



Call:H2020-ICT-2016-2

Project reference: 760809

Project Name:

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

Deliverable D4.1

Preliminary results on multi-antenna access and link enhancements

Date of delivery: 30/04/2018

Version: 1.0

Start date of project: 01/06/2017

Duration: 24 months

Document properties:

Document Number:	D4.1
Document Title:	Preliminary results on multi-antenna access and link enhancements
Editor(s):	Hardy Halbauer (NOK-GE)
Authors:	Andreas Georgakopoulos (WINGS), Apostolos Voulkidis (WINGS), Aspa Skalidi (WINGS), Dong Min Kim (AAU), Gerhard Wunder (FUB), Elisabeth de Carvalho (AAU), Evangelia Tzifa (WINGS), Evangelos Kosmatos (WINGS), Ioannis Maistros (WINGS), Katerina Demesticha (WINGS), Kilian Roth (Intel), Konstantinos Tsoumanis (WINGS), Miao Honglei (Intel), Nurul Huda Mahmood (AAU), Panagiotis Demestichas (WINGS), Panagiotis Vlacheas (WINGS), Paraskevas Bourgos (WINGS), Renato Abreu (AAU), Stelios Stefanatos (FUB), Vera Stavroulaki (WINGS), Yiouli Kritikou (WINGS), Mohamad Assaad (CNRS), Salah Eddine Hajri, Juwendo Denis (CNRS), Ali Maatouk (CNRS), Wenjie LI (CNRS), Hardy Halbauer (NOK-GE), Paolo Baracca (NOK-GE), Rana Ahmed Salem (NOK-GE), Silvio Mandelli (NOK-GE), Yejian Chen (NOK-GE), Shangbin Wu (SEUK), Yue Wang (SEUK), Martin Schubert (HWDU), Samer Bazzi (HWDU), Ronald Böhnke (HWDU), Onurcan Iscan (HWDU), Wen Xu (HWDU), Marcin Pikus (HWDU), Martin Kurras (HHI), Zoran Utkovski (HHI), Johannes Dommel (HHI), Daniyal Amir Awan (HHI), Martin Kasparick (HHI), Stefan Cerović (Orange), Raphaël Visoz (Orange), Yi Yuan (Orange), Wassim Tabikh (Orange), Stéphane Paquelet (b<>com), Luc Le Magouarou (b<>com), Matthieu Roy (b<>com), Jean Dion (b<>com)
Contractual Date of Delivery:	30/04/2018
Dissemination level:	PU ¹
Status:	Final
Version:	1.0
File Name:	ONE5G D4.1_final.docx

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

PU = Public

Abstract

This report summarizes the intermediate work results of WP4 “Multi-antenna access and link enhancement”. The investigated technologies are grouped into “Future-proof multi-service access solutions”, “Massive MIMO enablers towards practical implementation” and “Advanced link management solutions assuming CRAN/DRAN deployments and/or massive MIMO”. The relation to current 3GPP standardization is highlighted, and the relation to the use cases of WP2 as well as the selected technologies for the proof of concepts conducted by WP5 are discussed.

Keywords

New radio (NR), mMTC, URLLC, eMBB, massive MIMO, CRAN, DRAN, beamforming, non-orthogonal access

Executive Summary

Report D4.1 summarizes the intermediate status of **WP4 “Multi-antenna access and link enhancement”**. The investigated technologies are briefly introduced and preliminary results, findings and conclusions are presented. The goal of this work package is the development and validation of lower layers techniques (mainly PHY/MAC) that support the three service categories eMBB, mMTC, and URLLC. A particular emphasis is put on implementation efficiency. The work is grouped according to the tasks of the work package:

T4.1 Future-proof multi-service access solutions

The emerging 5G service categories ultra-reliable low latency communications (URLLC) and massive machine type communications (mMTC) introduce new design challenges. This task investigates different solutions for efficient multi-service access. In particular, we address challenges in providing fast and reliable access, supporting a massive number of low-payload connections, and scheduling different services with different KPIs concurrently. Uplink grant-free access is assumed. The “Megacity” scenario is the main target in this task. However, many of the developed technology components can also provide gains for “Underserved areas”, e.g. for low-cost, low-energy mMTC.

The solutions presented in this task are grouped into two clusters: “*Design of non-orthogonal multiple access*” and “*Solutions for URLLC services*”.

Non-orthogonal Multiple Access (NOMA) allows multiple users to access the radio network using overlapping radio resources in the time and frequency domain, and/or user separation via advanced signal processing at the network side in the power domain. There exist different NOMA schemes with different design targets. Three such schemes for grant free access are presented in this report. Generally, the benefits are increased cell throughput with improved cell-edge performance for eMBB, and higher connection density with reduced signaling overhead and power consumption for mMTC.

In Section 2.1.1 we study NOMA based on interference cancellation, which exploits that the SINR from multiple transmitters are sufficiently different at the receiver. However, clustering is difficult when users are in RRC_IDLE state. For this case, we develop uplink clustering techniques based on machine learning. The analysis shows a significant reduction of the collision probability.

In Section 2.1.2 we combine coded random access based uplink transmission schemes at packet-level coding with NOMA on the physical layer. Such combination allows for resolving collisions by UE-specific signatures using advanced receivers, thus improving the decoding probability of grant-free uplink transmissions.

In Section 2.1.3 we focus on enhancements for Non-Orthogonal Coded Access (NOCA). We develop a non-linear receiver with scalable complexity and enhanced performance. Gaussian Approximation (GA) is applied for NOCA to realize soft PIC, to balance the performance-complexity trade-off. Detailed system level simulations show that IDMA exhibits better convergence capabilities by means of better randomization due to user-specific interleaving with relatively higher complexity. On the other hand, NOCA outperforms IDMA with very few iterations due to efficient MMSE filtering with relatively lower complexity.

Solutions for URLLC services. URLLC is a new service class introduced by 3GPP in 5G NR. The main URLLC requirements are high reliability (99.999%) at 1 ms user plane latency. This report presents various solution approaches for URLLC in the uplink. Eliminating competition among multiple users in the access layer, introducing new packet structure and receiver design for reliable short packet transmission in the physical layer, and optimizing the frame structure for bi-directional TDD are the key design goals.

