the required inter-cell signaling over-head, TDD frame adaptation flexibility, and the achievable URLLC outage performance. Our results show a no-table URLLC outage improvement compared to standard dynamic TDD setups, e.g., 80 percent URLLC outage latency reduction.

Backed by our former conclusions, the last part of the PhD project demonstrates the potential of adopting a machine learning (ML) algorithms for real-time selection of the TDD radio frame structure. A simple, but efficient, Q-reinforcement-learning (QRL) approach for distributed online TDD frame optimization is proposed. First, a QRL network is utilized to estimate the near-optimal numbers of downlink and uplink transmission opportunities for a balanced traffic handling. A secondary QRL instance is selects the corresponding downlink and uplink symbol structure that minimizes the directional URLLC tail latency. The QRL-based solution is evaluated for both macro networks and newly emerging indoor industrial wireless deployments with dense small cell layouts. The proposed solution offers a significant URLLC outage gain in terms of autonomization of the TDD frame design on a real-time basis, URLLC outage latency reduction, and CLI-avoidance.

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-695-3

AALBORG UNIVERSITY PRESS