Call:H2020-ICT-2016-2

Project reference: 760809

Project Name:

E2E-aware Optimizations and advancements for Network Edge of 5G New Radio (ONE5G)

Deliverable D4.2
Final Results on Multi-Antenna Access and Link Enhancements

Date of delivery:	31/05/2019	Version:	0.1
Start date of project:	01/06/2017	Duration:	25 months

Document properties:

Document Number:	D4.2
Document Title:	Final results on multi-antenna access and link enhancements
Editor(s):	Luc Le Magoarou (B-COM)
Authors:	Renato Barbosa Abreu (AAU), Vincent Angilella (CNRS), Mohamad Assaad (CNRS), Daniyal Awan (HHI), Alexandru-Sabin Bana (AAU), Paolo Baracca (NOK-GE), Alexis Bazin (Orange), Samer Bazzi (HWDU), Ioannis-Prodromos Belikaidis (WINGS), Gilberto Berardinelli (AAU), Ronald Böhnke (HWDU), Elisabeth de Carvalho (AAU), Stefan Cerovic (Orange), Yejian Chen (NOK-GE), Panagiotis Demestichas (WINGS), Juwendo Denis (CNRS), Jean Dion (B-COM), Johannes Dommel (HHI), Jochen Fink (HHI), Andreas Georgakopoulos (WINGS), Imène Ghamnia (Orange), Salah Eddine Hajri (CNRS), Hardy Halbauer (NOK-GE), Najeeb Ul Hassan (HWDU), Onurcan Iscan (HWDU), Bruno Jahan (Orange), Martin Kasparick (HHI), Evangelos Kosmatos (WINGS), Martin Kurras (HHI), Luc Lemagoarou (B-COM), Wenjie Li (CNRS), Hao Lin (Orange), Ali Maatouk (CNRS), Nurul Huda Mahmood (AAU), Silvio Mandelli (NOK-GE), Honglei Miao (Intel), Jimmy Jessen Nielsen (AAU), Stéphane Paquelet (B-COM), Marcin Pikus (HWDU), Kilian Roth (Intel), Matthieu Roy (B-COM), Rana Ahmed Salem (NOK-GE), Martin Schubert (HWDU), Aspa Skalidi (WINGS), Vera Stavroulaki (WING), Stelios Stefanatos (FUB), Galini Tsoukaneri (SEUK), Koen de Turck (CNRS), Zoran Utkovski (HHI), Raphael Visoz (Orange), Shangbin Wu (SEUK), Gerhard Wunder (FUB), Wen Xu (HWDU), Yi Yuan (Orange).
Contractual Date of Delivery:	31/05/2019
Dissemination level:	PU[1]
Status:	Final
Version:	1.0
File Name:	ONE5G_D4.2_final.docx

Abstract

This is the second and final public report of the ONE5G work package WP4 "Multi-antenna access and link enhancement". D4.2 gives an overview of the overall project results, from June 2017 to May 2019, with a slight focus on the new results achieved after the intermediate report D4.1. The research results cover the following three areas 1) “Future-proof multi-service access solutions”, 2) “Massive MIMO enablers towards practical implementation”, and 3) “Advanced link management solutions assuming CRAN/DRAN deployments and/or massive MIMO”. We summarise the main outcomes and we discuss potential impacts regarding the 5G evolution.

Keywords

5G New Radio (NR), mMTC, URLLC, eMBB, massive MIMO, CRAN, DRAN, beamforming, non-orthogonal access, grant-free random access.

[1] CO = Confidential, only members of the consortium (including the Commission Services)

PU = Public

Executive Summary

Deliverable D4.2 summarises the final status of ONE5G Work Package WP4 “Multi-antenna access and link enhancement”. The investigated technologies are summarized and the main results, benefits, and conclusions are presented. The objective of this work package is the development and validation of lower layers techniques (mainly PHY/MAC) that support the three service categories enhanced Mobile Broad Band (eMBB), massive Machine Type Communications (mMTC), and Ultra-Reliable Low Latency Communications (URLLC), for the two main scenarios in ONE5G, namely “Megacities” and “Underserved Areas”. Moreover, a special emphasis is placed on implementation efficiency. An overview of the individual technologies is given at the end of each section. Finally, the impact on the 3GPP standardization is highlighted (see Section 5 and Appendix A).

The results are organized according to the tasks of the work package:

T4.1 Future-proof multi-service access solutions:

This task investigates and develops fast and reliable access solutions with a particular focus on IoT-type services which require URLLC and mMTC. The main benefits are: 1) Non-Orthogonal Multiple Access (NOMA) resolves collisions for random access channels or uplink transmission with configured grant over shared resources (“grant free access”). This avoids the delay and overhead associated with scheduling and increases the number of supported users per cell. 2) Joint user activity and data detection can provide additional gains, especially for short packets. 3) NOMA enables the coexistence of services with different delay and reliability constraints like eMBB and URLLC on the same resources.

T4.2 Massive MIMO enablers towards practical implementation:

This task is aiming at efficient implementations of massive MIMO (mMIMO), with emphasis on low complexity, improved flexibility, low cost, and high performance. This includes hybrid and digital beamforming solutions and its application to beam management, decentralized beamforming, and antenna deployments. We propose optimized array formats, which can be flexibly adapted to changing service scenarios. Also, a flexible and fast reconfigurable hardware architecture is developed for multi-service transmission. Beamforming solutions for in-band backhaul and multicast transmission are proposed. Moreover, the crucial aspect of Channel State Information (CSI) acquisition is addressed. This involves improved training and feedback designs for high-quality CSI with acceptable overhead, as well as schemes for mitigating pilot contamination. Finally, robust solutions for fronthaul constraints and restricted training length are developed.

T4.3 Advanced link management solutions for interference coordination and avoidance, based on the assumption of Cloud Radio Access Network (CRAN) / Distributed Radio Access Network (DRAN) deployments and/or massive MIMO:

This task addresses challenges related to multi-node designs including CSI acquisition and feedback, efficient signalling, cell-less designs, interference management, fronthaul design, resource allocation, scheduling, and functional split. CRAN networks with multi-link cooperative transmission offer superior spectral efficiency and flexibility, which can be exploited in many ways, for eMBB, URLLC, mMTC, and even coverage enhancement.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

3GPP	3rd Generation Partnership Project
5G	Fifth Generation
5G-NR	5G – New Radio
AAoA	Azimuth Angle of Arrival
ADC	Analogue to Digital Converter
ANM	Atomic Norm Minimization
AoA	Angle of Arrival
AP	Access Point
AR/VR	Augmented Reality / Virtual Reality
ARPU	Average Revenue Per User
ARQ	Automatic Repeat Request
AWGN	Additive White Gaussian Noise
BB	Base-Band
BBU	Base-Band Unit
BER	Bit Error Rate
BH	Back-Haul
BLC	Bit Level Combining
BLER	Block Error Rate
BPDN	Basis Pursuit Denoising
BPSK	Binary Phase Shift Keying
BS	Base Station
CAPEX	CAPital Expenditure
CBT	Cell Boarder Throughput
CC	Chase Combining
CCDM	Constant Composition Distribution Matcher
CFR	Channel Frequency Response
CIC	Cascaded Integrator Comb
CIR	Channel Impulse Response
CLI	Cross-Link Interference
CMOS	Complementary Metal Oxide Semiconductor
CRA	Coded Random Access
CRAN	Cloud Radio Access Network
CRC	Cyclic Redundancy Check
CS	Compressive/Compressed Sensing
CSE	Cell Spectral Efficiency

CSI	Channel State Information
CSIT	Channel State Information at the transmitter
CU	Central Unit
D2D	Device to Device
DC	Difference of Convex
DFE	Digital Front End
DFT	Discrete Fourier Transformation
DL	Down-Link
DM	Distribution Matcher
DMRS	DeModulation Reference Signals
DnF	Decode and Forward
DoA	Directions of Arrival
DoF	Degrees of Freedom
DPS	Dynamic Point Selection
DRAN	Distributed Radio Access Network
DtF	Detect and Forward
EAoA	Elevation Angle of Arrival
EC	European Commission
ECC	Ergodic Channel Capacity
EE	Energy Efficiency
EESM	Exponential Effective SNR Mapping
ETF	Explicit Time domain Feedback
EWSMSE	Expected Weighted Sum Mean Squared Error
FBL	Finite Block Length
FDD	Frequency Division Duplex
FEC	Forward Error Correction
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
F-NCJT	Fully-Overlap Non-Coherent Joint Transmission
FPC	Fractional Power Control
FSA	Frame slotted ALOHA
GA	Genetic Algorithm
GAMP	Generalized Approximated Message Passing
GF	Grant Free
gNB	next Generation NodeB
H2020	Horizon 2020
HARQ	Hybrid Automatic Repeat Request

HPBW	Half-Power BeamWidth (HPBW)
HPPP	Homogeneous Poisson Point Process
HW	Hard-Ware
IAB	Integrated Access Backhaul
ICT	Information and Communication Technologies
IDMA	Interleave Division Multiple Access
IFFT	Inverse Fast Fourier Transformation
IoT	Internet of Things
IPR	Intellectual Property Rights
IR	Incremental Redundancy
ISD	Inter Site Distance
JSDM	Joint Spatial Division and Multiplexing
KL	Kullback Leibler
KPI	Key Performance Indicator
LDPC	Low Density Parity Check
LMMSE	Linear Minimum Mean Squared Error
LNA	Low Noise Amplifier
LOS	Line Of Sight
LTE	Long Term Evolution
MAC	Medium Access Control
MAMRC	Multiple Access Multiple Relay Channel
MBB	Mobile Broad Band
MBMS	Multimedia Broadcast Multicast Services
MC-CDMA	Multi-Carrier Code-Division Multiple Access
MCDM	Multi-Composition Distribution Matcher
MEC	Multi-access Edge Computing
MIMO	Multiple Input Multiple Output
ML	Maximum Likelihood
mMIMO	Massive Multiple Input Multiple Output
MMSE-SIC	Minimum Mean Squared-Error with Successive Interference Cancellation
mMTC	Massive Massine Type Communications
MMV	Multiple Measurement Vectors
mmWave	millimetre Wave
MPR	Multi-packet reception
MRT	Maximum Ratio Transmission
MSE	Mean Squared Error
MTC	Machine Type Communication

MU	Multi User
MU-MIMO	Multi User MIMO
MXL	Matrix Exponential Learning
NB-IoT	Narrow-Band Internet of Things
NCJT	Non Coherent Joint Transmission
NE	Nash Equilibrium
NF-NCJT	None-Fully-Overlap Non-Coherent Joint Transmission
NFV	Network Function Virtualization
NOCA	Non-Orthogonal Coded Access
NOMA	Non-Orthogonal Multiple Access
NORA	Non-Orthogonal Random Access
NP-hard	Non-deterministic Polynomial-time hardness
NR	New Radio (3GPP Release 15)
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency-Division Multiple Access
OL	Overloading Factor
OMA	Orthogonal Multiple Access
OMAMRC	Orthogonal Multiple Access Multiple Relay Channel
OMP	Orthogonal Matching Pursuit
OOB	Out Of Band
OST	One Step Thresholding
PA	Power Amplifier
PAS	Probabilistic Amplitude Shaping
PDCCH	Physical Downlink Control Channel
PDSCH	Physical Downlink Shared Channel
PER	Packet Error Rate
PHY	PHYsical Layer
PSCM	Probabilistically Shaped Coded Modulation
PSK	Phase Shift Keying
PTR	Phase Tracking Reference Signal
PUSCH	Physical Uplink Shared Channel
QAM	Quadrature Amplitude Modulation
QnF	Quantize and Forward
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
QuaDRiGa	QUAsi Deterministic RadIo channel GenerAtor
RACH	Random Access Channel

RAN	Radio Access Network
RB	Resource Blocks
RBP	Restless Bandit Problem
REG	Resource Element Group
RF	Radio Frequency
RMNSQE	Root Mean Normalized Squared Quantization Error
RRH	Remot Radio Head
RRM	Radio Resource Management
RSMA	Resource Spread Multiple Access
RTT	Round Trip Time
RU	Radio Unit
Rx	Receiver
RZF-CI	Regularized Zero Forcing with Controlled Interference
SB	Sub-Band
SBC	Spatial Basis Coverage
SCMA	Sparse Code Multiple Access
SCPTM	Single-Cell Point-to-Multipoint
SIC	Successive Interference Cancellation
SINR	Signal- to- Interference- plus- Noise Ratio
SLC	Symbol Level Combining
SNR	Signal- to- Noise Ratio
SU	Single User
SV	Singular Vector
TDD	Time Division Duplex
TeC	Technology Components
TRP	Transmission Reception Point
Tx	Transmitter
UCA	Uniform Cylindrical Array
UCI	Uplink Control Information
UE	User Equipment
UL	Up-Link
UMa	Urban Macro
UMi	Urban Micro
UPA	Uniform Planar Array
URLLC	Ultra-Reliable Low Latency Communications
WB	Wide-Band
ZP CSI-RS	Zero Power Channel State Information Reference Signals

1 Introduction

This deliverable D4.2 summarises the final results of ONE5G, work package WP4. The work was carried out from June 2017 to May 2019. The progress of the work corresponds to the work plan as set out in the project proposal.

The overall objective of WP4 is to propose and investigate new techniques and enhancements at lower layers (PHY/MAC) to address the needs for various services (eMBB, mMTC, URLLC), with focus on the two main ONE5G scenarios “Megacity” and “Underserved Areas”. The results are expected to have an impact on the evolution of 5G, either for future product development or 3GPP standardization. In particular, WP4 aims to improve the performance at a link (and multi-link) level, by developing advanced grant-free access schemes, large-scale antenna solutions (massive MIMO) and innovative technologies for CRAN/DRAN.

The document is structured as follows. The sections 2, 3 and 4 correspond to the tasks as described in our project proposal, respectively:

- T4.1 “Future Proof Multi-Service Access Solutions”,
- T4.2 “Massive MIMO Enablers towards Practical Implementation”, and
- T4.3 “Advanced Link Management Based on CRAN/DRAN, and massive MIMO”,

Each section is structured in subsections, which address particular technical challenges (topics) related to the 5G long term evolution. Different aspects of each topics are investigated jointly by several partners. At the end of each section we summarize the results and highlight the main benefits of each proposed technology.

Section 5 gives an overview on the main areas of innovation with a discussion on the expected impact.

Finally, Section 6 gives an overall conclusion and outlook on future work.

2 Future Proof Multi-Service Access Solutions

In response to the growing demand of multi-service mobile communication from vertical sectors such as Factories of the Future, Automotive, Smart Cities, Energy and others, the 5G ecosystem has introduced two new service classes, namely URLLC and mMTC [NGMN15]. URLLC services target highly reliable communication with very low latencies, whereas mMTC services supports the growing Internet-of-Things (IoT) applications characterized by a very large number of low cost devices operating with sporadic traffic over limited spectral resources

This section presents the studies carried out in ONE5G for the sake of addressing the challenges of URLLC and mMTC service classes from a physical layer and MAC perspective. Both, the "*Megacity*" and "*Underserved Areas*" scenarios [ONE17-D21] are addressed, though the emphasis is on the former.

The TeCs investigated in this task are grouped into three technology clusters. Section 2.1 addresses Non-Orthogonal Multiple Access (NOMA), which is designed to improve resource utilization by concurrently serving a large number of mMTC users over a limited bandwidth. In particular, NOMA schemes for uplink and downlink transmissions, receiver design for Grant Free (GF) NOMA access, performance comparison of various NOMA schemes and multiplexing of eMBB and mMTC services using NOMA are studied. Different enablers for URLLC, such as resource-efficient GF access, beamforming design and interference-aware frame design are presented in Section 2.2. Finally, Section 2.3 focuses on the reliability aspects of mMTC for sporadic transmission of short packets. The proposals include redesigning the packet structure, and reliable transmission schemes in a CRAN setting.

2.1 Non-Orthogonal Multiple Access

NOMA has gained high interest in academia and standardization [3GPP-181403] as an alternative to Orthogonal Multiple Access (OMA) for 5G-NR to support massive connectivity and increase the achievable throughput. The underlying principle behind NOMA is to loosen the paradigm of orthogonal transmissions by allowing different users (or layers) to concurrently share the same physical resources, in time, frequency and space. The 3GPP NOMA study item [3GPP-181403] has been completed in December 2018. However, the work has not continued because the discussion was still controversial and also other priorities prevailed.

NOMA techniques can be roughly categorized by the way the UE specific signatures are generated. Examples for signatures includes spreading codes, or interleaver sequences (concatenated with low-rate error-correcting codes). In addition, an optimized power allocation can be used to improve the performance of receiver algorithms such as, e.g., successive interference cancellation (SIC). The objectives of this topic within ONE5G is to tailor transmission procedures as well as specific processing steps at both transmitter and receiver side.

In Section 2.1.1, a link-level performance comparison of several NOMA candidates, in particular Non-Orthogonal Coded Access (NOCA) and Interleave Division Multiple Access (IDMA), is conducted. Both fall into the category of dense sequence-based NOMA. In contrast, Section 2.1.2 elaborates on signature designs for low-density spreading NOMA, where the information-bits are carried by specific designed sequences with high number of non-zero elements. A user grouping scheme for contention based uplink NOMA transmission with the goal of reusing preambles allocated to different groups is proposed in Section 2.1.3. NOMA in the context of grant-free access is investigated in Section 2.1.4, where specific NOMA techniques are exploited at the receiver side to resolve collisions in the contention based uplink. Finally, Section 2.1.5 studies multicarrier sequence-based NOMA schemes that allow to superpose eMBB and MTC services on the same resources.

2.1.1 Link Level Comparison of NOMA Solutions

Among the NOMA candidate solutions for both grant-based transmission and grant-free transmission with collision in 3GPP RAN1, the NOMA user signatures are characterized by user-specific spreading, scrambling, interleaving, and sparse signal structure. In this final deliverable, we specifically select four representative NOMA solutions to cover the major NOMA user signatures. They are Non-Orthogonal Coded Access (NOCA) [ZMZ+16], [WZM+17], [Che18a], [CW18], [ACC+18] and Interleave Division Multiple Access (IDMA) [PLW+06], [CW18], [Che18b], which are user-specific spreading and interleaving based solutions, respectively. Furthermore, Resource Spread Multiple Access (RSMA) [CSS+17] exploiting user-specific scrambling and Sparse Code Multiple Access (SCMA) [NB13] exploiting the sparse signal structure are selected for link level comparison. Overloading, which is the requirement of NOMA system, here simply refers to the ratio between user number and information rate reduction, introduced by spreading, repetition or sparse signal structure [CW18]. In Figure 2-1, the generic transmitters of the selected NOMA systems are presented.

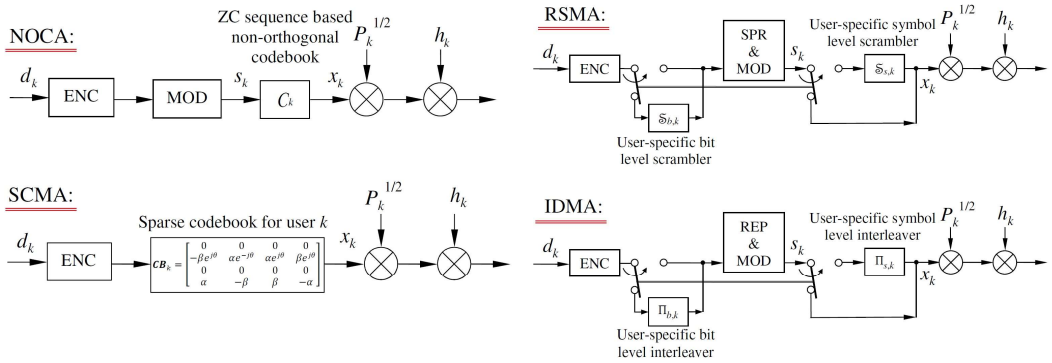


Figure 2-1: The generic transmitters of the selected NOMA systems.

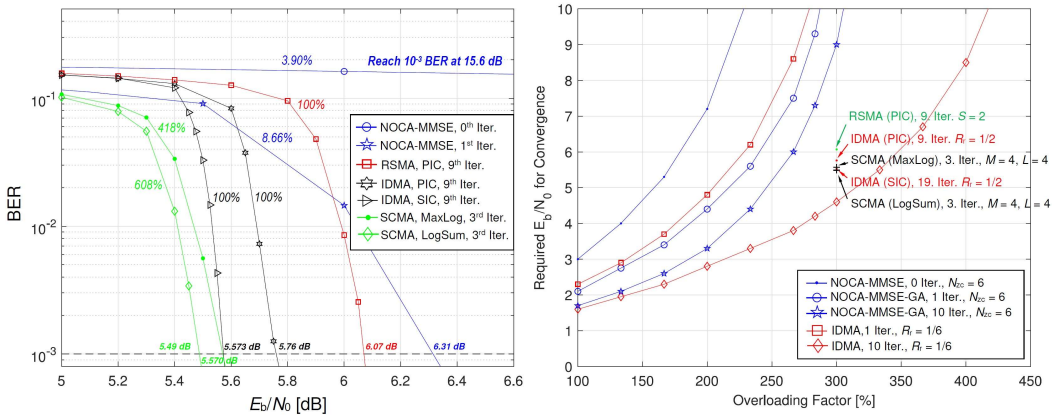


Figure 2-2: Link level comparison of NOCA and NOMA.

In Figure 2-2, the BER performances of NOCA, RSMA, IDMA and SCMA are presented on the left-hand side. The overloading factor of NOCA is 250%, and the overloading factor of the other schemes is 300%, because only the simulations for NOCA consider user signature collision. Further, the computation complexity IDs, by accumulating and quantizing the number of numerical operations [CW18], are also plotted on each curve, by assuming that the complexity of IDMA is 100%. On the right-hand side, the overloading characteristic of the selected NOMA schemes is illustrated. We make the following observations.

- NOCA is a very flexible approach with low and scalable complexity. At relatively low overloading, NOCA delivers similar performance as other candidates. The NOCA

simulations enable completely random user signature selection, while the other NOMA solutions are supported by centralized scheduling. Hence, NOCA is very appropriate for overloaded mMTC.

- RSMA might be the most compliant scheme to the current standard, due to the well adopted scrambling mechanism, although RSMA is slightly outperformed by SCMA and IDMA.
- SCMA turns out to be a powerful scheme to deliver good performance. High complexity and relatively low flexibility to scale the complexity make the degrees of freedom relatively low for SCMA to trade-off the performance and complexity.
- IDMA is another flexible approach to provide scalable performance. The detection philosophy of Gaussian approximation follows the natural probabilistic approaches, which makes IDMA outperform many other NOMA solutions, if considering same complexity. Regarding to the practical implementation, efficient user-specific interleaving generation mechanism is necessary.

For future work, investigations should evaluate and compare MU-MIMO and MIMO-NOMA for eMBB, URLLC and mMTC traffic, in order to exhibit their advantages, in terms of different parameter settings. Some initial results can be found in [Che19].

The presented contribution is linked to the 5G-NR work on Evaluation methodology.

2.1.2 Regular Signature Design for Low-Density Spreading NOMA

Low-density code-domain (LDCD) - NOMA is a prominent sub-category of signature-based multiplexing, which relies on low-density signatures (LDS). In general, sparse codes comprising a small number of non-zero elements are employed for modulating each user symbol over shared physical resources. With sparse signatures, significant receiver complexity reduction can be achieved by utilizing belief propagation (BP), i.e. message passing algorithm (MPA), which enables user separation with moderate complexity even when the received powers are comparable. One critical design parameter for LDCD NOMA system with K users sharing N resources in a non-orthogonal fashion is the sparsity of the signatures, i.e. the number of non-zero entries in a binary $N \times K$ signature matrix **F**. The sparse code design can be either regular, where each user occupies a fixed number of resources, and each resource is used by a fixed number of users; or irregular, where the respective numbers are only fixed on average. However, only some results exist in literature considering the general structure of the signature matrix, i.e. specific properties of, e.g. number- and position of non-zero entries within **F**. In [SZS17], a *closed-form analytical expression* is derived for the limiting empirical spectral distribution of a spreading (signature) matrix with *regular random user signatures*. As a result, a *regular* spreading is found to be superior in terms of sum-spectral efficiency compared with irregular user-resource allocation as well as compared to dense randomly-spread NOMA.

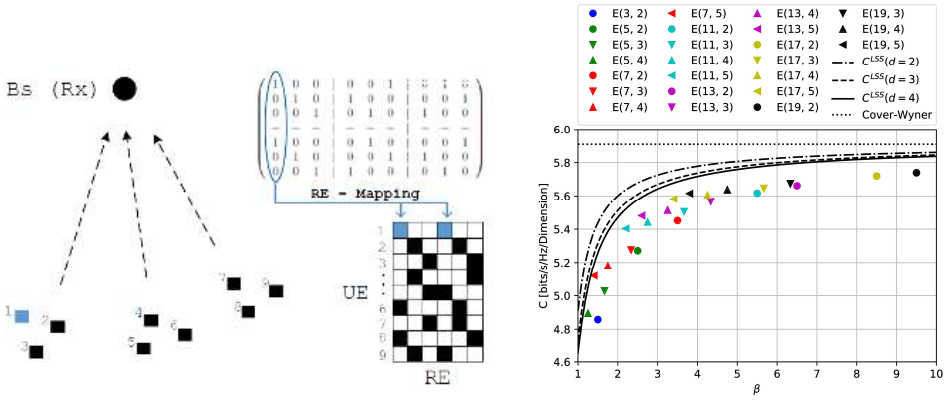


Figure 2-3: System Model [left] and Spectral Efficiency per Dimension for selected code constructions versus overload (β) [right].

The proposed construction presents a sub-class of the special type of partial geometries with quasi-cyclic structure, as used for LDPC. The properties of this graph are closely related to the performance of multi-user decoding at the receiver, i.e. based on Message Passing. Further, the proposed construction can generate signature matrices fulfilling the regular condition for a wide range of numerologies, i.e. number of resources, number of users, overloading, number of non-zero elements per signature. A numerical example is depicted in Figure 2-4, where in total 25 usrs, each transmitting packets of 250 bits in total 12.500 channel uses (CUs). The proposed construction allows all user similarly to transmit its data at different overload regimes, with a trade-off between signature lengths (in this example bandwidth) and overload (ratio of number of supported users to the signature length).
Hence, compared to orthogonal multiplexing, as considered for Rel.15, the proposed constructions support the trade-off between QoS requirements such as latency, reliability and spectral efficiency in a flexible manner (even in the overloaded case), with low-complexity receiver architecture. This makes it appropriate for both unscheduled transmissions targeting mMTC scenarios, as well as scheduled transmission targeting eMBB and URLLC scenarios. The joint effects of sparse signature design and forward-error correction in finite-size systems is characterized by the interplay between the system parameters such as sparsity, system load, and channel coding rate. Further, regular sparse construction gives signatures with small density, support decoding algorithms with low computational complexity [DUT+19] and can be combined with other (sparse) NOMA schemes like SCMA.
This contribution can be linked to the 5G-NR work on Transmitter-side processing and NOMA related procedures.

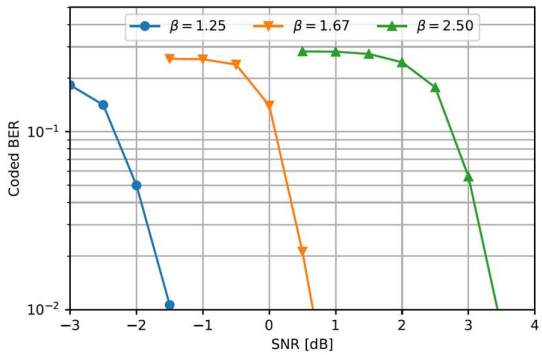


Figure 2-4: Performance comparison of 25 users (250 bits) with MPA / LDPC [DUT+19] for different overload factors and fixed number of channel uses.

2.1.3 Contention Based Uplink NOMA Transmission

An increased number of collisions at the Random Access Channel (RACH) incurs when numerous devices attempt to establish a connection simultaneously can result in increased network access delays and network outages, especially when the system is under heavy traffic load. Previous work has proposed the concept of NOMA at the RACH, by using a Successive Interference Cancellation (SIC) based Non-Orthogonal Random Access (NORA) [LLZ+17] scheme. NORA allows UEs to connect to the network, even when a collision occurs, provided that their transmissions are received by the BS with significant time difference. Such a NORA concept has been investigated in previous ONE5G reports [ONE18-D41], wherein the benefit of NORA in terms of reducing the probability of collision has been studied through theoretical analysis, when UEs are clustered according to their power differences.

The work is concerned with placing the UEs in different logical zones at the initial access stage, so that preambles can be reused among zones, if their power difference is significant to separate colliding preambles. The different zones can be defined in different domains (e.g. time domain, power domain), or in combination of multiple domains. Initially, the BS calculates minimum separation distance d_{sep} in the chosen domain, which can be experimentally updated each time

multiple UEs collide. The cell area is then split in different zones based on d_{sep}, and preambles can be re-used among zones that are further apart than d_{sep}, see Figure 2-5. The limits of the zones are broadcasted in SIB2, and devices are required to decode it before attempting to connect in order to determine the logical zone they belong to.

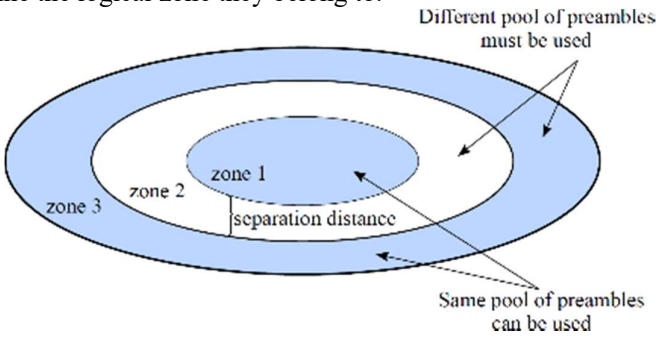


Figure 2-5: Illustration of a cell with 3 zones, where preambles can be reused among them

To further reduce the collisions within the same zone, we propose a probabilistic preamble selection scheme based on Reinforcement learning [TWW+19]. Essentially, the BS observes the preamble usage of the immediate past (observing window) and generates Preamble Usage Reports (PURs) for each zone. UEs then select their preamble probabilistically based on the PUR of their zone. The length of time in the past that the BS observes the preamble usage is dynamically updated using a Reinforcement Learning approach in order to decrease the RACH collisions.

We performed several experiments with different number of devices, and the Uniform and Beta distributions that specify when the devices must start transmitting their data [3GPP-37.868]. Specifically, devices must start transmitting their data within the first 60 and 10 seconds with the Uniform and Beta distributions respectively. In our experiments we compared our proposed approach against the currently used scheme in Rel. 15 that only employs SIC-based NORA which we consider as our baseline, our proposal with the best static observing, and our proposal with reinforcement learning. Our results (Figure 2-6) show that our approach outperforms both the currently used schemes, as well as the approach that uses a static observing window.

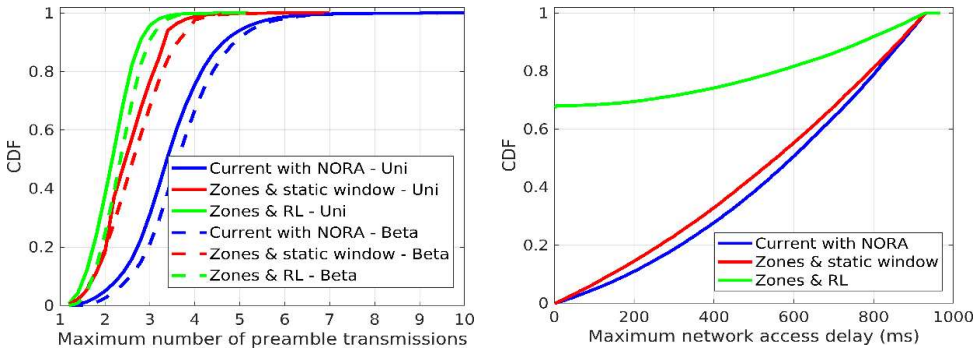


Figure 2-6: (left) CDF of the maximum number of preamble transmissions required for establishing a connection, (right) CDF of the average network access delay

We can see (Figure 2-6) that our proposed approach requires a lower number of maximum preamble transmissions to establish a connection with both start-time distributions, compared to the currently used approach where no zones are implemented (blue lines), as well as the proposed approach when using a static observing window determined using a greedy algorithm (red lines). Specifically, our approach decreases the number of preamble transmissions by ~30%, compared to the currently used approach that simply implements NORA. The decreased number of preamble transmissions then results in significant decrease in the maximum network access delay, which is up to ~57% when 50000 devices are used.

To summarize, probabilistic preamble selection with reinforcement learning results in significantly lower number of collisions during the preamble transmission stage. That decreases the number of required preamble transmissions to establish a connection and decreases the network access delay. This also translates to an increased system load. This contribution can be linked to the 5G-NR work on Transmitter-side processing and NOMA related procedures.

2.1.4 Enhanced Grant-Free Access with Advanced Receiver

GF transmission is an enabling technology for 5G as it allows to reduce latency for URLLC (cf. Section 2.2) and signalling overhead in mMTC scenarios (cf. Section 2.3). The performance of classical contention based access protocols like slotted ALOHA is usually limited by packet collisions: if multiple UEs transmit in the same slot (corresponding to a set of time-frequency resources), the messages cannot be decoded in general. Therefore, several improved Coded Random Access (CRA) schemes have been proposed recently [PSL+15]. The basic idea consists in repeating packets in multiple slots, and resolving collisions through SIC of already decoded packets. The number of repetitions can be optimized for a given system load using tools from coding theory such as density evolution [Liv11]. Some further enhancements are possible by replacing the simple packet repetitions with more general packet erasure codes [PLC15]. Such CRA schemes can asymptotically approach a normalized throughput of 1 packet/slot for the collision channel model, the same as with orthogonal resource reservation.
A key assumption in most publications on CRA is that packets received without collision can always be correctly decoded, which is not guaranteed for transmission of short packets over mobile radio channels. On the other hand, treating packet collisions as erasures may be too pessimistic considering multi-user detection at the receiver. Several NOMA schemes proposed for 5G make use of sparse resource allocation patterns (see, e.g., [DWY+15] or [DWD+18] for an overview), which bears a close resemblance to CRA methods, where each active UE selects a

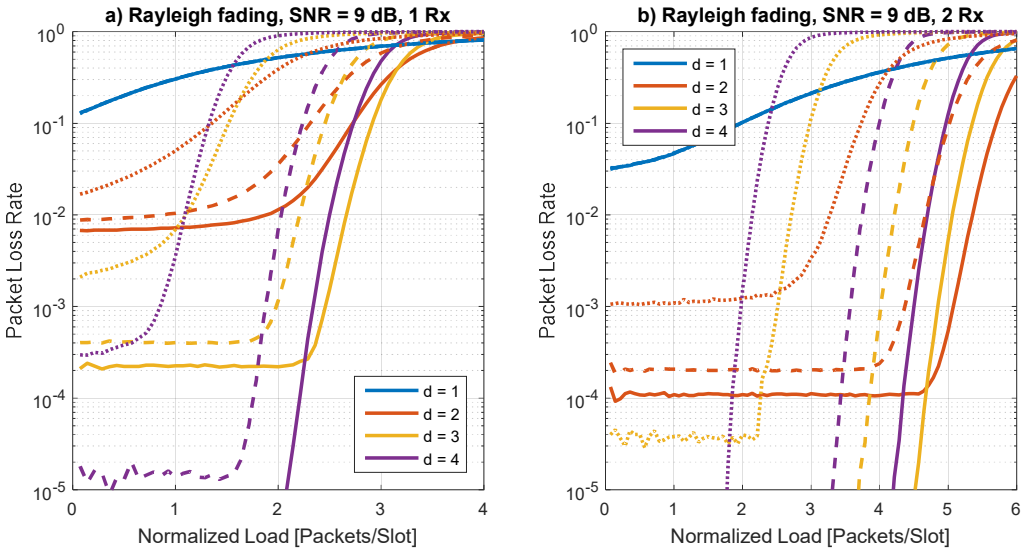


Figure 2-7: Comparison of grant-free access using *d* slots for each packet and SIC at the receiver with selection combining (dotted), Chase combining (dashed), and low-rate channel coding (solid).

subset of the available slots for transmission. However, NOMA additionally allows combining received signals from multiple slots and decode a message even if no slot is free from interference. Furthermore, near-far effects resulting from different path loss or fading coefficients can be exploited to improve the performance of SIC.

In order to illustrate the potential gains of NOMA with slot combining compared to CRA, we consider a scenario where each active UE selects d out of 14 slots, each comprising 240 resource elements in the time-frequency domain, to transmit a message consisting of 256 bits. The channels are Rayleigh block fading per slot, which is a suitable model, e.g., in combination

with frequency hopping. For simplicity, we use Polyanskiy's normal approximation [PPV10] to determine the packet error rate and assume that interference from successfully decoded packets can be perfectly cancelled at the receiver. The following transmit and receive strategies are compared in Figure 2-7:

- Selection combining (dotted): Packets are repeated in d slots, and the receiver selects the slots with the best SINR for decoding.
- Chase combining (dashed): Packets are repeated in d slots, and the receiver performs maximum ratio combining of all slots before decoding.
- Low-rate channel coding (solid): Packets are encoded over d slots, and the receiver considers all slots jointly for decoding.

Note that the special case $d = 1$ corresponds to slotted ALOHA. The first scheme based on packet repetitions with slot-wise decoding, which is representative for CRA, provides gains only for relatively low system loads with few active UEs and fails to achieve the strict reliability requirements of packet loss rates below 10^{-5} for URLLC with a single receive antenna in Figure 2-7a. On the other hand, NOMA with low-rate channel coding provides significant gains and achieves a target packet loss rate of 10^{-5} even for overloaded systems. As illustrated in Figure 2-7b, the maximum system load can be increased to more than 4 packets/slot using two receive antennas. In this case, the best performance is achieved for $d = 3$. Observe that the curves exhibit a threshold behaviour: The interference can be completely removed with high probability up to a certain system load, after which the packet loss rate sharply increases. Larger values of d result in a lower error floor, but also reduce the load threshold due to the additional interference. This clearly demonstrates the advantage of a sparse resource allocation for NOMA with SIC at the receiver. To conclude, block-wise sparse NOMA based on low-rate channel codes can more than double the supported system load compared to conventional coded random access schemes based on packet repetitions and slot-wise decoding. The results are published in a joint ONE5G paper [MAB+19]. This contribution can be linked to the 5G-NR work on Receiver-side processing and NOMA related procedures.

2.1.5 NOMA multiservice underlay communication

This section focuses on multicarrier sequence-based NOMA schemes that allow to superpose different services (eMBB and MTC) on the same resources. The main concept consists in superposing two sets of orthogonal waveforms, namely OFDMA as the first signal set and MC-CDMA as the second signal set, which turns out to be attractive for serving mMTC traffic in future 5G networks on top of eMBB services. Interestingly, the power imbalance required in power domain NOMA appears naturally as the MTC signals are spread in the frequency domain and have therefore a power much smaller than the power of OFDMA signals. The principle of such NOMA schemes is illustrated in Figure 2-8 which shows an OFDMA system with N subcarriers and a subcarrier spacing of $1/NT$ Hz, where NT is the OFDM symbol period. The second set of M users use spreading sequences of length N. For more details, see [CMK+18].

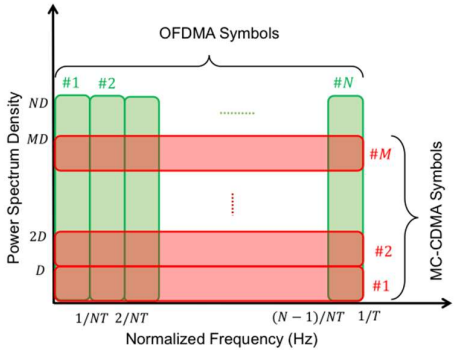


Figure 2-8: Principle of Frequency domain NOMA.

The performance of the proposed NOMA scheme with iterative successive interference cancelation (SIC) was evaluated on AWGN channels using 16-QAM modulation for the OFDMA users and QPSK for the MC-CDMA users. We present in Figure 2-9 the results of 5 IC iterations with N = 256 and M = 44 corresponding to an overload factor of 17.2%. We can see that the BER curves corresponding to the soft-decision detector converges rapidly to the theoretical interference-free BER curves and virtually coincides with it at BER values below 10^{-3} while a BER floor appears for the hard decisions detector at those values [MCK+18].

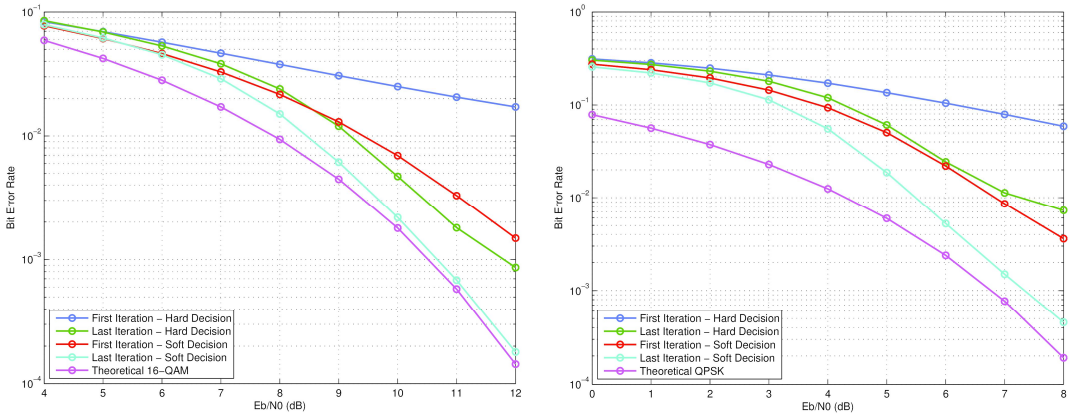


Figure 2-9: Iterative interference receiver simulations results.

To circumvent the complexity of the iterative receiver, we consider a variant of our NOMA scheme, which allows using ML detector and thus avoiding the use of complex interference cancellation receivers. We partition the set of N subcarriers into N/K groups of K subcarriers each and to limit the spreading of the MC-CDMA symbols to one group for K = 4 (an overload factor of 25%). ML detection is then applied at the receiver to find the set of OFDMA/MC-CDMA symbols combination that minimizes the cumulative squared error between the estimated and received signals. To further reduce the complexity, it is shown that the ML receiver, conditioned on the value of the MC-CDMA symbol, can be reduced to a threshold detection using appropriately shifted set of thresholds. One can refer to [CMK+18] for more details. For performance evaluation, the presented scheme was simulated using an AWGN channel. In the simulations, 16-QAM modulation was used for the OFDMA users, QPSK was used for the MC-CDMA users, and the spreading length was K = 4, as it was discussed previously. It can be seen that the BER results for users are very close to the theoretical interference free BER curves. The distance between the two curves is only 0.5 dB at the BER of 10^{-6} and the two curves must asymptotically (at vanishing BER values) coincide due to the equal minimum Euclidean distance. Consequently, the proposed multiservice NOMA scheme with ML detection allows superposing MTC and eMBB services on the same resources, by achieving a channel overload factor of 25%. Furthermore, the ML detection scheme presents itself as a way to achieve good performance at a much reduced complexity than the SIC receiver, which is widely used in NOMA schemes. Results also show that the proposed iterative interference cancellation approach reduces the SNR degradation virtually to zero, which is not possible with single-shot SIC receivers.
This contribution can be linked to the 5G-NR work on Transmitter-side processing and Receiver-side processing.