In Section 2.2.1 we study grant free (GF) access for URLLC. Since GF schemes simplify user scheduling and resource allocation, they emerge as a candidate solution, which have the potential of reducing the latency with respect to traditional grant-based (GB) approaches of the LTE radio standard. Detailed system-level simulations are used to assess the performance. Up to 50% improvement in the supported load, that meets the URLLC requirements, is observed by optimizing the power control setting for uplink GF access.

In Section 2.2.2 we consider a massive access scenario where the MTC devices sporadically transmit short data packets. We optimize the throughput-latency-reliability trade-off by revisiting the design of frame structure conventionally consisting of preamble, metadata and data. Doing so results in an improvement in the probability of correct user activity detection.

In Section 2.2.3, the design of a bi-directional frame structure for dynamic TDD is considered. Dynamic TDD allows switching the transmission direction between uplink and downlink at every TTI, thus lowering the latency. However, cross-direction interference is a strong limiting factor in this case, especially due to the disparity in the transmission powers in the two directions. The designed frame structure facilitates instantaneous transmission in TDD mode by reducing the out-of-band emission. Instead of using guard bands between neighbouring sub-bands corresponding to different numerologies, we introduce a novel precoder to mitigate the crosstalk interference in the baseband.

T4.2 Massive MIMO enablers towards practical implementation

This task is aiming at efficient implementations of massive MIMO, with emphasis on low complexity, improved flexibility, and high performance. This includes hybrid and digital beamforming solutions, optimized array formats, and flexible and fast reconfigurable hardware architecture solutions. Finally, the task addresses the crucial aspect of CSI Acquisition (pilot and feedback design for mMIMO).

Flexible hardware architecture and multi-service support. In this activity we investigate optimized implementations and enablers for massive MIMO.

In Sections 3.1.1 and 3.1.2, we study the impact of the array shape on system performance. The deployment of antennas for massive MIMO has a huge impact on the performance, however it is often neglected in most of the applications. Two technologies provide performance evaluations on this topic. First, depending on the user distribution, the deployment of antenna arrays, e.g. as [8x4] or [1x32] array can result to in sum-spectral efficiency gains of more than 100% compared to sub-optimal deployment. Second, using a cylindrical antenna deployment provides coverage gains for 50% of the users compared to sectorized planar array as shown in [KMT+18]. For usage in cellular deployments, ONE5G recommends using joint precoding over all antennas instead of sectorizing the cylindrical array.

In Section 3.1.3, flexible and fast configurable hardware architectures that considers a wide range of standards is proposed. The reconfigurable hardware architecture proposed by ONE5G combines FEC features from 4G LTE, WiFi and 5G New Radio with respect to CRC, Coding, HARQ, Segmentation, and modulation without performance loss in the respective technology.

Section 3.1.4 provides an analytic MIMO performance prediction that takes into account channel hardening. The advantage of this work is that computational complex numerical simulations can be avoided allowing for faster development and evaluation of new MIMO techniques.

Efficient CSI acquisition in TDD/FDD and feedback compression in FDD. The work in this section has shown that it is possible to achieve a sufficient channel estimation quality for a wide range of practical scenarios, while at the same time drastically reducing the computational complexity. We also showed that the chosen schemes are in addition robust to additional non-linear impairments often encountered in consumer grade wireless systems.

In Section 3.2.1, we study physical models and algorithms for FDD channel estimation based on statistical tools. The enhanced estimation algorithm is compared to other approaches such as

Linear Minimum Mean Square Estimation (LMMSE). A bound is derived which can be used to adjust the number of virtual paths to be estimated, for a given data rate constraint.

In Section 3.2.2, we propose a Compressed Sensing (CS)-inspired channel estimation algorithm that explicitly takes into account the hierarchical sparsity property. We evaluate properties and performance of proposed schemes by means of analysis and simulation. A conventional CS algorithm, which ignores the hierarchical sparsity property, is shown to require a significantly larger minimum pilot overhead.

In Section 3.2.3, we jointly investigate DL training and achievable rates in multiuser (massive) MIMO systems with correlated channels. We address the open problem of choosing a suitable training duration to ensure a bounded achievable rate loss to the one achieved with perfect CSI at the BS.

In Section 3.2.4 we consider the CSI feedback in a TDD massive MIMO system. We show how the systems performance can be improved by efficiently grouping the different users in a MU-MIMO setup. For the FDD CSI feedback we show how the codebook-based CSI feedback present in 3GPP NR Rel. 15 can be extended to support a more explicit feedback of the channel and the interference at a moderate overhead. This enables future enhancements of the system regarding multi point transmission and reception.

In Section 3.2.5 we consider advanced radio concepts such as multi-TRP coordination, as envisioned in NR phase II. In this context, explicit CSI knowledge at the BS side is highly beneficial. In this work, we exploit the underlying time domain sparsity of the channel for explicit time domain feedback. An enhanced scheme is proposed which feedbacks the most significant taps of the sparse time domain channel impulse response back to the gNB.

In Section 3.2.6 we jointly investigate UL channel estimation and MIMO detection regarding robustness. The results for the computational complexity show a 16% improvement of the proposed scheme over MMSE in the case that MMSE equalizer cannot be reused for multiple OFDM symbols. In the case that this matrix can be reused 14 times, the complexity advantage is reduced to 10%.

Analog and/or Digital Beamforming/Precoding. Hardware complexity reduction is a main enabler for future massive MIMO systems. In this activity we study low-complex and flexible beamforming solutions.

In Section 3.3.1, ONE5G achieves complexity reduction by using 1-bit phase shifter with a hybrid SLNR precoding scheme with less than 5% performance loss compared to full digital precoding.

Section 3.3.2 investigates the impact of array size and architecture on the spectral efficiency and user throughput under different deployment conditions and identifying most efficient array architectures. The work has not been finished yet.

In Section 3.3.3, massive MIMO is used to improve the performance of multicasting for V2X. Three strategies are considered: 1) unicast beam for every UE, 2) one beam per group, 3) single multicast beam for all UEs. The best performance is achieved by the heuristic grouping strategy, which combines the advantages of the other two (extreme) cases. With a simple heuristic grouping strategy for 2), which groups together UEs transmitting over the same DFT beams gains of more than 100% are achieved compared to the 1) and 3).

In Section 3.3.4, optimal precoders in the case of imperfect CSI at the BSs are studied by solving the expected sum rate stochastic maximization problem using the difference of convex functions approach. The proposed approach provides three times sum-rate gain compared to classical precoders and higher sum-rate than the other approaches.

In Section 3.3.5, functionality split is one of the enablers to deploy massive MIMO RRHs, where the sum-rate performance is constraint by the fronthaul capacity. For limited fronthaul links ONE5G shows that trace-weighted analog beamformer always outperforms state of the art equal combining [PKC+17].

In Section 3.3.6, the flexibility of massive MIMO facilitates the use of in-band backhaul in underserved areas. In such a scenario, it is shown that the degrees of freedom can be used for efficient interference reduction to achieve reliable backhaul links up to 5 km. As a complementary technique, ONE5G proposes to use Probabilistically Shaped Coded Modulation (PSCM) [PX17] for additional SNR gain.

Complementing the hardware complexity reduction from Sec. 3.3.1, complexity reduction can also be achieved with full digital precoding using low resolution ADCs, shown in Sec. 3.3.7. In some scenarios digital precoding with low resolution ADCs is more energy efficient and achieves a higher sum-rate than hybrid precoding systems.

T4.3 Advanced link management solutions for interference coordination and avoidance, based on the assumption of CRAN/DRAN deployments and/or massive MIMO

This task addresses challenges related to multi-node designs including CSI acquisition and feedback, efficient signaling, cell-less designs, fronthauling, resource allocation, scheduling, and network slicing. These studies are highly relevant to eMBB and mMTC use cases, as the CRAN, under cooperative transmissions, can naturally accommodate a huge number of devices with maximum data rates. The solutions presented in this section are grouped into two clusters: “Physical layer techniques and procedures for CRAN/DRAN” and “Resource allocation and traffic management in CRAN/DRAN”.

Physical layer techniques and procedures for CRAN/DRAN, as presented in Section 4.1, addresses physical layer challenges towards achieving the full premise of sophisticated cooperative/coordinated CRAN transmissions. Since cooperation/coordination among links depends crucially on the availability and quality of global CSI at the CU, the focus is mostly on the design of low-overhead, scalable mechanisms for CSI acquisition and feedback.

Section 4.1.1 investigates the reference signal (RS) framework, considered in NR, related to CSI acquisition and feedback for cooperative transmissions. The limitations of the current RS design regarding the feedback of wideband and subband information of selected beams are identified and a solution towards reducing quantization errors is described.

Section 4.1.2 investigates the issue of feedback signalling overhead in NR-CoMP and identifies the limitations of the current framework in implementing the multiple possible transmission modes considered in NR-CoMP. A preliminary discussion on potential approaches for improving performance in terms of feedback overhead is provided.

In Section 4.1.3 the feasibility of low-overhead, global CSI acquisition at the UE side of a downlink, dense, and wide area CRAN deployment is investigated. Using a stochastic geometry model for the system topology, analytical results show that accurate CSI acquisition with very limited overhead can indeed be achieved, however, only under operational conditions with sufficiently large path losses and/or blockage effects (e.g., megacity applications and/or mmWave transmissions)

Section 4.1.4 proposes a novel receiver processing approach based on non-linear filter design and machine learning methods that enables efficient, joint processing of uplink signals by multiple RRHs. Numerical results demonstrate improved performance to previously proposed approaches, with respect to fronthaul link rate requirements and overhead required to train the system.

Resource allocation and traffic management in CRAN/DRAN, as presented in Section 4.2, focuses on the design of higher layer mechanisms and investigates topics ranging from user scheduling/clustering to network slicing, which are particularly challenging in a CRAN scenario due to the multitude of nodes involved.

Section 4.2.1 proposes a scheduling algorithm for partitioning UEs in groups such that the channels of the UEs in each group have small spatial correlation, which is a desirable property as it leads to minimal interference by means of precoding. A convex relaxation of the optimal

problem formulation is proposed, with numerical results demonstrating that the proposed solution results in a system performance equal to a conventional one with double the number of RRHs, thus resulting in significant cost and power savings.

The problem of optimal functionality placement on a given physical substrate is considered in Section 4.2.2. In particular, given a set of distributed units (DUs) and a CU, the optimal functional split, i.e., distribution of communication protocol functions among DUs and CU, is obtained as a solution of a large size mixed integer problem. Using a simulated annealing algorithm to efficiently solve the problem, various examples are provided, demonstrating that the optimal split depends is highly depended on traffic demands. The problem and proposed solutions are highly relevant to network slicing and CRAN functional split, currently investigated in 3GPP.

Section 4.2.3 evaluates the performance of a number of transmission modes currently considered in NR-CoMP. An extensive simulation study indicates that the optimal mode is highly dependent on operational conditions and performance metric considered. This, in turn, suggests an adaptive mode selection approach, aided by a flexible signalling framework, which will be developed in the course of the project.

Section 4.2.4 addresses the problem of cross-link interference (CLI) management in NR duplexing (dynamic TDD) due to mismatched subframe directions. A solution to this problem is proposed allowing the receiver of a link to acquire knowledge of the cross-link interference channel(s) and, in turn, suppress the interference by advanced interference rejection algorithms. System level simulations indicate gains of up to 30% for the downlink where the effect of CLI is most significant.

In Section 4.2.5, centralized scheduling strategies for cooperative incremental redundancy retransmissions in the slow-fading Multiple Access Multiple Relay Channel (MAMRC) are investigated for the Underserved Area scenario. The goal is to maximize the long-term aggregate throughput by applying the proper centralized scheduling strategy of the sources, under a fairness constraint. The performance of three proposed selection strategies is evaluated by simulations depicting close to the theoretical optimal performance.