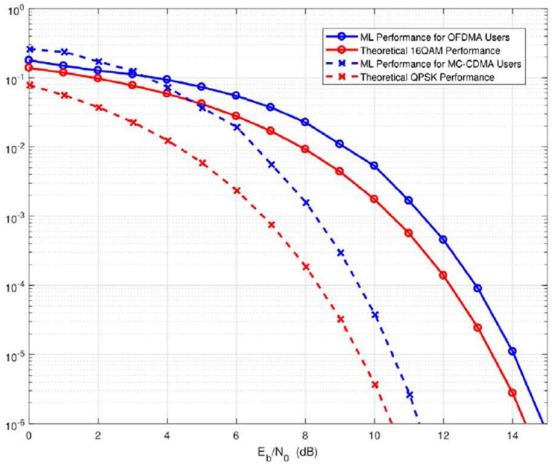


Figure 2-10: BER of the proposed scheme with ML receiver.

2.2 URLLC Enabled by GF Access, HARQ, and Frame Design

URLLC is a new service class introduced by 3GPP in 5G-NR Rel. 15, targeting highly reliable and very fast communication of typically short packets [3GPP-38.913]. In 5G-NR Rel. 15 the basic support for URLLC was introduced with TTI structures for low latency as well as methods for improved reliability. In 5G-NR Rel. 16 [3GPP-38.824] further use cases (factory automation, transport industry and electrical power distribution) with tighter requirements have been identified as important for NR evolution.

In 3GPP Rel-16, discussions on 5G-NR URLLC are grouped into different work items, as follows:

(1) Layer-1 enhancements, including potential improvement to Physical Downlink Control Channel (PDCCH), Uplink Control Information (UCI), Physical Uplink Shared Channel (PUSCH), and scheduling/HARQ/CSI processing timeline,
(2) Uplink inter-UE transmission prioritization and multiplexing, and
(3) Enhanced UL GF transmissions.

This section presents five technical solutions for URLLC, in agreement with the study directions of 3GPP Rel-16, While GF uplink transmissions allow for faster access to the medium – thereby reducing the latency, they might potentially affect the reliability as a result of collisions arising from unscheduled transmissions. Methods to improve the reliability of GF transmissions without compromising the latency advantages are studied in Subsection 2.2.1. Subsection 2.2.2 considers low latency MTC services and investigates the reliability of different preamble detection schemes when multiple BSs collaborate. The design of efficient transmission schemes tailored specifically towards URLLC services considering the DL transmission direction for BSs equipped with massive MIMO antennas serving users with single/few receive antennas is investigated in Subsection 2.2.3. Subsection 2.2.4 proposes a flexible numerology for URLLC through a bi-directional frame structure which duplexes the radio resources in both time and frequency. This allows additional freedom in resource allocation with the goal of reliability enhancement and latency reduction through flexibility in switching between uplink and downlink directions. Finally, enhanced HARQ techniques for URLLC applications are investigated in Subsection 2.2.5.

2.2.1 URLLC Uplink Grant Free Access

The grant free access enabled in 5G NR Rel-15 through the configured grant functionalities enables low latency transmission in the uplink. However, it comes with the cost of inefficient usage of resource, if dedicated resources are allocated per UE, or risk of reduced reliability in case the resources are shared by multiple UEs. The following studies includes enhacements for achieving GF reliable communication with improved resource utilization.

The contribution of such studies is linked to the 5G-NR Release 16 work on enhanced uplink grant-free transmissions.

Power control enhancements for uplink grant-free URLLC

Power control is a mechanism that allows to manage the levels of both intra- and inter-cell interference. Fractional power control is typically applied for high system throughput. However, for URLLC, the main target is to meet the strict requirement, such as, $1\text{-}10^{-5}$ transmission success probability within 1 ms. In this work, power control enhancements for grant-free URLLC are investigated. The target is to optimize power control settings considering URLLC performance indicators. In addition to the standard open loop power control available in NR Rel-15, it is also investigated whether applying a power boosting for retransmissions can improve the success rate within the considered latency constraint. System-level simulations are used to assess the performance, based on guidelines in 3GPP TR38.802.

The power control is given by: $P[dBm] = min\{P_{max}, P_0 + 10log_{10}(M) + \alpha PL + g(\Delta_{pb})\}$, where P_{max} is the maximum transmit power, P_0 is the target receive power per resource blocks (RB), M is the number of RBs, α is the fractional path loss compensation factor, PL is the path loss and $g(\Delta_{pb})=PB_{step}\cdot\Delta_{pb}$ gives a power boosting step for each retransmission with index Δ_{pb}.

The evaluation shows that full path loss compensation has better outage performance and less sensitivity to P_0 setting than fractional path loss compensation. This is because $\alpha<1$ incurs a penalty to the cell edge URLLC UEs. The value of P_0 should be optimized for URLLC to cope with interference and fading margins. Figure 2-11 shows a summary of the achived URLLC load according the evaluated power control configurations. The outage capacity gain when power boosting retransmissions are applied is approximately 20%. Higher gains should be achieved in scenarios where UEs are not power limited. It is important to mention that the error rate of the first transmission should be low (such as, 10^{-3}), hence avoiding excessive interference from boosting retransmissions. More details are available in [AJB+18], [JAB+17].

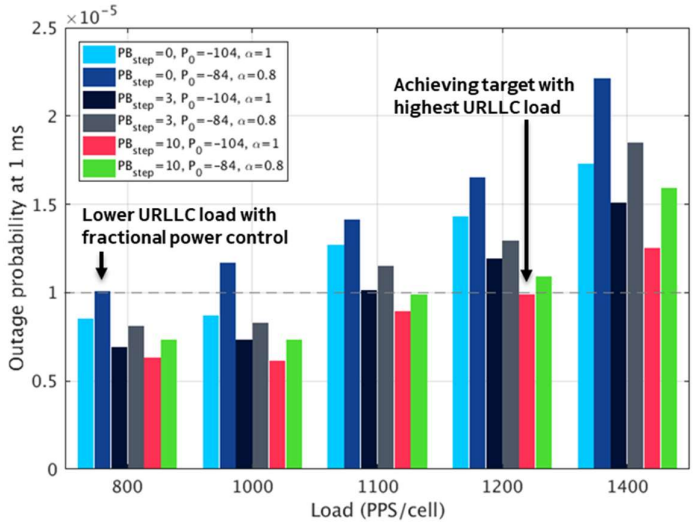


Figure 2-11: Achieved URLLC load in packets per second (PPS) for different power control settings (3D UMa, 32 B packet, 10 MHz, MMSE-IRC with 2 antennas, 143 µs mini-slot [AJB+18]).

Resource configuration for uplink grant-free URLLC

For URLLC with sporadic traffic, the grant-free resources and transmission parameters are assigned to multiple users through RRC signalling. The achievable URLLC load using grant-free transmissions in shared resources is limited by intra- and inter-cell interference. This part of the work describes a Radio Resource Management (RRM) solution which encompasses multiple grant-free allocations associated with Modulation Coding Scheme (MCS) and power control settings, and a selection method.

Users in favourable channel conditions, e.g. in cell centre, have the potential to apply configurations using lower transmission bandwidth, with higher MCS and compensate with higher power spectrum density. This reduces the overlapping with transmissions from users in worse channel condition, as in cell edge, which tend to operate using maximum transmit power. Multiple configurations can be provided to the UEs by RRC and the switching between configurations can be signalled through PDCCH.

The performance evaluation was conducted through system level simulations considering an urban macro scenario. The solution using two grant-free configurations, with MCS QPSK1/8 and QPSK1/2, is compared with a semi-static solution using single grant-free configuration, with MCS QPSK1/8. The results show that a gain of ~90% in URLLC achievable load can be obtained with the multiple grant-free configurations and selection method. Further details on power control settings and MCS selection criteria are provided in [JAB+18].

Blind retransmission over shared pre-scheduled resources

5G-NR supports blind repetitions defined by the *repK* parameter of the configured grant [3GPP-38.321]. With this, the UE proactively retransmits a packet replica, instead of waiting for feedback following each transmission attempt, which is subject to delay and transmission errors. The improved latency and reliability with blind repetitions comes at a cost of poor resource efficiency, since a repetition is issued regardless of the status of the previous transmission attempt. In short, blind repetitions drain channel capacity, whereas resource efficient stop-and-wait protocols lead to higher latency [BMA+18].

Herein, we propose a scheme in which blind repetitions are performed with improved resource efficiency and low delay penalty. As in coded-random access schemes [PSL+15], we consider SIC to remove the interference of decoded replicas. However, we also consider a set of dedicated resources for the initial transmission to reduce the latency of the iterative process. A group of N UEs are configured by the BS with dedicated resources for their initial transmissions. The BS also configures a resource pool of size M shared by these UEs to perform blind repetitions. In total a UE can perform T transmission attempts using its dedicated and the shared resources. At each retransmission the UE can select (randomly or according to a sequence) among the M resources. The BS attempts to decode each transmission over the configured resources and store the successful ones. The decoded ones are reconstructed and subtracted from the shared pool. The initial transmissions on dedicated resources have a high success probability. Therefore, the probability that multiple non-decoded replicas collide in the shared pool is low; thus, facilitating the SIC process.

The study shows that the proposed scheme is up to 23% more resource efficient than single shot transmissions for $N \geq 10$ UEs sharing the retransmission resources as illustrated in Figure 2-12. And it leads to ~57% lower latency compared with feedback-based retransmissions [ABJ+18].

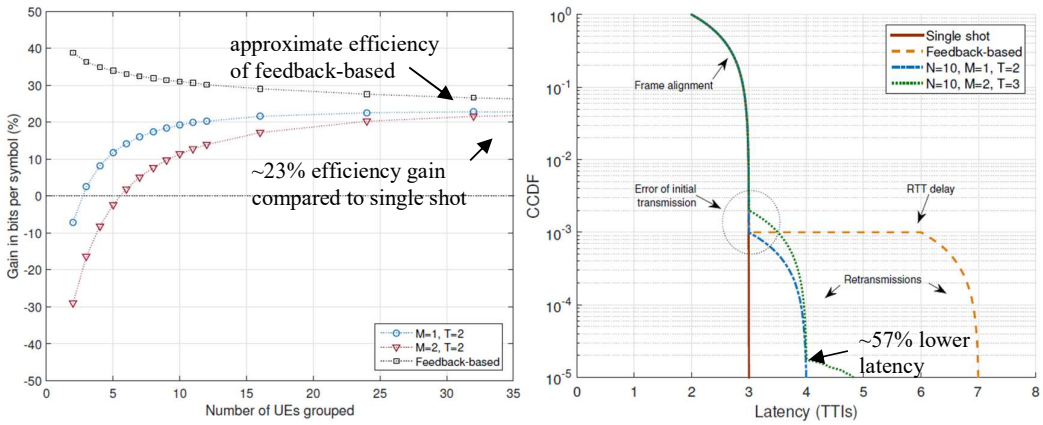


Figure 2-12: Resource efficiency compared to single shot (left), and latency comparison (right).

Multiplexing of grant-free URLLC and eMBB in uplink

5G should support heterogeneous services which have different characteristics in terms of traffic type and QoS requirements. URLLC services are usually characterized by sporadic small packets which should be transmitted with strict latency and reliability requirements. eMBB, on the other hand, demands high throughout transmissions for high data volumes. Efficient multiplexing methods are required for maximizing the capacity for these services togheter.

In this study we evaluate strategies for multiplexing grant-free uplink URLLC and eMBB traffic. The BS does not know a priori when a grant-free transmission will occur, so two options are considered. It either allocates separate resources/bands for URLLC and eMBB to avoid their mutual interference, as supported in NR Rel-15. Or it allows both traffic to overlay in the available band and exploit the receiver capability for decoding simultaneous transmissions. MMSE receivers are considered with and without SIC for removing URLLC interference over eMBB (URLLC is decoded first).

An analytical approach is presented for evaluating the supported load depending on the allocation strategy for different operation regimes. The evaluation is conducted considering NR assumptions. The results show that, in high SNR and using MMSE with SIC, the highest URLLC and eMBB loads can be achieved using overlaying allocation. Otherwise, the use of separate bands for each traffic should be preferred. Further details are available in [AJB+19]. And in [AJK+19], the impact of eMBB overlaying with URLLC is evaluated in system level.

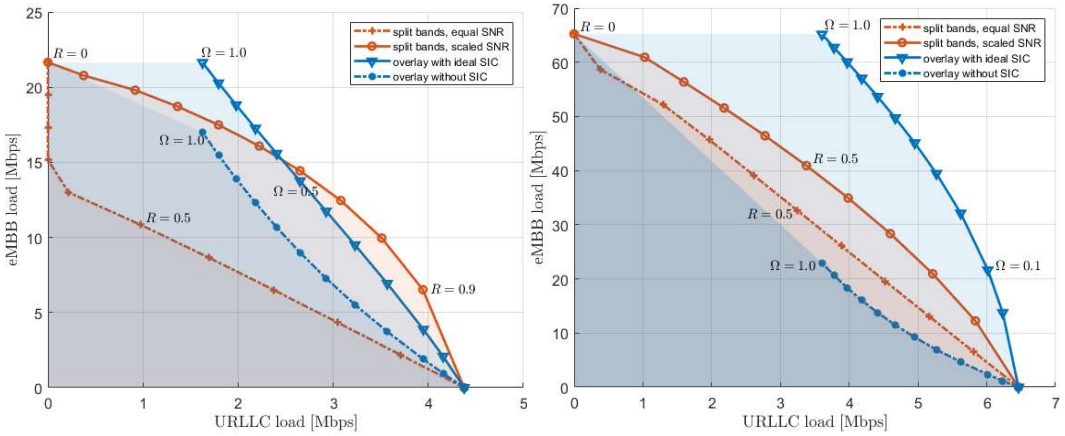
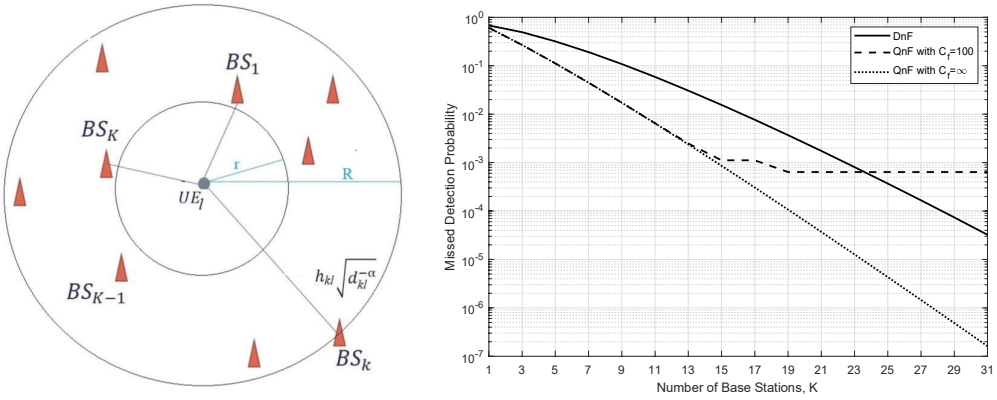


Figure 2-13: Achievable loads for URLLC and eMBB for different allocation strategies, in SNR of 0 dB (left), and SNR of 10 dB (right) over 10 MHz of bandwidth. R is bandwith split between URLLC and eMBB, Ω is the average receive power from eMBB user over the one from URLLC user. MMSE with 4-antenna, 50 URLLC UEs generating packets of 32B, and 2 full buffer eMBB UEs.

2.2.2 Preamble Detection using Multiple Base Stations

For the uplink access procedure in LTE where non-orthogonal preamble sequences are used, it was proposed to improve the user activity detection by employing massive MIMO [LY18] or having multiple BSs in a CRAN systems with limited-capacity fronthaul [USDP17]. The basic idea is to utilize the spatial diversity at the receiver side to improve preamble detection probability. Inspired by this idea, we consider a contention-based, narrow-band random access for low latency MTC use case and investigate different preamble detection schemes when multiple BSs collaborate. Specifically, the considered technique allows to exploit the additional information from multiple bases stations to improve the reliability of preamble detection, compared to local detection. Furthermore, the developed model can be used to optimize the length of the used preamble sequence, so as to minimize latency. The considered system model is illustrated in Figure 2-14 a). We compare two baseline preamble detection schemes, DnF and QnF [USDP17]. The DnF scheme does local detection and uses 1-bit result feedback from which the BS applies majority voting, whereas the QnF scheme quantizes the received signal and does centralized detection in a gNB Central Unit (CU). We consider their detection performance when there is limited feedback capacity between the BSs and the CU. We extend the work in [USDP17] by considering the trade-off between number of collaborating base stations and number of quantization bits. A detailed description of the work is provided in [TWK+19]. The following numerical results illustrate the relative merits of the schemes.



a) System model **b) Missed detection performance**

Figure 2-14: a) Base stations in between distance *r* and *R* collaborate in detecting preamble sent from UE; b) Missed detection probability versus total number of BSs K, for different values of feedback link capacity C_f.

The results in Figure 2-14 b) show the missed detection probability of the two schemes versus the number of collaborating BSs, where the baseline is K=1. With an increasing number of BSs, the feedback link carries more information until it becomes saturated. It can be seen that when the capacity of the feedback link C_f is sufficiently large, i.e. when the number of BSs K≤6 for C_f=30, K≤8 for C_f=40 and when $C_f \rightarrow \infty$, the QnF scheme outperforms the DnF scheme. In these cases, the QnF detection algorithm is able to fully take advantage of the observations from multiple BSs, whereas the DnF algorithm does local detection.

However, for limited-capacity backhaul, the QnF scheme saturates the feedback channel at some point, which prevents the use of additional BSs. This results in the detection performance of QnF exhibiting an error floor, as K increases and eventually the DnF scheme outperforms the QnF. This is because a reduced number of quantization levels and the cumulated quantization error significantly degrade the detection performance of QnF, while DnF benefits from local detection at the BSs without loss of feedback information. Other results of this work in [TWK+19] show that the QnF scheme is able to achieve the same missed detection probability as DnF with lower transmit power, as long as the feedback link has sufficient capacity.

To conclude, we have shown that centralized preamble detection schemes can be exploited to improve detection probability manyfold, by exploiting multiple base stations for detection. The Quantize-and-Forward scheme is preferable over the Detect-and-Forward scheme in deployments where a number of BSs are at approximately the same distance to the UE so that the same quantization steps are applicable for all BSs and as long as the backhaul capacity is sufficient. In our simulation study we found that three orders of magnitude in improved missed detection probability was achievable with the use of 20 and 25 base stations for the QnF and DnF detection schemes, respectively.

The contribution of this study can be linked to the 5G-NR Release 16 work on Layer-1 enhancements (improvement to scheduling/random access procedures).

2.2.3 Advanced Beamforming Designs to Enable New Services and Network Functionalities

This work item investigates transmitter design for enabling URLLC in Megacity scenarios (for V2X and factories), where large antenna arrays are assumed at the transmitter. This work is in line with the WP4 objective of "optimizing control signalling overhead for high reliability and reduced latency, especially in the case of short packets where the ratio of control information versus data is high", studied in a massive MIMO context.

Massive antenna arrays at the transmitter provide several advantages tailored to URLLC, such as: high quasi-deterministic SNR links, which are quasi-immune to fading; and high spatial multiplexing capability. The benefits of massive antenna systems are generally conditioned on the acquisition of the CSI [PNS+18]. In a mobile environment constrained by channel coherence time, and extreme latency requirements, acquiring instantaneous CSI becomes a severe limitation.

For this reason, we investigate beamforming schemes that are robust to uncertainties in the estimation of the instantaneous CSI. Those schemes rely partially on long-term statistics of the channel, here the Singular Vectors (SV) of the covariance matrix at the transmitter. As the performance of pure statistical beamforming is insufficient to ensure URLLC, the knowledge of the singular vectors must be complemented by instantaneous channel estimation. We estimate the projection of the channel into the singular vectors, allowing to coherently combine multiple singular vectors to send a stream of data.

Firstly, we consider a single terminal and compare maximum ratio transmission with transmission along all of the singular vectors. We consider a TDD system where the channel is estimated based on uplink symbols. The channel estimate is then used to form the beam for downlink transmission. In the precoding using singular vectors, we project the channel estimate into the subspace spanned by the singular vectors of the covariance matrix and perform maximum ratio transmission based on the projected channel estimate. The effect of the projection is to remove the estimation errors that are outside of the subspace, which brings a significant improvement in the packet error rate. For more details, the reader can refer to [BXD+18], and for several other beamforming methods and procedures tailored to URLLC to [PSN+18].
In Figure 2-15a), we show the outage probability for both schemes in terms of the training length. For a fixed packet length of 28 symbols, we have a varying training length which reduces the effective number of symbols that should carry 256 bits, but at the same time increases the instantaneous estimation accuracy. The outage probability is the probability that the post-processing SNR is smaller than a threshold (computed as the minimal information-theoretic SNR to transmit 256 bits error-free). The number of transmit antennas is fixed to 100. The SNR per link is 4dB. The channel is the sum of 20 paths, each defining a fading coefficient. All the paths have the same average power. We notice first that there is an optimal number of training symbols that minimize the outage probability. Furthermore, transmission along the singular vectors (all SVs) provides considerable benefit, outperforming significantly MRT by up to 2 orders of magnitude in terms of outage probability. Figure 2-15b) compares the

performance obtained when multiplexing two users, either in time (TDMA-allSVs) or spatially using zero-forcing (ZF). It can be seen that, despite the considerable performance increase of the scheme relying on all SVs compared to MRT, TDMA is surpassed by spatial multiplexing with ZF. This is due to the fact that in TDMA, one user would experience lower latency, whereas the second, a higher latency, therefore dictating the total latency shown in the plot. In summary, utilizing the channel structure when performing coherent beamforming provides higher reliability and lower latency in the single user case. For multi-user, zero-forcing outperforms TDMA with coherent BF and second order statistics. The contribution of such studies can be linked to the 5G-NR Release 16 work on Layer-1 enhancements – improvement CSI processing timeline.

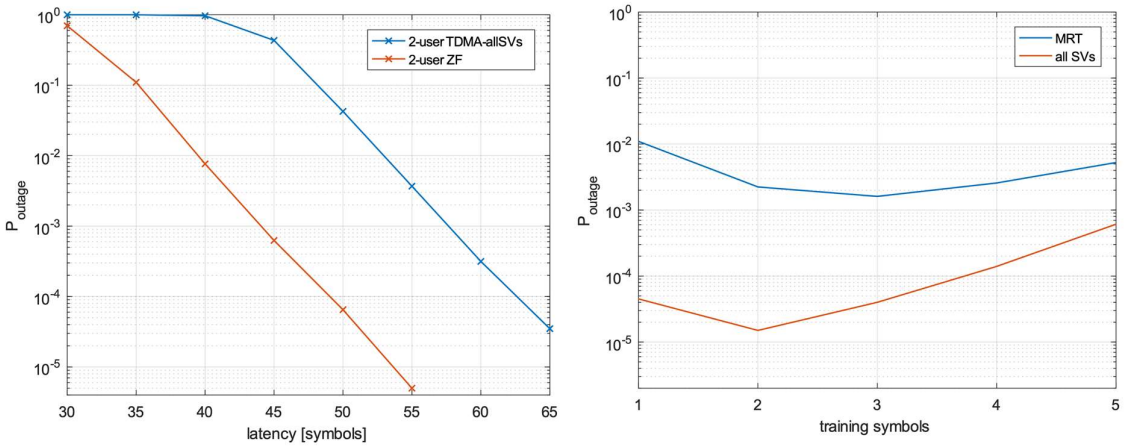


Figure 2-15: a) Outage vs. training duration; b) Outage vs latency with optimal training length depending on the latency constraint; with M=100 BS antennas and p=20 SVs.

2.2.4 Interference Mitigation for Bi-Directional URLLC

One of the key enablers of achieving URLLC is to use flexible numerology, which can be tuned to maximize the reliability of the wireless link. Also, in order to achieve extreme low latency, we want to be able to switch between uplink and downlink directions in a flexible manner.

A major step towards a more flexible numerology is adjacent-channel full duplex, employing bi-directional frame structure which duplexes the radio resources in both time and frequency, as shown in Figure 2-16. With this scheme, we gain an extra degree of freedom compared to classical FDD and TDD. Under low latency constraints, large variations in the *bi-directional* data rates can be supported using time and frequency duplexing with flexible numerology.

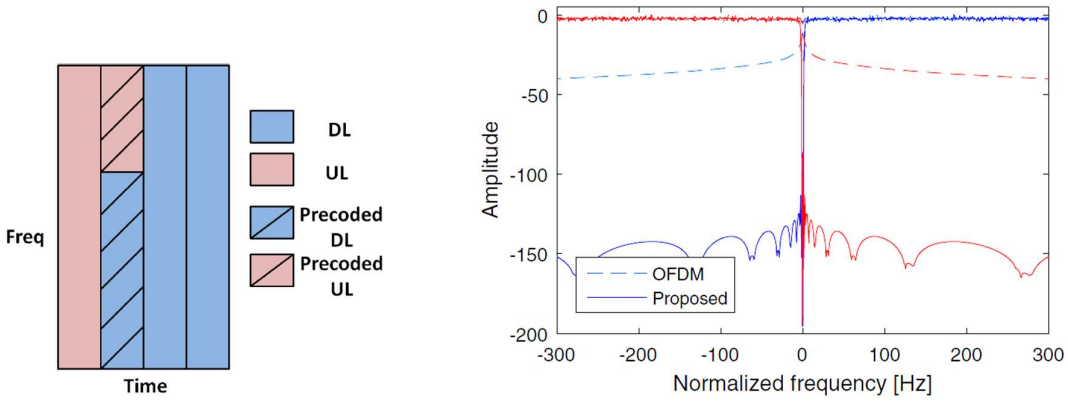


Figure 2-16: Adjacent services utilizing different numerologies.

To maintain high spectral efficiency, the bi-directional traffic flows are packed close to each other in the frequency domain, which leads to considerable amount of crosstalk. Instead of using guard bands between adjacent subbands, we introduce a novel precoder to mitigate the crosstalk interference in baseband. Using the precoder, the Out-of-Band (OOB) emissions are suppressed by around 100 dB compared to the non precoded solution, as shown in Figure 2-16. The precoder is derived from a static interference matrix, hence does not need to be updated regularly. The results are described in [ONE18-D41] and published in [IX17].

The contribution of this study can be linked to the 5G-NR Release 16 work on UL inter-UE transmission prioritization and multiplexing.

2.2.5 HARQ Investigations regarding URLLC

In previously published work [BMS+18] we have shown that, in case of very low SINR, the models and the performance of Chase Combining (CC) and Incremental Redundancy (IR) HARQ were behaving in a seemingly atypical way. In particular, the gain of IR with respect to CC, was shown to disappear if very low SINR is encountered, where the combined SINR of packets requiring a 2nd retransmission is similar for both schemes. Note that in non-critical cases, i.e. packets requiring only 1 retransmission, IR has a combined SINR gain with respect to CC.

While with eMBB traffic the attention is typically focused on expected values and/or average performance, for URLLC traffic the systems must be designed to be robust to rare, but not impossible deep fading realizations (i.e. happening with probability 10^{-5}).

Therefore, we have analyzed the performance of different HARQ schemes, in conjunction with LDPC channel codes, following 3GPP's specifications in TS 38.212. We focused on the comparison between the two most employed HARQ schemes: Chase Combining (easier to implement, low complexity) with both Symbol Level Combining (SLC) of according modulation symbols and Bit Level Combining (BLC) of according soft bits, as well as Incremental Redundancy (better performance in terms of BLER): see Figure 2-17 for an implementation block scheme with the three considered HARQ techniques.

SLC and BLC have the same performance in terms of BLER when QPSK, i.e., modulation order Q=2, is used; at the first HARQ Retransmission they both deliver a 3 dB SINR gain. On the other hand, when Q > 2 SLC outperforms BLC (see in Figure 2-18 the SLC vs BLC SINR gain). The loss of performance of BLC in comparison to SLC is more pronounced at low code rates and increases over the subsequent HARQ rounds. Moreover, we have observed that BLC's loss of performance grows using a higher modulation order and bigger transport block size. According to our model based on ML estimation, SLC needs to compute an estimate of the transmitted symbol based on the received symbols and the channel realizations. On the other hand, BLC simply sums up the received LLR corresponding to the same coded bits, leading to lower number of required operations. In terms of memory requirement, SLC and BLC need the same number of units for storing the received signal when Q = 2. When higher modulation orders are used, the memory requirement of BLC is larger than the SLC's one. In conclusion, when a QPSK modulation is used, BLC is the preferable combining scheme for CC, since it has the same performance in terms of BLER and memory requirement w.r.t. SLC but lower computational complexity. On the other hand, for higher modulation orders SLC brings significant benefits in terms of performance w.r.t. BLC.

Figure 2-19 shows the gain of IR vs CC. The most relevant parameter is the code rate R. The mother code rate gives an indication of the minimum selectable code rate in order not to incur in any repetition; in our simulations, we used BG2 of 3GPP TS 38.212, with mother code rate 0.2. We have also proved that IR's gain w.r.t. CC increases with the modulation order: we have obtained significantly higher gains using 64-QAM and 16-QAM compared to QPSK. Furthermore, we have compared transmissions with different transport block sizes suitable for

URLLC (32 and 200 bytes): the results show that for smaller blocks of bits the gain of IR w.r.t. CC is higher.

Further results can be found in Appendix B.5, plotting BLER vs SNR for QPSK and 16-QAM with R=0.4 and describing a new method for the evaluation of URLLC link level performance in a realistic system environment.

Note that, with the current state-of-the-art 3GPP LDPC channel codes, there is no gain of IR w.r.t. CC for URLLC-type applications: the two HARQ schemes have equal performance for SINR values that are meaningful for URLLC. Since CC has easier implementation and lower computational complexity, it seems that there is not a big gain in enforcing IR with the current LDPC codes when one is interested in high reliability regions. With our analysis, we have understood that channel coding and HARQ are crucial procedures when it comes to reliability enhancement, thus, we suggest not to abstract them in system level simulations when dealing with the strict reliability requirements of URLLC. A new approach, which unifies system-level and link-level simulations, is needed in order to meaningfully design a system that can meet the reliability requirement of URLLC. The contribution of this study can be linked to the 5G-NR Release 16 work on Layer-1 enhancements – improvement to HARQ processing timeline.

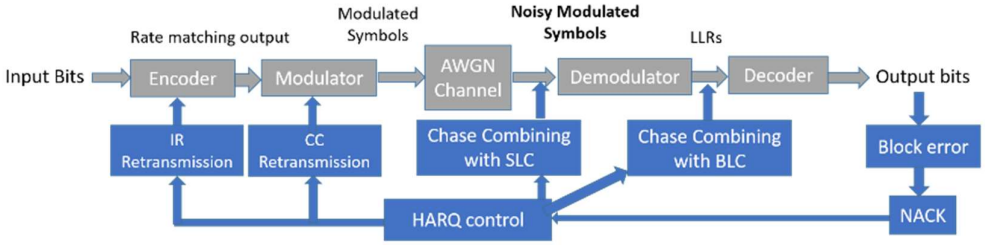


Figure 2-17: An overview of where HARQ manages operations with the three considered schemes, CC-BLC, CC-SLC, and IR.

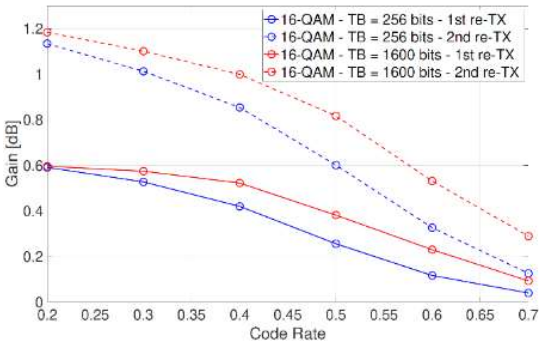


Figure 2-18: SLC vs BLC SINR Gain vs Code Rate.

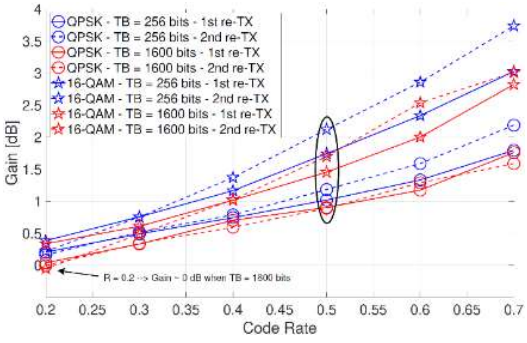


Figure 2-19: IR vs CC-SLC SINR Gains vs Code Rate.

2.3 Massive MTC Support of Reliable Short Packet Transmissions

Massive MTC is characterized by several essential aspects, including:

i) small size of the payload which leads to FBL effects (reduction of the maximum achievable rates with respect to Shannon rates);

ii) massive random access: though only a fraction of users is active at any given time, the number of instantaneously active users may exceed the overall message blocklength;

iii) uncoordinated access: users may access the channel without any prior resource requests to the network infrastructure in a GF fashion.

Most of the massive MTC research focuses on how to maximize the number of supported MTC connections. This section focuses instead on reliability aspects, i.e. on solutions for improving the reliability of short packet transmissions in massive MTC scenarios. In particular, the two solutions presented in this section face the problem of how to ensure successful transmission of short packets under latency and/or reliability constraints, in case of a large number of devices.

Short packet transmissions without grant are likely to access the channel by using a slotted ALOHA procedure. Multi-packet reception is foreseen as a major ingredient for improving capacity for slotted ALOHA. In the solution presented in section 2.3.1, the possibility of retrieving multiple simultaneous transmissions enables by multi-packet reception, has also the promise of a latency decrease and reliability improvement. In particular, the impact of different superslot sizes is investigated.Targeting reliable communication in massive connectivity scenarios with (very) short messages, the solution approach in Section 2.3.2 investigates the performance of a GF transmission scheme based on structured (sparse) superposition coding, in a practical CRAN with architectural constraint in the form of capacity-restricted fronthaul links. The proposed solution elucidates the merits of multi-connectivity in the uplink and local (RU) processing for the reliability performance under this architectural constraint. The proposal fits new conceptual approaches in 5G, i.e. application-specific RU-CU functional splits.

2.3.1 Short Packet Transmission with Reliability-Latency Constraints

This work investigates the transmission of short packets under latency-reliability constraints. The contribution focuses on the grant-free access (also known as configured grant in the latest 3GPP technical reports) scenario. The reservation-based access is considered to be a main contributor to latency, especially for short packet transmissions. Therefore, the throughput of grant-free access with reliability-latency constraints is investigated in this work item.

Multipacket reception (MPR) based on a frame slotted ALOHA (FSA) protocol is analysed in [GKS19] from a throughput under latency-reliability constraints perspective, in a scenario where the number of arrived users or only the arrival distribution are known.

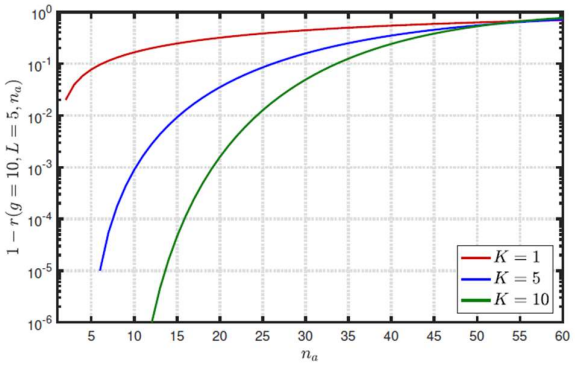


Figure 2-20: Reliability-latency of K-MPR frame slotted ALOHA with varying number of users on the X-axis, and varying K (superslot size) (Figure from [GKS19])

The reliability-latency performance of the K-MPR frame slotted ALOHA is shown in Figure 2-20. A K superslot is defined as the concatenation of multiple basic slots, and is dimensioned such that there are up to K simultaneous transmissions occurring, and all of them are successfully received. Increasing the superslot size K has been noticed to improve the reliability. For instance, in the case of 15 active users, increasing K from 5 to 10 provides roughly an improvent of two orders of magnitude in reliability. The limitation of the reliability when the number of users reaches toward 50 is given by the total resource grid size, which for this particular case is 5x10. The results also point out that the throughput can be improved if the knowledge of arriving packets can be accurately estimated. Moreover, even if there is a

considerable estimation error, when high reliability is targeted, the throughput performance still improves.

2.3.2 Reliable Schemes for Short-Packet-Transmission in Massive MTC

The contribution addresses a massive mMTC scenario where a large number of low-complexity devices communicate by transmitting sporadically short messages over shared (scarce) physical resources. We investigate the performance of a transmission scheme which integrates initial access and data transmission, enabled by a Bayesian receiver architecture which performs joint user activity detection and data decoding in a multiuser scenario. The proposed framework has several defining features: i) the message of each user is very short, such that there is no dedicated metadata to detect the activity. In fact, the metadata is extremely minimized and we do not assume that the user sends a dedicated ID, but it is identified based on the codebook that is allocated to the user; ii) no a priori channel state information (CSI) is assumed and the users do not invest resources to enable channel estimation at the receiver, such that the communication is genuinely noncoherent.

We address the challenge of massive connectivity in a CRAN scenario, characterized by a hierarchical structure in which part of the processing functionalities of the RUs are migrated to a centralized cloud processor or CU. In the CRAN scenario each user is served simultaneously by several RUs, and joint processing is performed at the CU. While the spatial diversity and the joint processing in CRANs have positive impact on the reliability, a major architectural constraint of CRANs is imposed by the capacity-limited fronthaul links connecting the RUs with the CU.

In this contribution, the performance of the transmission scheme for massive random access is investigated under this architectural constraint. In particular, a non-coherent, grant-free transmission scheme is proposed where the encoding is performed based on a Gabor frame structure. The transmission scheme is an instance of the structured superposition coding approach addressed in [ONE18-D41]. The fronthaul processing is a form of Detect and Forward (DtF), where local detection at the RUs is performed, and the local estimates are then compressed and merged at the central processor via the capacity-limited fronthaul links. In the particular example, the local detection can be performed with DFT/FFT operations and can be implemented on the low-PHY layer.

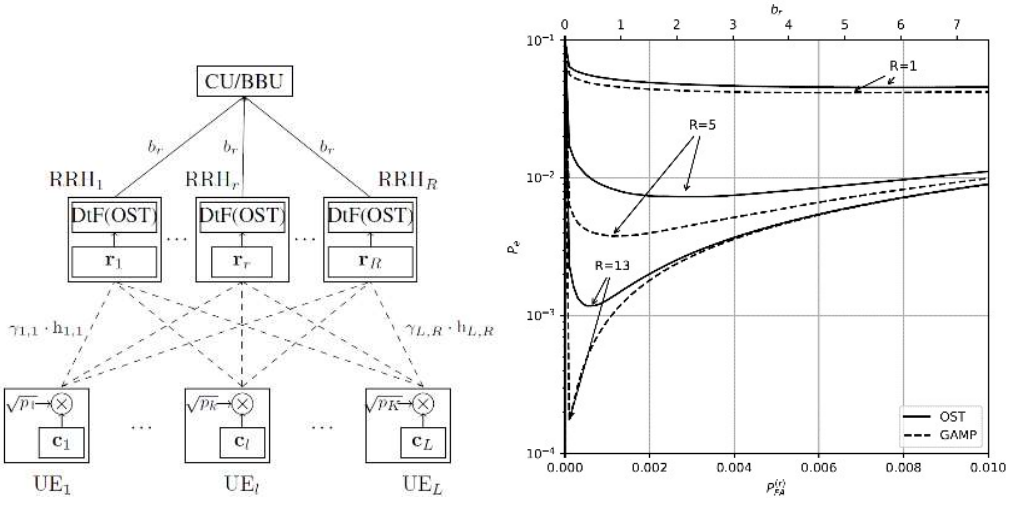


Figure 2-21: a) CRAN architecture with fronthaul limitations; b) Overall error performance with DtF fronthaul processing. The error probability, which accounts for both detection and decoding errors, is plotted as function of the probability of false alarm in the local (RU) detection step.

With DtF, the overall probability of error at the CU is a function of the local (RU) probability of false alarm, which is directly related to the fronthaul rate. Figure 2-21 depicts the overall error performance as function of the required fronthaul rate b_r and the local probability of false alarm P_{FA} in a system with 911 system users with activation probability $\rho = 0.1$ and per-user SNR=5dB. The number of RUs is a parameter. Two detection/decoding algorithms are compared: a simple One Step Thresholding (OST) scheme and a more sophisticated Bayesian inference algorithm based on Generalized Approximated Message Passing (GAMP) (see [AUP+18] for more details).The error probability can be reduced as the number of local radio units increases.

The proposed form of local detection (preprocessing at the RUs) may be interpreted as a form of functional split where only low phy-layer functionality (FFT-like processing) is performed at the RUs, and the remaining phy-layer (and also higher-layer) functions are implemented at the CU/BBU pool. This functional split, which is taylored to the random access mMTC scenario with sporadic transmissions, fits new conceptual approaches in 5G, where different functional splits based on the application are foreseen.

We note that the proposed approach could be particularly beneficial for industrial environments, where a potentially massive number of devices transmit sporadically short messages such as alarms or sensory measurements. In this context, the framework represents a rather extreme example of integration of control and data, suited for this scenario.

2.4 Conclusion

This task proposes various multi-service access solutions for 5G-NR and beyond. Investigations in this task are grouped into three technology clusters in line with study items for 3GPP Rel. 16, namely NOMA, URLLC solutions and mMTC support for sporadic transmission of short packets. The proposed transmitter and receiver side processing solutions for NOMA result in high normalized throughput and reduced access delays in congested conditions, while accommodating an overload factor greater than one. In particular, NOCA appears to be a flexible scheme with scalable complexity suited for overloaded scenarios, while solutions like RSMA, SCMA can achieve higher overloading factors at the cost of higher complexity. Reduced access delays can instead be achieved by preamble reuse among different coverage zone and reinforcement learning mechanisms for preamble selection. In terms of URLLC solutions, the proposed GF random access enhancements lead to an improvement in the supported URLLC load in the network, while fulfilling the targeted latency and reliability. Such enhancements include power control and resource configuration mechanisms, along with retransmission techniques. The value of using advanced receivers such as SIC and ML is assessed, also for overlay transmission of different service types. We further demonstrate that the performance of mMTC transmissions can be improved by redesigning the packet structure and introducing better preamble detection schemes. Moreover, a reliable transmission scheme for critical massive connectivity with CRAN architecture that can significantly reduce the error probability is also proposed.

Though most of the proposed solutions target the “Megacity” scenario in particular, many of the developed technology components are also applicable to “underserved areas”. For example, greater than one overload factor with NOMA schemes and GF access are found to be resource-efficient, especially for sporadic traffic scenarios – and hence are well suited for underserved areas. Similarly, the proposed eMBB and URLLC/mMTC multiplexing solution will allow serving eMBB users and deployed sensors (e.g., for farming applications) simultaneously.