Contributions to other work packages

Finally, in Chapter 5, we show how the WP4 results contribute to other work packages, in particular WP2 (use cases, and system level simulations) and WP5 (proof of concepts). We briefly elaborate on how WP4 contributes to our two main scenarios, “Underserved Areas” and “Megacities”. We give an overview on the main technical contributions, and we list the associated use cases and KPIs. A link-to-system model is presented as an interface to the WP2 system level simulator. Furthermore, we present seven technologies as candidates for WP5 implementation.

Table of Contents

Executive Summary	4
Table of Contents	9
List of Figures	11
List of Tables	14
List of Acronyms and Abbreviations.....	15
1 Introduction.....	20
2 Future proof multi-service access solutions.....	20
2.1 Design of non-orthogonal multiple access.....	20
2.1.1 Contention-based uplink RACH for NOMA.....	22
2.1.2 Enhanced grant-free access with advanced receivers.....	23
2.1.3 Link-level comparison of NOCA and IDMA.....	23
2.2 Solutions for ultra-reliable and low latency communication.....	25
2.2.1 Access Solutions for URLLC.....	26
2.2.2 Short-Packet Transmission for URLLC Applications.....	28
2.2.3 Interference mitigation for bi-directional URLLC.....	32
3 Massive MIMO enablers towards practical implementation	34
3.1 Flexible hardware architecture and multi-service support.....	34
3.1.1 Impact of array shape in different deployments.....	35
3.1.2 Sector and Beam Management with Cylindrical Arrays.....	36
3.1.3 Flexible and fast reconfigurable HW architecture for multi-service transmission.....	38
3.1.4 MIMO performance prediction.....	40
3.2 Efficient CSI acquisition in TDD/FDD and feedback compression in FDD.....	42
3.2.1 Parametric channel estimation for massive MIMO.....	43
3.2.2 Hierarchical sparse channel estimation for multiuser massive MIMO with reduced training overhead.....	44
3.2.3 On the amount of downlink training for FDD correlated massive MIMO scenarios.....	46
3.2.4 Improving CSI acquisition through spatial multiplexing – TDD/FDD.....	47
3.2.5 Efficient feedback schemes for more accurate CSI and advanced precoding.....	50
3.2.6 Joint investigation of UL channel estimation and massive MIMO detection regarding robustness.....	52
3.3 Analog and/or Digital Beamforming/Precoding.....	53
3.3.1 Genetic algorithm assisted hybrid beamforming for wireless fronthaul.....	54
3.3.2 Hybrid array architectures covering different deployment scenarios.....	55
3.3.3 Multicast Beamforming.....	57
3.3.4 Decentralized beamforming algorithms.....	58
3.3.5 Massive MIMO with Hybrid Analog-Digital Precoding in a CRAN Architecture.....	60
3.3.6 Enhanced Backhauling.....	61
3.3.7 A Comparison of Hybrid Beamforming and Digital Beamforming with Low-Resolution ADCs for Multiple Users and Imperfect CSIR.....	65
4 Advanced link coordination based on CRAN/DRAN and massive MIMO.....	66
4.1 Physical layer techniques and procedures for CRAN/DRAN.....	67
4.1.1 Reference signal framework in new radio for cooperative transmission.....	68
4.1.2 Efficient signalling in massive MIMO multi-node networks.....	69
4.1.3 Compressive channel estimation in CRAN.....	71
4.1.4 Nonlinear mechanisms in cell-less systems.....	73

4.2	Resource allocation and traffic management in CRAN/DRAN	74
4.2.1	Architecture optimization for Cell-less mMIMO systems.....	75
4.2.2	Optimised functionality placement and resource allocation in a CRAN/DRAN context	78
4.2.3	Centralized and distributed multi-node schedulers for non-coherent joint transmission	80
4.2.4	NR duplexing with CRAN and network coordination.....	81
4.2.5	Centralized scheduling on the uplink of the Multiple Access Multiple Relay Channel (MAMRC).....	83
5	Use cases, system-level evaluation and proof of concept	85
5.1	Proof of Concept (PoC)	85
5.2	Connection to WP2 use cases	85
5.3	Contribution to ONE5G scenarios	87
5.4	Link-to-System Model.....	87
6	Conclusions and future work.....	90
7	References.....	92
8	Appendix.....	99
8.1	Sector and Beam Management with Cylindrical Arrays	99
8.2	Hierarchical sparse channel estimation for multiuser massive MIMO with reduced training overhead	101
8.3	CSI feedback for FDD massive MIMO	102
8.4	Genetic Algorithm Assisted Hybrid Beamforming for Wireless Fronthaul	105
8.5	Hybrid array architectures covering different deployment scenarios	106
8.6	Decentralized beamforming algorithms.....	107
8.7	Massive MIMO with Hybrid Analog-Digital Precoding in a CRAN Architecture ...	108
8.8	Description of the RZF-CI precoder	109
8.9	Nonlinear Mechanisms in Cell-less Systems	110
8.10	Architecture optimization for Cell-less mMIMO systems.....	114
8.11	Optimised functionality placement and resource allocation in a CRAN/DRAN context.....	117
8.12	Centralized and distributed multi-node schedulers for non-coherent joint transmission	118
8.13	NR duplexing with CRAN and network coordination.....	119

List of Figures

Figure 2-1 Overall work plan for NOMA SI.....	21
Figure 2-2 Illustration of coded random access with successive interference cancellation for three UEs (represented by rows) and three slots (represented by columns). The packets are decoded and cancelled in the given order.	23
Figure 2-3 Generic NOCA receiver structure	24
Figure 2-4 NOCA and IDMA comparison, 10 users, repetition rate $\frac{1}{4}$ (IDMA), spreading factor 4 (NOCA), WSSUS channel (60km/h), BPSK	25
Figure 2-5 URLLC Uplink Grant-Free Transmission with reactive HARQ and power boosting on the retransmissions. P is the transmit power without boost and g() indicates the requested boost.	27
Figure 2-6 Outage at 1ms for different power control configurations and loads.	28
Figure 2-7 Packet structure and Packet error probability (PER).....	29
Figure 2-8 Graphical model representation for the Bayesian inference-based decoder.....	31
Figure 2-9. Left: Probability of correct user activity detection (for all active users) versus false alarm rate; Right: Single user performance: block error probability (BER) vs. total received SNR, for joint user activity detection and data decoding.....	31
Figure 2-10 Adjacent services utilizing different numerologies	32
Figure 2-11 Spectra related to different subcarrier spacing of two numerologies (Numer.1 and Numer.2).....	33
Figure 2-12 Left: Out-Of-Band emissions for precoded transmit signal and guard band inserted with different overhead. Right: average Out-Of-Band interference vs data rate reduction for precoded and guard band inserted.	34
Figure 3-1 CDF of the UE SINR for different UPA shapes.....	36
Figure 3-2 Cell border throughput versus cell spectral efficiency for different UPA shapes	36
Figure 3-3: Wideband SINR due to sectorization with UPAs.....	37
Figure 3-4: Single user spectral efficiency with MRT precoding without interference.	37
Figure 3-5: Daytime scenario.	37
Figure 3-6: Evening time scenario.	37
Figure 3-7: Impact from Inter-Sector Interference. IF in the legend stands for “Interference” ..	38
Figure 3-8: Sum rate over 3 sectors each serving 30 users in comparison with joint precoding over all antennas serving 90 users.	38
Figure 3-9 Rx Digital Front-End.....	39
Figure 3-10 Forward Error Correction enabler	40
Figure 3-11 : Illustration of the channel hardening measure on a simple ray-based channel model. Small-scale and large-scale contributions are highlighted.	
Figure 3-12: Performance of the physical channel estimator for 16 and 256 antennas with respect to the number of virtual paths.	44
Figure 3-13 a) Mean squared error (MSE) performance of proposed algorithm (HiHTP) and conventional algorithm (IHT) for the single user case. (b) MSE performance of proposed algorithm for the multiuser case.....	46

Figure 3-14 (left) Effect of c on the training duration (right) Effect of correlations on the training duration	47
Figure 3-15 Comparison of sum spectral efficiency vs. SNR.....	49
Figure 3-16 Comparison of Jain's fairness index vs. SNR	49
Figure 3-17 Spectral Efficiency of the proposed clustering scheme.....	50
Figure 3-18 Explicit CSI Feedback System Model.....	51
Figure 3-19 Limited CSI Knowledge Causes Loss of Sparsity.....	51
Figure 3-20 OMP Flowchart	51
Figure 3-21 Normalized spectral efficiency for the different feedback schemes.....	52
Figure 3-22 Simulations results for un-coded (left) and coded (right) BER for MMSE and Sequential DCD MU-MIMO detection with full channel knowledge (Ideal) and real channel estimation (ChEst).....	53
Figure 3-23 Mean remote node sum rate comparison of digital/hybrid SLNR/ZF beamforming schemes ($NT = 8, K = 3, NR, l = 1 \forall l$).....	54
Figure 3-24 Beam patterns of digital SLNR (left) and hybrid SLNR beamforming (right).	55
Figure 3-25: Hybrid array architecture.....	56
Figure 3-26: Spectral efficiency versus the number of simultaneously served UEs, each served with 2 x-polarized MIMO layers, with different array sizes	56
Figure 3-27 Multicast beamforming. The same signal is transmitted over each beam.....	57
Figure 3-28 Comparison of unicast and multicast MIMO transmission	58
Figure 3-29 Sum rate comparison: Correlated low rank matrices, 1 antenna/user,4 users/cell, 8 antennas per base station, 2 cells, 25% estimation error	59
Figure 3-30 Sum rate comparison: Correlated low rank matrices, 2 antennas/user,4 users/cell, 8 antennas per base station, 2 cells, 25% estimation error.....	60
Figure 3-31: Sum-rate vs. the number of active RF chains.....	61
Figure 3-32 Illustration of (a) a wired BH and (b) the proposed in-band wireless BH with mMIMO.	62
Figure 3-33 (a) BER on the UE side as a function of the SNR and (b) normalized received power on the BS 2 as a function of the distance between the BS 1 and the UE.	63
Figure 3-34 The PAS system investigated	64
Figure 3-35 Frame error rate (FER) for two DM architectures: BL-PAS = BL-DM+PAS, SL-PAS = symbol-level CCDDM+PAS.	64
Figure 3-36 Number of operations in terms of MAC (multiply-accumulate) operations for the CCDDM (blue) and the BL-DM with parallel processing (violet). The processing of B_1, B_2, B_3 , can be done in parallel, thus resulting in a lower number of MAC/symbol than symbol-level CCDDM. However, the sum of MAC/symbol is higher for B_1, B_2, B_3	64
Figure 3-37 Avg. sum-spectral efficiency for different receiver configurations and ADC resolution 1-8 bit different number of RF chains M_{RFE}	65
Figure 3-38 Spectral and energy efficiency of digital beamforming hybrid beamforming with 64 receive antennas and 4 users at different SNR with ADC resolution 1-8.	66
Figure 4-1 Different resource allocation schemes result in different interference scenarios.	70
Figure 4-2 Illustration of the interference scenarios on different PRBs.....	71