Table 2-1. -Summary of key recommendations and benefits in terms of Future Proof Multi-Service Access Solutions

Feature	Recommendation	Benefits
2.1.1 Link Level Comparison of NOMA solutions	Both NOCA and IDMA are appropriate for mMTC type traffic. Both NOMA schemes allow relaxed scheduling and control.	NOCA is low-complexity. It supports random user-specific spreading codes selection and exhibits high robustness against user signature collision. IDMA is with increased but affordable complexity. It can support asynchronous communication. In comparison to 3GPP Rel-15, one of the most important benefits of NOCA and IDMA is that the supported user number can be 5 to 10 fold. NOCA can typically achieve 250% overloading and IDMA can achieve even higher overloading for asynchronous traffic.
2.1.2 Non-Orthogonal Multiple Access and Code Design	Low-density spreading NOMA with iterative near-optimal multiuser detection benefit from structure of the underlying factor graph. We propose a flexible (regular-sparse) code construction.	The proposed signature design allows to flexible trade different QoS requirements at high overload (more than 250% compared to 3GPP Rel 15) with low-complex receiver architectures.
2.1.3 Contention based Uplink NOMA transmission	SIC-based NOMA approached can benefit from splitting the coverage area in different zones and re-using the preambles among those zones. Further improvements can be achieved when RL is used for preamble selection.	Our scheme decreases the number of collisions in the RACH by ~ 30%, and the network access delay by ~ 57%, compared to the RA process with NORA of Rel 15.
2.1.4 Enhanced Grant-Free Access with Advanced Receiver	Grant-free access reduces the delay for URLLC and signalling overhead for mMTC. We propose to use a block-wise sparse NOMA scheme to mitigate the interference caused by packet collisions.	Block-wise sparse NOMA based on low-rate channel codes can more than double the supported system load compared to conventional coded random access schemes based on packet repetitions and slot-wise decoding.
2.1.5 NOMA multiservice underlay communication	Superposing eMBB and mMTC on the same resources can be performed by superposing two sets of orthogonal waveforms, namely OFDMA as the first signal set and MC-CDMA as the second signal set,	Our results show that the proposed multiservice NOMA scheme with ML detection allows superposing MTC and eMBB services on the same resources, by achieving a channel overload factor of 25%. This helps increasing the number of served MTC devices by 25%.
2.2.1 URLLC Uplink Grant Free Access	RRM principles for GF: 1) GF URLLC should be aided by mini-slot repetitions and HARQ with short RTT. 2) For periodic traffic, dedicated resources can be used for initial transmission and shared resources for repetitions, aided by SIC. 3) To improve outage capacity for sporadic URLLC, use full pathloss compensation, optimized P0 and robust MCS adapted based on coupling gain. 4) GF URLLC and eMBB can use	The proposed grant-free design enables URLLC with improved resource utilization compared with Rel-15. For deterministic traffic, the shared retransmission scheme leads to 23% improvement in resource efficiency. Power boosting retransmission allows at least 20% higher outage capacity in UMa. The use of multiple GF configurations with multiple MCS shows ~90% higher achievable load for URLLC. And the multiplexing of eMBB and

	overlaying allocations when employing MMSE+SIC and for low URLLC load, while separate resources should be used for stricter URLLC requirements.	URLLC using overlaying allocation allows to reach almost 100% resource utilization compared with ~35% if the bandwith part is only used for sporadic URLLC.
2.2.2 Preamble Detection using Multiple Base Stations	Centralized preamble detection schemes can be exploited to improve detection performance. The Quantize-and-Forward scheme is preferable over the Detect-and-Forward scheme in deployments where a number of BSs are at approximately the same distance to the UE so that the same quantization steps are applicable for all BSs and as long as the backhaul capacity is sufficient.	The schemes improve detection reliability, hereby lowering access latency. In our simulation study we found that three orders of magnitude in improved missed detection probability was achievable with the use of 20 and 25 base stations for the QnF and DnF detection schemes, respectively.
2.2.3 Advanced Beamforming Designs to Enable New Services and Network Functionalities	Large antenna arrays at the BS are assumed, and a technique consisting in refining the instantaneous channel estimation based on the long-term channel structure is proposed. The trade-off between how many training symbols are required for each scheme is shown in an URLLC context.	Utilizing the channel structure when performing coherent beamforming provides higher reliability (up to two orders of magnitude compared to MRT not utilizing it) in the single user case. For multi-user, zero-forcing outperforms TDMA with coherent BF and second order statistics.
2.2.4 Interference Mitigation for Bi-Directional URLLC	The proposed bi-directional frame design based on adjacent-channel full duplex allows for flexible duplexing of radio resources in both time and frequency.	A precoding scheme is proposed which achieves suppression of OOB emissions by around 100 dB compared to the non precoded solution for adjacent channel full duplex
2.2.5 HARQ Investigations regarding URLLC	3GPP LDPC codes may be used for URLLC, but the current base graphs are not optimized for very low code rates demanded. Moreover, when investigating URLLC, it may be advisable to do that in a complete system level and link level tool, allowing to simulate all effects dealing with such high reliability requirements.	Deep investigation of HARQ performance with current codes, providing insights on performance with multiple CC and IR schemes, and their sensitivity with respect to all MCS parameters: CC-SLC seems preferable to IR (similar performance with low code rates, but lower complexity) and to CC-BLC (providing a gain up to 1.2 dB).
2.3.1Short Packet Transmission with Reliability-Latency Constraints	The throughput under reliability-latency constraints is investigated in this work item. Since short packet transmissions are considered, the grant-free scenario is considered, and the number of arrived users or the arrival distribution are used in order to maximize the throughput.	Increasing the superslot size K has been noticed to improve the reliability. The results also point out that the throughput can be improved if the knowledge of arriving packets can be accurately estimated. For instance, in the case of 15 active users, increasing K from 5 to 10 provides roughly an improvent of two orders of magnitude in reliability. Moreover, even if there is a considerable estimation error, when high reliability is targeted, the throughput performance still improves.
2.3.2 Reliable Schemes for Short-Packet-Transmission in Massive	In the mMTC scenario with sporadic transmission of short messages, reliable operation can be achieved by sparse superposition coding and receiver diversity (C-RAN architecture), in	The transmission scheme trades the number of users that can be simultaneously supported with the message length. When the messages are short, the number of simultaneously active users that can be

MTC	combination with advanced Bayesian receivers. Architectural constraints in the form of fronthaul limitations should be accounted for.	supported is 2-5 times higher than the one typically considered in mMTC simulations. In the CRAN scenario, the advanced Bayesian receiver yields higher reliability (up to an order of magnitude) when compared to simpler, correlation based receiver.

3 Massive MIMO Enablers towards Practical Implementation

The objective of this task is to minimize the gap between theoretical work and the real-world application of massive MIMO. The focus is on the application of massive MIMO in cellular networks. We propose enabling technologies, which can be categorized into *implementation-oriented* and *deployment-oriented* ones. One primary benefit of Massive MIMO is to increase the spectral efficiency by higher-order spatial multiplexing, thus the main focus here is on eMBB use cases in the Megacity scenario. However, massive MIMO beamforming gains are also used for underserved areas, e.g. by providing access links or in-band backhaul over larger distance than without massive MIMO.

Implementation-oriented enablers are:

- "Efficient pilot and Feedback Schemes for CSI Acquisition with Reduced Overhead" in Section 3.2 deals with the major "show-stopper" of massive MIMO in FDD systems [BLM16], the pilot and feedback overhead. Thereby, the focus is on algorithms for CSI acquisition and overhead reduction, e.g. by compression.
- "Distributed Beamforming" in Section 3.4 focuses on distributed massive MIMO where two complementary options to reduce the required data-exchange overhead are studied; one is robust decentralized beamforming algorithm and the other is functional split between baseband processing and RRHs.
- "Beam Management" in Section 3.5, addresses challenges related to beam-management in hybrid architectures as an enabler for future mmWave access.
- "Efficient Implementation: Hybrid Array Designs, Forward Error Correction, and Digital Frontend" in Section 3.6, which focuses directly on massive MIMO related implementation aspects to reduce complexity by solutions for fast and reconfigurable architectures, low resolution analogue to digital converters and hybrid-beamforming algorithms for wireless fronthaul.

Deployment oriented enablers are:

- "Pilot contamination mitigation" in Section 3.1 is relevant especially to massive MIMO TDD systems that are still sum-throughput limited due to pilot contamination [FRZ+17].
- "Massive MIMO Techniques for Flexible Access and Backhaul, and Multicast Transmissions" in Section 3.3 which enables massive MIMO also for other applications than data transmission in eMBB, namely in-band backhaul and corresponding signal shaping for underserved areas and massive multicast beamforming.
- "Optimized Array Formats and Capacity Analysis" in Section 3.7 is focused on the deployment of antennas, in particular to find more suitable array formats for various environments and applications.

Given the research nature of this project, the majority of the technologies described in the following aim at the long term evolution of 5G, meaning that the solutions/algorithms will not work on current 5G NR, Relase 15/16. On the other hand, a couple of the implementation oriented techniques can already be applied to nowadays networks, as they are subject to the

operators/network vendors implementation, The same is true for several deployment oriented solutions, where standards do not impose restrictions.

3.1 Pilot Contamination Mitigation

While massive MIMO provides high multiplexing gain, its performance critically depends on acquiring accurate CSI at the transmitter, which is then used to encode the transmitting signals and null the interference at the receivers [NAA+14]. For TDD systems, and by assuming channel reciprocity, CSI can be obtained by estimating the channels using predefined orthogonal pilot sequences, one for each user. The pilot sequences length hence scales linearly with the number of users if mutual orthogonality is required. When the number of users is high and large pilot sequence lengths cannot be afforded, the same pilot sequences are reused by multiple users simultaneously, which results in channel estimation errors and therefore in interference among the users. This phenomenon is called pilot contamination. For FDD, the number of pilots in the downlink scales with the number of BS antennas and therefore the reuse of pilots results in channel estimation errors and hence in performance reduction.

In this section, the problem of pilot contamination is tackled from different angles and using different approaches. In Section 3.1.1, the problem of pilot contamination in TDD systems is addressed by exploiting the spatial structure of wireless channels. The pilots are assigned to the users depending on the similarity of their channel covariance information, through the formulation of optimization frameworks that take into account the intra and inter cell interference. In Section 3.1.2, the time correlation of user channels is exploited in order to reduce the pilot contamination. By considering the channel memory, the CSI of only a fraction of users is refreshed at each time instant and therefore the pilots are assigned to only a fraction of users at each time. While the aforementioned works consider fixed and equal transmission powers among the users, the work in Section 3.1.3 investigates the reduction of pilot contamination through the use of open-loop uplink Fractional Power Control (FPC), which is adopted in 5G standard and already developed for LTE. This work allows understanding the potential gain of FPC and how its parameters should be optimized in different scenarios. While the focus in the aforementioned works is on TDD systems, Section 3.1.4 considers the use of Joint Spatial Division and Multiplexing (JSDM) technique [NAA+14] in FDD, which allows to reduce the CSI feedback overhead. This, however, requires non-trivial clustering of the users in realistic scenarios and per-cluster precoding process. This section deals with this issue and proposes new user clustering and scheduling schemes by using tools from graph theory.

3.1.1 TDD: Improving CSI Acquisition through Spatial Multiplexing

5G-NR will rely heavily on multi-antenna systems and beamforming techniques. The latter are especially critical for high carrier frequencies in order to combat the poor path-loss conditions. Consequently, obtaining accurate CSI estimates is of paramount importance as it conditions the accuracy of the subsequent beamforming. In TDD massive MIMO systems, CSI estimates are obtained using channel reciprocity and UL training with the needed time-frequency training resources depending on the total number of user antennas. Owing to the limited channel coherence time, the channel training procedure is restricted and the same pilot sequences need to be reused which results in pilot contamination. In order to efficiently implement TDD massive MIMO and leverage its critical beamforming capacities, a practical solution to mitigate pilot contamination should be developed. One way to address the latter issue is to exploit the spatial structure of the wireless channels [SG76]. Indeed, based on the narrow angular spreads of the different incident signals at the base station, the non-overlapping Directions of Arrival (DoA) of different channels can be leveraged in order to achieve orthogonal transmission in the spatial domain. This will improve the accuracy of CSI estimation and the subsequent beamforming. In this work, we exploit the spatial structure of the different wireless channels in order to address the issues of intra and inter-cell pilot contamination.

We choose to decouple the two problems and address them successively. In order to deal with intra-cell copilot interference, we construct groups based on the users' spatial signature, which refers to the main DoAs of the channel at the base station. The spatial signatures can be obtained after a dedicated, low-overhead training period or deduced from the channels covariance matrices, if available. In each cell, any given group is formed such that it contains users assigned the same pilot, however, with their channels having minimum overlapping spatial signatures and that provides a maximum coverage of the angular space. The latter conditions maximize the exploitation of the available Degrees of Freedom (DoF) while minimizing the impact of pilot contamination. The proposed approach is referred to as spatial basis coverage user selection. We provide two formulations of the grouping problem, and we propose two grouping algorithms, namely Power Agnostic Spatial Basis Coverage (Pag-SBC) and Power Aware Spatial Basis Coverage (Paw-SBC). Pag-SBC uses two nested greedy phases where a maximum number of users with minimum overlapping in their spatial signatures are assigned to each group. Paw-SBC further exploits the knowledge of the channel gain and its energy in each direction by solving several instances of a knapsack problem. We go a step further by investigating the guaranteed performance of the two approximation algorithms. The details of the algorithms in addition to the performance guarantee proofs can be found in [HA18].

Based on the constructed groups, we address inter-cell interference through an efficient cross-cell pilot allocation. Indeed, the spatial diversity of the interference signals can also be exploited. We propose a graphical framework based on copilot groups spatial signatures. Using this information, the network is able to allocate specific UL reference signals to groups, such that cross-cell interference can be efficiently managed. The resulting pilot allocation problem is formulated as a max-τ-cut problem [GZJ+17], which enables us to use a low complexity algorithm that provides the $\left(1-\frac{1}{\tau}\right)$ approximation of the optimal solution, where τ is the number of pilot sequences. The results show a high gain, in terms of spectral efficiency, as compared to the conventional massive MIMO system where the clustering of the users is not used. An example of the results is shown in Figure 3-1, in which the aforementioned three proposed approaches are denoted respectively by "Pag SBC", Paw SBC", and "Paw+max-τ-cut". Results show that our proposed approach can effectively double the spectral efficiency for a given outage probability. This high gain is due to the fact that in the conventional MIMO scheme, the interference is very high in the considered scenario and no smart interference management is used while in our proposed scheme a smart clustering and pilot allocation is used, which largely reduces the interference.

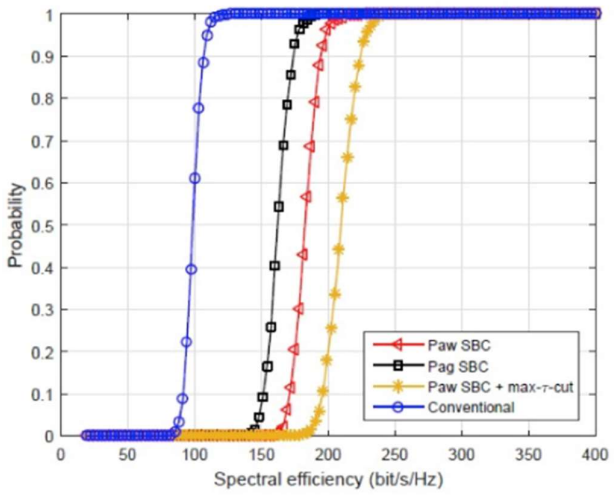


Figure 3-1: Comparison of CDFs of achievable spectral efficiency with 5 users per copilot group and SNR=10 dB.

3.1.2 Pilot Allocation taking into Account Markovian Channel Model and Traffic Patterns

While in the previous sections the focus is on the spatial correlation of the users as a mean to reduce the impact of pilot contamination, we consider here a different approach by taking into account the time correlation of the users' channels and their traffic patterns. In fact, ignoring the traffic pattern of the users will result in underutilizing the system resources, as the pilot sequences may be allocated to users having low amount of data to be transmitted. In addition, if, for a given user, the channel coefficients change slowly over time, it is highly inefficient to acquire its CSI at each time instant and one can use the corresponding pilot for other users. In these cases, exploiting the channel memory to optimize the allocation of pilots for CSI acquisition is of paramount importance.

To model the channel memory, we consider that channels evolve according to a Markovian stochastic process. Markovian modelling of the wireless channel is commonly used in the literature to incorporate memory, e.g. [EA07], [JV12]. In more detail, the channel between a user and the BS is modelled as a K-state Markov chain. Time is slotted and users are synchronized. We denote by $X_n(t) \in \{h_1, \dots, h_K\}$ the channel state of user n at time slot t. The state of the channel remains the same during a time slot and evolves according to the probability transition matrix $P_n = \left(p_{n,i,j}\right),\ i,j \in \{1, \dots, K\},$ where $P_{n,i,j} = P\left(X_n(t+1) = h_j | X_n(t) = h_i\right)$. Channels are assumed to be independent and non-identical across users, i.e., two different users may have different probability transition matrices. We consider that there are M pilot sequences and N users in the systems with $M < N$. The problem is then how to allocate efficiently M pilots to only M users at each time (in order to avoid pilot contamination) and to use for the remaining users the previously estimated CSI. One can see that as far as the old CSI remains close to the current channel coefficient, the performance of the system will be improved as compared to the case where the pilot sequences are allocated without taking into account the time evolution of the channel. The objective is then to allocate the pilots in such a way to keep the CSI of the users close to the real ones and hence to optimize the network throughput. This allocation problem can be formulated as a Restless Bandit Problem (RBP). RBPs are a generalization of Multi-Armed Bandit Problems [Whi88], sequential decision-making problems that can be seen as a particular case of Markov Decision Processes. RBPs are known to be PSPACE-hard and therefore finding the optimal solution for such problems is out of reach.

In this work, we show that the pilot allocation problem can be formulated as a Partially Observable Markov Decision Process and develop a suboptimal solution using the Lagrangian relaxation approach. We then prove that the optimal solution of the resulting relaxed problem is a threshold-based policy. This allows the development of a heuristic pilot allocation policy for the original stochastic problem based of the framework introduced by Whittle in [Whi88]. The obtained allocation policy is then of type Whittle Index Policy. We prove that Whittle Index Policy is asymptotically optimal in the many users regime (i.e., as the number of users and the number of available pilots grow large) and observe its remarkably good performance even for smaller and moderate number of users. One can refer to [LAD+18] for more details. In Figure 3-2, we consider a network composed of one cell with massive MIMO and 40 users. We plot the spectral efficiency achieved by our policy for 50 and 150 antennas and compare the results to a conventional MIMO scheme where all users send their pilots all the time. Results show a gain of 14% achieved by our scheme.

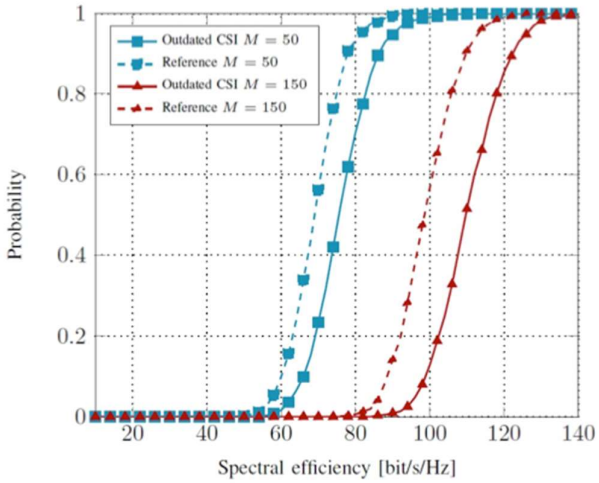


Figure 3-2: Spectral efficiency for the proposed feedback policy and conventional massive MIMO.

In addition to the aforementioned work, we have also investigated the interaction between pilot contamination, dynamic traffic arrivals and power control. We have developed in [AHB+18] a throughput optimal policy that allocates the power to the users in such a way to stabilize the queues whenever possible. The developed policy consists in allocating the power in solving a convex optimization framework in which the objective function is a proportional fairness like function at the SINR level (instead of the rate level). On top of this power control, the pilot scheduling can performed using the max weight policy. W refer to [AHB+18] for more details.

3.1.3 Fractional Power Control to Mitigate Pilot Contamination in 5G Massive MIMO

A balanced combination of uplink power control [SFK15] and pilot coordination among the cells [GCL+18] seems necessary to limit the negative impact of pilot contamination on massive MIMO system performance. In this work, we evaluate in realistic scenarios the open-loop uplink FPC adopted in 5G standard (and already developed for LTE) to understand: a) how the FPC parameters should be optimized in different scenarios and b) what the potential gains are. Therefore, we perform extensive system-level simulations by adopting the three-dimensional spatial channel model [3GPP-38.901] proposed by 3GPP: simulations have been performed by assuming 19 sites, with 3 sectors per site, and wraparound; however, neither joint transmission nor coordinated beamforming is assumed among the BSs and each UE is served just by its anchor BS, i.e., the inter-cell interference limits the system perfomance. With open-loop FPC, the power in logarithmic scale used by a certain UE k to transmit its pilot sequence can be written as

$$P_k = min\{P^{(UE)}, P_0 + 10log_{10}N + \alpha L_k\}, \quad (3\text{-}1)$$

where $P^{(UE)}$ is the maximum UE transmit power, N is the number of RBs, L_k is the large scale fading attenuation between UE k and its anchor BS, P_0 is used to control the SNR target on a RB, and α, with $0 \leq \alpha \leq 1$, is the fractional compensation factor of the large scale fading attenuation. The main objective of our contribution consists of numerically optimizing the parameters α and P_0 to maximize the performance.

In Figure 3-3 we show the Cell Spectral Efficiency (CSE) a) and the Cell Boarder Throughput (CBT) b) for different values of α and P_0 in an UMa, where BSs are equipped with 128 antennas, serve their UEs by using zero-forcing, and a pilot reuse 3 is implemented. More details about the simulation setup can be found in [BGG+18]. Here, we compare FPC against a baseline with no power control (noPC) where all the UEs transmit at the maximum power. First, we observe that there is an optimal α for a given value of P_0 which might be even different depending on the KPI: for instance, $\alpha = 0.5$ maximizes both the CSE and the CBT with $P_0 =$

60 dBm, whereas with $P_0 = 100$ dBm, $\alpha = 0.9$ maximizes the CSE and $\alpha = 0.8$ maximizes the CBT. Then, we also observe that the gain achieved by FPC when compared to the baseline is very limited in the CSE (in the order of 10%), but is quite significant in the CBT (up to 350%), confirming that an optimized FPC is very beneficial in mitigating pilot contamination. Finally, we notice that while the CSE is maximized with the combination $\alpha = 0.5$ and $P_0 = 60$ dBm, better CBT performance can be obtained with higher values of α and lower values of P_0. Additional results considering different beamforming criteria, reuse factors and BS array sizes can be found in [BGG+18], which show that the gain achieved by FPC: a) increases when the number of BS antennas increases, b) is usually higher with MRT when compared to ZF, and c) is usually higher with pilot reuse 1 when compared to pilot reuse 3.

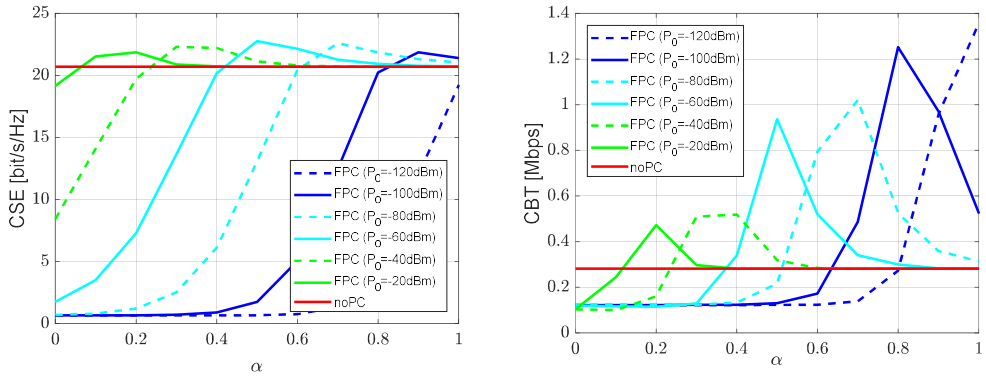


Figure 3-3: CSE a) and CBT b) for different values of P_0 and α with zero forcing, pilot reuse 3 and BSs equipped with 128 antennas.

3.1.4 FDD: Improving CSI Acquisition through Spatial Multiplexing

In FDD, the amount of CSI feedback from each single-antenna UE to the base station scales linearly with the number of BS antennas. This can be very restrictive, since it results in limiting the number of scheduled users. Fortunately, the CSI feedback overhead can be reduced by transmitting the reference signals using an appropriate precoding. Furthermore, the overhead can be drastically reduced by using Joint Spatial Division and Multiplexing (JSDM) technique in the downlink [NAA+14]. The main idea of JSDM consists of: *i*) partitioning the users into groups, where users within each group have, ideally, the same channel covariance eigenspace; and *ii*) using a two-stage precoding scheme (outer and inner precoders). Using this technique, the reference signals will be beamformed using the outer precoder. An example of users grouping is provided in Figure 3-4. In realistic scenarios, the users in each group may not have identical channel covariance eigenspaces and the orthogonality conditions of the eigenspaces between the groups cannot be met as well. Dealing with user scheduling and clustering is therefore of interest as it can help in boosting the performance of the system. We therefore propose a new user-clustering scheme along with a new similarity metric of the user covariance eigenspaces. Using Graph Theory, we also developed a new user scheduling policy for CSI feedback that provides promising gains as compared to the existing work in this area [NAA+14], [Sun17]. For example, simulations show a CSI feedback reduction of approximately six to seven-folds, compared with the full CSI case in [NAA+14]. More details on the proposed solution and the obtained results can be found in [MHA+18],[MHA+19],[ONE18-D41].

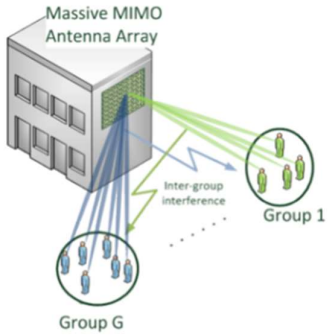


Figure 3-4: Example of Clustering in massive MIMO.

3.2 Efficient Pilot and Feedback Schemes for CSI Acquisition with Reduced Overhead

As mentioned in Section 3.1, classical CSI estimation techniques (e.g. least squares) require training overheads that scale with the number of transmit antennas and are not suitable for massive MIMO. Classical CSI feedback techniques require similar overheads, as well. In order to remedy these problems, we focus on a single-cell scenario in this section and present training and feedback designs for single-user as well as multi-user massive MIMO systems that exploit the underlying physical channel properties, i.e. the sparsity of the channel (in the path, time, or covariance eigen-domain). Such designs reveal the true dimensions of the channel and allow for a (possibly large) reduction of training as well as feedback overheads compared to classical techniques. In contrast to Section 3.1, we do not consider pilot reuse as a possible solution to reduce training overheads and thus the pilot contamination problem does not exist.

5G-NR has defined different pilot (training sequence) densities in time as well as frequency domains for up to 32 antenna ports. Nonetheless, even the densest pilot configuration may not be sufficient for accurate channel estimates when classical techniques are employed, if more antenna ports were to be used in the future. Exploiting the underlying channel sparsity presents a more viable and long term solution. Further, even if the current 5G-NR pilot densities are enough, there are no thorough guidelines on how to choose a suitable density, except those that consider user mobility. In light of this, the performed analytical and numerical investigations of this section provide valuable conclusions and guidelines for choosing suitable pilot/training densities. A similar argumentation holds for 5G-NR CSI feedback.

In Section 3.2.1, we discuss the validity of the planar wave assumption in massive MIMO systems and its impact on parametric channel estimation methods. Section 3.2.2 introduces the concept of *hierarchically* sparse channels in the path domain and presents a novel Compressive/Compressed Sensing (CS) channel estimation algorithm for UL wideband multiuser systems that reveals important insights on the training durations needed for accurate channel estimation. While Section 3.2.2 proposes angle and delay solutions from a predefined grid, Section 3.2.3 relaxes this assumption and suggests continuous solutions using the recently proposed concept of atomic norm minimization. Section 3.2.4 discusses limited channel covariance feedback in multiuser systems and analyses how the number of fed back covariance eigenvectors by the users should be adapted to the SNR in order to realize the DL channel estimation schemes considered in [ONE18-D41]. Finally, Section 3.2.5 exploits channel sparsity in the time domain to reduce the CSI feedback overhead and presents an orthogonal matching pursuit algorithm to accurately recover the (compressed) CSI at the BS, along with a complexity analysis.

3.2.1 Parametric Channel Estimation for Massive MIMO

Our main objective is to study physical channel models and their limitations. In previous contributions [LP18a] and [LP18b] we studied physical models assuming the distance separating the transmitter and receiver is large enough compared to their respective sizes, so that the spherical wavefront can be well approximated by planes. In that particular case (which is sufficient in most practical situations), the optimal number of propagation paths to estimate was sought for typical environments, showing that this number is, in practice, much smaller than the number of physical paths.

Recently, we were interested in the plane wave assumption and its validity. This problem is of interest because antenna arrays composed of a very high number of antennas can become so large that the plane wave assumption is not valid unless a very large distance separates the transmitter and the receiver (this is illustrated in Figure 3-5).

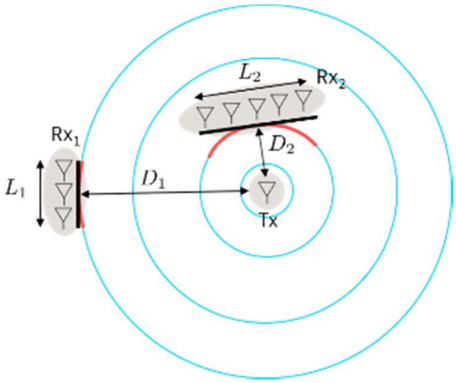


Figure 3-5: Illustration of a situation in which the plane wave assumption does not apply. The receiver Rx1 is far enough from the transmitter Tx with respect to its size, so that the spherical wavefront (in red) is well approximated by a plane (in black). This is not the case for receiver Rx2

We characterized precisely the limits of the plane wave assumption and designed estimation algorithms taking into account the curvature of the wavefronts by adapting classical channel estimation algorithms (adding a distance estimation step) in [LLP18]. This allowed to extend the validity of physical model based methods to smaller distances between transmitter and receiver, which is particularly interesting in a micro-cell scenario. Such results are promising because taking into account the curvature of the wavefronts could yield:

1. Better detection of uplink pilots,
2. Optimized analog beams for hybrid systems,
3. Shorter pilot sequences for downlink channel estimation.

As an illustration of the method's potential, we performed simulations using the Quadriga channel simulator [JRB+18] in a MU-MIMO scenario (see [LLP18] for a complete description). The results in terms of channel estimation error are shown on Figure 3-6 below. They indicate that taking into account the wavefronts curvature (curves corresponding to (S_{joint}) and (S_{seq})) allows to get better estimation quality compared to the classical least squares and methods based on the plane wave assumption (curve corresponding to (S_{PWM})). Moreover, we proposed efficient estimation algorithms (S_{seq}), so that the complexity is not increased a lot compared to these simpler methods.

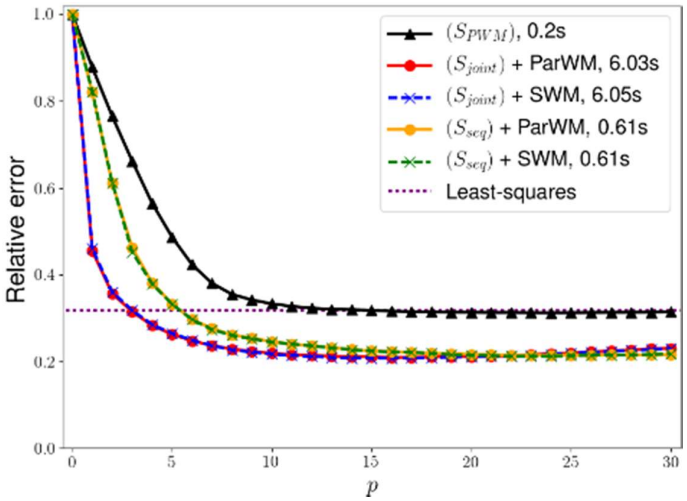


Figure 3-6: Comparison of classical least-squares channel estimation, plane wave assumption based channel estimation and channel estimation taking into account wavefront curvature. The quantity p is the number of estimated paths.

3.2.2 Hierarchical Sparse Channel Estimation for Multiuser Massive MIMO with Reduced Training Overhead

Towards the objective of efficient massive MIMO CSI acquisition, a novel, low-overhead channel estimation algorithm for TDD wideband massive MIMO was proposed based on a, so called, hierarchically-sparse representation of the channel. This representation not only captures the fundamental sparsity properties of the channel (i.e., limited number of propagation paths) but also takes into account physical propagation properties, i.e., no two paths can have the same Angle-of-Arrival (AoA) (for a sufficiently large antenna array). With this representation, a novel channel estimation algorithm was proposed along with a rigorous analysis of its performance. This analytical performance characterization also serves as an objective function to be optimized by the training sequence design.

Unfortunately, identification of the optimal training sequence is difficult. However, the analysis naturally suggests a specific training sequence design, which, even though suboptimal, is analytically and numerically verified to achieve excellent performance. In particular, it was shown that the number of required pilot subcarriers to be shared by a group of UEs scales independently of the number of channel paths (per UE) and number of active UEs, as the bandwidth and number of antennas increase. This is demonstrated in the numerical example of Figure 3-7 where the channel estimation MSE ("HiIHT" curves) is depicted as a function of the pilot overhead (number of pilot subcarriers to total number of subcarriers) for various number of channel paths. Even though the actual MSE depends on the number of paths, the pilot overhead threshold above which reliable estimation is achieved is always the same. In contrast, conventional algorithms that do not take into account the hierarchical sparstity property ("IHT" curves) both requires almost an order of magnitude greater pilot overhead and this overhead is also heavily dependent on the number of channel paths.

This result reveals that massive MIMO enables the shifting of the training overhead from the frequency to the spatial domain, effectively allowing for a more efficient bandwidth utilization, an effect that is not possible in conventional MIMO transmissions. Detailed presentation of the algorithm and its analysis can be found in [ONE18-D41], [WRF+18a], and [WSF+18], and an extension to a setup considering joint channel and data detection is presented in [WRF+18b]. These references demonstrate the superiority of the proposed algorithm compared to

conventional compressive sensing approaches that fail to exploit the hierarchical sparsity property of the massive MIMO channel.

The proposed algorithm and analysis is valid for all use cases considered in ONE5G as the need for reduced training overhead is universal. This need is more pronounced in applications with high mobility and/or small data payload as in mMTC, considered in the core use cases #1, #2, and #3 in [ONE17-D21].

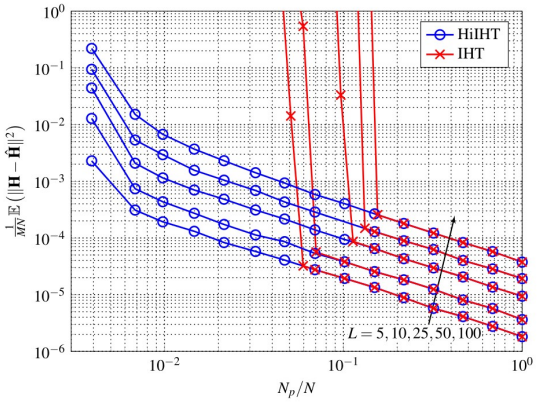


Figure 3-7: Channel estimation MSE achieved by the proposed algorithm ("HiIHT") and conventional algorithm ("IHT") as a function of the pilot overhead for various number, L, of channel paths

3.2.3 Wideband Massive MIMO Channel Estimation via Atomic Norm Minimization

The task of estimating the channel matrix $\boldsymbol{H}$ corresponding to a wideband, massive MIMO channel can be regarded as the estimation of L superimposed, two-dimensional harmonic signals. In this case, L is the number of propagation paths and the two-dimensional "frequency" of each path corresponds to its AoA-delay tuple. One of the most recent tools in CS is that of Atomic Norm Minimization (ANM), which deals exactly with the problem of estimating superimposed harmonic signals from the noisy observation of a small number of entries of $\boldsymbol{H}$ [CF14], [TBS+13], [CC15], [YXS16], making it a promising candidate for massive MIMO channel estimation. Unfortunately, direct application of the ANM approach results in a complexity proportional to MN, where M, N are the number of (receive and transmit) antennas and OFDM subcarriers, respectively, rendering it impractical for the massive MIMO setting.

Motivated by this observation, a novel, low-complexity ANM-based channel estimator was proposed towards reliable channel estimation with a small number of pilot subcarriers (training overhead). The fundamental idea is to decouple the AoA and delay domains and treat them sequentially and independently by means of an ANM-based estimator operating on an equivalent MMV observation model for each domain. The full details of the algorithm can be found in [SBW18]. Its complexity scales only linearly to the system dimension, $M + N$, and therefore it can be applied to the massive MIMO setting. Its performance is shown to come close to an analytical performance lower bounded that holds for any unbiased estimator [SBW18].

As an example of the potential of the proposed algorithm, Figure 3-8 depicts the channel estimation MSE achieved by the proposed and alternative algorithms for a system with $M = N = 100$ and with a variable number $N_p \leq N$ of randomly distributed pilot subcarriers. A sparse channel with $L = 3$ paths, delay spread equal to 25% of the useful OFDM symbol duration, and an average subcarrier SNR of 10 dB was considered. The performance results were obtained after averaging over many randomly and independently generated channel realizations. It can be seen that the proposed algorithm achieves a very small MSE for a training

overhead less than 15% of the system bandwidth and its performance closely follows the universal lower bound. The standard LMMSE estimator has significantly poorer performance as it fails to capture the channel sparsity property. The performance of the BPDN algorithm as well as a previously proposed ANM-based estimator [WXT17], [CXY17] are also shown in the figure (the latter performance is only for channel realizations with non-closely located channel paths in the AoA-delay domain as the algorithm fails otherwise). It can be observed that both perform worse than the proposed algorithm.

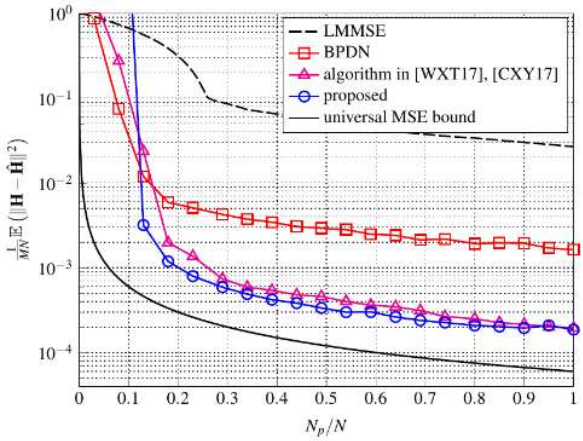


Figure 3-8: MSE performance of various channel estimation algorithms as a function of training overhead.

3.2.4 On the amount of DL training in correlated massive MIMO channels

In [ONE18-D41], we considered a multiuser, massive MIMO scenario with spatial channel correlations, and addressed the problem of choosing a suitable training duration T that maintains a *bounded* rate loss due to channel estimation errors. This goes one step beyond the work in [BX18a] which only considered the effect of T on the channel estimation MSE. [ONE18-D41] showed that T can be much smaller than the number of BS antennas M in highly correlated scenarios. A critical assumption was that the BS had perfect knowledge of the channel covariance matrix C_k to user k. We now relax this assumption and show that (almost) the same achievable rate performance can be obtained if the K users feed back only a subset of dominant covariance eigenvectors to the BS. The subset is chosen carefully according to the SNR in the data phase. Let us define $\Phi_k(S)$ to be the channel estimation error covariance matrix for a given choice S of the training matrix, and revisit results from [BX18b], [BSL+2019] as they will be helpful in developing the adaptive covariance feedback scheme.

Theorem [BX18b] *Let $\lambda_1(\Phi_k(S))$ be the largest eigenvalue of $\Phi_k(S)$ and $c = O(1)$ be a constant. Then the condition*

$$\lambda_1\big(\Phi_k(S)\big) \leq c\, SNR^{-1}, \tag{3-2}$$

is sufficient to maintain a bounded ergodic rate loss, i.e., $E[\Delta_k] < log_2(1+c)$ as $SNR \rightarrow \infty$.

Further, it was shown that in order to obtain a training solution that satisfies condition (3-2), a per-user search can be initially performed. This takes place by searching for the T_k dominant eigenvectors of C_k that, when trained, ensure condition (3-2) is satisfied for user k. This gives K different range spaces that are further processed in a second step. There, any possible overlaps between the found range spaces are exploited to reduce the training overhead. This results in a training solution satisfying

$$range\ (S) = range\big(U_1^{1:T_1}, \ldots, U_K^{1:T_K}\big) \tag{3-3}$$

A Limited Channel Covariance Feedback Scheme: In case the BS has no channel covariance knowledge, user k can feed back the dominant T_k eigenvectors to the BS, based on which Equation (3-3) can be approximated. Feedback/quantization format remains to be discussed. A practical option is vector quantization, where the quantization codebooks consist of vectors. A challenging problem here is how to choose the codebook entries, since those may not be optimal for all possible eigenvector realizations. Fortunately, it has been shown in [BX18b] that when the BS is equipped with an $M_r \times M_c$ UPA, eigenvectors converge to columns of the $dftmtx(M_c) \otimes dftmtx(M_r)$ matrix as M_c and M_r grow without bounds for any covariance realization ($dftmtx(M_c)$ refers to the unitary DFT matrix of size $M_c \times M_c$). This provides a simple yet close to optimal solution for massive MIMO systems, as DFT based codebooks can be used for eigenvector feedback.

On the Adaptive Feedback Rate: Recall from Equations (3-2) and (3-4) that T_k increases with the SNR. Consequently, a higher SNR increases the required covariance feedback rate. Nonetheless, this should constitute no problem since the covariance changes (and thus needs to be fed back) on a much slower scale, compared to, e.g., the instantaneous channel.

Numerical Results: A Laplacian angular spectrum model with horizontal and vertical deviations of 10˚ and 2˚ respectively is used to generate the covariance matrices of the $K = 8$ users, with mean azimuth angle of arrivals {-52.5˚, -37.5˚,…, 52.5˚}. The BS is equipped with a 16x16 UPA, and is mounted at an elevation of 50 m. Figure 3-9 (left) shows the achievable sum rate with zero-forcing precoding for $c = 1$. The proposed scheme where the BS obtains partial covariance information via user feedback suffers only negligible losses compared to the case with perfect covariance information over most of the SNR range, and is furthermore able to achieve high sum data rates that follow the perfect CSI sum rates. Figure 3-9 (right) plots the number of fed back eigenvectors of different users, and shows its increasing behaviour with respect to the SNR. As the users have different covariance matrices, a different number of eigenvectors is fed back per user. These numbers are much smaller than M, even for large SNR values.

We note that the presented analysis does not take into account the effect of antenna mutual coupling on channel power (and correspondingly on the eigenvalues of channel covariance matrices). In practice, the coupling arising from standard half wavelength antenna spacings can be neglected. For smaller antenna spacings, the framework in [LNB+18] can be used to complement the presented analysis.