Figure 4-3 MSE of the global CSI estimate for the CS-based channel estimator as a function of the training sequence length Np . The analytical expression for the MSE of the oracle estimator is also shown.	73
Figure 4-4 (left) Performance of D&F Vs. Q&F with $M = 3$ antennas at each RRH and $K = 6$ devices transmitting in the uplink. The results are shown for an increasing number of RRHs denoted by $R \in \{2, 3, 4\}$ and 100 training samples. (right) Performance of D&F (with $R = 3, Bq = 4$) Vs. Single BS (CS) [ACY+18]. The results are shown for an increasing cluster size $K \in \{4, 6, 8, 10\}$	74
Figure 4-5 Comparison of CDFs of normalized large-scale correlation for $K = 20, C = 4$	77
Figure 4-6 Average Downlink Throughput versus the number of APs and different K values ..	77
Figure 4-7 Function Split between central and distributed unit [3GPP-38801]	78
Figure 4-8 Traffic distribution percentage to distributed units for different QoS requirements .	79
Figure 4-9 Percentage of each function split option.....	80
Figure 4-10 CDFs of user throughputs of different network coordination schemes (5 users per TRP)	81
Figure 4-11 UL DMRS locations of the interfering UE (left) and DL ZP CSI-RS (right) for the desired UE	82
Figure 4-12 Median DL throughput comparison (left) and median UL throughput comparison (right).....	82
Figure 4-13 Orthogonal Multiple Access Multiple Relay Channel (OMAMRC) with feedback.	83
Figure 4-14 Long-term aggregate throughput of different strategies with slow link adaptation and symmetric link scenario.	84
Figure 5-1 Spectral Efficiency of the proposed clustering scheme	89
Figure 8-1: User distribution	100
Figure 8-2: Sectorized UPA with colour coded sectors.	101
Figure 8-3: Sectorized UCA with colour coded sectors.....	101
Figure 8-4 Example support of matrix Wt for the case of three users in total with only the first and third active.	102
Figure 8-5 Array types: x-pol elements arranged in subpanels with different size.....	106
Figure 8-6 System model.	109
Figure 8-7 (left) Bit Error Rate for PLAF and MMSE-SIC for $M=3$ and $K=5,4,3$. (right) Bit Error Rate for PLAF and NLAf for $M=3, K=5$, and active device %=100, 75 and 60.	111
Figure 8-8 Centralized and distributed schedulers (left) and examples of PRB allocations of DPS/F-NCJT/NF-NCJT (right).....	118
Figure 8-9 Procedure of the proposed UE-to-UE CLI management.....	120

List of Tables

Table 2-1. Mapping between ONE5G and 3GPP on NOMA topics.....	21
Table 3-1. FEC parameters in LTE and NR systems.	39
Table 4-1 Simulation parameters	76
Table 4-2 Comparison between different network coordination schemes.	81
Table 5-1. Candidate technologies for implementation in WP5	85
Table 5-2. Overview on WP4 technologies and their contribution to WP2 use cases	85
Table 8-1: Simulation assumptions	101
Table 8-2: Array configurations used for simulations.....	106
Table 8-3: System simulation parameters	106
Table 8-4. Key evaluation assumptions.	119
Table 8-5. Key evaluation assumptions for NR duplexing.	120

List of Acronyms and Abbreviations

3GPP	3rd Generation Partnership Project
5G	Fifth Generation
ACK	Acknowledgement
ADC	Analog-Digital Converter
AP	Access Point
APSM	Adaptive Projected Subgradient Method
AR	Augmented Reality
ARPU	Average Revenue Per User
AWGN	Additive White Gaussian Noise
BBU	Baseband Unit
BER	Bit Error Rate
BH	Backhaul
BL-DM	Bit-Level Distribution Matcher
BLER	Block Error Rate
BP	Basis Pursuit
BPSK	Binary Phase Shift Keying
BS	Base Station
CCDM	Constant Composition Distribution Matcher
CDF	Cumulative Distribution Function
CF	Cell Free
CFO	Carrier Frequency Offset
CFR	Channel Frequency Response
CIR	Channel Impulse Response
CJT	Coherent Joint Transmission
CLI	Cross Link Interference
CoMP	Coordinated Multipoint
CP	Cyclic Prefix
CPRI	Common Public Radio Interface
CPU	Central Processing Unit
CQI	Channel Quality Indicator
CRA	Coded Random Access
CRAN	Cloud Radio Access Network
CRC	Cyclic Redundancy Check
CS	Compressive (Compressed) Sensing
CSCB	Coordinated Scheduling Coordinated Beamforming
CSI	Channel State Information
CSIR	CSI at Receiver
CSI-RS	CSI Reference Signal
CSIT	Channel State Information at Transmitter
CU	Central Unit
DAC	Digital to Analog Converter
DBF	Digital Beamforming
DC	Direct Current

DCD	Dichotomous Coordinate Descent
DCI	Downlink Control Information
DFE	Digital Front End
DFT	Discrete Fourier Transform
DFT-s-OFDM	DFT-spread-OFDM
DL	Downlink
DM	Distribution Matcher
DM-RS	Demodulation Reference Signal
DPS	Dynamic Point Selection
DRAN	Distributed Radio Access Network
DU	Distributed Unit
EBF	Eigen-Beam-Former
EC	European Commission
eMBB	Enhanced Mobile Broadband
eMBMS	Evolved Multimedia Broadcast Multicast Services
ESE	Elementary Signal Estimator
ETF	Explicit Time Domain Feedback
EWSMSE	Expected Weighted Sum Mean Squared Error
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FE	Functional Entity
FEC	Forward Error Correction
FER	Frame Error Rate
F-NCJT	Fully Overlapped NCJT
FTP	File Transfer Protocol
GA	Gaussian Approximation
GB	Grant Based
GF	Grant Free
gNB	basestation in NR
GoB	Grid of Beam
HARQ	Hybrid Automatic Repeat reQuest
HBF	Hybrid Beamforming
HiHTP	Hierarchical HTP
HiHT	Hierarchical IHT
HTP	Hard Thresholding Pursuit
HW	Hardware
IC	Interference Cancellation
IDMA	Interleave Division Multiple Access
IF	Interference
IHT	Iterative Hard Thresholding
IPR	Intellectual Property Right
IQ	In phase/Quadrature phase
IRC	Interference Rejection Combining
ISD	Inter-Site Distance
JD	Joint Detection