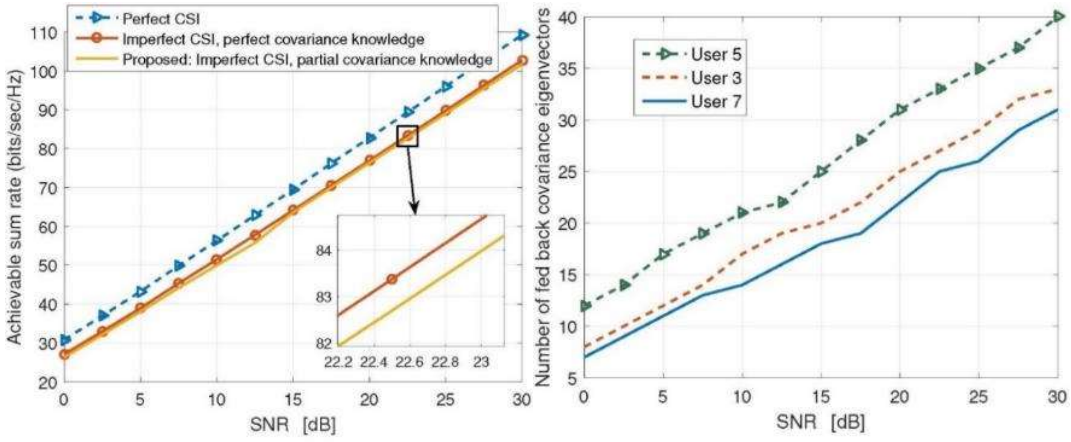


Figure 3-9. Left: Achievable sum rates, right: number of feedback eigenvectors with respect to the SNR.

3.2.5 Efficient Feedback Schemes for more Accurate CSI and Advanced Precoding

In this work, in order to exploit the underlying time domain sparsity of the channel, we consider explicit CSI feedback based on time domain channel impulse response [AVW18]. We refer to it as Explicit Time domain Feedback (ETF). We feedback the most significant taps of the sparse Channel Impulse Response (CIR) back to the gNB. In [AJW19], a time domain based compression scheme, which used Orthogonal Matching Pursuit (OMP) for sparsity recovery, showed around 14% gain in spectral efficiency vs state-of-the-art Rel. 15 NR type II CSI codebook. For successful reconstruction of the DL CSI at the gNB, the positions of the significant taps has to be available at the gNB. Fortunately, the tap positions change very slowly compared to the channel tap values. However, the gNB can't learn this piece of information on its own, there has to be a UE feedback with the tap location information. In this work, we present an efficient way for the UE to report the tap location information as shown also in [AJW19]. A similar method was also used in [ATM19], to feedback compressed CSI for NR type II CSI in Release 16.

Assuming a setup where where M is the number of transmit antenna ports, L is the number of beams per polarization, N is the number of UE receive antennas, N_a is the number of active subcarriers. L_{range} is the considered channel delay range and $f_{OS.}$is the oversampling factor used inside the basis matrix $\boldsymbol{F}_{N_a \times f_{OS} L_{range}}$. $\boldsymbol{F}_{N_a \times f_{OS} L_{range}}$ is built as a submatrix of the oversampled $N_{FFT} \times f_{OS.} N_{FFT}$ DFT matrix. The rows of $\boldsymbol{F}$, correspond to the locations of the active subcarriers, whereas the columns of $\boldsymbol{F}$ are the channel support search range. The output of the sparsity recovery algorithm is the $N_s \times 1$ channel support vector, which needs to be fed back to the gNB. $\boldsymbol{C}_{N_a \times N_s}$ is the compression matrix applied finally to the CFR matrix, where N_s is the channel support, i.e. the number of significant taps.

At the UE, the time domain channel taps matrix

$$\boldsymbol{G}_{N_s \times 2LN} = \boldsymbol{C}_{N_s \times N_a} \boldsymbol{H}_{N_a \times 2LN}$$

Out of all the coefficients in $\boldsymbol{G}_{N_s \times 2LN}$, only the strongest k_{taps} quantized coefficients are to be fed back to the gNB. The locations of the significant taps can be fed back to the gNB via a bit-map of size $2LNN_s$. Given that tap locations are discovered in descending order of power, the first column in the bitmap corresponds to the strongest tap location. Hence, to reduce the bitmap size, we can assume that all the coefficients on the first x_f columns are going to be selected in the set of k_{taps} quantized coefficients. That reduces the bitmap size to $2LN \times (N_s - x_f)$, i.e. saving $2LNx_f$ as shown in Figure 3-10. The blue cells correspond to the locations of the significant taps which will be quantized and fed back to the gNB. The gNB assumes the rest of the coefficients to be zero.

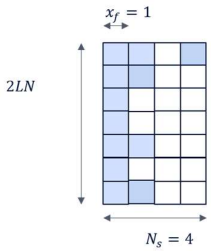


Figure 3-10: Bit-map for feeding back tap location information

In Table 3-1, the geometric mean spectral efficiency (normalized with respect to NR type II CSI) vs total UL overhead is depicted for NR type II CSI and ETF with long term update of tap location information and W_1 at 10ms and 50ms. The simulation is made assuming max rank of 2 layers, $M = 64$ antenna ports, $L = 4$ transmit beams per polarization, $N = 2$ UE receive antennas, UE speed of 3kmph, and a carrier frequency of 2GHz. ETF used $N_s = 4$ and $k_{taps} = 64$.

Scheme	Normalized geometric mean	UL Overhead [bits]
NR type II CSI	1	548
ETF $N_s = 4$ $x_f = 4$ update=10/10	1.096	512
ETF $N_s = 4$ $x_f = 4$ update=10/50	1.081	461

Table 3-1. Normalized geometric mean vs UL overhead for rank=2

As shown in Table 3-1, ETF can achieve 8% higher spectral efficiency compared to NR type II CSI while saving 16% of the UL overhead.

3.3 Massive MIMO Techniques for Flexible Access and Backhaul, and Multicast Transmissions

In a 5G system, a base station providing not only access transmission but also backhaul transmission can be envisioned. One example is the Integrated Access Backhaul (IAB) node, in particular in mmWave band. It is naturally understood that the radio coverage reduces as the carrier frequency increases. Thus, the practical issue for the 5G operators is the increased CAPEX for establishing the traditional X2 interface and S1 interface with, typically optical fibres. To solve this issue, the wireless backhaul link becomes a cost-efficient solution. In another example, in some low Average Revenue Per User (ARPU) areas, the backhaul link based on wired connection is often unaffordable, calling for enhancements on the radio backhaul solution. In current 3GPP NR Rel. 16, a study item focusing on the IAB design has opened. However, up to now, the main deployment scenario is for mmWave deployment, covering a dense urban area, while a rural low ARPU scenario has not yet been considered.

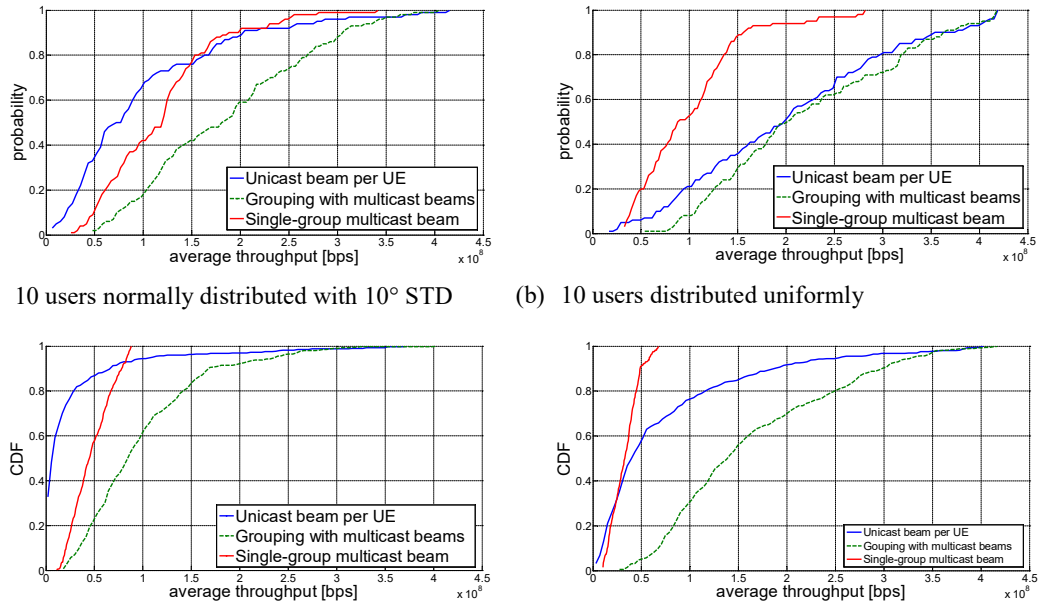
In this section, we present three technical clusters: first, a massive multicast beamforming technique is presented, which aims to exploit massive MIMO beamforming for multicast services. Second, a technical solution directly dedicated to rural area considering low ARPU constraints is presented. Different from 3GPP IAB scope, the main challenge in this context is that the backhaul distance is way beyond what has been targeted in 3GPP IAB. Thus, special and powerful beamforming should be given for backhaul, taking advantage of massive antennas; while to serve the normal access UEs, less costly omnidirectional antennas are considered. Moreover, to further drive the backhaul link performance to its limit, we present an enhanced modulation scheme, named Probabilistically Shaped Coded Modulation (PSCM), which is well known for its remarkable added-value via shaping gain, leading to an increased throughput compared to the classical QAM modulation. However, the benefit from PSCM can be guaranteed under a relatively stable channel condition. Given the channel characteristic of backhaul link between two high height fixed nodes, the channel condition is rather favourable for such modulation, thus, PSCM in the backhaul context becomes plausible.

3.3.1 Multicast Massive MIMO

The 3GPP Rel-14 study item "Scenarios and Requirements for Next Generation Access Technologies" [3GPP-38.913] noted that “The new radio access technology (RAT) shall leverage usage of RAN equipment (hard- and software) including e.g. multi-antenna capabilities (e.g. MIMO) to improve Multicast/Broadcast capacity and reliability”. Current standards like Multimedia Broadcast Multicast Services (MBMS) or Single-Cell Point-to-Multipoint (SCPTM) do not employ MIMO. It is expected that future systems will make extensive use of massive MIMO antennas for all kinds of services, in particular for new automotive and IoT use cases (both industrial and mMTC). The evolution is leading towards a more flexible and dynamic integration of unicast and multicast, which includes the concurrent delivery of both unicast and multicast/broadcast services to the users [3GPP-38.913]. Because of the priority of other

activities a draft study item "NR mixed mode broadcast/multicast" (RP-180669) has been postponed and might be reconsidered in Rel-17.

ONE5G has studied multicast and unicast beamforming strategies for the case that identical content is to be distributed over a local network (same cell or small cluster of cells) via unicast and multicast transmissions sharing the same spectrum. We assume partial CSI obtained by beamformed reference signals on the basis of a DFT grid of beams. If the UEs have different angles of departure, then this compromises the array gain. Alternatively, the multicast message can be transmitted with full array gain via unicast links that are multiplexed in space, frequency, or time (known as "beam sweeping"). We have studied this trade-off for different UE distributions, numbers of UE, and grouping strategies. The simulation results are shown in Figure 3-11, as reported in [ONE18-D41]. The achievable gains depend on the UE distribution and also on the number of UEs. Neglectible multicast gains are observed for 10 UEs that are uniformly distributed. But for more UEs or non-uniform distributions, the throughput can be increased many times over.



(a) 10 users normally distributed with 10° STD (b) 10 users distributed uniformly

(c) 40 users normally distributed with 10° STD (d) 40 users distributed uniformly

Figure 3-11: Comparison of unicast and multicast MIMO transmission

These ONE5G results highlight the importance of combining multicast beamforming with suitable grouping and scheduling strategies. The integration of unicast and multicast transmissions will help realizing local multicast transmission for efficient distribution of identical content to groups of UEs. Adding massive MIMO is the natural next step in the evolution of multicast services with superior efficiency in terms of energy and spectrum.

3.3.2 Wireless Backhaul for Coverage Enhancement in Low ARPU Networks

In low ARPU areas, extending the coverage of the network is a challenging task as there are severe constraints in terms of cost. This is particularly the case for the BH link between the BSs and the core network when using a classical optical fibre communication. Nevertheless, a wireless link between two BSs, the first one acting as a relay for the second one, would be an appropriate solution to reduce the cost in low ARPU networks. The combination of in-band wireless BH and mMIMO system is proposed for this study. Using an in-band wireless BH, the access links and the BH link are multiplexed in the same frequency band and the use of sub-6

GHz frequency bands is possible, leading to a low path loss. Additionally, a mMIMO system is an adaptive solution, able to provide high data rates by increasing the SINR. As in [LZL15], we propose to multiplex the BH link and the access links in the space domain rather than in the time domain and to improve the spectral efficiency of the system. The spatial interference on the UE side is managed by a mMIMO precoding technique. We focus on the downlink of a TDD OFDM system. The proposed solution is illustrated by Figure 3-12.

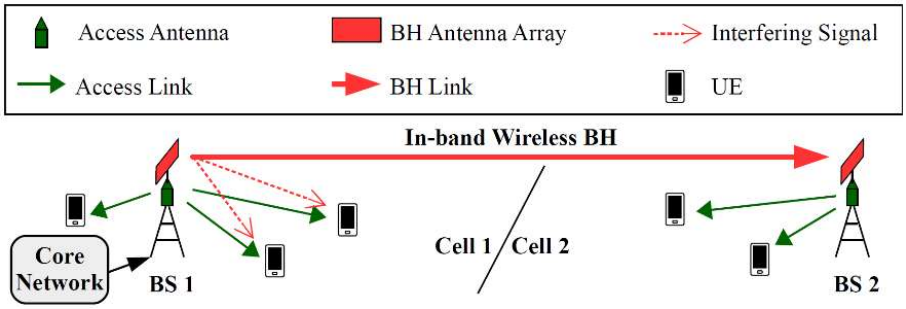


Figure 3-12: Illustration of the proposed in-band wireless BH link with mMIMO.

For this project, a new mMIMO precoding technique, called *Regularized Zero Forcing with Controlled Interference* (RZF-CI), has been proposed in order to maximize the received power on the receiving BS, while limiting the interference power level on the UE side to a predetermined value. Additional details can be found in [ONE18-D41].

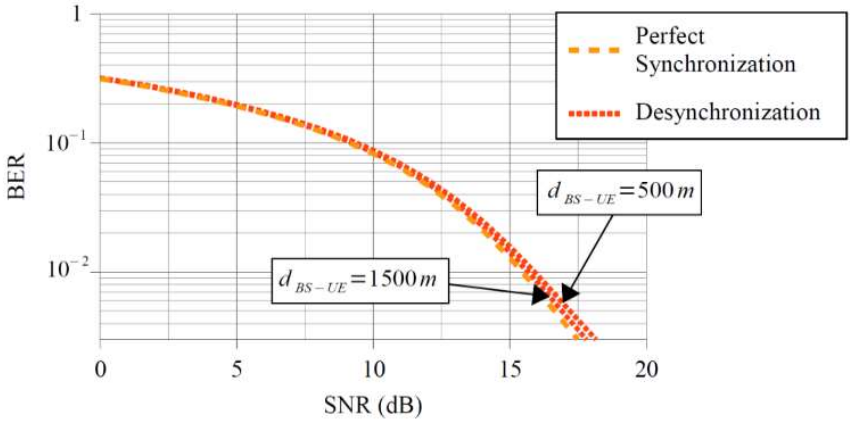


Figure 3-13: BER on the UE side with the RZF-CI precoder as a function of the SNR (dB), considering a perfect time synchronization and considering a large time desynchronization (38 µs). The distance between the BS 1 and the considered UE is d_{BS-UE}=500m and d_{BS-UE}=1500m.

The study in [ONE18-D41] considers a perfect time synchronization between the access link and the BH link. However, such a synchronization is not feasible in a real system because of hardware limitations. Therefore, the impact of a time desynchronization between the access link and the BH link has been evaluated thanks to an analytical study and to simulations.

The analytical study shows that a time desynchronization between the BH link and the access link induces a spreading in the frequency domain of the interference on the UE side, coming from the BH link. However, when the BH link and the access link are asynchronous, only the edge sub-carriers of the considered UE induce additional interference power from the BH link. In order to evaluate this impact, simulation results are given by Figure 3-13. The simulation parameters are similar to those used in [ONE18-D41] but consider an eventual time desynchronization between the BH link and the access link. These results show that the impact of a time desynchronization between the BH link and the access link on the BER on the UE side is negligible and thus an asynchronous BH link can be considered.

3.3.3 Signal Shaping for MIMO Backhaul Channels

Probabilistically Shaped Coded Modulation (PSCM) is a promising technique that can improve the coded modulation performance significantly, as shown in [IX18] and [IBX19] for AWGN channels and in [IBX18] for fading channels. Moreover, it may also be used for high throughput applications, as discussed in [PX17] and [IBX18]. PSCM, e.g., Probabilistic Amplitude Shaping (PAS) [BSS15], is a good candidate for backhaul channels for two reasons: backhaul links often use large QAM constellations to achieve high throughput (large constellation allows to achieve higher shaping gain), and the backhaul link is static and does not vary much over time, i.e., no Doppler effect, line of sight (this reassembles to some extent AWGN where signal shaping is shown to perform well) [HXK18].
PSCM shaping systems are built by a serial concatenation of a Distribution Matcher (DM) and a channel encoder at the transmitter and a channel decoder and an inverse DM at the receiver. The throughput of the state-of-the-art DMs and inverse DMs is low compared to the throughput of channel coders (e.g., LDPC coders), and constitutes a limitation for the throughput of the whole system. We develop a Multi-Composition Distribution Matcher (MCDM) using multi-composition codes, which improves the performance of distribution matching for shorter sequences and is suitable for parallelization based on independent processing of sub-sequences. That is, the data sequence to be shaped can be divided into smaller sequences which are shaped independently. Previous distribution matching techniques, e.g., Constant Composition

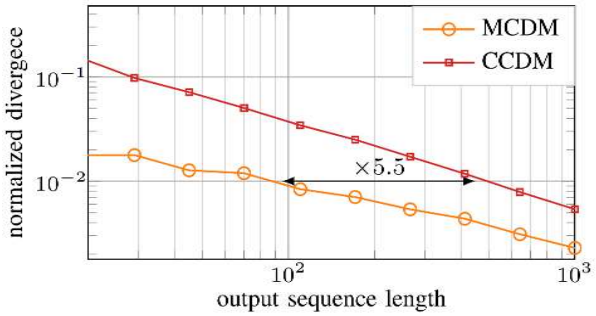


Figure 3-14: Information divergence at the output of MCDM and CCDM vs block-length.

Distribution Matcher (CCDM) [SB16], would lead to large performance loss when using short sequences (see Figure 3-14). MCDM achieves the same divergence for sequences ca. 5 times shorter than the CCDM, thus a parallel processing of 5 sequences is possible without performance degradation. The performance gain can be used to reduce the processing time. This parallel processing gain can be directly turned into a throughput gain. The factor depends on the target distribution. Binary MCDMs can be combined with the parallelization technique [PX17]. PSCM has been contributed as a modulation scheme for NR (see [3GPP-1700076]), therefore the investigation of distribution matching schemes is relevant for the 3GPP standardization. More details can be found in [PX17].

3.4 Distributed Beamforming

This technology cluster investigates two distributed beamforming designs for MIMO systems. Section 3.4.1 considers a CRAN architecture with massive MIMO RRH, and proposes a hybrid beamforming architecture where the analogue functionalities are kept at the RRHs and the baseband digital functionalities are migrated to the BBU. The proposed beamforming design does not require instantaneous CSI and instead relies on the second-order statistics of the channel to determine the analogue beams. Thereby, the complexity is reduced due to smaller number of RF chains and due to the smaller number of instantaneous channel coefficients needed to be estimated. The second TeC in Section 3.4.2 considers a TDD system and studies optimal precoders in the case of imperfect partial CSI at the BSs by solving the expected sum rate stochastic maximization problem using the difference of convex functions approach.

3.4.1 Beamforming Design and Function Split for Partially Centralized RAN with Massive MIMO RRH

In a CRAN architecture with massive MIMO RRH, the core objective is to keep minimal functionality at the local RRHs in order to i) ease the deployment costs, scalability and flexibility, ii) alleviate the transport load between the RRHs and the centralized BBU/cloud and iii) facilitate CSI acquisition. We study a hybrid beamforming architecture where the analogue beamforming functionalities are kept at the RRHs and the baseband digital beamforming functionalities are migrated to the BBU [KPD+17].

One major advantage of this architecture is that the complete instantaneous channel is not needed at the BBU in order to design the analogue and digital beamformers, as this would be a heavy requirement in a distributed architecture. Instead, we exploit second-order statistics (channel covariance matrix) to determine the analogue beams and the instantaneous channel seen from the analogue beams to design the digital beams.

The key result consists of the optimization of the hybrid analogue/digital beamforming as follows:

a) Analogue beamforming and the number of analogue beams are determined according to second-order statistics (covariance matrix) estimated at the BBU.
b) Digital beamforming is designed based on the low-dimensional (instantaneous) effective channel state information observed through analogue beamforming.
c) Hardware constraints, i.e., limited-fronthaul and quantization noise, are taken into consideration.

In terms of the performance gain, we compare the proposed scheme to a fully digital implementation with unlimited fronthaul capacity [KPD+17]. As expected there is a performance degradation when the number of RF chains is low: between 4-16% loss for a number of 16 RF chains. The degradation comes from the hybrid implementation as well as the fact that the design of the analogue beamforming is based on second-order statistics. On the other hand, for a low fronthaul capacity, it is better to activate a small number of RF chains: e.g., for RRHs with 64 antennas, and fronthaul capacity equal to 200 bps, 8 RF chains should be activated.

In terms of the complexity vs cost trade-off, the reduction in cost comes from the implementation via hybrid analogue and digital beamforming instead of a fully digital implementation, i.e. it comes from the reduction of the number of RF chains. Moreover, the number of instantaneous channel coefficients to be estimated per RRH at the BBU is reduced from the number of antennas at the RRH times the number of users to the number of RF chains times the number of users.

3.4.2 Beamforming Algorithms for System Utility Optimization toward Massive MIMO

This section is devoted to the beamforming design for the optimizations on the sum rate maximization and on the minimum user rate maximization (rate balancing) respectively. For the sum rate maximization, we will consider two types of partial Channel State Information at the transmitter (CSIT):

1) with the knowledge of the channel estimate and the channel estimation error covariance and 2) with pathwise channel information. For the rate balancing, we provide a new algorithm with perfect CSIT.

We consider an optimization with sum rate maximization as system utility and a TDD system where the downlink CSITs come from uplink channel estimations by reciprocity. We study linear precoders in the case of imperfect CSI at the BSs by solving the expected sum rate stochastic maximization problem using the difference of convex (DC) functions approach (see

Appendix 8.6 of [ONE18-D41]). There are already other solutions for the imperfect CSI case; however, none of them properly take into consideration jointly the channel estimate as well as the channel estimation error covariance which results in suboptimal performance.

Figure 3-15 shows the sum rate versus SNR for a MIMO system composed of 2 cells, with 8 two-antenna users in total and 8 antennas per BS, where, the transmit covariance matrices considered are low rank matrices (i.e. rank equal to 2). In Section 3.3.4 of [ONE18-D41], more explanation about this figure and the curves is given. As seen in the figure, the proposed approach provides higher sum-rate than the other approaches. Through the DC approach, our new objective function leads to the solution of the eigenmatrix of some matrices. We proposed to calculate the expected values of these matrices. The gain over the other approaches comes from exploiting the channel covariance information not only in terms of interference, but also in terms of signal power.

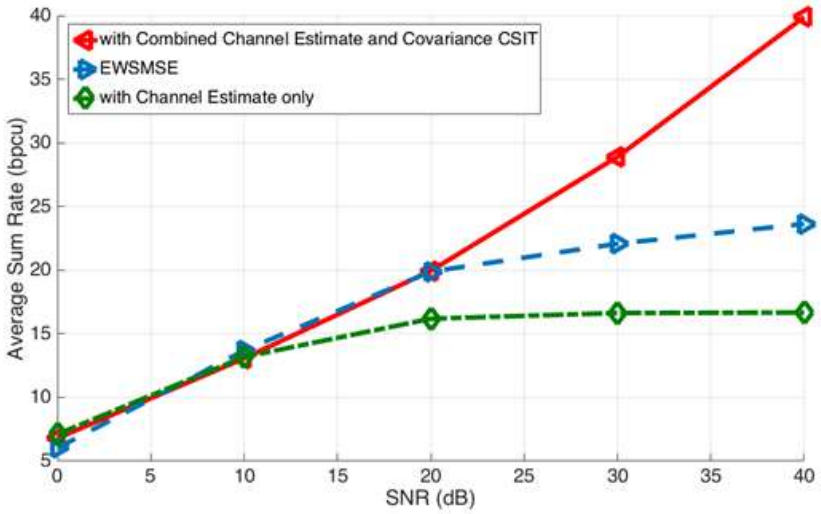


Figure 3-15 Sum rate comparison: Correlated low rank transmit covariance matrices, 2 antennas/user, 4 users/cell, 8 antennas per BS, 2 cells, 75% and 25% of channel estimate and estimation error.

The Multi-User downlink, particularly in a Multi-Cell Massive MIMO setting, requires enormous amounts of instantaneous CSIT (iCSIT). In [KTS+17] we have focused on exploiting channel covariance CSIT (coCSIT) only. In particular multipath induced structured low rank covariances are considered that arise in Massive MIMO and mmWave settings, which we call pathwise CSIT (pwCSIT). The resulting non-Kronecker MIMO channel covariance structures lead to a split between the roles of transmitters and receivers in MIMO systems. For the beamforming optimization, we considered a minorization approach applied to the Massive MIMO limit of the Expected Weighted Sum Rate. Simulations indicate that the pwCSIT based designs may lead to limited spectral efficiency loss compared to iCSIT based designs, while trading fast fading CSIT for slow fading CSIT. We also pointed out that the pathwise approach may lead to distributed designs with only local pwCSIT, and analyze the sum rates for iCSIT and pwCSIT in the low and high SNR limits. A simulation result is shown in the Figure 3-16 for 2 cells, 2 user/cell, 3 paths/user, M=10 transmit antennas and N=4 receive antennas.

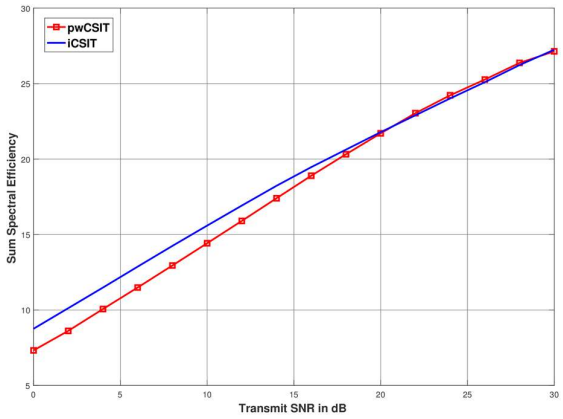


Figure 3-16 Expected sum rate comparison for M=10, N=4.

We consider an optimization with user rate fairness as system utility henceforth. Ensuring user fairness is one important criterion in designing wireless networks. Most of the fairness problems, which are closely related min-max or max-min optimization problems, are nonconvex and cannot be solved directly. In this context, several works (see the references in [GSY19]) have studied balancing optimization problems w.r.t. SINR and MSE. In this work [GSY19], we focus on user rate balancing in a way to maximize the minimum (weighted) rate among all the users in the network, in order to achieve network-wide fairness.

The user-rate balancing problem is studied in [RHL11], but without explicit precoder design. Our solution exploits the relation between user rate (summed over its streams) and a weighted sum MSE in a way to transform the max-min user rate into a min-max matrix-weighted user MSE. But also another ingredient is required: the exploitation of scale factor that can be freely chosen in the weights (user priorities) for the weighted rate balancing (see Appendix B.3).

The performance of our proposed scheme ("max-min user rate"), is evaluated through numerical simulations in a multiple-input-multiple-output broadcast channel scenario. We compare it to the min-max user MSE approach in [SSB08] where the MSE balancing problem is solved.

Figure 3-17 plots the minimum achieved per user rate using our max-min user rate approach with equal user priorities and the min-max user MSE [SSB08] w.r.t. the Signal to Noise Ratio (SNR). We observe that our approach outperforms significantly the per-user MSE balancing optimization, and the gap gets larger with more streams. In Figure 3-18, we illustrate how rate is distributed among users according to their priorities. We can see that, the rate is equally distributed between the users with equal user priorities, whereas with different user priorities, the rate differs from one user to another accordingly. Furthermore, the Sum Rate (SR) reaches its maximum when user priorities are equal, as the channel statistics are identical for each user.

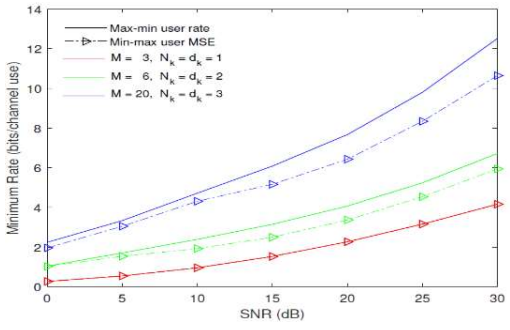


Figure 3-17 Minimum rate in the system VS SNR: K = 3 users, M number of transmit antennas, N_k number of receive antennas of user k, $d_k$number of streams for user k.

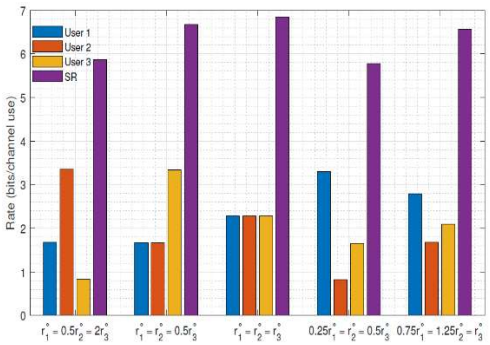


Figure 3-18 Rate distribution among users: K = 3, SNR= 10 dB, M = 6, $N_k = d_k = 2$ and r_k^o is the priority for user k.

3.5 Beam Management

Beam management procedures are applied to assign antenna beams to UEs. These antenna beams can be taken from a predefined grid of beams, or can be generated flexibly. The beams represent an additional spatial dimension of resources. For the detection of the most appropriate beams to connect BS and UE various procedures can be used. Beam selection, however, is impacted by channel quality, mobility, and available beam set. The design of beam management procedures therefore requires prediction of performance and assessment of selection accuracy under different channel conditions.

In the following subsections different enablers for the application of beam management procedures are proposed. Section 3.5.1 addresses the use of low resolution ADCs for mmWave digital beamforming. The investigations verify with link level simulations the reliability of analytical performance prediction for beam management design for different ADC resolutions. Section 3.5.2 focuses on the beam management requirements under different SNR conditions to select the correct beam in a fast and reliable way. For given misdetection probabilities the requirements for beam training time intervals and SNR ranges have been analysed.

3.5.1 Joint Investigation of UL Channel Estimation and MIMO Detection Regarding Robustness

Evaluation of digital beamforming systems with low resolution ADC in the past have mainly concentrated on either signal processing aspects, achievable rate or BER analyses. There have been analyses considering uplink wideband systems [MCL+17], multiple users in the uplink [JDC+17], or the effects of imperfect CSI at the receiver [LTS+17]. Different signal processing issues such as channel estimation [AAL+14] and MIMO equalization [WLW15], [MCL+17] have been analysed individually without considering all aspects of a receiver combined.

In this work we want to bridge the gap between a link level based simulation and an information theory based evaluation. The main target is to show that indeed we can approach the attractive data rates promised by analytic methods using standard signal processing techniques. In [RST16] the authors show a similar comparison for a point-to-point LTE system. Due to the limited space not all details are given. In this work we only describe the overview and the differences with our theoretical evaluation in [RPS+18] and the link-level evaluation in [RN18]. To make the situation comparable the same simulation assumptions were used for both cases. This comparison shows that indeed the data rates predicted for mmWave digital beamforming with low resolution ADCs translates well into link-level simulation results. At low per-antenna SNRs where such systems will likely operate, the performance of the link-level simulation can achieve the same data rate at about 4 dB lower SNR compared to the theoretical prediction. At high SNRs, the additional limitations of the constellation used in this work leads to a maximum data rate that is substantially lower compared to the theoretical prediction. The details of this evaluation can be found in [RPS+218].

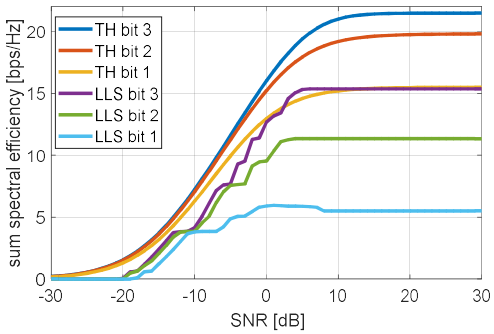


Figure 3-19: Comparison of link level simulation (LLS) and analytical (TH) rate results for a 4 user mmWave multipath system with imperfect CSI.

3.5.2 Channel Quality Estimation Sequence Design for Beam Management

Analogue or hybrid beamforming are considered to be possible solutions to reduce the power consumption of mmWave analogue front-ends. These solutions are based on the concept of phased array antennas. In this type of systems the signals of multiple antennas are phase shifted, combined and afterwards converted into the analogue baseband followed by an A/D conversion. To utilize the full potential of the system, it is essential that the beams of Tx and Rx are aligned.

In this case, we define the configuration of the analogue combiner/precoder as a single beam. A search procedure is used to align the beams of Tx and Rx [AD12], [OLL+16]. This beam search procedure does either utilize beams of different width with additional feedback or many beams of the same width with only one feedback stage [PDG+16]. In both cases, the beams with specific width, maximum gain and flatness need to be designed.

For the data communication following the beam-training it is essential that the optimal beam is selected. Selecting a non-optimal beam would lead to substantial performance degradation. In this context, we wanted to research the dependency of the channel quality measure on the sequence length and the SNR. This work provides an analysis of this situation and the results could serve as rough guidelines for selecting sequence length for practical systems. An example for selecting wrong beam depending on the power difference between the beams is shown in Figure 3-20 a). In this case, the parameter Δα defines the receive power difference in the digital baseband, after analogue combining between the two best beams in a codebook. The SNR is the one after the analogue combining for the beam with the highest power.

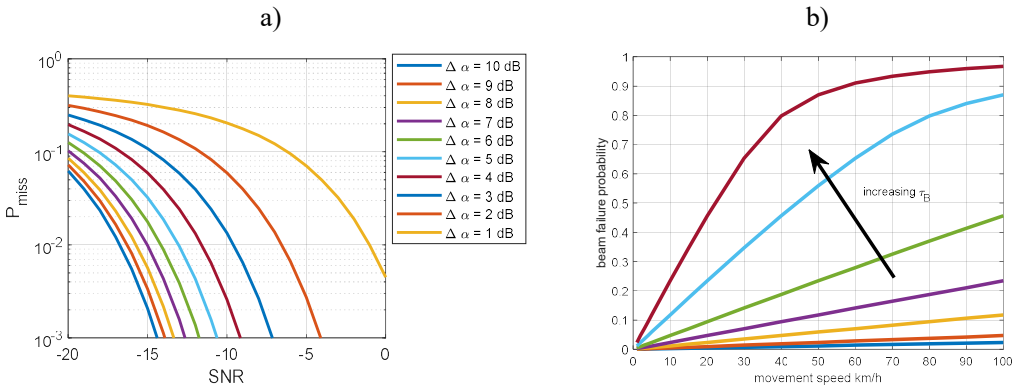


Figure 3-20: a) Probability of selecting the wrong beam with a sequence length of 573 and b) probability that of beam failure assuming maximum movement speed of 30 km/h and 10 m channel large scale parameter decorrelation distance and beam training interval τ_B of {10, 20, 50, 100, 200, 500, 1000} ms.

Another important aspect of beam-training is the time interval during which the training needs to be performed. This depends on the assumptions of the maximum mobility and the area that needs to be covered. For this evaluation we assume that the maximum speed of a device would be 30 km/h. This speed is modeling a streetcar, bus or car in a dense urban environment. We model a system where the mmWave is providing a high data rate connection to the vehicle, and the connectivity for the users inside the vehicle is provided via for example Wi-Fi or another technology. In the results in Figure 3-20 b) we see that with an upper bound of 10 % on the beam failure between two beam-trainings this time interval should be limited to about 100 ms.

3.6 Efficient Implementation: Hybrid Array Designs, Forward Error Correction, and Digital Frontend

In the past, energy efficiency of mobile networks has mainly been considered by the industry or academic research [EAR11-D42], [EAR11-D32], [GRE15], however, now it is even considered in 3GPP. A study item on the topic of UE power saving has started during the 4th quarter of 2018 [3GPP-181463]. For future mmWave systems power consumption is a major challenge for both device and access point [RRE14]. The successful deployment of truly massive arrays will depend on the development of new power-efficient implementations. Thus, overcoming this bottleneck would be an important step towards large scale adoption of future mmWave eMBB systems. Also, 3GPP NR is targeting a wide variety of services [NGMN15]. Supporting the requirements of all services with one hardware requires a flexible signal processing architecture.

In this section we target physical layer enhancements regarding flexibility and power efficiency. In Section 3.6.1 the size of the array and the related beam shapes are adapted to the number of simultaneously served UEs, this ensures the UE performance while power consumption of the gNB in low load condition is reduced. In Section 3.6.2 we show how a signal processing hardware architecture supporting multiple services with different requirements can be designed. Section 3.6.3 shows how to design mmWave hybrid beamforming precoders. As one of the major challenges for mmWave is the power consumption, Section 3.6.4 is comparing different receiver front-end architectures in terms of their energy efficiency.

3.6.1 Hybrid Array Architectures for Different Deployment Scenarios

Hybrid array architectures allow efficient hardware implementation which can minimize the power consumption needed to serve a certain number of MIMO layers. The power consumption of a massive MIMO base station is determined by the hardware driving a high number of antenna ports and antenna elements. The number of simultaneously transmitted MIMO layers has only a small impact on power consumption. In case of a low number of UEs (=simultaneous MIMO layers) the spectral efficiency is lower, but the power consumption does not scale in the same way. As a result the energy efficiency will be low ([ONE18-D41] section 3.3.2.). In recent investigations [HWW+18] we analyzed how the combination of the number of MIMO layers, the number of antenna ports and active transmitting elements, together with the specific array shape, influence the achievable performance and related power consumption.

The array shapes A, K and L and the related average UE throughput versus the number of simultaneously served UEs is shown in Figure 3-21. In this example an average UE throughput of 20 Mbit/sec with 12 UEs (point a) and array A can be maintained with eight active UEs with array K (point b). With only four active UEs even array L is sufficient (point c) to maintain the same average UE throughput.

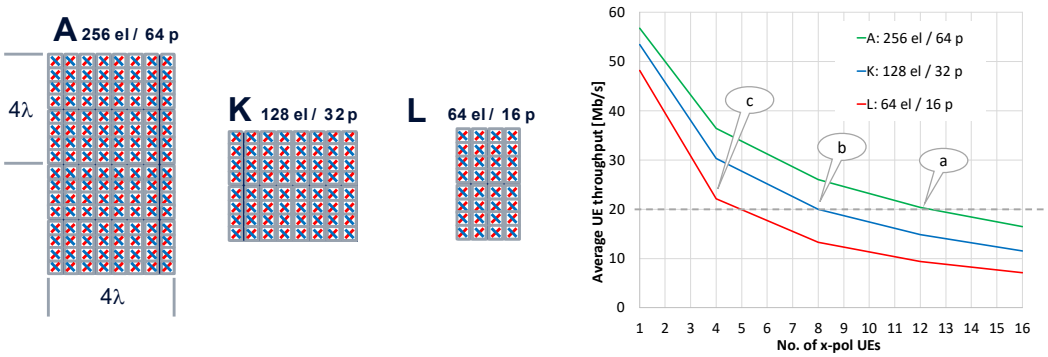


Figure 3-21: Array shapes A, K and L, and average UE throughput vs. no. of simultaneous UEs.

The applied simple power consumption model shows that the power consumption is strongly reduced when using the smaller arrays. The model comprises Tx and Rx conversion units (scaling with number of antenna ports), PA on transmit side and Low Noise Amplifier (LNA) on receive side (scaling with number of antenna elements), and filter and splitter losses. In Figure 3-22, the relative power savings for the different array types are shown and the contributions of the different building blocks are indicated. The PA is the most dominant contribution to the overall power consumption

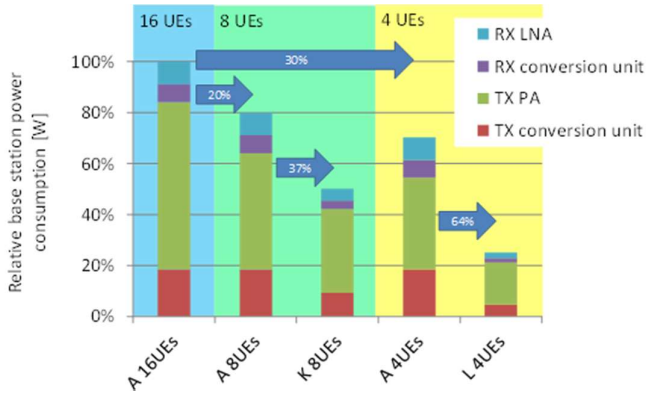


Figure 3-22: Relative power consumption for array types A (=100%), K and L and no. of UEs

Although with the large array A the power consumption reduces for 8 and 4 simultaneous UEs (20% and 30%, resp.), the additional power reduction using the smaller arrays K and L is significantly larger (additional 37% and 64%, resp.). The array adaptation according to the traffic conditions can increase energy efficiency of a practical system significantly.

3.6.2 Flexible and Fast Reconfigurable HW Architecture for Multi-Service Transmission

NR 5G aims at addressing, for the same user, several kinds of services simultaneously. This goal requires having a very fast reconfigurable HW architecture allowing a very flexible HW implementation. As already introduced in [ONE18-D41], two main HW components are optimized to be implemented for real time communications in the framework of the ONE5G WP5. The first one is the Digital Front End (DFE) that adapts the data stream to the antenna taking into account the RF impairments. The second one is the Forward Error Correction (FEC) that adapts the codeword length to the service to be transmitted while keeping the background compatibility with the LTE standard. The strength of the FEC component is to be able with the same instance to deal with either LTE (turbo code) or NR 5G (LDPC) with a same architecture (only by playing with its reconfiguration).

The DFE is studied at the reception level. It targets the filtering of the RF bands and the reduction of the sampling rate. According to the underserved areas verticals, that needs long

range coverage and low data rate, the DFE studies are focused on NB-IoT standard. The reception part aims at the selection of the NB-IoT carriers from the ADC samples. This operation is summarised in Figure 3-23. It encompasses the following processes.

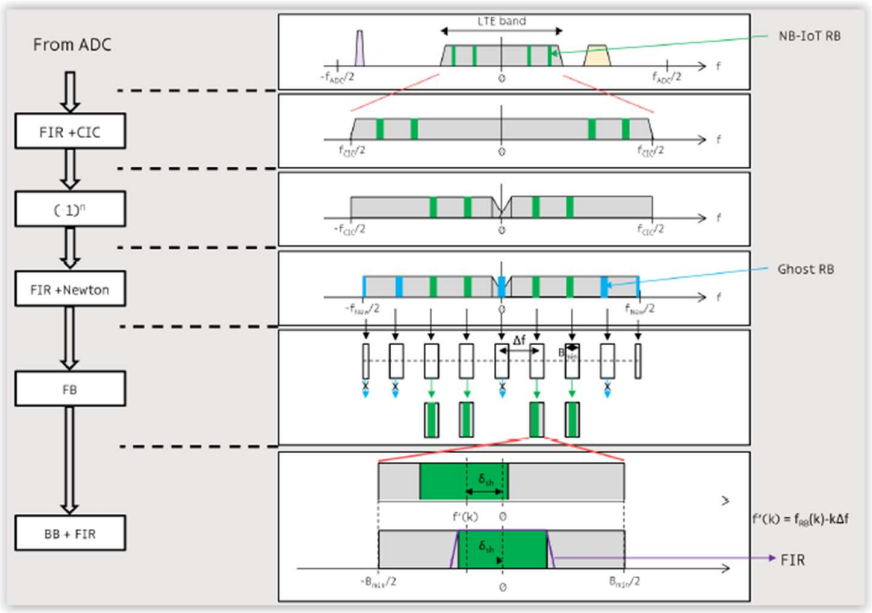


Figure 3-23: NB-IoT Digital Front End – Rx side

- The ADC signal is scaled with an Automatic Gain Control;
- The signal is filtered on two stages;
  - A first combination of a Cascaded Integrator Comb (CIC) and a FIR filter adapts the bandwidth to the LTE band;
  - A second operation provides a fine adaptation of the sampling rate;
- Each NB-IoT resource block is extracted and converted in Base-Band;
- Each NB-IoT channel is filtered.

Forward Error Correction

The FEC is studied at the reception level (decoding part). It lowers the probability of errors at the reception. According to the smart-Megacity verticals, that is more high-data rate oriented, the studies are focused on multiple services aggregating different channels (5G-NR PUSCH, 5G-NR PDSCH) and ensuring 100 Mbps throughput, and max 1 ms latency duration, so that it is adapted to massive MIMO demonstrator.

The design switches its configuration in few clock cycles that induce low latency at the physical layer level. Such a process allows using the same hardware circuitry to provide an aggregation of service with no degradation of the Quality of Experience.

The FEC encompasses both transmission and reception part:

- The dataflow coding (according to [3GPP-38212])
- The system rate adaptation strategy (rate matching or Modulation and coding scheme) according to [3GPP-38214]
- The mapping of the stream (BPSK, QPSK, 16QAM, 64QAM)

This study is WP5 oriented. Several Physical layers (WiFi, NR) have been merged in one OFDM system for real time communication. Some performance curves are available in [ONE19-D52]. The FEC component aggregates mapped symbols and decodes it according to the corresponding communication system. In standalone, each system reaches hundreds of Mbps (depending on the configuration of the system).

3.6.3 Genetic Algorithm Assisted Hybrid Beamforming for Wireless Fronthaul

The application of wireless fronthaul provides enhanced flexibility to support the 'Live Event Experience' use case in WP2. This topic uses Genetic Algorithm (GA) to design the hybrid precoder for wireless fronthaul. The resolution-limited phase shifters in the analogue precoder are optimized using the GA, while the digital precoder is computed based on the maximum signal to leakage plus noise ratio criterion. Since the analogue precoder only uses one-bit or two-bit quantization, it largely reduces the complexity of system design. It has been demonstrated via simulations [ONE18-D41], [Wu18a] that even with one/two-bit resolution phase shifters, after approximately 150 iterations, the GA assisted hybrid beamforming can still achieve 85-90% performance of a fully digital beamforming system. The results benefit from the quasi-static properties of the fronthaul channels. The proposed method aligns with the WP4 objective of reducing the cost and saving energy in 5G system design. Details of the proposed design and results are given in [ONE18-D41], [Wu18a].

3.6.4 A Comparison of Hybrid Beamforming and Digital Beamforming with Low-Resolution ADC's for Multiple Users and Imperfect CSI

For the success of 5G it will be important to leverage the available mmWave spectrum. To achieve an approximately omnidirectional coverage with a similar effective antenna aperture compared to state-of-the-art cellular systems, an antenna array is required at both the mobile and base station. Due to the large bandwidth and inefficient amplifiers available in CMOS for mmWave, the analogue front-end of the receiver with a large number of antennas becomes especially power hungry. Two main solutions exist to reduce the power consumption: hybrid beam forming and digital beam forming with low resolution ADCs. In this work we compare the spectral and energy efficiency of both systems under practical system constraints. We consider the effects of channel estimation, transmitter impairments and multiple simultaneous users for a wideband multipath model. Our power consumption model considers components reported in literature at 60 GHz. In contrast to many other works we also consider the correlation of the quantization error, and generalize the modelling of it to non-uniform quantizers and different quantizers at each antenna. The result shows that as the SNR gets larger the ADC resolution achieving the optimal energy efficiency gets also larger. The energy efficiency peaks for 5 bit resolution at high SNR, since due to other limiting factors the achievable rate almost saturates at this resolution. We also show that in the multi-user scenario digital beamforming was for all evaluated cases more energy efficient than hybrid beamforming. In addition, we show that if mixed ADC resolutions are used we can achieve any desired trade-off between power consumption and rate close to those achieved with only one ADC resolution. Details of the evaluation can be found in [RPS+18].

3.7 Optimized Array Formats and Capacity Analysis

The performance of multi-user massive MIMO systems highly depends on the UE distribution, the propagation environment, and the antenna geometry. Wider antenna arrays yield a finer angular resolution. Thus, depending on the deployment scenario, the chosen antenna array format can have a large influence on the capacity. ONE5G has analysed the system performance when using different types of antenna arrays, namely Uniform Planar Arrays (UPAs) and Uniform Cylindrical Arrays (UCAs) (the standardization process of 5G-NR is today highly focused on sectorised UPAs, which are also well studied in the literature). However, the array only covers a third of the angular span (120°) and the array gain is degraded at the boundaries of this range. Three sectorised antenna arrays are then necessary to cover all directions. Whereas UCAs provide a more uniform coverage of the deployment site and do not require sectorisation.

Section 3.7.1 provides simulations based on the 3GPP channel model [3GPP-38.901] using different formats of UPAs. The results suggest that, for UMa and UMi scenarios, better UE separation is achieved using wider antenna arrays, i.e. more rows than columns of vectors.

In the previous deliverable [ONE18-D41], extensive simulations showed a capacity gain using UCAs compared to sectorised planar arrays. This work is extended in Section 3.7.2 using analytic formulas instead of computational intensive simulations.

Section 3.7.3 considers the impact of different array formats on channel hardening by analytical means and especially highlights the impact of the angular spread in the elevation domain as well as the relation to the inter-antenna spacing.

The general trend towards flexible networking is conflicting with today's rigid sectorisation that dictates certain preferred directions, generates inter-sector interference and requires procedures for inter-sector handoff. Using antenna arrays adapted to the deployment scenario can then greatly simplify both signal processing and radio resources management. Indeed, planar arrays are well-suited for direction specific scenarios whereas circular ones are more efficient in direction agnostic scenarios with widely variable UE distributions.

3.7.1 Impact of Array Format in Different Deployments

In massive MIMO systems, the shape of the BS array must be adapted to the deployment scenarios and to the UE distribution. We performed extensive system simulations considering the three-dimensional spatial channel model proposed by 3GPP [3GPP-38.901] to understand which BS array configuration maximizes system performance, with a specific focus on UPAs. Numerical results suggest the usage of wide arrays, i.e., with more columns than rows, instead of tall arrays: more specifically, simulations with 32 cross-polarized antennas at the BS show that a UPA with just 1 row (and 32 columns) outperforms UPAs with 2, 4 and 8 rows. That happens because in 3GPP modelling UEs are distributed over the azimuth rather than in the elevation domain: indeed, the BS-UE LOS link spans an azimuth range of 120°, but an elevation range of just 25° and 40° in UMa and UMi scenarios, respectively. This work has been finalized in the previous deliverable and more details can be found in [ONE18-D41].

3.7.2 Sector and Beam Management with Cylindrical Antennas

ONE5G has compared cylindrical antenna arrays (UCAs) and triple-sectorized uniform planar arrays (UPAs) for spatial multiplexing downlink multiple-user transmission [KMT+18], [ONE18-D41]. Assuming independent precoding and transmission per sector for both the UPAs and the UCA in a cellular outdoor environment, the UPA outperforms the UCA due to the inherent interference coordination of the UPA geometry. However, this changes when joint precoding and transmission over all antennas of the UCA and sectorised UPAs is considered, then the UCA outperforms the UPAs. A legitimate concern of reviewers and partner was that by the single BS assumptions and precoding over all antennas, intra-sector and inter-cell interference power is zero and the gain from the UCA may vanish in a full cellular deployment. This issue was studied in the last part of the ONE5G project and the main outcomes are presented in the next paragraph while a peer-reviewed companion publication is in preparation.

The triple-sectorized UPAs and UCA are compared in a 21 BS scenario mainly following 3GPP system level simulation (SLS) assumptions. The full parameter list is given in Appendix B.1. We emphasize that the total number of antennas and transmit power is the same in both configurations. Also the half-power beamwidth (HPBW) of the UCA has been optimized beforehand. Figure 3-24 compares the UPAs and UCA by the CDF of the user received SINR. 30 users are spatially multiplexed by an MMSE precoder jointly over all antennas. It is observed that the UCA outperforms the sectorized UPA e.g. by 1.1 dB SINR gain at the 50^{th} percentile. Compared to single BS, the gain has reduced somewhat, however there are still several parameters for optimization in the UCA deployment that have not be addressed in this work, while the UPA deployment has been studied and optimized extensively in the past two decades.

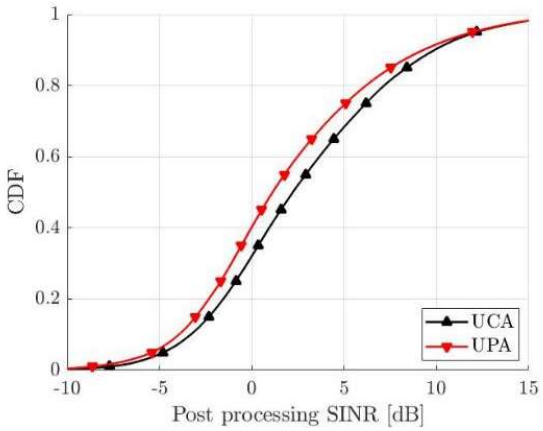


Figure 3-24: CDF of receive SINR in [dB] per RB per user comparing triple sectorized UPAs with UCA. There are 30 users spatial multiplexed in a full system level simulation. Other parameters are listed in Appendix B.1

The results of the system level simulations motivate further research into the topic of use-case customized design of antenna deployments to achieve certain target KPIs [ONE19-D32]. This topic has been ignored and is still neglected in 3GPP standardization, because it not only requires changes in the deployment of the antennas, but also other changes, e.g. in the codebook design, see [KMT+18], to show gains over existing deployments.

3.7.3 MIMO Performance Prediction

Channel Hardening

When increasing the number of antennas of the transmitters in a MIMO system, the underlaying normal distribution hypothesis of Gaussian (correlated or uncorrelated) models falls. That is why parametric ray-based models, that do not require Gaussianity, become increasingly relevant when considering wider antenna arrays and massive MIMO systems. One of the advantages of massive MIMO is called *channel hardening* [BHS17], [NL17] which consists in a reduction of multipath small-scale fading averaged over the multiple transmit and receive antennas, mitigating the small-scale variations of the channel capacity. In the literature, the channel hardening effect is characterized by the coefficient of variation of the channel gain. In ONE5G, we propose a new formula based on a realistic ray-based channel model. We have tested our formula against those of the literature (based on correlated or uncorrelated Gaussian channels, green and black in Figure 3-25). The channels were generated with the Quadriga simulator, a radio channel simulator developed by HHI. Our analytic formula is the closest to the simulation results. More details can be found in the previous deliverable [ONE18-D41] and in [RPM+18].

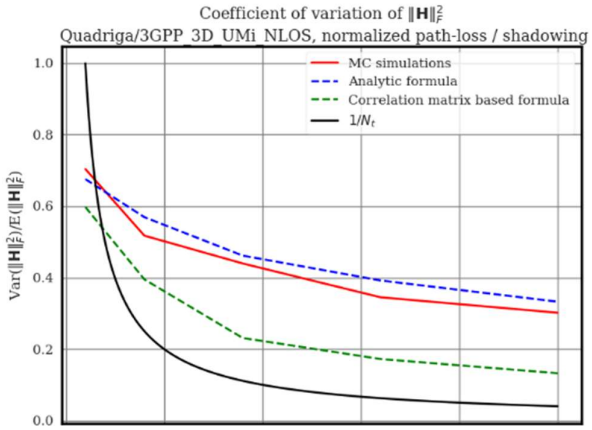


Figure 3-25: Channel Hardening evaluation on a realistic scenario.

Frequency Domain Degrees of Freedom of the MIMO Channel

We have analysed the number of time domain coefficients that are required to model the Channel Impulse Response (CIR) of a MIMO-OFDM channel based on two approaches. First each MIMO channel matrix coefficient is modelled as a taped delay line. This approach is more efficient for small-scale MIMO setups with reduced amounts of antennas. For wider arrays, we modelled each frequency-dependent cluster as a tapped delay line. In both cases we studied the number of taps required to achieve a certain error threshold and chose the more efficient one. The goal is to find out when a classical channel matrix approach is sufficient and when a novel approach exploiting the angular channel sparsity is preferable. We found out that the boundary is around $N_r N_t \approx 16$ for indoor scenarios which is easily attained if multiple receive antennas are used.

3.8 Conclusion

In this section massive MIMO enablers are addressed, differentiating between deployment-oriented and implementation-oriented ones. Deployment-oriented aspects such as pilot interference mitigation, user distribution and application adaptive antenna geometries, and solutions for in-band wireless backhaul provide significant gains compared to existing techniques or provide solutions that are not covered by current standards. These deployment-oriented enablers target both ONE5G scenarios, the "*Megacity*" and "*Underserved area*", while covering all the five use-cases in [ONE17-D21], for example the "Live Event Experience".

The implementation-oriented enablers focus on complexity reduction and efficiency improvements of massive MIMO. The presented techniques on training and feedback overhead reduction to obtain accurate CSI at the transmitter for phase adaptive downlink precoding show promising gains for multi-user data-transmission. These technique exploit the channel structural properties, e.g. sparse angle-delay channel representation or low-rank channel covariance matrix, in a variety of scenarios. Furthermore, precoding techniques for hybrid beamforming in the mm-wave frequency range are improved. Beside improvements of well-known techniques, also "disruptive" solutions on distributed beamforming and functional split of partially centralized RAN are studied in this section as attractive alternatives to the classical centralized approach. Finally, studies on efficient implementation show that massive MIMO gains can also be realized with low complexity solutions, e.g. low resolutions ADCs in the presence of imperfect CSI.

3GPP NR Rel. 15, the first standardized version of 5G, is only an intermediate step on the long journey that brings massive MIMO into reality. The massive MIMO enablers presented in this section go clearly beyond 3GPP and provide essential contributions along the massive MIMO journey.

Table 3-2. Summary of key recommendations and benefits in terms of Massive MIMO Enablers towards Practical Implementation

Feature	Recommendation	benefits
3.1.1 TDD: Improving CSI Acquisition through Spatial multiplexing	Clustering the users according to their spatial signatures and allocating the pilots to the formed clusters using the proposed spatial basis coverage allocation provides a huge performance improvement.	Results show that one can increase the spectral efficiency by 2/3 with respect to a baseline massive MIMO (with no interference management, random pilot assignment, pilot reuse in all cells).

3.1.2 FDD/ Improving CSI Acquisition through Spatial multiplexing	Grouping the users according to their channel covariance matrices and scheduling the groups for CSI feedback allows achieving a tremendous gain in network throughput	Results show that the achieved throughput can be doubled as compared to a baseline conventional massive MIMO (with no interference management).
3.1.3 Pilot Allocation taking into account Markovian Channel Model and Traffic Patterns	Scheduling the users for CSI acquisition taking into account the channel time correlation and the traffic arrival achieves an important performance gain.	Results show that one can increase the spectral efficiency by 14% with respect to a baseline massive MIMO where all users transmit their pilots all the time
3.1.4 Fractional Power Control to Mitigate Pilot Contamination in 5G Massive MIMO	Uplink FPC is fundamental to mitigate pilot contamination in massive MIMO systems.	Very high gains (up to 350%) can be achieved in the cell border throughput by using FPC when compared to noPC.
3.2.1 Parametric Channel Estimation for Massive MIMO	Using a physical description of the channel is beneficial for efficient channel estimation. It is possible to generalize steering vectors to take into account very large antenna arrays close to the users.	In average 40% decrease in mean squared error (MSE) compared to the classical least squares method and to the plane wave model in an UMi environment with a ULA of 256 antennas at a height of 5 meters and users randomly located in the adjacent street.
3.2.2 Hierarchical Sparse Channel Estimation for Multiuser Massive MIMO with Reduced Training Overhead	Training overhead reduction of uplink multiuser massive MIMO can be achieved via a sophisticated training design and computationally efficient channel estimation algorithms exploiting the higherarchical sparsity of the wireless channels.	For an asymptotically large number of antennas and bandwidth, the proposed algorithm achieves reliable channel estimation with a bandwidth overhead that is almost an order of magnitude smaller than that required by conventional approaches and is independent of the number of propagation paths (per user).
3.2.3 Wideband Massive MIMO Channel Estimation via Atomic Norm Minimization	For operational conditions with a limited number of propagation paths (e.g., mmWave communications), superesolution techniques such as atomic norm minimization can be applied to achieve near-optimal and low-overhead channel estimation in uplink massive MIMO.	Compared to standard LMMSE channel estimation and for a very sparse channel (3 paths), the proposed algorithm provides a channel estimate MSE that is orders of magnitude smaller and with less than 50% of the training overhead.
3.2.4 On the amount of DL training in correlated massive MIMO channels	Careful design of training sequences and their number according to the operating SNR and the channel (spatial) covariance may significantly reduce the DL training overhead in FDD multiuser (massive) MIMO scenarios. A small number of fed back channel covariance eigenvectors from the user to the BS helps the latter in this design.	Proposed scheme results in accurate channel estimates but with a large reduction in training overhead. This translates to larger effective throughput gains. Exact reduction depend on the covariance structure and number of users, but for typical covariance matrices and number of served users, the training overhead reduction can exceed 50% compared to state of the art methods.

3.2.5 Efficient Feedback Schemes for more Accurate CSI and Advanced Precoding	Explicit time domain based CSI feedback can provide forward compatibility to advanced MIMO concepts in future releases. In addition, due to the time domain sparsity, better overhead reduction can be achieved with time domain compression.	Proposed time domain based explicit CSI feedback scheme can achieve 8% higher spectral efficiency compared to Rel. 15 NR type II CSI while saving 16% of the UL overhead.
3.3.1 Multicast Massive MIMO	ONE5G recommends the use of beamforming based MIMO-multicast for group-wise transmission to spatially distributed UEs instead of time-shared unicast, at least in the V2X and IoT use cases (both industrial and mMTC).	The achievable multicast gains depend on the UE distribution and also on the number of UEs. It was shown that the throughput can be increased many times over.
3.3.2 Wireless backhaul for coverage enhancement in low ARPU network	ONE5G proposes a new precoding scheme for wireless backhaul link which takes into consideration also the potential interference coming from an independent but parallel access link.	The RZF-CI precoding scheme, as shown in D4.1, can provide better (up to 8 dB) link budget in terms of received SNR for wireless backhaul link compared to ZF precoding. At the same time, the loss is limited to less than 1.5 dB compared to the ideal MRT precoding. It is also confirmed in D4.2 that the performance is stable in both synchronized and non-synchronized situation between backhaul and access links. Thus, the RZF-CI scheme can enable cost reduction for the deployment
3.3.3 Signal Shaping for MIMO Backhaul Channels	Signal Shaping can improve the coded modulation performance for high order modulation for AWGN and fading channels, also in high throughput scenarios and short transmission frames.	Shaping achieves 1dB SNR gain. By reducing the sequence length, parallel processing of 5 sequences is facilitated without performance degradation. This leads to higher throughput and lower latency of the shaping encoder by a factor 5 compared to the state-of-art Constant Composition Distribution Matcher.
3.4.1 Beamforming Design and Function Split for Partially Centralized RAN with Massive MIMO RRH	The purpose of CRAN architecture is to keep minimal functionality at the RRH in order to i) benefit costs and scalability, ii) alleviate transport between RRHs and BBU, and iii) facilitate CSI acquisition. This can be done by performing analogue beamforming at the RRHs, based on second order statistics estimated at the BBU. The low-dimensional CSI observed through analogue beamforming is then employed for digital beamforming at the BBU.	The proposed scheme decreases the complexity due to reduction of RF chains and due to the smaller number of instantaneous channel coefficients to be estimated. However, it incurs a performance degradation in terms of sum-rate compared to a fully digital implementation, which can range between 4-16% for the case of 16 active RF chains.
3.4.2 Beamforming Algorithms for System Utility Optimization toward Massive MIMO	With partial CSIT under the sum rate optimization (SRO), we should take into account the channel estimation and the estimation error covariance together to make the beamforming design; and we can also use the local	With partial CSIT under SRO, the sum rate versus SNR performance has not a saturation floor with increase of SNR contrary to others existing techniques; by using the local pathwise CSIT, the CSIT

	pathwise CSIT. For the user rate balancing, a new proposed matrix-weighted user MSE approach provides beamforming expressions for MU-MIMO.	acquisition can slow down with limited spectral efficiency loss. For the user rate balancing, the balanced rate outperforms the minimum rate obtained by user MSE balancing of 15% with 15dB; and the implementation on software or HW of our algorithm is straightforward.
3.5.1 Joint Investigation of UL Channel Estimation and MIMO Detection Regarding Robustness	In the case of severe non-linearities of the analog front-end, always jointly (UL channel estimation and MIMO detection) design the receiver signal processing architecture.	Reduced computational complexity. Gain of about 20 % relative to linear receiver.
3.5.2 Channel Quality Estimation Sequence Design for Beam Management	Optimize the beam-training sequence length and as well as beam-training interval to adapt to the communication requirements.	The beam-training overhead is minimized. A gain of up to 40 % relative to 802.11 ad beamtraining procedure is possible.
3.6.1 Hybrid Array Architectures for Different Deployment Scenarios	Flexible adaptation of array shape and size according to variations in deployment and traffic load conditions.	Increased energy efficiency for massive MIMO operation in varying load conditions, applicable to various deploy-ment scenarios. More than 60% energy saving compared to full size array in low load conditions possible.
3.6.2 Flexible and Fast Reconfigurable HW Architecture for Multi-Service Transmission	Flexible hardware component is required to address a 5G Real time communication which addresses several services, with no disturbance. Developed algorithms take into account each service's specification.	The component runs on FPGA at 250 MHz. Near antennas, the DFE handles IQ samples with a reduced processing latency. The loopback Transmission / Reception is processed in 20 µs. The upper layer receiver, containing generic mapping and FEC deals with variable packets length for up to 300 Mbps. Thanks to pipelining, the context switches from 8 ns (when the block is unoccupied) to 524 ns (when emptying pipelines). Services are deserved in a transparent way to the user. Mutualization of the HW architecture for multiple services preserves HW resources by factor 2/3 with equivalent throughput and latency as dedicated one.
3.6.3 Genetic Algorithm Assisted Hybrid Beamforming for Wireless Fronthaul	To reduce hardware cost, it is recommended to deploy low-resolution limited-RF-chain hybrid beamforming antenna arrays with the proposed algorithm to wireless backhaul and fronthaul. Future research and development directions can include hardware implementation.	In SNR regime -15dB to 0dB, 2-bit resolution phase shifters can achieve 93% of fully digital beamforming performance. In SNR regime 5dB to 15dB, 2-bit resolution phase shifters can achieve 95% of fully digital beamforming performance.

3.6.4 A Comparison of Hybrid Beamforming and Digital Beamforming with Low-Resolution ADC's for Multiple Users and Imperfect CSI	As our evaluation showed that especially in the low per antenna SNR region the digital beamforming system has substantial spectral and energy efficiency benefits, this system architecture should also be considered for future mmWave communication systems.	Improved energy efficiency or coverage. Gain of 50 % relative to hybrid beamforming solutions.
3.7.1 Impact of Array Format in Different Deployments	With massive MIMO, the shape of the BS array must be adapted to the deployment scenarios and to the UE distribution.	In UMa (probably the most relevant scenario for massive MIMO) wide arrays strongly outperform tall arrays: for instance, a 1x32x2 array provides a gain up to 7 dB when compared to an 8x4x2 array in the median of the UE SINR.
3.7.2 Sector and Beam Management with Cylindrical Antennas	Widely used UPAs, e.g. in 3GPP, are not the best antenna deployment for some use-cases or scenarios. With massive MIMO, ONE5G recommends a use-case customized design of antenna deployments to achieve desired target KPIs for both cellular and non-cellular scenarios.	In single BS scenario the UCA provides a more homogeneous SNR/SINR in the horizontal-plane and improves reliability (5 % user throughput) compared to state-of-the-art (3GPP) sectorized uniform planar array (UPAs) by approximately 15 %. In multiple BS scenario (SLS) the average SNR is increased by 3 dB.
3.7.3 MIMO Performance Prediction	State of the art formulas for SNR variations prediction doesn't work well when considering realistic propagation scenarios. We introduced a new formula that accurately measures the variations of the SNR depending on both antenna array topologies and environment.	Up to 4 times gain in SNR variations prediction accuracy in simulated scenario over baseline correlation matrix approach. SNR is a key input for scheduling and resource allocation at system level. Accurate prediction of the variations of the SNR can enhance scheduling choices.

4 Advanced Link Management Based on CRAN/DRAN, and massive MIMO

The increasingly dense network infrastructure expected for "*Megacity*" scenarios, as well as advancements in cloud computing, strongly motivate CRAN deployments where, in principle, all transmission-reception points (TRPs) -equivalently referred to in the following as APs or RRHs- are controlled by a CU with every UE in the system served by multiple TRPs. Aided by network state information at the CU side regarding, e.g., link qualities and spatial traffic distribution, CRAN promises significant performance gains by means of sophisticated signalling and scheduling schemes.

The potential of CRAN has been recognised by 3GPP where the issue of functional split, i.e., which functions of the complete RAN protocol stack will reside at the CU side and which functions reside at distributed units (RRHs), has received significant attention [ONE18-D31]. However, CRAN operation introduces many challenges and there is currently no clear view on how the full potential of multi-connectivity gains can be reached. This is also reflected by the recent specification of the, so called, lower-layer functional split by 3GPP, which, instead of specifying a single architecture, allows for a multitude of options to be proprietary selected by operators and vendors [3GPP-38.816].

This section summarises the major technical achievements of ONE5G that address and provide solutions for some of the major challenges of CRANs, towards achieving its full multi-connectivity potential. Section 4.1 considers the issue of efficient CSI acquisition towards enabling sophisticated cooperative transmissions and tracking of interference levels in the network. Section 4.2 provides a set of tools towards reducing cross-link and/or cross-mode interference, an issue that is critical in dynamic TDD and sidelink (D2D) communications, mechanism that are considered in NR. Section 4.3 provides a set of efficient signal processing and scheduling solutions for harvesting the gains offered by cooperative transmissions, towards the ultimate goal of cell-less (cell-free) communications that eliminates the interference limitation experienced by the conventional cellular network. The important topic of functionality placement/split is treated in Section 4.4, where an efficient algorithm for optimal and dynamic allocation of computational resources among CU and RRHs is proposed.

4.1 CSI Acquisition for CRAN

The full potential of CRAN, enabled by advanced joint transmission and decoding as well as network-wise resource allocation and scheduling, can only be achieved when accurate global CSI is available. However, in dense CRAN deployments, estimation and feedback of CSI (in short, CSI acquisition) becomes challenging. This is because the number of TRPs associated with each UE can potentially be much greater than one or two (as in the case of conventional cellular networks). This naturally implies a significantly increased number of channels that need to be accurately estimated, and, in turn, an increased overhead dedicated for channel training and feedback purposes. Additional overhead may also be introduced by the need to measure and track interference levels experienced at the receiver side for optimal link adaptation.

In NR Rel-15, several advancements have been implemented regarding CSI acquisition procedures such as improved feedback codebooks for MIMO channels [3GPP-38.211] and the introduction of the quasi-co-location concept, which is particularly suited for network coordination/cooperation schemes [3GPP-38214]. However, even though Rel-15 specifications are, in principle, applicable in a CRAN setting, they are not efficient for obtaining global CSI as they are not optimized towards multi-connectivity scenarios where a UE can, in principle, associate to an arbitrary number of TRPs.

In this section, the challenge of CSI acquisition in CRAN deployments is investigated with the aim of both understanding fundamental system performance aspects as well as propose

improvements and new features to 3GPP specifications. In Subsection 4.1.1, an analytical characterization of FDD downlink CRAN performance with low-overhead training sequences is described, providing insights on the trade-off between performance and training overhead. Subsection 4.1.2 proposes a novel quantization scheme that significantly improves the current state of the art, as considered in 3GPP Rel. 15, for CSI feedback. Finally, Subsection 4.1.3 proposes a novel signalling approach utilizing the tools available in 3GPP Rel. 15, towards compensating cross-link interference by advanced receiver processing.

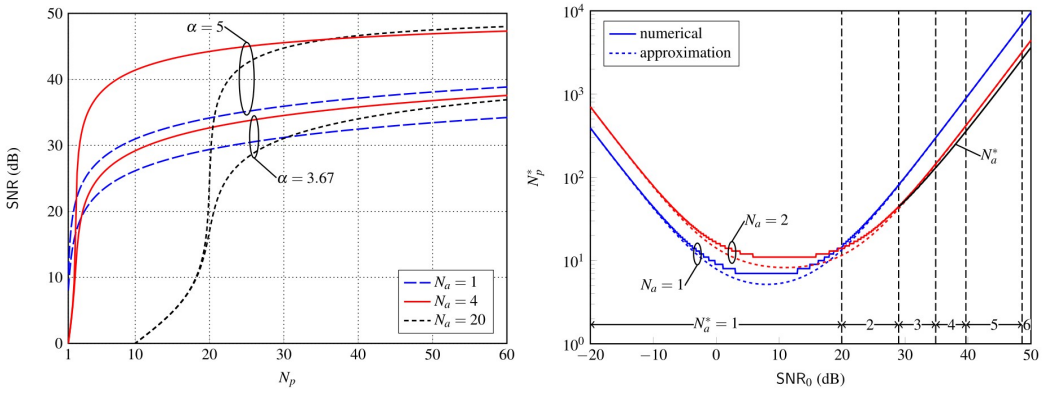
4.1.1 CRAN Performance under Low-Overhead Channel Estimation

In a dense CRAN deployment, the conventional training design based on orthogonal sequences may be unacceptable as it requires an overhead that is proportional to the number of RRHs in the system. This observation naturally suggests the investigation of the potential of low-overhead (and, necessarily, non-orthogonal) training sequences for CSI acquisition. An analytical performance characterization of channel estimation performance under this approach was provided in [ONE18-D41] and [SW18], taking also into account the randomness of RRH positions in a practical deployment. However, that study focused only on the channel estimation Mean Square Error (MSE), which does not provide a clear representation of the system performance during the data transmission stage.

Towards obtaining insights on how channel estimation errors affect the data transmission performance, the downlink performance of a FDD CRAN deployment serving a random UE is investigated. In particular, it is assumed that during the data transmission stage, the UE is served by a cluster of its $N_a \geq 1$ closest (in distance) RRHs, with each RRH transmitting the same symbol but without any precoding. This transmission scheme is chosen due to its tractability (no joint data precoding is required) and is also considered for Multimedia Broadcast multicast service Single Frequency Network (MBSFN) transmissions in 3GPP [3GPP-38.300]. In addition, it is assumed that the UE has no prior information about channels and RRH positions, reflecting a scenario where the overhead to obtain this information cannot be afforded. The effective SNR is considered as a figure of merit, which captures the effect of channel estimation errors in data decoding at the UE side. Under certain reasonable assumptions on channel model and training sequence design, the average (over channel fading and RRH positions) effective SNR is analytically computed. This analytical formula enables the (joint) optimization of critical design aspects such as RRH cluster size N_a and training overhead N_p (in channel uses).

Figure 4-1 a) shows the average SNR achieved with various cluster sizes N_a as a function of the training overhead N_p and for two values of the propagation path loss factor α. The results shown correspond to an RRH density and transmit power that result in an average SNR_0=40 dB under perfect CSI and $N_a = 1$. As expected, the larger the training overhead, the greater the SNR due to the more accurate channel estimates. Note that even with small N_p, cooperative transmissions ($N_a > 1$) are preferable over conventional cellular operation ($N_a = 1$), with the achieved SNR gain more prominent for large path loss factors. Figure 4-1 b) shows the minimum training overhead N_p^*, required to achieve a performance that is only 1 dB smaller compared to SNR_0 (SNR achieved under perfect CSI and $N_a = 1$) for various values of N_a. The path loss factor was set to $\alpha = 3.67$. It can be seen that for small SNR_0, the conventional cellular network with $N_a = 1$ requires the smaller overhead. However, for large SNR_0, minimum training overhead is achieved with $N_a > 1$, with the cluster size requiring the least training overhead (N_a^*) increasing as SNR_0 increases.

The full details of this work along with additional insights can be found in [SW19].



(a) (b)

Figure 4-1: a) SNR performance as a function of training overhead for various values of cluster size and path loss factor, b) minimum training overhead required to achieve an SNR that is 1dB less than the case of ideal CSI and association with only the closest RRH.

4.1.2 Enhanced CSI Feedback and Downlink Control Channel Transmission in NR

In this section, we present the recent development on enhanced CSI feedback and downlink control channel transmission schemes in 5G-NR. Specifically, amplitude quantization scheme based on NR Type-II codebook has been developed to improve CSI feedback quality due to amplitude quantization errors.

Amplitude quantization for Type-II codebook based CSI feedback

In 3GPP NR system, two codebook types, namely, Type-I and Type-II, have been standardized for the CSI feedback in the support of advanced MIMO operation [3GPP-38.211]. Both codebook types are constructed from a two-dimensional DFT based grid of beams, and enable the CSI feedback of the selected beams as well as PSK-based co-phase combining between two polarizations. Type-II codebook-based CSI feedback is more elaborate than that of Type-I, reporting both the Wide-Band (WB) and Sub-Band (SB) amplitude information of the selected beams. As a result, Type-II codebook-based CSI feedback provides a more accurate CSI at the transmitter side leading to an improved precoded MIMO transmission performance.

To reduce the CSI feedback signalling, Type-II codebook feedback utilizes 1 bit (corresponding to two quantization levels) for each SB amplitude in addition to the 3 bits (corresponding to eight quantization levels) considered for WB amplitude feedback. Typically, the WB amplitude value is obtained as a simple linear average of the quantized SB amplitudes. However, due to the coarse SB amplitude quantization, the resulting WB amplitude is prone to significant quantization errors.

In ONE5G (see [MMF18] for details), we developed two methods for joint WB and SB amplitude calculations. Specifically, given the observed SB amplitude vector, the optimal WB amplitude is derived to minimize the total squared quantization errors. Based on the assumption of quantized SB amplitudes being i.i.d. random variables, a sub-optimal method is also developed to reduce the complexity of the optimal WB quantization scheme. The performance of both methods ("optimal"/"simple") as well as that of the conventional method which is based on the mean of SB amplitudes are shown in Figure 4-2 and Figure 4-3 as a function of the SB amplitude variance. The SB amplitude variance represents the level of frequency selectivity. Each figure corresponds to two different cases of minimum SB amplitude whose value ensures that all the generated SB amplitudes are above a certain threshold. It can be seen that the proposed optimal method universally achieves the best quantization performance among the three methods. Specifically, it is observed from Figure 4-2 and Figure 4-3 that the root mean normalized squared quantization Error (RMNSQE) can be reduced by ~50% by the developed

optimal method compared to the conventional method. However, this comes at the cost of a relatively large computational complexity. The lower-complexity sub-optimal method performs worse but it does significantly outperform the conventional method for highly frequency-selective channels, i.e. large SB variance region. Please refer to [MMF18] for more details.

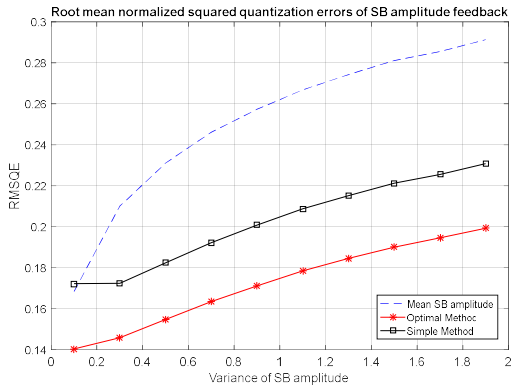


Figure 4-2: RMNSQE Comparison, Minimum SB amplitude: 1.

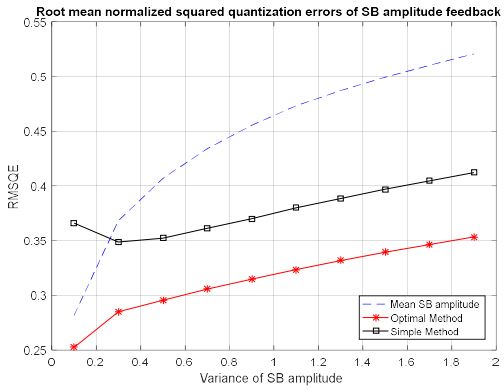


Figure 4-3: RMNSQE Comparison, Minimum SB amplitude: 2.

Downlink control channel transmission in NR

As an important enabler for MIMO transmission in NR, physical downlink control with advanced transmission schemes has also been thoroughly analysed in this section.

Due to its crucial function in a scheduled system, PDCCH is a core element to enable all physical layer data transmissions including advanced MIMO schemes. Recently, configurable distributed PDCCH with the intention to cope with different scenarios has been developed in 3GPP. To have a comprehensive understanding of respective technical advantages and potential scenario dependent limitations, detailed performance analysis and evaluations of configurable distributed PDCCH are thoroughly studied in ONE5G [MF18]. In particular, Exponential Effective SNR Mapping (EESM) has been employed as the performance metric of configurable distributed PDCCH in different scenarios. It is demonstrated from EESM results that configurable distributed PDCCH offers additional degree of freedom for the trade-off between achieved frequency diversity and channel estimation gain by adjusting resource bundling level in terms of the size of REG bundle, according to the channel and interference scenario experienced by the control channel transmission.

Specifically, different interference scenarios, i.e., frequency-flat interference and frequency-selective interferences are investigated in [MF18]. It is observed from simulation results that large REG bundle size provides best overall EESM performance in frequency-flat interference scenario where channel estimation performance plays the dominant role on the reception performance of PDCCH. However, in frequency-selective interference scenarios, where diversity transmission is more beneficial for the reliable reception of PDCCH. As a consequence, small and medium REG bundle sizes furnish better EESM performance and are more preferable. This clearly motivates the configurability of REG bundle size for PDCCH operating in different interference scenarios

4.1.3 CSI Signalling for NR Network Coordination and Duplexing

Towards supporting the high data rates requirements of the ‘Outdoor hotspots and smart offices with AR/VR and media applications’ and ‘Live event experience’ use cases in WP2 [ONE17-D21], NR network coordination and duplexing are essential techniques. As these techniques require precise interference measurements to assist (optimal) resource allocation by the coordinating TRPs, efficient CSI signalling is of critical importance.

As an important baseline network coordination scheme, Non Coherent Joint Transmission (NCJT) [3GPP-36.741], which is shown in Figure 4-4 (Left), allows multiple TRPs to individually transmit data streams to the target user equipment (UE) with minimum information exchange among them. However, the legacy CSI feedback procedure (i.e., transparent NCJT mode) shown in Figure 4-4 (Upper right) has the problem known as rank explosion, i.e., the total number of streams sent from TRPs exceed the number of antennas of the UE. In order to solve this problem, we proposed an explicit signalling scheme, informing the UE of the NCJT mode as shown in Figure 4-4 (Lower right). In this case, when UE reports rank and precoding matrix indicators to TRPs, the UE will take the total number of streams into account to avoid rank explosion. Another benefit of the non-transparent NCJT mode is that, in NCJT, full overlapping may not always be the case and the numbers of resource blocks allocated from two TRPs might be different. As a result, the Phase Tracking Reference Signal (PTRS) frequency density might also be different. A non-transparent NCJT mode enables a UE to be adaptive to different PTRS density configurations [R1-1801970].

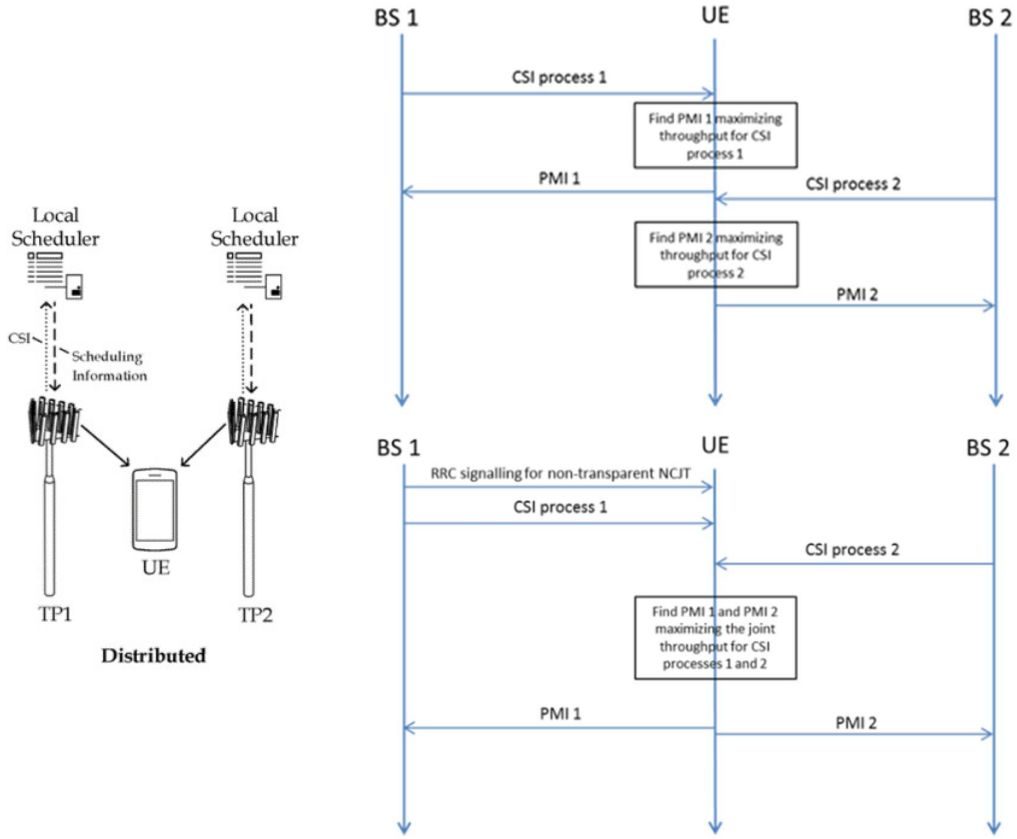


Figure 4-4: Left: NCJT with distributed schedulers. Upper right: signalling procedure for transparent NCJT. Bottom right: proposed signalling procedure for non-transparent NCJT.