JNCC	Joint Network-Channel Coding
JSDM	Joint Spatial Division and Multiplexing
JT	Joint Transmission
KPI	Key Performance Indicator
LDPC	Low Density Parity Code
LMMSE	Linear MMSE
LOS	Line of Sight
LS	Least Squares
LTE	Long Term Evolution
MA	Multiple Access
MAMMOET	Massive MIMO for Efficient Transmission
MAMRC	Multiple-Access Multiple Relay Channel
MAP	Maximum a posteriori
MBB	Mobile Broadband
MCS	Modulation and Coding Scheme
MIMO	Multiple-Input Multiple-Output
MISO	Multiple Input Single Output
ML	Maximum Likelihood
mMIMO	Massive MIMO
MMSE	Minimum Mean Square Error
mMTC	Massive Machine Type Communication
MRC	Maximum Ratio Combining
MRT	Maximum Ratio Transmission
MS	Mobile Station
MSE	Mean Squared Error
MTC	Machine Type Communication
MU-MIMO	Multi User MIMO
NCJT	Non Coherent Joint Transmission
NF-NCJT	Non-fully-overlapped NCJT
NGFI	Next Generation Fronthaul Interface
NLAF	Non-linear filter
NOCA	Non-Orthogonal Coded Access
NOMA	Non-Orthogonal Multiple Access
NP	nondeterministic polynomial time
NR	New Radio
OFDM	Orthogonal Frequency Division Multiplexing
OLPC	Open Loop Power Control
OMP	Orthogonal Matching Pursuit
OOB	Out of Band
PAS	Probabilistic Amplitude Shaping
PBCH	Physical Broadcast Channel
PDCCH	Physical Downlink Control Channel
PDCP	Packet data compression protocol
PDP	Power Delay Profile
PDSCH	Physical Downlink Shared Channel

PER	Packet error probability
PHY	Physical Layer
PIC	Parallel Interference Cancellation
PLAF	Partially Linear Filter
PMI	Precoding Matrix Index
PoC	Proof of Concept
POCS	Projection Onto Convex Sets
PPS	packets per second
PRB	Physical Resource Block
PSCM	Probabilistically Shaped Coded Modulation
PTM	Point to multipoint
PT-RS	hase tracking reference signal
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
QAM	Quadrature Amplitude Modulation
QCLed	Quasi-Colocated
QoS	Quality of Service
QP	Quadratic Program
QPSK	Quadrature Phase Shift Keying
RACH	Random Access Channel
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block
RF	Radio Frequency or Radio Frontend
RI	Rank Indicator
RIP	Restricted isometry property
RKHS	Reproducing Kernel Hilbert Space
RLC	Radio Link Control
RMa	Rural Macro cell
RRC	Radio Resource Control
RRH'	Remote Radio Head
RRM	Radio Resource Management
RS	Reference Signal
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RZF	Regularized zero-forcing
RZF-CI	Regularized Zero Forcing with Controlled Interference
SC-FDMA	Single-Carrier FDMA
SC-PTM	Single Cell Point To Multipoint
SDF	Selective Decode-and-Forward
SDR	Semi-definite Relaxation
SDU	Service Data Unit
SE	Server Entity
ShDec	Shaping Decoder
ShEnc	Shaping Encoder

SIC	Successive Interference Cancellation
SIMO	Single Input Multiple Output
SINR	Signal-to-Interference-plus-Noise Ratio
SIR	Signal to Interference Ratio
SLNR	Signal-to-Leakage-plus-Noise Ratio
SL-PAS	Symbol-level CCDDM+PAS
SNR	Signal-to-Noise Ratio
SPARCS	Sparse Regression Codes
SRS	Sounding reference signal
SSB	Synchronization Signal Block
STD	Standard Deviation
SU-MIMO	Single User MIMO
SVD	Singular Value Decomposition
TCI	Transmission configuration information
TDD	Time Division Duplex
TeC	Technical Component
TP	Transmission Point
TRP	Transmission Reception Point
TTI	Transmission Time Interval
TX	Transmit, Transmitter
UCA	Uniform Cylindrical Array
UE	User Equipment
UL	Uplink
ULA	Uniform Linear Array
UMa	Urban Macro
UMi	Urban Micro
UPA	Uniform Planar Array
URLLC	Ultra-Reliable Low Latency Communications
V2X	Vehicular to Everything
VR	Virtual Reality
WP	Work Package
WSSUS	Wide-Sense Stationary Uncorrelated Signals
ZC	Zadoff-Chu
ZF	Zero Forcing
ZP	Zero Power

1 Introduction

This document presents intermediate results from ONE5G WP4. In particular, we show technical solutions for the various vertical 5G use cases identified in WP2 [ONE17-D21]. The focus is on PHY/MAC enabling technologies, with an emphasis on implementation aspects. We highlight the relation to the relevant technical fields in 3GPP standardization discussion, and we discuss our view on how WP4 technical contributions can impact current and future 3GPP releases (release 16 and beyond). For each of the technical contributions, an overview of its concept and initial results will be presented, and an outlook to the planned work in later stages of the project will be given. The report is complemented by a technical annex leading to more technical details, and references generated jointly in the context of the WP4 work or leading to the current 3GPP framework.

The remainder of this document is organized as follows. Section 2 presents future-proof multi-service access solutions proposed in ONE5G. Massive MIMO enablers towards practical implementation are described in Section 3. In Section 4, we investigate how to leverage massive MIMO in a multi-node/CRAN/DRAN network structure. In Section 5, we discuss how the results contribute to WP2 use cases and system level simulations, as well as to WP5 proof-of-concept. Conclusions will be drawn and future work will be highlighted in Section 6.

2 Future proof multi-service access solutions

The emerging 5G service categories *Ultra-Reliable Low-Latency Communication* (URLLC) and *massive Machine Type Communication* (mMTC) introduce new challenges in providing efficient end-to-end (E2E) connections for user, beyond what is possible in current wireless technologies. Providing fast and reliable access and supporting a much larger number of users are two key driving factors, with different requirements specific to the different service-classes. For this we address various future-proof multi-service access solutions. The “Megacity” scenario is the main target in this task. However, many of the developed technology components can also provide gains for ‘underserved areas’. We have grouped those investigations into the following two clusters: “*Design of non-orthogonal multiple access*” and “*Solutions for URLLC services*”. In all cases, a major design target has been to ensure efficient co-existence of the new 5G service categories (mMTC and URLLC) with enhanced mobile broadband (eMBB) services. The main focus is on KPIs measuring access rate (e.g., number of connected devices per area), high reliability and low latency.