In duplexing, also known as dynamic TDD, one of the main performance limiting factors is the Cross-Link Interference (CLI), which occurs when there are misaligned subframes among independent (not coordinated) transmission links established in close distance. Figure 4-5 (left) shows an example of CLI, when TRP1 is in uplink mode and TRP2 is in downlink mode, each serving different UEs. The "interfering UE" uplink signal is interfering the downlink reception of the "desired UE". Similarly, downlink signal by TRP2 is interfering the uplink reception of TRP1. In order to mitigate CLI, ONE5G proposed a novel framework for TRP coordination, treating all interferers as (virtual) TRPs. Under this approach, the victim UE (or TRP) can perform measurement on the interfering channel. The proposed procedure is shown in Figure 4-5 (right). TRP1 informs TRP2 about the DMRS configuration of the "interfering UE" via backhaul. Then, TRP2 assigns ZP CSI-RS to the "desired UE" to allow it to perform

measurement of the interfering UE channel. After measuring the interference channel, the desired UE can use advanced receivers (e.g., interference rejection combining) to suppress the impact of interference on the downlink signal. With the proposed procedure, simulations in [ONE18-D41] have shown that the downlink throughput can improve by 30%.

Both CSI signalling procedures in Figure 4-4 and Figure 4-5 are used in Section 4.2.1 and Section 4.2.2 for interference management in NR network coordination and duplexing.

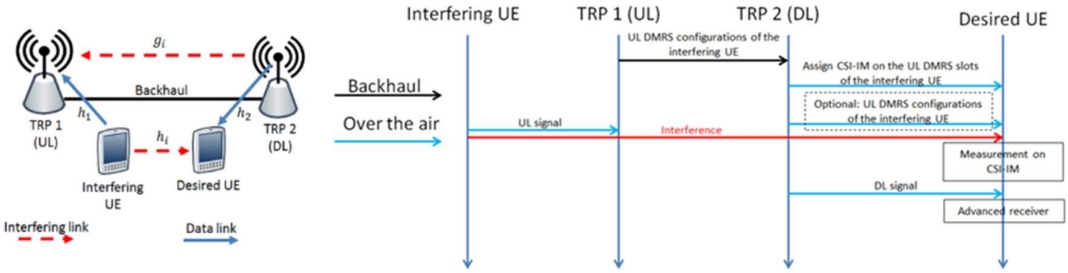


Figure 4-5: Left: diagram of UE-to-UE CLI. Right: proposed signalling procedure for UE-to-UE CLI mitigation.

4.2 Interference Management

The increasing network density both in terms of infrastructure nodes and UEs results in multiple transmissions occurring simultaneously on the same time/frequency/spatial resources, hence resulting in significant interference. As a result, interference management plays an important role in network design with 3GPP providing various mechanisms to this end. This section investigates centralized and distributed network coordination in CRAN deployments towards reducing interference generated by independent (i.e., non-cooperative) transmissions in the network. This case arises in schemes such as (a) NCJT, where multiple TRPs serve a UE, each via an independent data stream, (b) integrated access backhaul (IAB), where access links and backhaul links may occur at the same resources causing interference, as well as (c) D2D enabled networks with D2D and cellular links cross-interfering.

This section presents the work in ONE5G on interference management, both from the perspective of fundamental principles, as well as from a practical, NR-aligned, design perspective. Section 4.2.1 proposes novel network coordination schemes for NCJT exploiting the tools available by the state of the art (3GPP Rel. 15). Section 4.2.2 applies the coordination framework described in Section 4.1.3 to a network with integrated access and wireless backhaul, a concept recently introduced in NR, showing significant performance improvements. Section 4.2.3 investigates optimal scheduling in D2D-enabled CRAN networks towards reducing cross-mode interference. Finally, Section 4.2.4 presents a scheme towards distributed network coordination via limited data exchange among communication nodes.

4.2.1 Centralized and Distributed Multi-Node Schedulers for Non-Coherent Joint Transmission

Both the CRAN and DRAN architectures are in principle capable of providing robust and high data rate connections for the ‘Outdoor hotspots and smart offices with AR/VR and media applications’ use case in WP2 [ONE17-D21]. 3GPP has proposed a number of network coordination schemes to support CRAN and DRAN. In this subsection, three of these 3GPP schemes, namely, (a) DPS with a centralized scheduler, (b) F-NCJT with a centralized scheduler, and (c) NF-NCJT with a distributed scheduler, were investigated. It should be noted that the NCJT scheme has been enhanced in ONE5G with the improved CSI signalling proposed in Section 4.1.3. System level simulations in [Wu18b] have shown that DPS performs well for cell-edge UEs whereas NF-NCJT achieves highest median throughput. The scheduling gain in NF-NCJT outweighs the interference rejection gain in F-NCJT. We conclude that, in CRAN and

DRAN, it is necessary to allow dynamic switching among various network coordination schemes for different services. Also, the sharing of antenna layout information and covariance matrix structures can improve system performance. Detailed descriptions and results can be found in [ONE18-D41], [Wu18b], [WZ19].

4.2.2 NR duplexing with CRAN and network coordination

In the 'Outdoor hotspots and smart offices with AR/VR and media applications' and 'Live event experience' use cases in WP2 [ONE17-D21], it is expected that the deployed 5G network will be dense. However, this high density implies a proportional increase of the transport network (e.g. fibre backhaul), as well. For this reason, the concept of IAB in 3GPP enables flexible and very dense deployment of NR cells without the need for densifying the transport network proportionately. In an IAB network, both access and backhaul/fronthaul links will share the same set of radio resources. In a typical 3GPP IAB setting, the cell of a NR gNB (macro TRP) consists of three sectors, and each sector is served by three IAB-enabled nodes (micro TRPs). The gNBs are interconnected via fibre backhauling and the IAB nodes of each cell are connected to the gNB via wireless fronthaul. An example of an IAB network can be found in Appendix B.2.

To allow for different services, an IAB network can operate with NR duplexing mode to flexibly satisfy uplink and downlink traffic. However, this flexibility comes at the cost of an IAB node potentially experiencing CLI when downlink fronthaul and uplink access transmission occur at the same time. These CLI effects are shown in Figure 4-6. Coordination of gNBs aided by the CSI framework described in Section 4.1.3 provides powerful tools to handle CLI in this scenario as well. As described in Section 4.1.3, this framework can be used to mitigate CLI in IAB networks by configuring ZP CSI-RS for a victim node to perform measurements on interference. The CLI measurement details and signalling procedures can also be found in [R1-1810866].

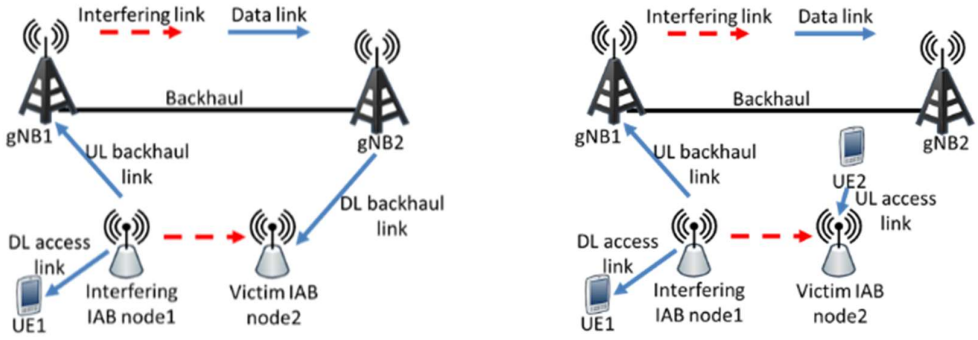


Figure 4-6: Left: CLI interfering downlink backhaul link of the victim IAB node; Right: CLI interfering uplink access link of the victim IAB node.

A system level simulation was performed to study NR duplexing with an IAB-enabled CRAN architecture in the proposed network coordination. The simulation assumed a heterogeneous network with an ISD of 200 meters and an operating carrier frequency of 4GHz. Seven sites (macro-TRP cells) were considered with three sectors per site and three IAB nodes per sector. In total 63 UEs were considered in both cases with and without IAB. The without-IAB case considers macro-TRP cells only. Taking into account signalling overhead, the throughput of a UE is upper bounded by the minimum of the information-theoretic throughput between the gNB and the IAB node, and the information-theoretic throughput between the IAB node and the UE, i.e.,

$$R_{\mathrm{UE}} \leq \min\{R_{\mathrm{gNB-I}} \quad , R_{\mathrm{IAB-UE}}\}. \tag{4-1}$$

In practice, signalling overheads have to be taken into account. In the simulation results, first, it can be observed in Figure 4-7 that the IAB network with NR IAB, duplexing and network

coordination significantly outperforms conventional networks with no IAB, both in terms of cell edge and median throughputs. Second, the application of duplexing mainly provides a balance between DL and UL throughputs, instead of throughput boost in both DL/UL. Last, the proposed CLI management in Section 4.1.3 increases both DL and UL throughputs in IAB. More detailed results can be found in Appendix B.2.

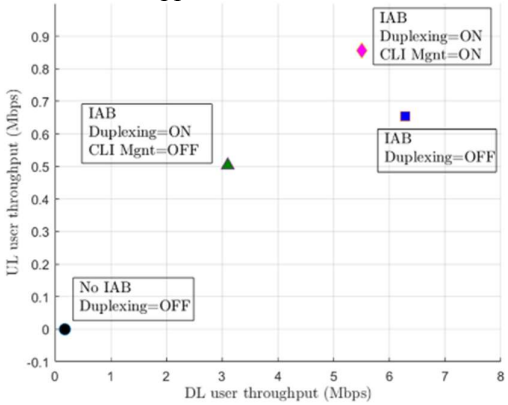


Figure 4-7: Median user DL/UL throughput comparison between with IAB and without IAB, in terms of different duplexing and CLI management settings.

4.2.3 User and Resource Scheduling in network massive MIMO with Underlay D2D

We consider in this section a distributed MIMO system composed of multiple APs and UEs, each having one antenna. Each UE can be served by a large number of APs that form a distributed massive MIMO system. If all UEs are active simultaneously, the system will suffer from pilot contamination as in a conventional massive MIMO where the antennas are co-located. Therefore, by grouping the users and serving them on orthogonal time-frequency resources the impact of pilot contamination will be reduced, as the same pilot will be reused by groups using orthogonal resources and orthogonal pilots will be used inside each group. Furthermore, the throughput of the system can be further improved if a smart allocation of APs to groups is performed. A group will be assigned a subset of APs and the remaining APs will be allocated to other groups, which will result in optimizing the usage of resources (APs) in the system. In addition, we consider that the system supports Device-to-Device (D2D) communications employing the same time-frequency resources as the massive MIMO UEs. The D2D links and the groups of UEs will interfere between each other if they use the same frequency time resources. The goal of this work is to study how to group the UEs in order to reduce the pilot contamination problem and the mutual interference with the D2D links, while still enjoying spatial multiplexing gains. In more detail, the objective is to group the users and to assign APs to groups in such a way to maximize the total throughput of the system. The groups of UEs can then be served on different time and frequency resources.

This problem can be cast as an NP-hard integer program. The computation time of the optimal solution is therefore prohibitive even for very small number of antennas and users. We therefore propose an approximation algorithm that has a lower complexity.

We compare this approximated solution to the one where all APs and all UEs are active. We consider a scenario with 15 UEs, 15 D2D and a varying number of APs. The “all active” solution describes the situation where all APs and UEs are selected. The approximated solution selects a subset of all APs and UEs in order to increase the total throughput. Results, depicted in Figure 4-8, show that our scheme achieves a throughput gain of 20%, especially with a high number of antennas. This can be explained by a higher need to mitigate the interference with a large number of APs.

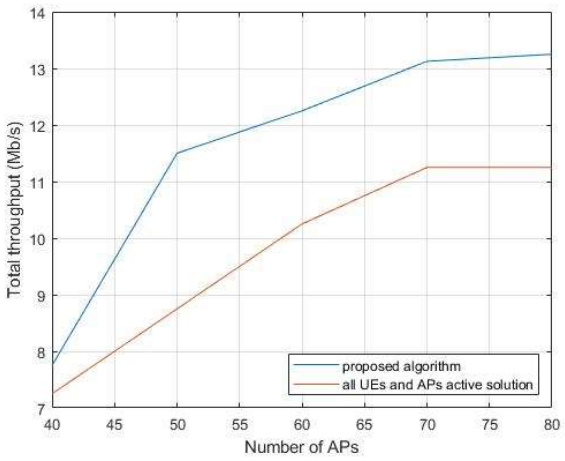


Figure 4-8: Throughput achieved by the proposed solution and by the all APs active scheme

4.2.4 CSI Acquisition and Interference Management using Matrix Exponential Learning

In cooperative CRAN systems, the optimization of the system performance can be achieved by advanced MIMO (beamforming) and interference management schemes. Since transmitters typically cannot afford to exchange too much information, the interference management should be performed in a decentralized way. Game theory along with machine learning tools seems to be a good approach towards this end. Recently, an interesting approach has been developed in [MBN+17] to deal with the problem of beamforming and interference management in networks composed of multiple transmitter-receiver pairs by using Game theory and Matrix Exponential Learning (MXL) technique, which is a variant of gradient-based approach. This technique is attractive due to its fast convergence to a Nash Equilibrium (NE). Although MXL allows the transmitters to decide their action (e.g. beamforming) locally, this technique also suffers from significant signalling overhead. In fact, as most of the gradient-based methods, the gradient matrix of the utility function to be optimized (e.g. throughput) should be estimated by each receiver and then sent back to transmitters as signalling information. This can be a huge burden for the network, especially with a large number of UEs, as the size of each gradient matrix is proportional to the number of RRHs and the number of OFDM subcarriers.

For the above reason, our main goal is to investigate a modified MXL-based algorithm, which requires less amount of signalling overhead and ensures the convergence to NE. Two possible strategies are considered: *i*) MXL-I, where each receiver feedbacks at each iteration only some elements of each gradient matrix instead of the full gradient matrix and the non-received elements are considered as 0 by transmitters so that the associated elements of the action matrix do not update; *ii*) MXL-S, where each receiver only sporadically feedbacks the whole gradient matrix, instead of doing so at each iteration, thus not all of the transmitters need to update their action at the same time. Both variants have very low complexity and can be implemented in practice. We have analysed these two variants of the MXL algorithm and shown that they still converge to the same NE almost surely as the original MXL algorithm. In both settings, we also derive the evolution of the average quantum Kullback-Leibler (KL) divergence [Ved02], which quantizes the distance between an arbitrary action matrix and the NE. We have obtained the upper bounds of the KL divergence to show the convergence rate of the proposed algorithms. The theoretical results can be found in [LAA+18], [LA18].

The proposed algorithms are applied to solve the Energy Efficiency (EE) maximization problem in a multicarrier multi-user MIMO network. For each link, the EE is defined as the ratio between the Shannon-achievable rate and the total power consumption (including transmission and circuit consumption) per transmitter. To provide a numerical example, we consider a simple

scenario of 9 transmitters and 9 receivers. During the communication, 3 OFDM subcarriers are used. Each transmitter is equipped with 4 transmission antennas and each receiver has 8 reception antennas. Numerical results are presented in Figure 4-9 showing the quantum KL divergence obtained by different MXL algorithms, in two different situations where the channel is static or i.i.d. stochastic. For MXL-I, each element of a gradient matrix is sent with a fixed probability p_I, whereas in MXL-S, receivers feedback the whole gradient matrix with a fixed probability p_S at each iteration. Note that the original MXL corresponds to the case where $p_S = p_I = 1$. We can see that the KL divergence decreases and tends to 0 in all cases, which means that the proposed variants of MXL algorithms converge to NE as the original MXL. It is quite reasonable that the algorithms converge slower under stochastic channel then under static channel. For the same level of traffic, we find that MXL-S converges faster than MXL-I. In fact, MXL-S converges slightly slower compared with the original MXL, even if the signalling information is reduced by half, i.e., $p_S = 0.5$. In Figure 4-10, we compare the average evolution of EE obtained by various MXL schemes. The results further justify our concerns. Notice that since EE is quite sensitive to the stochastic channel, the curves are still noisy even by averaging over 500 simulations.

In practical uplink communication systems, the total amount of feedback information that can be sent by all users is limited at a single time-slot. Our work demonstrates that, using the reduced amount of gradient information, the proposed joint feedback and beamforming algorithms still converge to the same optimum as the full feedback case, but with a slight decrease of convergence speed (more time is needed to achieve the optimum). Simulation results further show, for the considered scenarios, that sporadically transmitting a complete feedback for each user is a better strategy than always transmitting incomplete feedback.

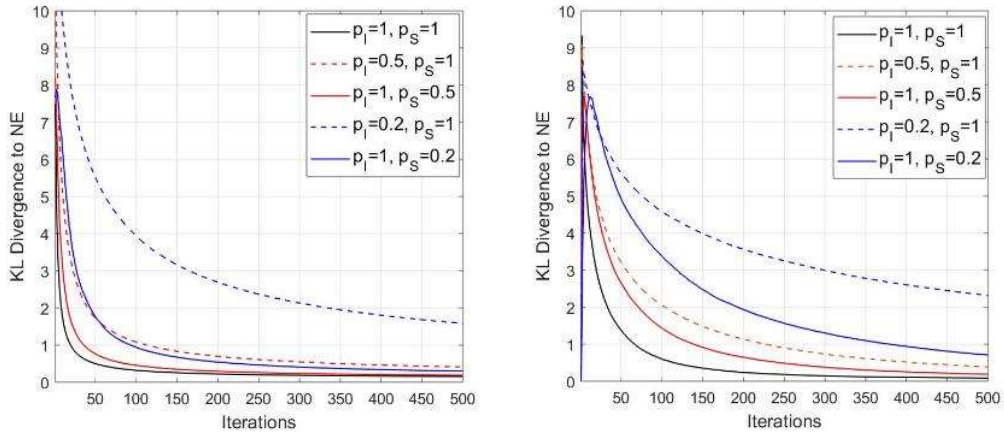


Figure 4-9: KL divergence to NE, average results from 100 simulations using static channel (left) and i.i.d. stochastic channel (right) by : (i) MXL-I with $p_I \in \{0.2, 0.5\}$ and $p_S = 1$; (ii) MXL-S with $p_S \in \{0.2, 0.5\}$ and $p_I = 1$; (iii) original MXL $p_I = p_S = 1$.

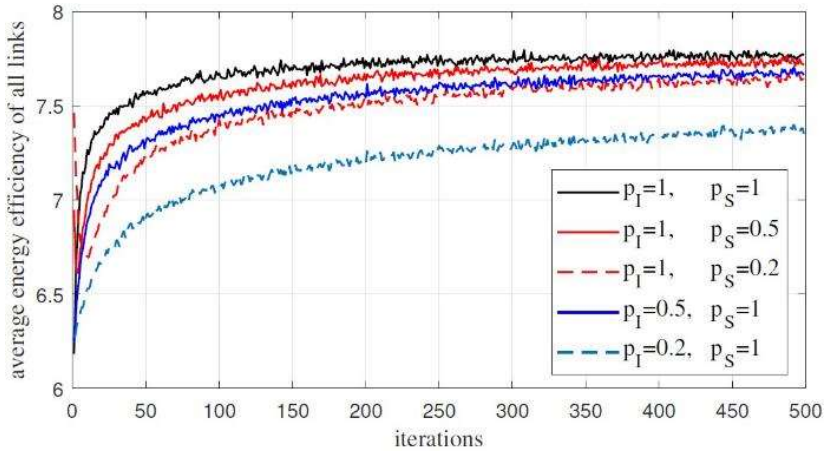


Figure 4-10: For a stochastic channel, evolution of average energy efficiency of all nodes, average results from 500 simulations by : (i) MXL-I with $p_I \in \{0.2, 0.5\}$ and $p_S = 1$; (ii) MXL-S with $p_S \in \{0.2, 0.5\}$ and $p_I = 1$; (iii) original MXL $p_I = p_S = 1$.

4.3 Cell-Less Communication

Cell-less (or cell-free) communication refers to a scenario where a high density of TRPs (RRHs) is deployed over a geographical area with all TRPs in the CRAN network cooperating at the signal level by jointly precoding downlink signals and detecting uplink signals. An example is shown in Figure 4-11 where a random UE within the cell-less deployment is shown to be served by three TRPs, while the other UEs in the system are also served in the same time-frequency resources by means of joint power control and precoding. This advanced multi-connectivity approach inherently eliminates fundamental limitations of the conventional cellular network, such as poor performance of cell-edge UEs, and can, in principle, harness the well-known benefits of network MIMO. This global cooperation approach appears as a natural candidate towards improving performance in new vertical applications such as V2X and industrial control networks, which have extreme requirements in terms of, e.g., reliability and latency, that the conventional network architecture cannot deliver.

However, the benefits of cell-less communication have been mostly demonstrated under idealistic assumptions such as availability of global CSI, ideal backhaul/fronthaul and uncorrelated propagation conditions. A multitude of challenges are still open, which explains why the recent 3GPP releases as well as operators and vendors currently focus on much simpler concepts such as dual connectivity instead of large scale cell-less operation.

This section provides five studies towards addressing some of the major challenges of cell-less communications. Section 4.3.1 proposes a novel UE scheduling algorithm towards improving the, so called, favourable propagation conditions that ultimately leads to a reduction of the number of TRPs required to achieve a certain performance level. Section 4.3.2 proposes a low-complexity algorithm that efficiently tackles the extremely challenging problem of joint UE-to-TRPs association, power control and precoding (beamforming). A similar setting and optimization problem is considered in Section 4.3.3 for multicast transmissions. Section 4.3.4 considers an uplink scenario, taking into account non-ideal fronthaul links. A novel detection scheme based on non-linear filtering at the RRHs and set-theoretic-aided data fusing is proposed, demonstrating significant gains in terms of looser fronthaul links requirements and training overhead. Finally, in Section 4.3.5, a simplified system model of uplink transmission by multiple UEs to a central node, aided by multiple RRHs acting essentially as relays is considered. Assuming access and fronthaul transmissions on the same resources, a novel scheme based on network coding is proposed that maximizes spectral efficiency.

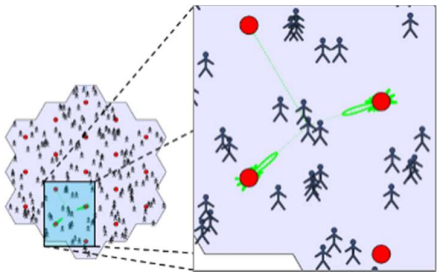


Figure 4-11: Cell-less multi-connectivity beamforming.

4.3.1 User Scheduling in Cell-Less Massive MIMO Systems

Favourable propagation is an important feature of cell-less systems with massive MIMO equipped APs. It emerges from the observation that, with a (very) large number of APs, channel vectors between the users and the access-points become high dimensional random vectors that are (asymptotically) mutually orthogonal under propagation conditions with sufficient scattering [GAO+15]. Leveraging favourable propagation, cell-less, massive MIMO networks can provide UEs with higher spectral efficiency by using simple and low-complexity signal processing techniques such as conjugate beamforming to jointly encode downlink signals and matched filtering to jointly decode uplink signals.

In [ONE18-D41], [HJA18], we demonstrated how to improve favourable propagation for cell-less massive MIMO with a finite number of single-antenna APs by resorting to advanced UE grouping and then scheduling the groups on different resources. We extended the scheme of [ONE18-D41] to the case of multiple antennas at each AP. In addition to that, and since the grouping problem in this case is NP-hard, we developed a new low-complexity UE grouping and scheduling algorithm. In fact, while in the work in [ONE18-D41] the scheduling was based on a semidefinite relaxation approach, which has a complexity that scales up with the number of UEs, we leverage here the principle of sequential convex approximation (SCA) to provide a lower complexity algorithm that can efficiently find a local optimal solution to the formulated NP-hard problem. Figure 4-12 depicts the cumulative density function (CDF) of the proposed SCA-based solution for M=150 and K=20, where M is the number of APs and K the number of users. Each AP is equipped with 10 antennas. It can be observed that the local optimal solution found by the SCA method outperforms the conventional approach (i.e., with no scheduling). The aforementioned results suggest that in cell-less massive MIMO systems, instead of serving all the users simultaneously on the same frequency resources, a substantial gain can be obtained by grouping the users and scheduling the groups on different resources.

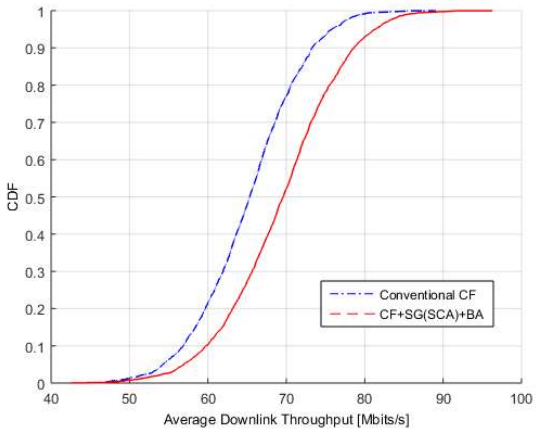


Figure 4-12: CDF of Average downlink throughput

4.3.2 Multi-connectivity beamforming for extreme reliability and massive multiple access

IMT-2020 contains three usage scenarios eMBB, URLLC, and mMTC, which are mutually exclusive because of the inherent tradeoffs between latency, reliability, throughput, and connection density [ITU2410]. The simultaneous support of these requirements requires new technologies with significantly higher spectral efficiency. In ONE5G we have studied a dense, cell-less network, where each UE is associated with one or more TRP, as illustrated in

Figure 4-13. Each TRP is equipped with multiple antennas, and the signals from different TRPs are combined coherently. The objective is to support a non-negligible data throughput (in the range of tens of Mbps) for a large number of UEs, with low latency and extreme reliability. Such ambitious requirements are needed, e.g. for advanced automotive and industrial use cases, such as UC1, UC2, and UC6 defined by WP2 [ONE17-D21].

In order to support massive multiple access, we assume that many UEs are scheduled on the same time/frequency resource and spatial multiplexing with SIC is employed. In particular, our communication scheme is *non-orthogonal, which means* that the number of UEs can be larger than the number of antennas. This is motivated by information-theoretical results which emphasize the role of non-orthogonal multiple access in achieving the network capacity [SVLLLP+17]. Also, we make the reasonable assumption that users are pre-scheduled, i.e., a group of UEs is characterized by channels that are similar in terms of path loss and the channel matrix is well conditioned, i.e., UEs are separable in the space domain. For example, the technique from Section 4.3.1 can be used.

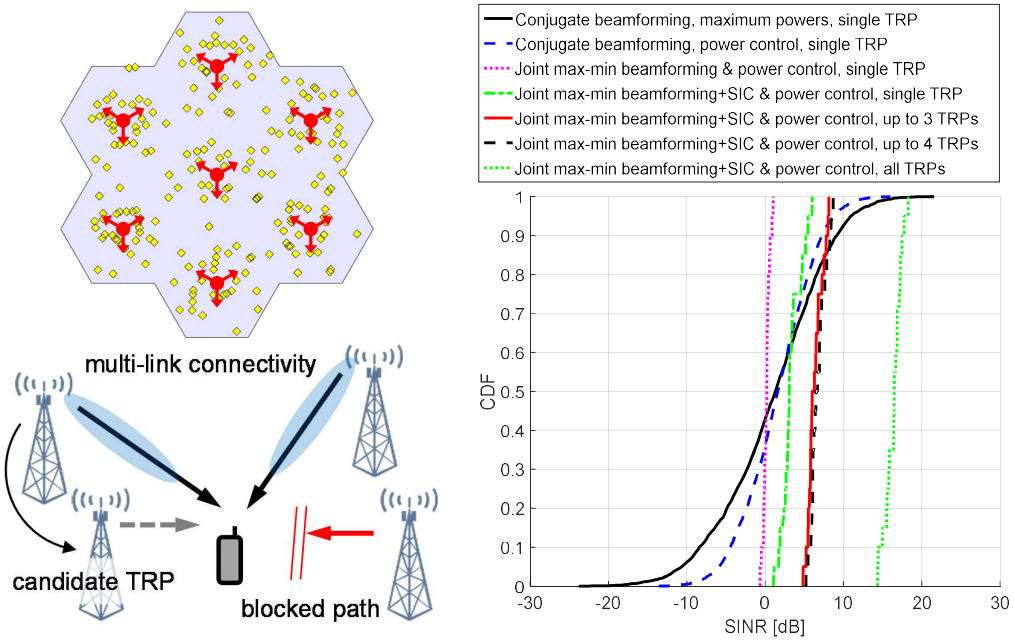
In order to control the interference between all links, we propose *max-min-SINR* power allocation and dynamic TRP association based on CSI knowledge. By maximizing the worst SINR among all users, the system can avoid outages due to interference and shadowing. Another benefit is a more uniform distribution of the capacity over the service area (avoiding "cell edge" effects). For flexible TRP association, we assume that each UE is always connected via a subset of n TRPs. The optimal TRP associations are computed along with the beamformers and transmission powers. This increased flexibility makes identification of the optimal system configuration difficult. The max-min problem is a large-dimensional, non-linear, mixed integer program which is generally very difficult to solve, especially since there is no obvious convexity which can be exploited.

The globally optimal solution (see [SXB19] for details) exploits properties of non-linear interference functions. The unique Pareto-optimal solution that balances all SINR at the same level is computed iteratively, in a computationally efficient and partly-decentralized manner, by alternatingly updating beamformers, TRP association and power allocation. For TRP association, we exploit the sparsity of the multi-TRP channel, which allows for a significant reduction of the search space.

Figure 4-13 shows the results from system level simulations. For single user conjugate beamforming and SNR-based TRP association, we observe the typical SINR distribution caused by mutual interference and individual coupling losses. Since 10 UEs per TRP (equipped with 8 antennas) are transmitting on the same resource, the interference is exceedingly large and many users have an SINR below the detection threshold. The SINR distribution can be improved by introducing conventional power control (second curve), i.e. users with strong channels reduce their transmission power, which in turn reduces the interference levels at other users. This is compared with the proposed max-min SINR balancing scheme, which jointly optimizes all beamformers, powers, and TRP associations such that all SINR are balanced at one level. If this level is too low then the scheduling algorithm must reconfigure the groups. We compare the single-link case (each user served by only 1 TRP) with multi connectivity (3, 4, and 21 TRPs). It is observed that 3 TRPs can already achieve a large portion of the theoretical upper bound. This has the advantage that a large part of the necessary cooperation effort can be restricted to a local

area. With respect to our baseline (conjugate beamforming) we observe a gain of around 4 dB for the 50th percentile. But more important is the reliability gain. We demonstrate that an interference-saturated scenario with 10% of users below -5dB can be turned into an equalized SINR distribution where all UEs achieve +5dB.

The proposed scheme is not only spectrally efficient, it also ensures energy efficient operation for both network and device. This is achieved by the combination of beamforming (focused energy in angular domain), power allocation (reduced transmission power) and dynamic association of densely deployed TRPs (reduced path loss). This must be set against an increased power consumption by the network hardware. In conclusion, ONE5G proposes MIMO-based dynamic multi-connectivity as a way to enable advanced usage scenarios that go beyond simple URLLC, eMBB, and mMTC. In combination with non-orthogonal multiple access, we can drastically increase the number of co-channel users, while ensuring low latency, reliability and throughput requirements. More details are given in [SXB19].



Direction	uplink	#UE per resource and TRP sector	10	ISD	100m
Bandwidth	20MHz	#antennas/TRP (azimuth only)	8	TRP combining	coherent
Carrier freq.	28GHz	Antenna type	ULA	Max. Tx power	30dB
Channel model	UMi Street Canyon	CSI	perfect	Noise figure	5dB

Figure 4-13: Multi-connectivity beamforming enables extreme reliability

4.3.3 RRH selection for multicast communication in cell-less systems

Multi-connectivity and virtual cells are promising concepts for enabling seamless connectivity in dense heterogeneous networks. For example, multiple RRHs can provide individual virtual cells for each user, by assigning individual (virtual) cell IDs. In this way, the virtual cell appears like a base station from the user perspective, but it can be passed through the network along with the movement of the user to avoid frequent (user sided) handovers. This benefit, however, comes at the cost of an increased computational effort that has to be spent for maintaining the virtual cell for each user. One way to reduce the computational overhead is to assign one virtual cell to a group of users. This may be particularly useful for users moving along similar routes, and requiring identical information from the network, because it allows to jointly optimize the beamformers for each group.

In Appendix B.4, we describe a RRH selection mechanism for multicast communication in cell-less systems. In the following, we compare the performance of the SATURATE algorithm for

solving $(B7.2)$ with a naïve greedy maximization of the non-submodular objective, and the optimal solution found by exhaustive search. In addition, we provide an upper bound on the actual worst-case multicast SNR, by solving the semidefinite relaxation (SDR) of the max-min fair multicast beamforming problem for the resulting subset of relays with a standard interior point solver. Moreover, we provide the worst-case multicast SNR of beamforming vectors obtained with an iterative scheme based on [FCS19].

Simulations were performed in randomly generated 1km×1km scenarios with 20 RRHs, 5 users, 10 antennas per RRH using rayleigh fading channels, and power constraint at each RRH, assuming a log-distance pathloss model. In Figure 4-14, *Sat*, *greedy*, and *optimal* correspond to the SATURATE, naïve, and exhaustive RRH selection, respectively. *(utility)*, *SDR bound* and *MC-BF* indicate the conjugate beamforming upper bound, the SDR upper bound, and the worst-case SNR of beamforming vectors computed with the method in [FCS19], respectively. It can be seen that the SATURATE algorithm outperforms the naïve greedy RRH selection by a large margin, and achieves a utility close to the optimal value. Although maximizing an upper bound on the multicast capacity gives no guarantees for the achievable capacity itself, the simulation verifies that the worst-case SNR of beamforming vectors obtained with the proposed scheme follows this upper bound with a rather small gap.

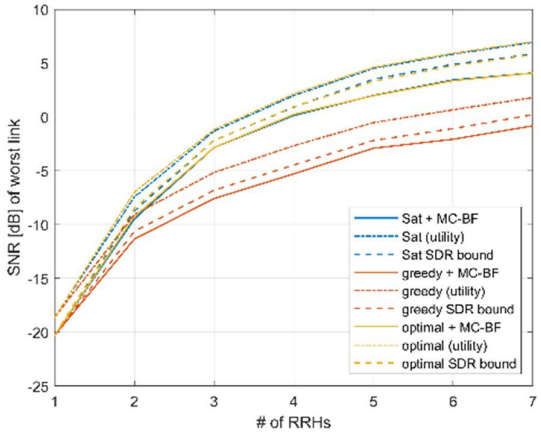


Figure 4-14: Performance of RRH selection and subsequent joint multicast beamforming.

Moreover, we note that the SATURATE algorithm approximates the RRH selection problem in a greedy manner, such that its computational complexity allows an online adaptation of the active set of relays.

4.3.4 Nonlinear Mechanisms in Cell-Less Systems

We developed a nonlinear machine learning-based Detect-And-Forward (D&F) symbol detection method for the uplink of cell-less systems with fronthaul capacity limitations. We focus on achieving low bit error rate (BER) for its practical importance from the viewpoint of reliability at a given transmission rate. Motivated by *cost* and *complexity* aspects, we consider a *capacity-limited fronthaul* and a *limited number of antennas* at each RRH. For local detection at an RRH, we developed a machine learning based non-linear detection method [ACY+18]. Likelihood values associated with the local detection are fused/combined at the CU. To this end, we develop a low-complexity set-theoretic method to estimate likelihood (density) functions [ACS2+18].

Figure 4-15 (*left*) Compares the developed D&F (solid lines) strategy and the conventional centralized joint processing of Quantize-And-Forward (Q&F) for un-coded QPSK modulation.

Figure 4-15 (*right*) The comparison is performed between D&F and the case with centralized single BS of [ACY+18]. See [ACY+18] and [ACS2+18] for details.

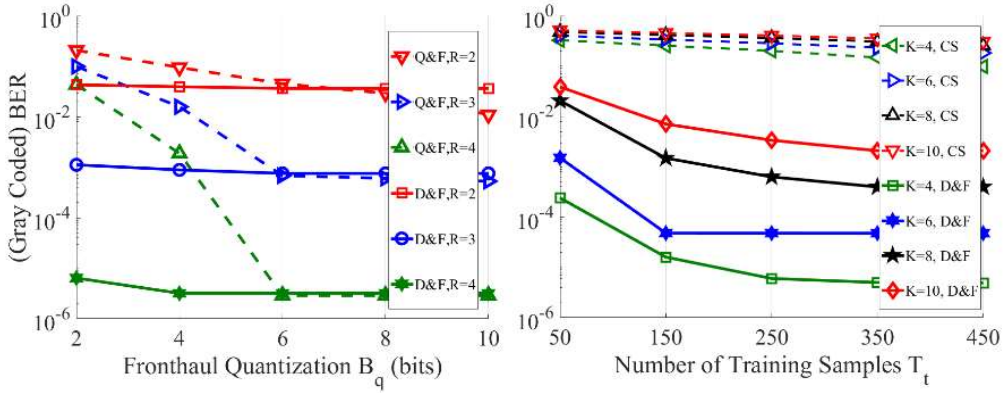


Figure 4-15: (left) Performance of D&F and Q&F schemes with 3 antennas at each RRH, K = 6 devices, R RRHs and 100 samples for training of detection filters. (right) Performance of a single BS/RRH (centralized CS) and D&F using 4-bit quantization, K devices and 3 RRHs.

For more details and results see [ONE18-D41],[ACY+18],[ACS2+18],[ACS1+18]. We have demonstrated that our nonlinear detection method combined with the set-theoretic likelihood estimation method can be used in capacity-limited cell-less systems with many users. The training overhead can be reduced by exploiting the spatial diversity inherent in such systems.

4.3.5 Centralized Scheduling for the Uplink Multiple Access Multiple Relay Channel (MAMRC)

The performance of three different cooperative HARQ protocols is investigated for the slow-fading orthogonal MAMRC (OMAMRC). The goal is to identify the protocol which offers the best performance-complexity trade-off (a detailed analysis can be found in our work [CVM18]. (M,L,1)-MAMRC is considered, where $M \geq 2$ independent sources communicate with a single destination using a help of L half-duplex dedicated relays, which apply Selective Decode-and-Forward protocol (Figure 4-16 a)).

Transmissions are divided into consecutive frames, consisting of time-slots grouped into two phases. All sources transmit successively for the first phase. The second phase consists of a limited number of time slots for retransmissions (Figure 4-16 b)). In each time slot of the second phase, the destination schedules one node to (re)transmit, conditional on the knowledge of the decoding set of each node and a partial knowledge of the quality of all links. The scheduling strategy proposed in [ONE18-D41], [CVM+18], which is shown to achieve a performance close to the optimal one (based on the exhaustive search approach and evaluated by performing MC simulations), is used. Limited feedback broadcast control channel and forward coordination control channels are available. A novel low overhead control signalling exchange protocol between the destination and the other nodes, adapted to the used node selection strategy, is proposed (Figure 4-17 a)). The three considered HARQ protocols are characterized as follows. The first one consists in sending Incremental Redundancy (IR) on all the source messages decoded correctly by the scheduled node (MU encoding), while the second one helps a single source (SU encoding) chosen randomly from its decoding set (but not decoded by the destination). The third one is of the CC type, where the selected node repeats the transmission (including modulation and coding scheme) of one source chosen randomly in its decoding set. It allows MRC at the destination of all the transmissions related to a given source. SU encoding and decoding is well mastered in terms of code construction (rate compatible punctured codes) and, clearly, less complex than MU encoding and iterative joint decoding. On the other hand, the CC approach can be considered as having a similar complexity to Single User IR HARQ.

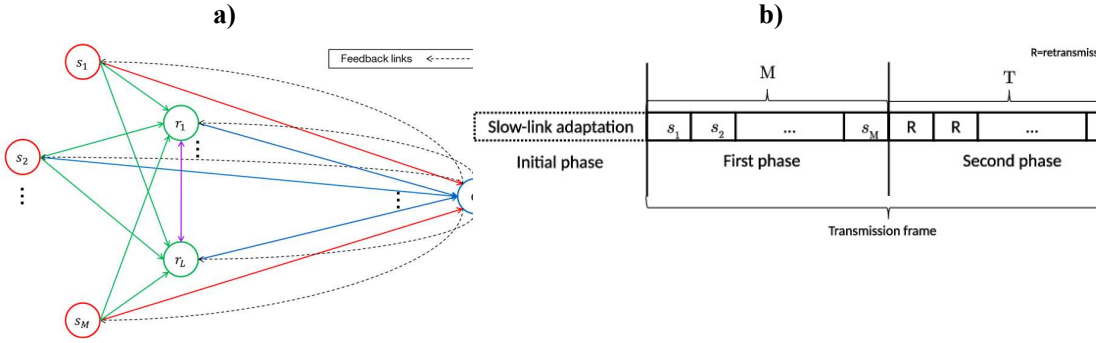


Figure 4-16: a) OMAMRC b) Transmission frame: initial, first and second phase.

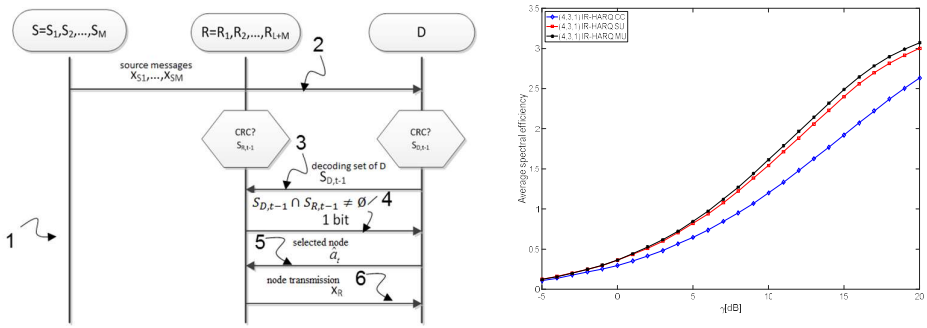


Figure 4-17: a) Novel low overhead control signalling exchange protocol; b) Average spectral efficiency for different HARQ protocols in (4,3,1)-MAMRC

Figure 4-17 b) shows the average spectral efficiency for different HARQ protocols for (4,3,1)-MAMRC network as a function of γ, which is the average SNR (over fading states) of the link between source s_1 and the destination. Link configuration is asymmetric, and the average SNRs of other links in the network are equal to or lower than γ. We observe that the IR-HARQ with MU encoding provides the highest average spectral efficiency. IR-HARQ with SU encoding performance is not far behind, providing slightly lower average spectral efficiency. It can be explained by the fact that "only" four sources are present in the network, so there is often a case where the selected node in the second phase cooperates with exactly one source, even if MU encoding is employed. Naturally, CC-type of HARQ has a noticeably worse performance compared with two IR based protocols for each value of average SNR, since the channel quality remains constant during both first transmission and following retransmissions (in which case IR-HARQ always performs better than CC-HARQ). We also show that as the number of sources in the network grows the difference in performance between the MU and SU encoding slowly grows. We conclude that IR-type of HARQ with SU encoding offers the best trade-off between performance and complexity for a small number of sources in our setting.

4.4 Functionality Placement in Service-Oriented NFV RAN

Various studies have taken place with respect to flexible functionality and functionality placement in the radio access network. For instance, as authors mention in [HR17], in order to mitigate the fronthaul requirements imposed by the CRAN architecture, several functional splits, each characterized by a different demarcation point between the centralized and the distributed units, have emerged. 3GPP considers 8 functional split options in [3GPP-38.801]. For instance, when split option 3 is considered then the low RLC (partial function of RLC), MAC, physical layer and RF are in DU. PDCP and high RLC (the other partial function of RLC) are in the CU. When split option 4 is considered then the MAC, physical layer and RF functionalities are accommodated by the DU. PDCP and RLC are in the CU. As a result, different combinations are done for the 8 split options proposed by 3GPP. However, the

selection of the appropriate functional split needs to consider a number of parameters. As a result, algorithms are proposed for the flexible selection of the appropriate functional split by considering Integer Linear Programming for achieving optimized performance.

Framed in this context, there is obviously a need to evaluate the impact of the RAN functional decomposition, as applied to different cases and deployment scenarios, as it impacts transport configuration and as it evolves to contribute to an end-to-end dynamic and reconfigurable 5G architecture [NGMN18]. Therefore, in the following subsection 4.4.1, we present an expansion of our work from D4.1 [ONE18-D41] related to optimized functionality placement and resource allocation for CRAN/DRAN. We have modelled the optimal functional split as an Integer Linear Programming problem. In [ONE18-D41] we proposed a simulated annealing algorithmic approach to solve it. In this report, we describe the investigation of an alternative solution based on swarm intelligence. Also, we provide an overview of the achieved improvement.

4.4.1 Optimized Functionality Placement and Resource Allocation in CRAN/DRAN Context

This work item studies the flexible split of RAN functionality, towards selecting the optimal operating point between full centralization (CRAN) and local execution (DRAN), adapting to network characteristics as well as current service requirements. However, to achieve a more complete investigation of the search space and reduce the time required to reach a decision, we propose a second algorithm.

The new algorithmic approach leverages swarm intelligence. It is built upon Particle Swarm Optimization, which is a population-based stochastic optimization algorithm inspired by social behaviour of bird flocking or fish schooling. Comparing the two algorithms, PSO manages to reach the optimal solution in instances where SA (with the current parameter values) converges to a sub-optimal one. However the time needed for convergence is not always shorter for one of the two algorithms. Depending on the case the parameters of both algorithms are adjusted. However PSO is preferred because it also enables a potential future distributed version of the algorithm.

Furthermore, the cost function has been re-examined and calibrated. The factors contributing to the cost function address energy consumption, computational latency and communication latency. As an additional feature, a database that keeps previous performance measurements has been created, in order to improve the outcome prediction of each possible option.

The studied network architecture has been expanded to include cases of non-symmetrical networks that include MEC infrastructure in some locations. For such networks an alternative version of the approach was also tested. In that version, we assume image processing workloads with a varying size of images. The aim of the algorithm is to decide whether the workload will run locally or in one of the MECs available in the network and which. This case is also integrated with MEC-related work in WP5 and showcased in the context of a demonstration. The correctness of the operation of the proposed solution has been tested through functional tests with varying network parameters.

Some indicative results are shown below. For our tests we assume a Central Unit (CU) and 10 Distributed Units (DUs), a subset of which are enabled, serving stable total network traffic, while QoS requirements are imposed by constraining the traffic intensity of each queuing system. In the distribution of traffic to the distributed units for two different use cases with their respective QoS requirements is depicted. In case a), the QoS requirements are not as strict, so only 3 of the available DUs are activated, leading to reduced operational costs. However, in case b), QoS requirements are higher (resulting from a higher ratio of URLLC use cases). Therefore, more DUs are required in order to handle the traffic. These are proof of concept results, showing that the algorithm adapts to varying conditions and decisions intuitively show some improvement.

In addition, a numerical evaluation of the achieved improvement in comparison to a baseline approach is depicted in Figure 4-18. Considered tests differ in the service type mixture assumed in the network. As a result, the weights corresponding to a) Computational Latency, b) Data Transmission Latency and c) Energy consumption in the cost function studied by the algorithm adjust to the mixture. Starting from a mixture of 80% URLLC, 10% eMBB, 10% other services (Test case 1), towards a mixture of 10% URLLC, 80% eMBB, 10% other services (Test case 8), the cost function of our proposed solution is compared to that of a baseline centralised approach. The figure shows the percentage of the cost function decrease. We can see that we achieve about 30-50% improvement of the considered cost. The objective function and parameters used are also described in the respective section of D4.1 [ONE18-D41].

Figure 4-18: Decrease of cost function (%) for different service types mix

4.5 Conclusion

This section addressed some of the major challenges in CRAN operation. Aspects such as CSI acquisition, cross-link/cross-mode interference management, cooperative cell-less communications, optimal functionality placement, and relay-aided communications were investigated. These studies (a) provided fundamental theoretical insights demonstrating the potential of CRAN, (b) proposed efficient scheduling, resource allocation, and signal processing algorithms, as well as (c) improved the current state of the art available in NR, by proposing novel ways of exploiting as well as improving the tools available in NR. The proposed solutions are essential towards realizing a true large-scale, flexible CRAN deployment that is able to fulfill the stringent requirements expected for 5G "*Megacity*" scenarios.

Table 4-1. Summary of key recommendations and benefits in terms of Advanced Link Management Based on CRAN/DRAN, and massive MIMO

Feature	Recommendation	Benefits
4.1.1 CRAN Performance under Low-Overhead Channel Estimation	Design of critical parameters for downlink (FDD) CRAN operation such as training overhead and number of cooperating RRHs (cluster size) depend critically on the path loss conditions and should be optimized accordingly. Cooperative transmissions should be considered in propagation conditions with path loss factors close to or greater than 4.	For a propagation path loss factor equal to 3.67, a dense RRH deployment, and a fixed training overhead, cooperative transmissions can provide close to 4 dB SNR gain compared to conventional (non-cooperative) cellular operation. This gain increases for larger path loss factors.
4.1.2 Enhanced CSI	It is recommended to apply the proposed	The enhanced CSI feedback

Feedback and Downlink Control Channel Transmission	joint WB and SB amplitude quantization methods to achieve accurate CSI feedback based on Type-2 codebook in NR. For reliable reception of downlink control channel scheduling DL/UL data packet, configurable size of resource element group bundle using same precoder is also proposed to achieve good trade-off between diversity and beamforming gain for different channel conditions.	accuracy and reliable reception of control channel shall enhance the overall system throughput and spectrum efficiency. Specifically, the proposed optimal CSI feedback can reduce the CSI amplitude quantization error by ~50% compared to the conventional method for Rel-15.
4.1.3 CSI Signalling for NR Network Coordination and Duplexing	It is recommended to use the proposed signalling procedures in NR network coordination and duplexing, i.e., a non-transparent NCJT mode and cross-link interference management with zero-power CSI-RSs. These are proposed to 3GPP NR standardization in Tdocs.	Simulations show that 30% performance user downlink throughput gain with the proposed CLI management procedure.
4.2.1 Centralized and Distributed Multi-Node Schedulers for Non-Coherent Joint Transmission	A distributed network architecture is sufficient to the majority of scenarios. Therefore, it is recommended to use a distributed network as baseline with the ability to switch between various distributed and centralized network coordination methods. How to efficiently switch between network coordination schemes will be an important issue in standardization, impacting the signalling design.	Simulations show that 36% performance median user downlink throughput gain with NF-NCJT over Rel-15 baseline (DPS).
4.2.2 NR duplexing with CRAN and network coordination	It is recommended to reduce cross-link interference in duplexing and IAB with the NR network coordination framework. This principle will significantly impact the NR standardization.	In sub-6 GHz (2GHz), the aggregate result is that the network with IAB still outperforms the network without in terms of downlink throughput (twice the median). In above-6 GHz (30 GHz), the network with IAB significantly outperforms the network without in terms of downlink throughput (28 times the median).
4.2.3 User and Resource Scheduling in Network Massive MIMO with underlay D2D	We developed a user and resource (AP) scheduling in network massive MIMO with underlay D2D. Results show that a fraction of APs must be used.	Results show that scheduling achieves a throughput gain of 20% as compared to the conventional scheme.
4.2.4 CSI Acquisition and Interference Management using Matrix Exponential Learning	We developed two CSI feedback schemes with reduced signalling information for distributed MIMO. Results show that, for the considered scenarios, sporadically transmitting a complete feedback for each user is a better strategy than always transmitting incomplete feedback.	The proposed reduced feedback schemes do not affect the convergence of the system to Nash Equilibrium but at a high convergence time. The methods are then useful in wireless networks with limited feedback.
4.3.1 User Scheduling in Cell-less Massive MIMO Systems	Improving the performance of Cell-free massive MIMO can be achieved through Location and large-scale fading based user grouping, along with optimizing access point assignment and pilot allocation.	Results show a throughput gain of 18% compared to conventional schemes.

4.3.2 Multi-connectivity beamforming for extreme reliability and multiple access	We develop a max-min strategy for equalizing the SINR distribution for cooperative multi-link beamforming and dynamic TRP association. The objective is the support of extreme reliability requirements for a large number of high-throughput users. Also, we avoid cell edge effects and achieve a more uniform distribution of capacity over the service area.	With respect to the baseline technology (conjugate beamforming) we observe a gain of around 4 dB for the 50th percentile. But more important is the reliability gain. We demonstrate that an interference-saturated scenario with 10% of users below -5dB can be turned into an equalized SINR distribution where all UEs achieve +5dB.
4.3.3 RRH selection for multicast communications in cell-less systems	RRH selection for joint multicast beamforming is difficult because it involves two coupled NP-hard problems. In Section 4.3.3., we showed that RRH selection can be decoupled from the optimization of the beamformers, by maximizing an upper bound on the achievable multicast capacity. This allows the application of submodular optimization methods (e.g. the SATURATE algorithm) for selecting the RRHs.	Simulations verify that the optimality gap of SATURATE for maximizing the upper bound is at most 0.7dB. Moreover, the multicast SNRs resulting from the subsequent beamformer optimization are only 1-2dB below this upper bound. The proposed method outperforms the naïve greedy approach by up to 8dB.
4.3.4 Nonlinear Mechanisms in Cell-Less Systems	It is known that increasing spatial diversity increases the reliability of a wireless uplink. So, cell-less systems powered by robust and low complexity detection methods are recommended in dynamic wireless environments that require high reliability. Also, nonlinear detection methods can help keep the number of antennas at the receivers small. Explicit channel estimation (and associated errors) can be avoided by using machine learning based methods.	Simulations show that by using our distributed framework, reliability performance (BER) can be improved by upto an order of magnitude 4 (over centralized solutions at the lowest fronthaul capacity) depending on the number of RRHs. The training set sizes can be reduced by more than 50% by employing multi-connectivity.
4.3.5 Centralized Scheduling for the Uplink Multiple Access Multiple Relay Channel (MAMRC)	In slow-fading orthogonal MAMRC with small number of sources, we recommend the use of IR-HARQ scheme with Single User encoding over the IR-HARQ with Multi User encoding and CC-HARQ schemes, as it offers the best trade-off between performance and complexity.	Average spectral efficiency that can be obtained using IR-HARQ with SU encoding, whose code construction is well mastered (rate compatible punctured codes), is close to the one provided with IR-HARQ with MU encoding, where iterative joint decoding is used (which is more complex), the coding loss being no larger than 1dB.
4.4.1 Optimized Functionality Placement and Resource Allocation in CRAN/DRAN Context	Flexible functionality placement in the Radio Access Network can significantly reduce the cost in terms of latency and energy consumption at the expense of increased computational requirements. The selection of the appropriate centralization level is a challenging task, but even simple methods can offer considerable benefits.	The proposed approach finds the optimal decision regarding functionality placement. We achieve a 30-50% decrease of the considered cost function.

5 Summary of WP4 Main Results and Impact

The following table summarizes the technical achievements of WP4 by grouping them into 11 innovation areas, each addressing aspects that are highly challenging and, at the same time, highly relevant for future (5G) cellular systems. It is noted that the impact of the work in ONE5G WP4 is ensured by the numerous 3GPP documents produced as a result of the work.

For a more detailed overview on the WP4 results, see the summary tables at the end of each section (Subsections 2.4, 3.8, 4.5). Our detailed contributions to 3GPP are summarized in Appendix A.

Table 5-1. Summary of ONE5G WP4 main results

Main Result	Description	Impact
Enhanced NOMA schemes for improved capacity, reduced latency, and service coexistence	ONE5G has developed enhanced NOMA techniques for increasing the number of supported devices per cell, which is particularly important for mMTC. Our approach includes regular spreading matrices, spatial preamble reuse, and reinforcement learning for preamble selection. Also, we propose NOMA for service coexistence, particularly for sharing resources between different service types, e.g. eMBB and URLLC.	The results relate to the NOMA study item (TR-38.812, see list of Tdocs in Appendix A). The work has not led to a dedicated NOMA work item. Instead, it is expected that different aspects of NOMA will be continued in more specialized 3GPP studies, e.g. on random access, or URLLC/eMBB multiplexing, in Rel-17 and beyond. The ONE5G results on probabilistic preamble selection based on reinforcement learning can be included in future products (base stations).
Grant free solutions for URLLC	ONE5G has developed enhancements for increasing the overall URLLC load for uplink grant free (configured grant) transmission where radio resources are shared among multiple users, such that collisions can happen. The main innovative aspects are retransmissions, power control, and resource grid enhancements.	WP4 solutions have been contributed to Release 16 (TR 38.824). For details, see the list of Tdocs in Appendix A For future product development (base stations) we have developed enhanced receiver algorithms for transmission with configured grant under possible collisions. Part of the results are published in a joint paper [MAB+19].
High-quality CSI for massive MIMO and CRAN	ONE5G developed techniques for improving the CSI feedback quality for massive MIMO, either by improving the CSI feedback quality of NR procedures (Type-II codebook) or reducing the feedback overhead without cost in CSI quality. Regarding quality of the acquired CSI itself by means of training, advanced estimation algorithms requiring low training overhead were proposed exploiting structural properties of the wireless channel (e.g., sparsity). Also, novel signalling schemes were developed, building on procedures	The improved Type-II quantization scheme can be applied directly to current NR. Together with our results for optimized training and feedback overhead, this is to be considered in future 3GPP releases (Rel-17 and beyond). The proposed NR signalling procedure for multi-connectivity CSI acquisition is already considered as a candidate solution in 3GPP. For a detailed list of related 3GPP

	currently available in NR. These schemes allow for improved CSI quality in multi-connectivity (CRAN) scenarios with heavy cross-link interference such as dynamic TDD.	contributions see Appendix A. Part of the results are published in a joint paper [BSL+2019].
Low-complexity CSI acquisition and robust beamforming	Constraints on the fronthaul capacity and latency put a limit on the amount of training and the achievable accuracy of the CSI. Such limitations motivate the development of robust strategies for channel estimation, MIMO detection, and beam management.	The WP4 results suggest a flexible RS framework to be considered in future 3GPP releases (17 and beyond), with a training overhead that may change depending on the operational conditions and/or SNR performance requirements.
Pilot Contamination Mitigation	WP4 has developed efficient solutions that mitigate pilot contamination by utilizing power control and channel-correlation between multiple-users in both first and second-order statistics. Also, spatial multiplexing techniques for pilot reuse have been developed for TDD and FDD.	While power control is already supported in 5G, the utilization of correlated multiple user channels requires further signalling between users and users-to-network. This is to be considered in future 3GPP releases (Rel-17 and beyond).
Massive MIMO array designs and efficient implementation	Implementation aspects such as new sub-array structures, flexible and fast reconfigurable HW architecture, low resolution ADCs and wireless fronthaul have been addressed. Also, WP4 proposed optimized antenna geometries that can be better adapted to certain user distributions.	These results have impact on the product development, in particular the array design and the multi-service digital front-end, as highlighted in the innovation report. Our work on circular array geometries has been contributed to RAN1 (see Appendix A).
Massive MIMO Beamforming for Backhaul and Multicast	ONE5G proposes new algorithms to shape backhaul signals and to coordinate interference between access and backhaul. On top, further SNR gain is achieved by Probabilistic Amplitude Shaping. This is complemented by beamforming designs for point-to-multipoint multicast channels.	Beamforming is expected to play a crucial role in 5G, in particular for high frequencies and [3GPP-38.913]. Our results relate to TR38.874 and follow-up activities in 3GPP. Furthermore, multicast beamforming might become relevant for a possible future study item “NR mixed mode broadcast/multicast” (see RP-180669), which was postponed.
Flexible functional split in CRAN	Motivated by the flexibility offered in a CRAN scenario, an efficient algorithm allowing for optimal distribution of functions among centralized (BBU) and distributed units (RRHs) was developed. The algorithm adjusts to the current traffic type and user requirements and aims at achieving multiple objectives such as improved user experience and reduced power consumption.	The solution exploits the flexibility offered by the multiple functional splits planned to be supported by 3GPP (TR38.801). Application in specific scenarios may suggest/promote a subset of the functional split options currently considered by 3GPP.

Cell-free operation	For massive-MIMO enabled cell-less systems, novel scheduling schemes based on user grouping were developed that result in a reduction of pilot contamination effects with low-complexity receiver processing (matched filtering). Algorithms for joint power control and UE-to-RRHs association were also developed. For the case of an overloaded system (more users than antennas), non-linear detectors are employed, which adapt to the non-stationarities of the environment via a machine learning approach, outperforming conventional (linear) detectors.	The WP4 solutions offer significant gains compared to conventional (non cooperative) massive MIMO. They suggest new signalling in order to implement the user grouping, pilot allocation, and adaptation of non-linear detectors, to be considered in future 3GPP releases (17 and beyond).
Extreme reliability enabled by multi-link connectivity	New vertical industry applications, such as robotised automatic processes (Factory of the Future) and V2X, will impose extreme requirements on reliability (up to $1\text{-}10^{-9}$). ONE5G has developed solutions for exploiting the diversity offered by multi-link communication, with potentially multiple RRHs associated with any UE.	Multi-link connectivity is seen as an enabler for advanced services that require new degrees of reliability and spectral efficiency. In a broad sense, this is related to the current 3GPP discussion on "Multi Connectivity" (TS37.340). However, the proposed solutions are more far-reaching and are expected to impact future 3GPP releases (17 and beyond).
Interference Management	ONE5G develops enhanced interference management techniques for the interaction between underlay D2D and cellular users. Also, interference management solutions enabling IAB in NR, NR duplexing with CRAN and network coordination, as well as decentralized beamforming algorithms were proposed.	Interference management solutions suggest new signaling in 3GPP. Part of our solutions has been contributed to the ongoing Rel-16 discussion on "multi-panel/massive MIMO" (see Appendix A).

6 Conclusions and Outlook

This final report D4.2 summarises the results and recommendations of the ONE5G work package WP4 “Multi-antenna access and link enhancements”. The focus of the work is on the development of link-specific techniques (mostly PHY/MAC) for the 5G long-term evolution. The main technical results have already been concluded at the end of each section (see Subsections 2.4, 3.8, 4.5). An overview on the main innovative features and benefits has already been given in Section 5. In this section we give concluding remarks on the main WP4 innovation areas, along with possible challenges for future work. We also discuss the contributions to the two ONE5G scenarios, “Megacity” and “Underserved Areas”.

The following WP4 main results (see Table 5-1 in Section 5 for more details) have been achieved.

- Advanced NOMA schemes and Grant Free Access. The potential gains of non-orthogonal multiple access (NOMA) have been investigated within the project. It is well known that the capacity limits of multi-user systems can in general only be achieved with NOMA. In addition, NOMA enables the coexistence of services with different delay and reliability constraints like eMBB and URLLC on the same resources. A tradeoff between performance and receiver complexity is facilitated by the design of the UE-specific signatures. NOMA can also be used to resolve collisions for random access channels or uplink transmission with configured grant over shared resources (commonly referred to Grant Free Access). Such Grant Free Access solutions avoid the delay and overhead associated with scheduling and increase the number of supported users per cell. In that case, a joint user activity and data detection can provide additional gains, especially for short packets.
  NOMA is a design principle that has a wide range of possible applications ranging from eMBB to URLLC and short-packet mMTC services. The NOMA standardization in 3GPP is currently on hold because of controversial discussions and other priorities. However, we anticipate that the above-mentioned advantages of NOMA will eventually lead to a renewed interest, and research will continue even beyond 5G.
- Massive MIMO array designs and efficient implementation: "Massive MIMO" is well kown as a key enabler for increased spectral efficiency. However, implementation loss, cost, and power efficiency are critical aspects to be considered. Towards increasing the efficiency of massive-MIMO-enabled TRPs, ONE5G has developed low-complexity antenna arrays based on hybrid architectures, low-resolution digital-analogue conversion, and new array geometries. This offers opportunities for new generations of antenna designs, addressing various deployment scenarios including indoor and outdoor. This is complemented by a very fast reconfigurable hardware architecture that enables flexible support of different services. Future generations of 5G (and beyond) cellular networks are anticipated to develop towards distributed but spatially dense antenna deployments. In this context, baseband power consumption can become an issue. Also, research is expected to explore fundamentally new concepts, e.g. intelligent meta-surfaces.
- CSI acquisition and pilot design for massive MIMO and CRAN: Network densification is one main factor in further increasing the system spectral efficiency. However, a crucial aspect is the availability of global, high-quality CSI. We have advanced the state of art by proposing new schemes for pilot contamination reduction, reduced pilot overhead, and reduced feedback overhead for high quality “explicit” CSI, which is the prerequisite for the introduction of new advanced multi-user and CRAN multi-antenna schemes. Another important aspect is the development of signal processing techniques that are more robust to CSI imperfections. We anticipate that CSI acquisition will remain an important issue. The development is going in the direction of distributed and massive antenna deployments and the efficient acquisition and distribution of CSI

requires further investigation. This development will continue to impact future releases of 3GPP.

- Flexible functional split in CRAN: The challenge is to select the optimal operating point between full centralization (CRAN) and local execution (DRAN), adapting to network characteristics as well as current service requirements, while taking into account given limitations of fronthaul/backhaul links. 3GPP has defined multiple options for functional splits between central and distributed unit [3GPP-38.801]. ONE5G has contributed new technical solutions for joint RRM and cooperative multi-point transmission for certain functional splits. But more research is needed, e.g. concerning the integration of radio and fibre, as well as power efficiency aspects.
- Interference Management and Massive MIMO Beamforming: Future network architecture will consist of a heterogeneous deployment of macro cells with small cells and new network components like D2D. Also, in-band backhauling and multicast point-to-multipoint links will become an integral part of the system. The resulting challenge is potential mutual interference which needs to be coordinated and controlled. Interference Management is therefore expected to play an increasingly important role for the 5G long term evolution.
- Extreme reliability enabled by multi-link connectivity: The requirements for future automotive or digital factory services will be extreme. For example, a reliability up to 1-10^{-9} is expected. In CRAN architectures, this can be effectively supported by exploiting multi-link diversity. ONE5G has contributed to this objective in different ways (pre-amble detection for mMTC, RRH selection for multicast communications, as well as max-min fairness for massive MIMO).
- Cell-free operation: The rigid architecture of today's cellular systems does not scale well with the demand for new services and network flexibility. ONE5G has studied new technologies enabling cell-less designs, based on more flexible ways of dealing with interference. This is a long-term evolution path for 5G and even beyond.

For all these areas of innovation, we have assessed the requirements and applicability with respect to the current 3GPP Rel-16 study items. In some areas, our work has contributed to technical documents submitted to 3GPP meetings (see the list of TDocs in Appendix A). Other solutions are more focused on efficient implementation, which is not directly relevant for standardization, but for product development. Also, WP4 has studied technologies that go beyond the scope of current 3GPP concepts, but have a high potential for future standard extensions (e.g. cell-less communication).

Some of the presented technologies have been contributed to WP5. In particular, the following technologies have been implemented and demonstrated: "Non-linear mechanisms in cell-less systems" (PoC#2), "Forward Error Correction" (PoC#3), and "Multi-connectivity for reliability improvement" (PoC#1). This will be described in the final WP5 report. Also, selected technologies from WP4 (Sections 2.2.1, 3.7.2, and 4.4.1) have been integrated into the WP2 system level simulator (see the upcoming final report of WP2). Finally, joint papers on Grant Free URLLC [MAB+19] and CSI acquisition [BSL+2019] have been prepared by authors of WP4.

Furthermore, the results address the requirements posed by the ONE5G scenarios "Underserved Areas" and "Megacities". The applicability and benefits of the WP4 solutions for these scenarios are as follows.

For Underserved Areas, one main challenge is coverage. But cost and energy efficiency are equally important. They enable an economically viable deployment in remote areas with difficult environmental conditions. WP4 has developed several technologies for Underserved Areas. The coverage challenge is addressed by massive MIMO technologies, namely beamforming for improving the SNR. This can be combined with signal/constellation shaping, which adds additional SNR gain. Also the proposed CRAN solutions can contribute to the objective of coverage increase. This is a promising way to overcome the limitation of transmit

power imposed by the regulation authorities. By transmitting over multiple TRPs, the coverage can be effectively improved. Furthermore, non-orthogonal multiple access (NoMA) is a promising technology for low-cost MTC services. By employing grant-free access with advanced receivers, we can reduce both power consumption and signalling overhead. This is complemented by new HARQ strategies, which also reduce power consumption and signalling overhead. Another important aspect is the hardware architecture, which should be flexible, cost- and power efficient. WP4 has developed a reconfigurable baseband processing for multi-service support, hybrid array designs, and optimized array formats.

For Megacities, high area throughputs and connection densities are of highest importance. Also, machine-type services with high demands on reliability and latency play an important role. Towards these goals, WP4 has developed several innovative technologies. The throughput challenge is addressed by massive MIMO and CRAN designs. This includes resource allocation and traffic management, multi-connectivity, beamforming, low-overhead and accurate CSI acquisition, hardware optimization, D2D communications, interference management, and cell-less operation with high density TRPs. For massive access and URLLC, we obtain gains from NoMA and enhanced HARQ.

It is worth observing that technologies such as massive MIMO, NOMA and enhanced HARQ technologies are beneficial for both Underserved Areas and Megacities. While for Underserved Areas the focus is on low energy and low complexity, the focus for Megacities is on massive access and low latency. Generally, SNR gains can be exploited either for coverage or throughput. Likewise, multiuser access efficiency can be exploited for low-cost, low-energy designs or for supporting a large number of devices.

According to the quality and scope of the described innovations, we conclude that the technologies presented in D4.2 are well in line with the objectives and goals of ONE5G.

Acknowledgment
Contribution from the following colleagues is also acknowledged: Aikaterini Demesticha, Aimilia Bantouna, Apostolos Voulkidis, Vasilis Foteinos, Evangelia Tzifa, Ioannis Maistros, Konstantinos Tsoumanis, Kostas Trichias, Nelly Giannopoulou, Paraskevas Bourgos, Yiouli Kritikou, Ioannis Tzanettis, Georgios Makantasis, Despoina Meridou, Christos Ntogkas, Georgios Loukas, Vasilis Laskaridis, Ioanna Drigopoulou, Ilias Romas, Michail Mitrou, Kostas Tsagkaris (WINGS).

References

[3GPP-1700076] 3GPP R1-1700076, “Signal shaping for QAM constellations,” Huawei, HiSilicon, RAN1 NR Ad Hoc, Jan. 2017.

[3GPP-181403] RP-181403, “Revision of Study on 5G Non-orthogonal Multiple Access”, ZTE, June 2018.

[3GPP-181463] 3GPP, “RP-181463 New SID: Study on UE Power Saving in NR”, June, 2018.

[3GPP-36.741] 3GPP TR 36.741, “Study on further enhancements to Coordinated Multi-Point (CoMP) Operation for LTE,” V14.0.0, Mar. 2017.

[3GPP-37.868] 3GPP TR 37.868, “Study on RAN improvements for Machine-type Communications (Release 11)”, September 2009.

[3GPP-38.211] 3GPP TS 38.211, “Physical channels and modulation,” V15.3.0, Sep. 2018.

[3GPP-38.300] 3GPP TS 38.300, “E-UTRA and E-UTRAN Overall Description,” V15.4.0, Dec. 2018.

[3GPP-38.321] 3GPP TS 38.321, “Group Radio Access Network; NR; Medium Access Control (MAC) protocol specification”, V15.2.0, June 2018.

[3GPP-38.801] 3GPP TR 38.801, “Study on new radio access technology: Radio access architecture and interfaces”, Apr. 2017

[3GPP-38.816] 3GPP TR 38.816, “Study on CU-DU lower layer split for NR,” Dec. 2017

[3GPP-38.901] 3GPP TR 38.901, “Study on channel model for frequencies from 0.5 to 100 GHz (Release 14),” Dec. 2017.

[3GPP-38.913] 3GPP TR 38.913, “Study on Scenarios and Requirements for Next Generation Access Technologies,” V14.2.0, Mar. 2017.

[3GPP-38212] 3GPP TS 38.212, “Physical channels and modulation,” V15.0.0, Sep. 2017.

[3GPP-38214] 3GPP TS 38.214, “Physical layer procedures for data,” V15.0.0, Dec. 2017.

[3GPP-38.824] 3GPP TR 38.824, “Study on physical layer enhancements for NR ultra-reliable and low-latency case (URLLC) (Release 16),” V1.1.0, Feb. 2019.

[AAL+14] A. Alkhateeb, O. El Ayach, G. Leus and R. W.“Heath, "Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831-846, Oct. 2014. doi: 10.1109/JSTSP.2014.2334278.

[ABJ+18] R. Abreu, G. Berardinelli, T. Jacobsen, K. Pedersen and P. Mogensen, "A Blind Retransmission Scheme for Ultra-Reliable and Low Latency Communications," 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, 2018, pp. 1-5.

[ACC+18] Attar, M.H.; Chen, Y.; Cheema, S.A.; Wild, T.; Haardt, M.; “NOCA versus IDMA using UFMC for 5G Multiple Access,” in Proc. 22nd Int. ITG Workshop on Smart Antennas (WSA’18), Bochum, Germany, March 2018.

[ACY+18] D. A. Awan, R. L.G. Cavalcante, M. Yukawa, and S. Stanczak, “Detection for 5G-NOMA: An Online Adaptive Machine Learning Approach”,IEEE

International Conference on Communications (ICC), May 2018

[ACS1+18] D. A. Awan, R. L.G. Cavalcante and S. Stanczak, “A robust machine learning method for cell-load approximation in wireless networks”, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, Alberta, Canada. April 2018.

[ACS2+18] D. A. Awan, R. L.G. Cavalcante, Zoran Utkovski and S. Stanczak, “Set-theoretic learning for detection in cell-less C-RAN systems”. 6th IEEE Global Conference on Signal and Information Processing, California, USA, Nov. 26-18, 2018.

[AD12] IEEE Standard for Information technology–Telecommunications and information exchange between systems–Local and metropolitan area networks–Specific requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band, Std., Dec 2012.

[AHB+18] M. Assaad, S. Hajri, T. Bonald, A. Ephremides, “Power Control in Massive MIMO with Dynamic User Population,” in Proc. of IEEE Globecom WS-5GNR, Abu Dhabi,2019

[AJB+18] R. Abreu, T. Jacobsen, G. Berardinelli, K. Pedersen, I. Z. Kovács and P. Mogensen, "Power control optimization for uplink grant-free URLLC," 2018 IEEE Wireless Communications and Networking Conference (WCNC), Barcelona, April 2018.

[AJB+19] R. Abreu, T. Jacobsen, G. Berardinelli, K. Pedersen, N. H. Mahmood, I. Z. Kovács and P. Mogensen, "On the Multiplexing of Broadband Traffic and Grant-Free Ultra-Reliable Communication in Uplink," accepted in 2019 IEEE VTC Spring, Kuala Lumpur, April 2019.

[AJK+19] R. Abreu, T. Jacobsen, K. Pedersen, G. Berardinelli and P. Mogensen, "System Level Analysis of eMBB and Grant-Free URLLC Multiplexing in”Uplink," accepted in 2019 IEEE VTC Spring, Kuala Lumpur, April 2019.

[AJW19] R. Ahmed, K. Jayasinghe, and T. Wild, “Comparison of Explicit CSI Feedback Schemes for 5G New Radio,” in IEEE VTC Spring 2019.

[ATM19] R. Ahmed, F. Tosato and M. Maso , “Overhead Reduction of NR type II CSI for NR Release 16“, in WSA 2019.

[AUP+18] P. Agostini, Z. Utkovski, J. Pilz and S. Stanczak, "Scalable massive random access in C-RAN with fronthaul limitations", in Proc. 15-th International Symposium on Wireless Communication Systems (ISWCS), Lisbon 2018.

[AVW18] R. Ahmed, E. Visotsky and Thorsten Wild. "Explicit CSI Feedback Design for 5G New Radio phase II", WSA 22nd International ITG Workshop on Smart Antennas 2018.

[BGG+18] P. Baracca, L. Galati Giordano, A. Garcia-Rodriguez, G. Geraci, and D. Lopez-Perez, “Downlink Performance of Uplink Fractional Power Control in 5G Massive MIMO Systems”, in Proc. IEEE GLOBECOM Workshop on Emerging Technologies for 5G and Beyond Wireless and Mobile Networks, Abu Dhabi (UAE), Dec. 2018.

[BHS17] E. Björnson, J. Hoydis, L. Sanguinetti, "Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency", Foundations and Trends® in

Signal Processing, 2017.

[BLM16] Björnson, E.; Larsson, E. G. & Marzetta, T. L., “Massive MIMO: Ten Myths and One Critical Question”, IEEE Communications Magazine, 2016, v. 54, pp. 114-123.

[BMA+18] Gilberto Berardinelli, Nurul Huda Mahmood, Renato Abreu, Thomas Jacobsen, Klaus Pedersen, Istvan Z. Kovacs, Preben Mogensen, "Reliability Analysis of Uplink Grant-Free Transmission Over Shared Resources," in IEEE Access, vol. 6, pp. 23602-23611, 2018.

[BMS+18] L Buccheri, S Mandelli, S Saur, L Reggiani, M Magarini “Hybrid retransmission scheme for QoS-defined 5G ultra-reliable low-latency communications”, IEEE WCNC, 2018

[BSL+2019] S. Bazzi, S. Stefanatos, L. Le Magoarou, S. E. Hajri, M. Assaad, S. Paquelet, G. Wunder, and W. Xu, “Exploiting the massive MIMO channel structural properties for minimization of channel estimation error and training overhead,” IEEE Access, vol. 7, 2019.

[BSS15] G. Böcherer, F. Steiner, and P. Schulte, “Bandwidth efficient and rate-matched low-density parity-check coded modulation,” IEEE Trans. Commun., vol. 63, pp. 4651–4665, Dec 2015.

[BX18a] S. Bazzi and W. Xu, “Low-complexity channel estimation in correlated massive MIMO channels,” 22nd International ITG Workshop on Smart Antennas (WSA), Mar. 2018.

[BX18b] S. Bazzi and W. Xu, “On the amount of downlink training in correlated massive MIMO channels,” IEEE Trans. Signal Process., vol. 66, no. 9, pp. 2286–2299, May 2018.

[BXD+18] A.-S. Bana, G. Xu, E. De Carvalho, and P. Popovski, “Ultra Reliable Low Latency Communications in Massive Multi-Antenna Systems”, 2018 52nd Asilomar Conf. Sign., Syst., Computers, Oct 2018, Pacific Grove, CA, USA.

[CC15] Y. Chi and Y. Chen, “Compressive two-dimensional harmonic retrieval via atomic norm minimization,” IEEE Trans. Signal Process., vol. 63, no. 4, pp. 1030–1042, Apr. 2015.

[CF14] E. J. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. Communications on Pure and Applied Mathematics, 67(6):906–956, 2014.

[Che18a] Chen, Yejian. "Exploiting Gaussian Approximation for Non-Orthogonal Coded Access", in Proc. 2018 IEEE 87th Veh. Technol. Conf. (VTC'18 Spring), Porto, Portugal, June 2018.

[Che18b] Chen, Yejian; “Indoor Localization by Detecting 80 Key Bits With Downlink Interleave Division Multiple Access,” in Proc. 2018 IEEE Wireless Commun. Networking Conf. (WCNC’18), Barcelona, Spain, April 2018.

[Che19] Chen, Yejian; “Achieve High Spectral Efficiency for 5G: Multi-User MIMO versus NOMA,” in Proc. 2019 IEEE 89th Veh. Technol. Conf. (VTC’19 Spring), Kuala Lumpur, Malaysia, April 2019.

[CMK+18] E. Caliskan, A. Maatouk, M. Koca, M. Assaad, G. Gui,, “A Simple NOMA Scheme with Optimum Detection,” in IEEE Globecom 2018.

[CSS+17] Cao, Yiqing; Sun, Haitong; Soriaga, J.; Ji, Tingfang; “Resource Spread

Multiple Access – A Novel Transmission Scheme for 5G Uplink," in Proc. Veh. Technol. Conf. Fall (VTC'17 Fall), Sep. 2017.

[CVM18] S. Cerovic, R. Visoz and L. Madier, "Efficient Cooperative HARQ for Multi-Source Multi-Relay Wireless Networks," in Proc. IEEE STWiMob'18, Limassol, Cyprus, Oct. 2018.

[CVM+18] S. Cerovic, R. Visoz, L. Madier, and A. O. Berthet, "Centralized Scheduling Strategies for Cooperative HARQ Retransmissions in Multi-Source Multi-Relay Wireless Networks," in Proc. IEEE ICC'18, Kansas City, USA, May 2018.

[CW18] Chen, Yejian; Wild, Thorsten; "Exploring Performance and Complexity of Selected NOMA Candidates in 5G New Radio," in Proc. 2018 IEEE Global Telecommun. Conf. (Globecom'18) Workshop: Non-Orthogonal Multiple Access Techniques for 5G Workshop (NOMA5G), Abu Dhabi, UAE, Dec. 2018.

[CXY17] J.-F. Cai, W. Xu, and Y. Yang, "Large scale 2D spectral compressed sensing in continuous domain," in IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2017.

[DUT+19] J. Dommel, Z. Utkovski, L. Thiele ans S. Stanczak, "Regular Sparse Code Design: Building-Blocks for Efficient Non-Orthogonal Multiple Access", 2019, submitted for publication in Proc. of IEEE Globecom 2019.

[DWD+18] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen and L. Hanzo, "A Survey of Non-Orthogonal Multiple Access for 5G," IEEE Communications Surveys & Tutorials, vol. 20, no. 3, pp. 2294-2323, third quarter 2018.

[DWY+15] L. Dai, B. Wang, Y. Yuan, S. Han, C. I and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," IEEE Communications Magazine, vol. 53, no. 9, pp. 74-81, September 2015.

[EA07] A. Eyilmaz and R. Srikant, "Fair resource allocation in wireless netwoks using queue-length-based scheduling and congestion control," 2007, pp.1333–1344.

[EAR11-D32] EARTH "D3.2 Green Network Technologies", December, 2011.

[EAR11-D42] EARTH "D4.2 Green Radio Technologies", December, 2011.

[FCS19] J. Fink; R. L. G. Cavalcante, S. Stanczak "Multicast Beamforming Using Semidefinite Relaxation and Bounded Perturbation Resilience" ICASSP2019 (to appear)

[FRZ+17] Fodor, G.; Rajatheva, N.; Zirwas, W.; Thiele, L.; Kurras, M.; Guo, K.; Tolli, A.; Sorensen, J. H. & d. Carvalho, E., "An Overview of Massive MIMO Technology Components in METIS", IEEE Communications Magazine, 2017, 55 , 155-161.

[GAO+15] X. Gao, O. Edfors, F.Rusek, and F.Tufvesson, "Massive MIMO Performance Evaluation Based on Measured Propagation Data", IEEE Transactions on Wireless Communications, vol. 14, no. 7, pp. 3899-3911, July 2015

[GCL+18] L. Galati Giordano, L. Campanalonga, D. Lopez-Perez, A. Garcia-Rodriguez, G. Geraci, P. Baracca, and M. Magarini, "Uplink sounding reference signal coordination to combat pilot contamination in 5G massive MIMO," in Proc. IEEE Wireless Communications and Networking Conference (WCNC), Barcelona (Spain), Apr. 2018.

[GKS19] H. M. Gürsu, W. Kellerer, C. Stefanović, “On Throughput Maximization of Grant-Free Access with Reliability-Latency Constraints”, Accepted for publication in ICC’2019, available arXiv:1902.07933 [cs.NI], Feb. 2019

[GRE15] GreenTouch “GreenTouch Final Results from Green Meter Research Study”, Online, June, 2015.

[GSY19] I. Ghamnia, D. Slock and Y. Yuan-Wu, “Rate Balancing for Multiuser MIMO Systems,” to be in Proc. of Spawc, Cannes, July 2019.

[GZJ+17] Xie, F. Gao, S. Zhang, and S. Jin, “A unified transmission strategy for TDD/FDD massive MIMO systems with spatial basis expansion model”, IEEE Transactions on Vehicular Technology, Vol.: 66, Issue: 4, April 2017.

[HA18] S. E. Hajri and M. Assaad, “A spatial basis coverage approach for uplink training and scheduling in Massive MIMO systems," submitted to the IEEE Transactions Wireless Communications.

[HJA18] S. Hajri, Juwendo Denis and M. Assaad, "Enhancing Favorable Propagation in Cell-Free Massive MIMO Through Spatial User Grouping", submitted to SPAWC 2018.

[HR17] D. Harutyunyan and R. Riggio, "Flexible functional split in 5G networks," 2017 13th International Conference on Network and Service Management (CNSM), Tokyo, 2017, pp. 1-9.

[HWW+18] H. Halbauer, A. Weber, D. Wiegner, T. Wild, “Energy Efficient Massive MIMO Array Configurations”, in Proc. IEEE GLOBECOM Workshop on Green and Sustainable 5G Wireless Networks, Abu Dhabi (UAE), Dec. 2018.

[HXK18] N. Ul Hassan, W. Xu, and A. Kakkavas, “Applying Coded Modulation with Probabilistic and Geometric Shaping for Wireless Backhaul Channel,” in Proc. IEEE PIMRC'18.

[IBX18] O. Iscan, R. Boehnke and W. Xu, “Probabilistically Shaped Multi-Level Coding with Polar Codes for Fading Channels”, IEEE Global Communications Conference Workshops, Dec. 2018

[IBX19] O. Iscan, R. Boehnke and W. Xu, “Probabilistic Shaping Using 5G New Radio Polar Codes”, IEEE Access. DOI 10.1109/ACCESS.2019.2898103

[ITU2410] ITU-R M.2410-0 “Minimum requirements related to technical performance for IMT-2020 radio interface(s)”, Nov 2017.

[IX17] Mohamed Ibrahim, Wen Xu, “A Frame Structure with Precoding for Bidirectional Low Latency Applications”, Globecom URLLC Workshop Dec 2017.

[IX18] O. Iscan and W. Xu, “Polar Codes with Integrated Probabilistic Shaping for 5G New Radio”, IEEE 88th Vehicular Technology Conference, Aug. 2018

[JAB+17] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, P. Mogensen, I. Z. Kovács, T. K. Madsen, "System Level Analysis of Uplink Grant-Free Transmission for URLLC," 2017 IEEE Globecom Workshops (GC Wkshps), Singapore, December 2017.

[JAB+18] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, I. Z. Kovács and P. Mogensen, "Joint Resource Configuration and MCS Selection Scheme for

Uplink Grant-Free URLLC," 2018 IEEE Globecom Workshops (GC Wkshps), Abu Dhabi, December 2018.

[JDC+17] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson and C. Studer, "Throughput Analysis of Massive MIMO Uplink With Low-Resolution ADCs," in *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 4038-4051, June 2017.doi: 10.1109/TWC.2017.2691318.

[JRB+18] Jaeckel, S.; Raschkowski, L.; Börner, K.; Thiele, L.; Burkhardt, F. & Eberlein, E. „Quasi Deterministic Radio Channel Generator User Manual and Documentation“, http://quadriga-channel-model.de, Fraunhofer Heinrich Hertz Institute, 2018.

[JV12] P. Jacko and S. Villar, “Opportunistic schedulers for optimal scheduling of flows in wireless systems with ARQ feedback,” 24th International Teletraffic Congress, 2012.

[KMG+08] Krause, A.; McMahan, H. B.; Guestrin, C.; Gupta, A. "Robust submodular observation selection." Journal of Machine Learning Research 9.Dec (2008): 2761-2801.

[KMT+18] Kurras, M.; Miao, Y.; Thiele, L.; Varatharaajan, S.; Hadaschik, N.; Grossmann, M. & Landmann, M. “On the Application of Cylindrical Arrays for Massive MIMO in Cellular Systems”, 22th International ITG Workshop on Smart Antennas (WSA), 2018.

[KPD+17] D. M. Kim, J. Park, E. De Carvalho and C. N. Manchon, "Massive MIMO functionality splits based on hybrid analog-digital precoding in a C-RAN architecture," *2017 51st Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, 2017, pp. 1527-1531.

[KTS+17] C. T. Kurisummoottil, Wassim Tabikh, D. Slock and Y. Yuan-Wu, “Noncoherent Multi-User MIMO Communications Using Covariance CSIT,” in Proc. of Asilomar, Pacific Grove, CA, USA, Oct-Nov. 2017

[LA18] W. Li, M. Assaad, Matrix Exponential Learning Schemes with Low Information Exchange, 2018, https://arxiv.org/abs/1802.06652v2 [cs.IT].

[LAA+18] W. Li, M. Assaad, A. Ayache, M. Larranaga, Matrix Exponential Learning for Resource Allocation with Low Information Exchange, in proc. of IEEE SPAWC 2018.

[LAD+18] M. Larranaga, M. Assaad, A. Destounis, G. Paschos, “Asymptotically Optimal Pilot Allocation over Markovian Fading Channels”, IEEE Transactions on Information Theory, 64(7), 5395-5418, 2018.

[Liv11] G. Liva, “Graph-Based Analysis and Optimization of Contention Resolution Diversity Slotted ALOHA,” IEEE Transactions on Communications, vol. 59, no. 2, pp. 477-487, February 2011.

[LLP18] L. Le Magoarou, A. Le Calvez and S. Paquelet, Massive MIMO channel estimation taking into account spherical waves, arXiv preprint, arXiv:1811.05669, Nov. 2018.

[LLZ+17] Y. Liang, X. Li, J. Zhang, Z. Ding. “Non-Orthogonal Access (NORA) for 5G Networks”, arXiv preprint, arXiv:1705.01235v1, May. 2017.

[LNB+18] T. Laas, J. A. Nossek, S. Bazzi, and W. Xu, “On the impact of the mutual impedance of an antenna array on power and achievable rate,” 22nd International ITG Workshop on Smart Antennas (WSA), Mar. 2018.

[LP18a] L. Le Magoarou and S. Paquelet, Parametric channel estimation for massive MIMO, IEEE statistical signal processing workshop (SSP), 2018.

[LP18b] L. Le Magoarou and S. Paquelet, Bias-variance tradeoff in MIMO channel estimation, arXiv:1804.07529, 2018.

[LTS+17] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst and L. Liu, "Channel Estimation and Performance Analysis of One-Bit Massive MIMO Systems," in *IEEE Transactions on Signal Processing*, vol. 65, no. 15, pp. 4075-4089, 1 Aug.1, 2017. doi: 10.1109/TSP.2017.2706179.

[LY18] L. Liu and W. Yu, "Massive connectivity with massive MIMO part I: Device activity detection and channel estimation", IEEE Transactions on Signal Processing, vol. 66, pp. 2933–2946, June 2018.

[LZL15] B. Li, D. Zhu, and P. Liang, "Small cell in-band wireless backhaul in massive MIMO systems: a cooperation of next-generation techniques", IEEE Trans. Wireless Commun., vol. 14, no. 12, pp. 7057–7069, Dec. 2015.

[MAB+19] N. Huda Mahmood, R. Abreu, R. Boehnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink Grant-Free Access Solutions for URLLC Services in 5G New Radio", in Proc. 16th Internat. Symp. On Wireless Commun. Systems (ISWCS), Oulu, Finland, Aug. 2019.

[MBN+17] P. Mertikopoulos, E. V. Belmega, R. Negrel, and L. Sanguinetti, "Distributed Stochastic Optimization via Matrix Exponential Learning," IEEE Trans. on Signal Processing, vol. 65, pp. 2277–2290, May 2017.

[MCK+18] A Maatouk, E Caliskan, M Koca, M Assaad, G Gui, H Sari, "Frequency-Domain NOMA with Two Sets of Orthogonal Signal Waveforms", accepted in IEEE Communications Letters, 2018.

[MCL+17] C. Mollén, J. Choi, E. G. Larsson and R. W. Heath, "Uplink Performance of Wideband Massive MIMO With One-Bit ADCs," in *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 87-100, Jan. 2017.doi: 10.1109/TWC.2016.2619343.

[MF18] H. Miao and M. Faerber, "Configurable distributed physical downlink control channel for 5G new radio: Resource bundling and diversity trade-off," *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, Barcelona, 2018, pp. 368-372. doi: 10.1109/WCNCW.2018.8368995

[MHA+18] A. Maatouk, S. Hajri, M. Assaad, H. Sari, S. Sezginer, "Graph Theory Based to Users Grouping and Downlink Scheduling in FDD massive MIMO", in proc. of IEEE ICC, 2018

[MHA+19] A. Maatouk, S. Hajri, M. Assaad, H. Sari, "On Optimal Scheduling for Joint Spatial Division and Multiplexing approach for FDD Massive MIMO," IEEE Trans. on Signal Proc., 67(4), pp. 1006-1021, 2019.

[MMF18] H. Miao, M. Mueck and M. Faerber, "Amplitude Quantization for Type-2 Codebook Based CSI Feedback in New Radio System," accepted in EuCNC 2018.

[NAA+14] J. Nam, A. Adhikary, J. Y. Ahn, and G. Caire, "Joint spatial division and multiplexing : Opportunistic beamforming and simplified downlink scheduling," IEEE J. Sel. Topics Signal Process., vol. 8, no. 5, pp. 876-890, Oct. 2014.

[NB13] Nikopour, H.; Baligh, H.; “Sparse Code Multiple Access,” in Proc. 24th Annu. IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun. (PIMRC13), Sep. 2013.

[NGMN15] NGMN “NGMN 5G White Paper”, Online, February, 2015.

[NGMN18] NGMN “Overview on 5G RAN Functional Decomposition”, Feb. 2018.

[NL17] H. Q. Ngo and E. G. Larsson, “No downlink pilots are needed in TDD massive MIMO,” IEEE Trans. Wireless Commun., vol. 16, no. 5, pp. 2921–2935, May 2017.

[OLL+16] K. Oteri, C. Lin, H. Lou, R. Yang “IEEE 802.11-16/1447r1 further details on multi-stage, multi-resolution beamforming training in 802.11ay,” Nov. 2016.

[ONE17-D21] ONE5G Deliverable D2.1 “Scenarios, KPIs, use cases and baseline system evaluation”, Nov. 2017

[ONE18-D31] ONE5G Deliverable D3.1. “Preliminary multi-service performance optimization solutions for improved E2E performance”, Apr. 2018.

[ONE19-D32] ONE5G Deliverable D3.2. “Final system-level evaluation and integration and techno-economic analysis”, June. 2019.

[ONE18-D41] ONE5G Deliverable D4.1. “Preliminary results on multi-antenna access and link enhancements”, Apr 2018

[ONE19-D52] ONE5G Deliverable D5.2. “Final report on implementation and integration of PoC components into the PoCs and final PoC results”, June 2019

[PDG+16] J. Palacios, D. De Donno, D. Giustiniano and J. Widmer, “Speeding up mmWave beam training through low-complexity hybrid transceivers,” *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Valencia, 2016, pp. 1-7. doi: 10.1109/PIMRC.2016.7794709.

[PLC15] E. Paolini, G. Liva and M. Chiani, “Coded Slotted ALOHA: A Graph-Based Method for Uncoordinated Multiple Access,” IEEE Transactions on Information Theory, vol. 61, no. 12, pp. 6815-6832, Dec. 2015.

[PLW+06] P. Li, L. Liu, K. Wu, W. K. Leung, “Interleave-Division Multiple-Access,” IEEE Trans. Wireless Commun., Vol. 5, No. 4, pp. 938–947, Apr. 2006.

[PNS+18] P. Popovski, J.J. Nielsen, C. Stefanovic, E. de Carvalho, E. Strom, K.F. Trillingsgaard, A.-S. Bana, D.M. Kim, R. Kotaba, J. Park, R.B. Sørensen, “Wireless Access for Ultra-Reliable Low-Latency Communication: Principles and Building Blocks”, IEEE Network, vol. 32, No. 2, Mar. 2018

[PPV10] Y. Polyanskiy, H. V. Poor and S. Verdu, “Channel Coding Rate in the Finite Blocklength Regime,” IEEE Transactions on Information Theory, vol. 56, no. 5, pp. 2307-2359, May 2010.

[PSL+15] E. Paolini, C. Stefanovic, G. Liva and P. Popovski, “Coded Random Access: Applying Codes on Graphs to Design Random Access Protocols,” IEEE Communications Magazine, vol. 53, no. 6, pp. 144-150, June 2015.

[PSL+15] E. Paolini, C. Stefanovic, G. Liva and P. Popovski, “Coded random access: applying codes on graphs to design random access protocols,” in IEEE Communications Magazine, vol. 53, no. 6, pp. 144-150, June 2015.

[PSN+18] P. Popovski, Č. Stefanović, J. J. Nielsen, E. De Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, A.-S. Bana, "Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)", Submitted to IEEE Transactions on -Communications, available arXiv:1810.06938 [cs.IT], Oct. 2018

[PX17] M. Pikus and W. Xu, "Applying bit-level probabilistically shaped coded modulation for high-throughput communications," in Proc. IEEE PIMRC'17, Montreal, Canada, Oct. 2017.

[R1-1700076] 3GPP R1-1700076, "Signal shaping for QAM constellations," Huawei, HiSilicon, RAN1 NR Ad Hoc, Jan. 2017.

[R1-1801970] 3GPP R1-1801970, "Issues on PTRS", Samsung, RAN1 #92, Feb. 2018.

[R1-1810866] 3GPP R1-1810866, "CLI management in NR IAB", Samsung, RAN1 #94b, Oct. 2018.

[RHL11] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "Linear Transceiver Design for a MIMO Interfering Broadcast Channel Achieving Max-min Fairness," Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Nov 2011.

[RN18] K. Roth and J. A. Nossek, "Robust massive MIMO equilization for mmWave systems with low resolution ADCs," *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, Barcelona, 2018, pp. 113-118.

[RPM+18] M. Roy, S. Paquelet, L. L. Magoarou, and M. Crussiere, "MIMO Channel Hardening for Ray-based Models", in IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob, accepted), 2018, eprint arXiv:1804.07491.

[RPS+18] K. Roth, H. Pirzadeh, A. L. Swindlehurst and J. A. Nossek, "A Comparison of Hybrid Beamforming and Digital Beamforming With Low-Resolution ADCs for Multiple Users and Imperfect CSI," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 3, pp. 484-498, June 2018. doi: 10.1109/JSTSP.2018.2813973.

[RPS+218] K. Roth, H. Pirzadeh, A. L. Swindlehurst, and J. A. Nossek. "Are the Data Rates Predicted by the Analytic Analysis of Receivers with Low Resolution ADCs Achievable?" In European Conference on Networks and Communications (EuCNC) 2018, pages 378–9, June 2018.

[RRE14] S. Rangan, T. S. Rappaport and E. Erkip, "Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges," in *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366-385, March 2014. doi: 10.1109/JPROC.2014.2299397

[RST16] M. Rupp, S. Schwarz, and M. Taranetz, "The Vienna LTE-Advanced Simulators: Up and Downlink, Link and System Level Simulation", 1st ed. Springer Publishing Company, Incorporated, 2016, p. 187.

[SB16] P. Schulte and G. Böcherer, "Constant composition distribution matching," IEEE Trans. Inf. Theory, vol. 62, pp. 430–434, Jan. 2016.

[SBW18] S. Stefanatos, M. Barzegar Khalilsarai, and Gerhard Wunder, "Wideband massive MIMO channel estimation via sequential atomic norm minimization," in IEEE Global Conference on Signal and Information Processing (globalSIP), 2018, [Online]. Available: https://arxiv.org/abs/1806.00813

[SDL06] Sidiropoulos, Nikos D., Timothy N. Davidson, and Zhi-Quan Luo. "Transmit beamforming for physical-layer multicasting." IEEE Trans. Signal Processing 54.6-1 (2006): 2239-2251.

[SVLLLP+17] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor "Non-Orthogonal Multiple Access in Multi-Cell Networks: Theory, Performance, and Practical Challenges", IEEE Commun. Mag., Oct. 2017.

[SFK15] V. Saxena, G. Fodor, and E. Karipidis, “Mitigating pilot contamination by pilot reuse and power control schemes for massive MIMO systems,” in Proc. IEEE Vehicular Technology Conference (VTC Spring), Glasgow (Scotland), May 2015.

[SG76] S. Sahni and T. Gonzalez, "P-complete approximation problems," Journal of the Association for Computing Machinery, vol.23, No.3, pp.555-565, July 1976.

[SSB08] S. Shi, M. Schubert, and H. Boche, “Downlink MMSE Transceiver Optimization for Multiuser MIMO Systems: MMSE Balancing,” IEEE Trans. Signal Processing, vol. 56, no. 8, pp. 606–619, Aug 2008.

[Sun17] X. Sun et. al, “Agglomerative user clustering and downlink group scheduling for FDD massive MIMO systems”, in Prof. IEEE ICC’17, May 2017, France.

[SW18] S. Stefanatos and G. Wunder, “Performance limits of compressive sensing channel estimation in dense cloud RAN,” in IEEE International Conference on Communications (ICC), 2018, Kansas City, USA.

[SW19] S. Stefanatos and G. Wunder, “Non-Coherent Joint Transmission in Poisson Cellular Networks Under Pilot Contamination,” 2019, submitted to IEEE Trans. Wireless Commun. Available: https://arxiv.org/abs/1903.05864.

[SXB19] M. Schubert, R. Böhnke, W. Xu. “Multi-connectivity beamforming for enhanced reliability and massive access”, EuCNC 2019, ONE5G special session, June 2019.

[SZS17] Ori Shental, Benjamin M. Zaidel, and Shlomo Shamai. "Low-density code-domain NOMA: Better be regular." arXiv preprint arXiv:1705.03326 (2017).

[TBS+13] G. Tang, B. Bhaskar, P. Shah, and B. Recht, “Compressed sensing off the grid,” IEEE Trans. Inf. Theory, vol. 59, no. 11, pp. 7465–7490, Nov 2013.

[TWK+19] Ta, H., Wang, Z., Kim, S. W., Nielsen, J. J., & Popovski, P., ”Preamble Detection in NB-IoT Random Access with Limited-Capacity Backhaul”, Accepted In *2019 IEEE International Conference on Communications (ICC)*.

[TWW+19] G. Tsoukaneri, S. Wu and Y. Wang, “Probabilistic Preamble Selection with Reinforcement Learning for massive Machine Type Communication (MTC) devices”, Under submission in PIMRC 2019.

[USDP17] Z. Utkovski, O. Simeone, T. Dimitrova, and P. Popovski, “Random access in C-RAN for user activity detection with limited-capacity fronthaul”, IEEE Signal Processing Letters, vol. 24, no. 1, pp. 17–21, 2017.

[Ved02] V. Vedral, “The role of relative entropy in quantum information theory,” Reviews of Modern Physics, vol. 74, no. 1, p. 197, 2002.

[Whi88] P. Whittle, “Restless bandits: Activity allocation in a changing

world,"Journal of Applied Probability, vol. 25, pp. 287–298, 1988.

[WLW15] S. Wang, Y. Li and J. Wang, "Multiuser Detection in Massive Spatial Modulation MIMO With Low-Resolution ADCs," in *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 2156-2168, April 2015. doi: 10.1109/TWC.2014.2382098.

[WRF+18a] G. Wunder, I. Roth, A. Flinth, M. Barzegar, S. Haghighatshoar, G. Caire, and G. Kutyniok, "Hierarchical sparse channel estimation for massive MIMO," IEEE/ITG Workshop on Smart Antennas (WSA'18), Bochum, Germany, May 2018.

[WRF+18b] G. Wunder, I. Roth, R. Fritscheck, B. Gross, and J. Eisert, "Secure massive IoT using hierarhcial fast blind deconvolution," in IEEE Wireless Communications and Networking Conference (WCNC), 2018, Barcelona, Spain.

[WSF+18] G. Wunder, S. Stefanatos, A. Flinth, I. Roth, and G. Caire "Low-Overhead Hierarchically-Sparse Channel Estimation for Multiuser Wideband Massive MIMO," IEEE Trans. Wireless Comm., submitted, Available: https://arxiv.org/abs/1806.00815.

[Wu18a] S. Wu, "Genetic algorithm assisted hybrid beamforming for wireless fronthaul", in Proc. EuCAP'18, London, U.K., Apr. 2018, pp. 1—5.

[Wu18b] S. Wu and Y. Qi, "Centralized and distributed schedulers for non-coherent joint transmission", in Proc. Globecom'18, Abu Dhabi, U.A.E., Dec. 2018, pp. 1—6.

[WXT17] Y. Wang, P. Xu, and Z. Tian, "Efficient channel estimation for massive MIMO systems via truncated two-dimensional atomic norm minimization," in IEEE Intl. Conf. on Communications (ICC), May 2017.

[WZ19] S. Wu and X. Zhang, "A low-complexity antenna-layout-aware spatial covariance matrix estimation method", in Proc. WCNC'19, Marrakech, Morocco, Apr. 2019, pp. 1—6.

[WZM+17] Wang, Qing; Zhao, Zhuyan; Miao, Deshan; Zhang, Yuantao; Sun, Jingyuan; Zhong, Zhangdui; "Non-Orthogonal Coded Access for Contention-Based Transmission in 5G," in Proc. Veh. Technol. Conf. Fall (VTC'17 Fall), Sep. 2017.

[YXS16] Z. Yang, L. Xie, and P. Stoica, "Vandermonde decomposition of multilevel Toeplitz matrices with application to multidimensional super-resolution", IEEE Trans. Inf. Theory, vol. 62, no. 6, pp. 3685–3701, Jun. 2016.

[ZMZ+16] Z. Zhao, D. Miao, Y. Zhang, J. Sun, H. Li, K. Pedersen, "Uplink Contention Based Transmission with Non-Orthogonal Spreading," in Proc. Veh. Technol. Conf. Fall (VTC'16 Fall), Sep. 2016.

Appendix A: Contributions to standardization

The work of ONE5G has contributed to the following 3GPP meeting contributions (TDoc).

Title of the contribution	Meeting identifier	Date of meeting	TDoc number	ONE5G Technology
Sectorized uniform planar arrays versus stacked uniform circular arrays	RAN1 NR Ad Hoc Meeting #3	Sep-17	R1-1716629	Massive MIMO array designs
Discussion on transition between NR network coordination schemes	RAN1 NR Ad Hoc#3	Sep-17	R1-1715936	Interference Management
Preliminary system level evaluation for NCJT in NR	RAN1 NR Ad Hoc#3	Sep-17	R1-1715937	Interference management
Discussion on UE-to-UE cross-link interference management and measurement	RAN1 NR Ad Hoc#3	Sep-17	R1-1716034	Interference management
Discussion on transition between NR network coordination schemes	RAN1 #90b	Oct-17	R1-1717602	Interference management
CLI management in NR IAB	3GPP RAN1 #94	Aug-18	R1-1808774	Interference management
Discussion on UE assistance/reporting for NR	RAN1 #91, Dec. 2017	Nov-17	R1-1720287	CSI for network coordination
Considerations on NOMA transmitter	RAN1 #92	Feb-18	R1-1802027	NoMA
Receiver considerations for UL NOMA	RAN1 #92	Feb-18	R1-1802028	NoMA
Essential procedures to be discussed with NOMA	RAN1 #92	Feb-18	R1-1802029	NoMA
Considerations on NOMA evaluation	RAN1 #92	Feb-18	R1-1802030	NoMA
Receiver considerations for UL NOMA	RAN1#92bis	Apr-18	R1-1804463	NoMA
Procedures to be considered for NOMA operation	RAN1#92bis	Apr-18	R1-1804464	NoMA
Considerations on NOMA evaluation	RAN1#92bis	Apr-18	R1-1804465	NoMA
Consideration on NOMA study	RAN1#92bis	Apr-18	R1-1804466	NoMA
Considerations on NOMA transmitter	RAN1 #93	May-18	R1-1806930	NoMA
Receiver considerations for UL NOMA	RAN1 #93	May-18	R1-1806931	NoMA
Procedures to be considered for NOMA operation	RAN1 #93	May-18	R1-1806932	NoMA
Further considerations on NOMA evaluation	RAN1 #93	May-18	R1-1806933	NoMA

Initial link level simulation results for NOCA	RAN1 #93	May-18	R1-1806934	NoMA
Initial system level simulation results for NOCA	RAN1 #93	May-18	R1-1806935	NoMA
Considerations on NOMA transmitter	RAN1#92bis	Apr-18	R1-1804462	NoMA
Remaining details of NR-PDCCH structure	RAN1 NR Ad-Hoc#3	Sep-17	R1-1716305	CSI acquisition for CRAN
Remaining details of NR-PDCCH structure	RAN1 #90b	Oct-17	R1-1717378	CSI acquisition for CRAN
Remaining details of NR-PDCCH structure	RAN1 #91	Nov-17	R1-1720081	CSI acquisition for CRAN
Remaining details of NR-PDCCH structure	RAN1 NR AH#1801	Jan-18	R1-1800321	CSI acquisition for CRAN
Remaining details of NR-PDCCH structure	RAN1 #92	Feb-18	R1-1802406	CSI acquisition for CRAN
Issues on PTRS	RAN1#92bis	Apr-18	R1-1804367	Efficient signallings in massive MIMO multi-node networks
Issues on PT-RS Design	RAN1 #93	May-18	R1-1806724	Efficient signallings in massive MIMO multi-node networks
On Configured Grant enhancements for NR URLLC	RAN1 #94	Aug-18	R1-1808570	URLLC enabled by grant-free access
On Configured Grant enhancements for NR URLLC	RAN1 #95	Nov-18	R1-1813118	URLLC enabled by grant-free access
Solution for UL inter-UE multiplexing between eMBB and URLLC	RAN1 #95	Nov-18	R1-1813117	URLLC enabled by grant-free access
Solution for UL inter-UE multiplexing between eMBB and URLLC	RAN1-AH-1901	Jan-19	R1-1900931	URLLC enabled by grant-free access

Appendix B

B.1 Sector and Beam Management with Cylindrical Antennas

As channel model we use QuaDRiGa ("QUAsi Deterministic RadIo channel GenerAtor", Version 2.0, publicly available at http://quadriga-channel-model.de/). The System Level Simulation Assumptions are as follows.

Parameter	Value
Scenario	UMa LoS, following [38.901]
Carrier frequency	3.75 GHz
Bandwidth	9 MHz
Multiplexing	OFDM – 5G NR
Resource block config	12 subcarrier, 14 symbols
Subcarrier bandwidth	15 kHz
Symbol length	1/(15 kHz)
Realizations	500
Base station (BS)	
XY-Deployment	Hexagonal grid, see [38.901]
Inter side distance	300 m
Height	25 m
Trannsmit power	40 dBm
Antenna Array shape	UCA or 3-sectorized UPA [KMT+18] [1]
Element pattern	Patch, see [38.901]
Number of elements per array	$3 \cdot 64 = 192$
Number of vertical elements in array	8
Number of horizontal elements in array	24

[1]Note, there is no mechanical down tilt of antenna array, thus geometry is not optimized for ground-based users

User equipment (UE)	
XY-Deployment	Uniform independent random per sector
Number of UEs per sector	10 (30 per BS)
Height	1.5 m
Receiver noise	Termal noise (20°) + 9 dB noise figure
Number of elements per array	1
Element pattern	Omni
Velocity	3 km/h

Transmission scheme	
Direction	Downlink

Duplex	TDD
Channel knowledge at BS	perfect
Transmission mode	Multiple-user spatial multiplexing
User selection	Random
Number selected user	30
Precoding scheme	MMSE

B.2 NR Integrated Access Backhaul with Network Coordination

A typical 3GPP network deployment is shown in Figure B-1.

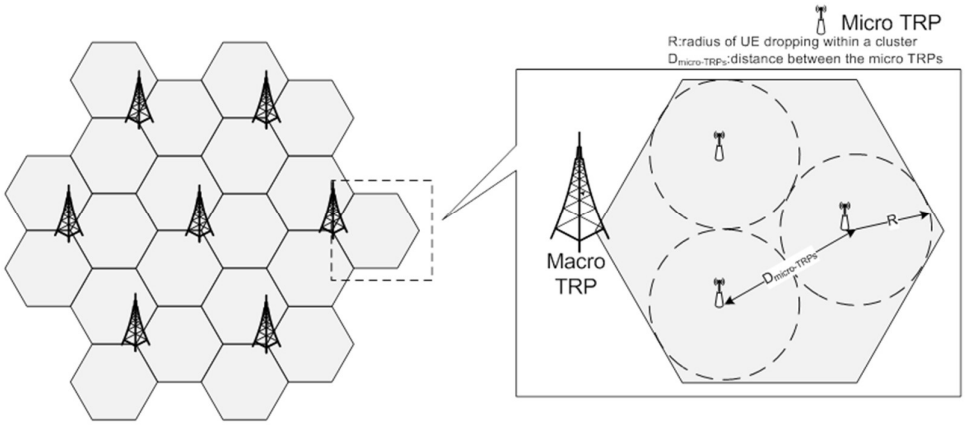


Figure B-1: An example of IAB network. Left: distribution of macro TRPs. Right: Distribution of micro TRPs in a sector.

It can be seen in Figure B-2 that in sub-6 GHz (2GHz), the network with IAB has worse geometry SINR than that without IAB, because IAB results in a denser deployment in an interference-limited scenario. In this case, each UE suffers more interference. However, each UE can be assigned with more PRBs due to less UEs are attached to a relay TRP (rTRP) node. The aggregate result is that the network with IAB still outperforms the network without in terms of downlink throughput (twice the median). In conclusion, in interference-limited scenarios, the gain due to more resources outweighs the loss due to higher interference.

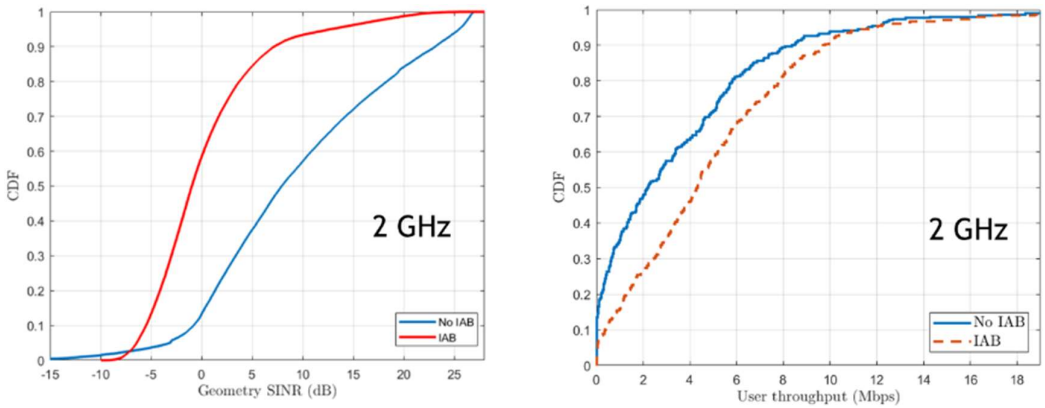


Figure B-2: Left: Geometry SINR comparison between with and without IAB at 2GHz; Right: User throughput comparison between with and without IAB at 2GHz.

We next consider the application of IAB in a noise-limited scenario. It can be seen in Figure B-3 that in above-6 GHz (30GHz), the network with IAB has much better geometry SINR than that without IAB. The reason for this is that coverage becomes more challenging in above-6 GHz

frequency bands and additional rTRPs help compensate the severe pathloss. Thus, the network with IAB significantly outperforms the network without in terms of downlink throughput (28 times the median). Therefore, the application of IAB is highly encouraged in noise-limited scenarios. Figure B-3 (right) shows poor cell-edge performance, because only single-hop rTRPs are considered in the simulation for simplicity. To overcome this, multi-hop rTRPs may need to be considered in practical scenarios.

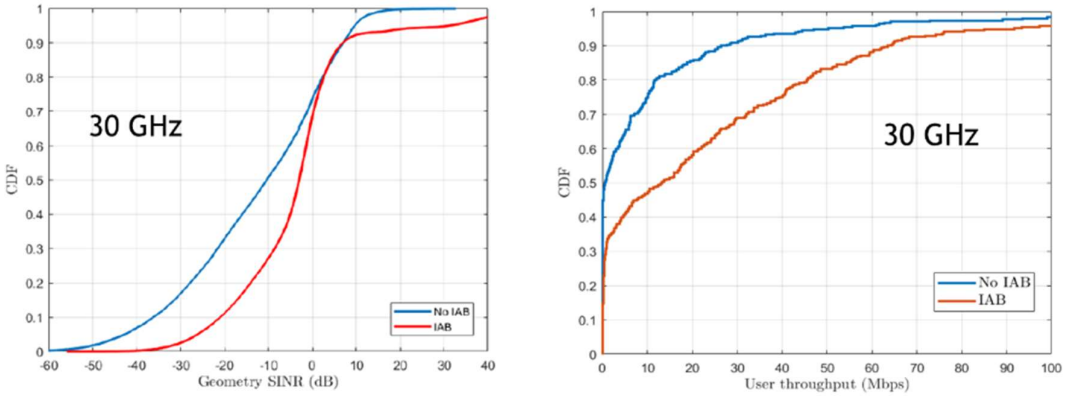


Figure B-3: Left: Geometry SINR comparison between with and without IAB at 30GHz; Right: User throughput comparison between with and without IAB at 30GHz.

B.3 User Rate Balancing

The user rate balancing problem is expressed as follows

$$\max_{\boldsymbol{G},\boldsymbol{F}} \min_k \frac{r_k}{r_k^o}$$

$$\text{s.t. } \mathrm{tr}(\boldsymbol{G}\boldsymbol{G}^{\mathrm{H}}) \leq P_{\max},$$

where $\boldsymbol{G}$ and $\boldsymbol{F}$ are respectively the global transmit and receive filtering matrices, r_k is the rate of user k and $P_{\max}$ is the total power constraint.

Using the rate definition from Lemma 1 in [RHL11], the max-min problem can be transformed into min-max matrix-weighted user MSE as follows

$$\min_{\boldsymbol{G},\boldsymbol{F}} \max_k \frac{\varepsilon_{w,k}}{\mathfrak{E}_k}$$

$$\text{s.t. } \mathrm{tr}(\boldsymbol{G}\boldsymbol{G}^{\mathrm{H}}) \leq P_{\max},$$

where $\varepsilon_{w,k} = \mathrm{tr}(\boldsymbol{W}_k \boldsymbol{E}_k)$ is the matrix-weighted per user MSE and $\mathfrak{E}_k = \ln \det(\boldsymbol{W}_k) + d_k - r_k^o$ is the target MSE, with $\boldsymbol{W}_k$ and $\boldsymbol{E}_k$ being the auxiliary weight matrix variable and the MSE matrix, respectively.

Considering the latter optimization problem and using MSE duality [SSB08], our general proposed framework is as follows

1. Initialize $\boldsymbol{G}$, $\boldsymbol{F}$ and $\boldsymbol{W}_k$; fix r_k^o and compute $\mathfrak{E}_k$
2. **repeat**
   a. for fixed $\boldsymbol{W}_k$, solve the matrix-weighted per user MSE balancing problem as in [SSB08], by using:

      - Perron-Frobenius theorem to find the optimal per user power in the virtual uplink

      - per stream MSE duality between the actual downlink and its virtual uplink channels

   b. update $\boldsymbol{W}_k$ and $\mathfrak{E}_k$
3. **until** required accuracy

B.4 RRH Selection for Multicast Communication in Cell-Less Systems

Consider the downlink in a CRAN scenario with remote radio heads (RRHs) $r \in \mathcal{S} = \{1, \dots, R\}$ and a group of users $k \in \mathcal{U} = \{1, \dots, K\}$ requiring identical information $x \in \mathbb{C}$ from the network. A subset $\mathcal{A} \subseteq \mathcal{S}$ of the RRHs serves the multicast group $\mathcal{U}$ in a cooperative manner, by jointly optimizing the beamforming vectors $\boldsymbol{w}_r \in \mathbb{C}^{N_r}$ of each RRH $r \in \mathcal{A}$. The receive signal for the kth user can be written as $y_k = \sum_{r\in\mathcal{A}} \boldsymbol{w}_r^H \boldsymbol{h}_{rk} + n_k$, where $\boldsymbol{h}_{rk} \in \mathbb{C}^{N_r}$ is the channel between RRH r and user k, N_r is the number of antennas of RRH r, and $n_k \in \mathbb{C}$ is the noise at user k.

Given this system, the goal is to maximize the achievable multicast rate (which is equivalent to maximizing the SNR of the worst user) subject to a cardinality constraint on the active set of RRHs as well as power constraints for the individual RRHs.

$$\underset{\substack{\mathcal{A}\subseteq\mathcal{S}\\ \boldsymbol{w}_r\in\mathbb{C}^{N_r}}}{\text{maximize}} \ \mathrm{SNR}^{\min}(\mathcal{A}, \{\boldsymbol{w}_r\}_{r\in\mathcal{S}}) = \min_{k\in\mathcal{U}} \left| \sum_{r\in\mathcal{A}} \boldsymbol{w}_r^H \boldsymbol{h}_{rk} \right|^2 \qquad (B7.1)$$

$$\text{s.t.} \ \ |\mathcal{A}| \le M; \quad (\forall r \in \mathcal{S}) \ \|\boldsymbol{w}_r\|^2 \le p_r$$

This problem is hard to solve, since both the combinatorial subset selection problem and the multicast beamforming problem [SDL06] are NP-hard. To alleviate this problem, we decouple the subset selection and beamforming problems. This can be done by first approximating a subset of the RRHs for the given cardinality that maximizes an upper bound on the achievable multicast capacity, and, subsequently, jointly approximating the beamforming vectors for this selected subset of relays.

To obtain an upper bound on the above objective, we define conjugate beamformers $\boldsymbol{w}_{rk} \triangleq \sqrt{p_r} \frac{\boldsymbol{h}_{rk}}{\|\boldsymbol{h}_{rk}\|}$ for every pair $(r, k) \in \mathcal{S} \times \mathcal{U}$, which yields

$$\mathrm{SNR}^{\min}(\mathcal{A}, \{\boldsymbol{w}_r\}_{r\in\mathcal{S}}) \le \min_{k\in\mathcal{U}} \left| \sum_{r\in\mathcal{A}} \boldsymbol{w}_{rk}^H \boldsymbol{h}_{rk} \right|^2 = \min_{k\in\mathcal{U}} \left| \sum_{r\in\mathcal{A}} \sqrt{p_r} \, \|\boldsymbol{h}_{rk}\| \right|^2 .$$

To maximize this upper bound over a set $\mathcal{A}$ with cardinality $\mathcal{A} \le M$ we pose the combinatorial problem

$$\underset{\mathcal{A}\subseteq\mathcal{S}}{\text{maximize}} \ \min_{k\in\mathcal{U}} f_k(\mathcal{A}) = \min_{k\in\mathcal{U}} \sum_{r\in\mathcal{A}} \sqrt{p_r} \, \|\boldsymbol{h}_{rk}\| \qquad \text{s.t.} \ \ |\mathcal{A}| \le M, \qquad (B7.2)$$

where $(\forall k \in \mathcal{U})$ $f_k(\mathcal{A})$ are submodular (in fact, modular) functions. Since the minimum of submodular functions is not submodular in general [KMG+08], approximating this problem with a greedy algorithm may result in bad performance. To alleviate this problem, the authors of [KMG+08] propose the submodular saturation (SATURATE) algorithm, which instead aims at maximizing the submodular surrogate function

$$F(\mathcal{A}) = \frac{1}{K} \sum_{k\in\mathcal{U}} \min(f_k(\mathcal{A}), c)$$

by performing a bisection search over the truncation levels c. This method exploits the facts that truncation is a concave function that preserves submodularity, and that $F(\mathcal{A}) \ge c \Leftrightarrow \min_{k\in U} f_k(A) \ge c$. Although the SATURATE algorithm relaxes the cardinality constraint in order to guarantee achieving the optimal objective, the authors report that the algorithm performs very well empirically, even without relaxing the cardinality constraint. Once a subset $\mathcal{A} \subseteq \mathcal{S}$ is obtained, beamforming vectors can be approximated e.g. according to [SDL06].

B.5 HARQ Investigations regarding URLLC

Additional link-level simulation results are shown in Figure B-4 and Figure B-5, where we plot for R=0.4 the BLER vs SNR for QPSK and 16-QAM, respectively.

As second part of our investigations, we estimated SINR joint distributions of first Tx and Retransmissions related to a realistic URLLC scenarios (see Figure B-6). We obtained this joint distribution out of a 3GPP compliant system level simulator and used it for generating time varying SINR samples as input to our link level simulation tool. We considered a UMa scenario, with FTP3 URLLC users and FB eMBB to have a worst-case scenario with full interference, parameters can be found in Table B-1. In this critical scenario, the experienced SINR values come down to -14 dB at the 10^{-5}-quantile for the URLLC UEs (see joint distribution in Figure B-6). For the considered scenario, the link-level simulations results in Figure B-7 show that we need very low code rates with QPSK to reliably transmit a packet within 1-2 Retransmissions.

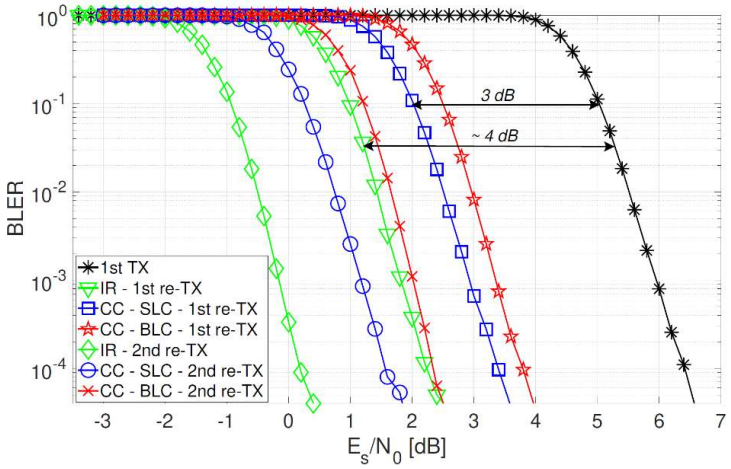


Figure B-4: QPSK R = 0.4 BLER Performance of different HARQ techniques vs SNR.

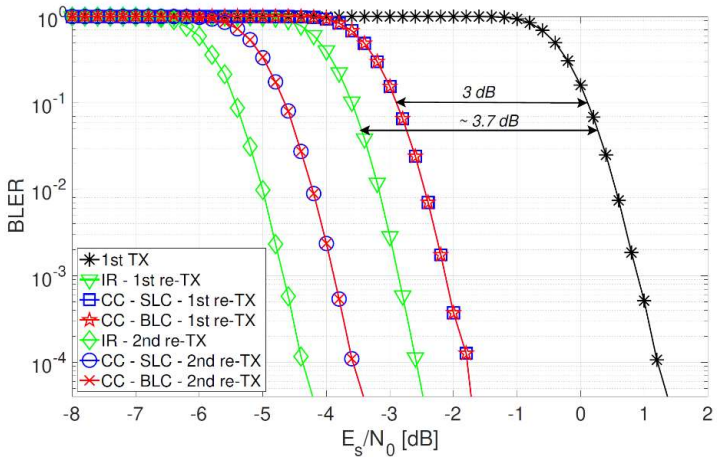


Figure B-5: 16-QAM R = 0.4 BLER Performance of different HARQ techniques vs SNR.

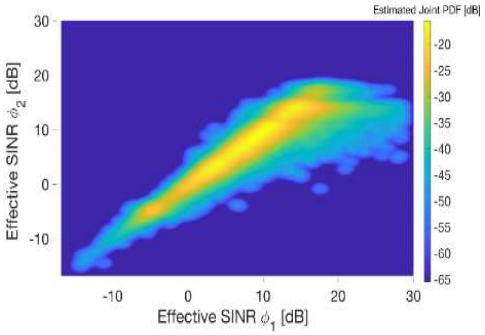


Figure B-6: Joint SINR distribution considered

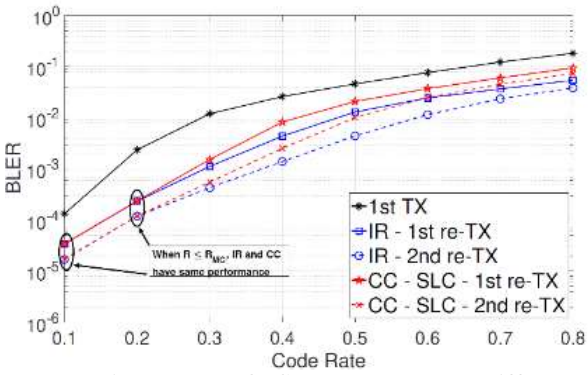


Figure B-7: QPSK BLER vs R, different Tx/Retx.

Table B-1. System level simulation parameters that generated the joint SINR distribution of Figure B-6 [TWK+19].

Parameter	Value
Network environment	3GPP UMa
Number of cells	21
ISD	500 m
Carrier configuration	20 MHz bandwidth at 2 GHz
Subcarrier spacing	15 kHz
Subcarriers per PRB	12
TTI	0.143 ms (2 OFDM symbols)
Antenna configuration	2×2 single-user single-stream MIMO
Transport block size	32, 50, or 200 bytes
HARQ scheme	Asynchronous with Chase Combining
Max number of HARQ retransmissions	6
BLER target 1st Tx	1%
Propagation model	3D
Traffic model	URLLC: FTP3 DL eMBB: full buffer
UE distribution	Uniformly distributed without mobility
Traffic composition	210 URLLC UEs + 105 eMBB UEs
Measured transmissions	10^6