Non-Orthogonal Multiple Access (NOMA), is designed to concurrently serve multiple users, for example, by exploiting the differences in their power levels through the use of successive interference cancellation (SIC) at the receiving end. NOMA schemes achieve spectral efficiency gains and improve user fairness compared to conventional multiple access schemes [DLC+2017]. Section 2.1 addresses the design of NOMA schemes. In particular, transmitter side signal processing techniques, receiver design, performance enhancement procedures and system level performance evaluation for NOMA are considered.

Transmission of short packets with fast access to the network is required in order to meet the target URLLC requirements of $1-10^{-5}$ reliability with 1ms user plane latency for a packet of 32 bytes payload [3GPP TR38913]. Section 2.2 studies URLLC solutions, such as multiple access enhancements, for short packet transmissions.

2.1 Design of non-orthogonal multiple access

NOMA schemes allow multiple users to access the radio network using overlapping radio resources in the time and frequency domain, and/or user separation via advanced signal processing at the network side in the power domain. The design of NOMA has gained significant

interest in the 3GPP study on NR technologies for the services envisioned for 5G systems [3GPP-38802]. Potential advantages include higher cell throughput with improved cell-edge performance due to increased capacity region for eMBB, higher reliability and lower latency through, e.g., frequency diversity gain and robustness to collisions for URLLC, and higher connection density with reduced signalling overhead and power consumption for mMTC [R1-1715576].

In the recent RAN1#92 meeting, NOMA has been extensively discussed. In particular, the following work plan, as shown in Figure 2-1, has been discussed [R1-1801414].

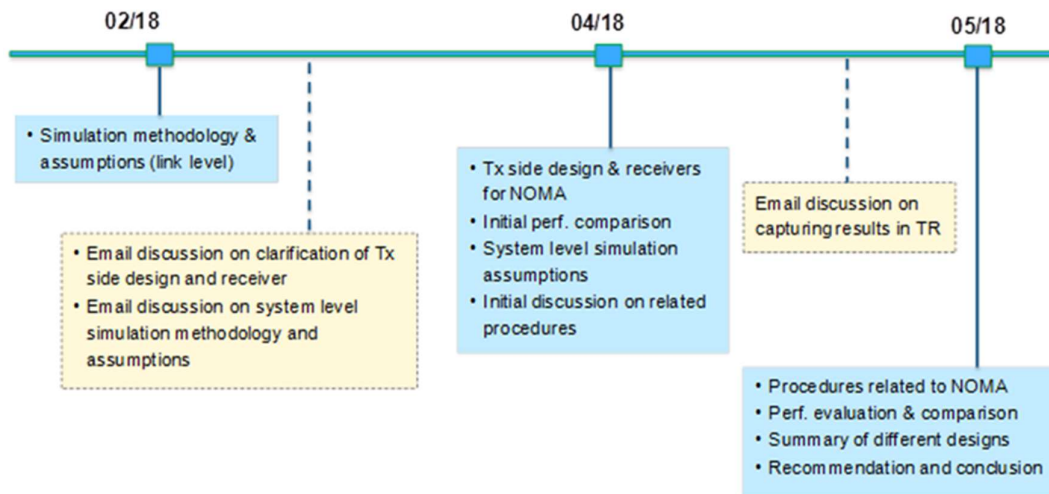


Figure 2-1 Overall work plan for NOMA SI

Detailed topics of study on NOMA have been listed in 3GPP [R1-1802005]. NOMA is one of the topics that are identified as enablers for mMTC, therefore enabling future-proof multi-service access, within 5G. The studies we have conducted in this area are detailed in the subsections that follow.

There exists a plethora of NOMA schemes, which mainly differ in the way dedicated user signatures are generated. Possible options include user equipment (UE)-specific channel coding, interleaving or scrambling on the bit level, and linear/non-linear spreading with potentially sparse resource allocation patterns on the symbol level. Furthermore, users can also be distinguished by their effective propagation channels depending on the location and power allocation. In general, advanced receivers based on successive/parallel interference cancellation (SIC/PIC) or joint detection (JD) are required to exploit the full potential of NOMA. Not all of these topics can be addressed by ONE5G. Constrained by the available resources, the project will address a subset of options. To also align with the time plan of 3GPP topics, the work under the studies of NOMA will be focused on the code and signature designs for NOMA, the design of NOMA receivers, as well as link-level evaluation and analysis of the system. Table 2-1 shows how the ONE5G activities are mapped to the 3GPP topics.

Table 2-1. Mapping between ONE5G and 3GPP on NOMA topics

3GPP [R1-1802005]	ONE5G NOMA studies
Transmitter side signal processing schemes for non-orthogonal multiple access	Non-orthogonal multiple access and code design, planned work in T4.1; Preamble signatures, planned work in T4.1
Receivers for non-orthogonal multiple access	Enhanced grant-free access with advanced receivers, Section 2.1.2.
Procedures related to the non-orthogonal multiple access [RAN1]	Contention based uplink RACH for NOMA, Section 2.1.1.

Link and system level performance evaluation or analysis for non-orthogonal multiple access	Link-level comparison of NOCA and IDMA, Section 2.1.3
---	---

2.1.1 Contention-based uplink RACH for NOMA

A large number of current studies on NOMA, which have been reported in the literature and in 3GPP, have been focused on the transmission and detection of NOMA signals with the devices being at the RRC_CONNECTED state, i.e., when the UE is already connected to a given cell. In particular, user clustering for downlink NOMA transmission has been reported. For example, in a downlink NOMA system, different UEs with distinct channel gains are grouped into clusters and scheduled accordingly, to facilitate the applications of SIC at the receiver. The centralized processing facilitates global SINR or channel information between the BS and each UE. However, the same approach cannot be applied to the uplink.

The current work is concerned with clustering the UEs at the initial access stage, such that preambles can be allocated according to the clusters, to reduce probability of collision during the transmission in a contention based random access. The use of NOMA at the initial access stage, i.e., at RRC_IDLE, can be extremely challenging, especially when clustering UEs according to the differences in the power domain is considered. At RRC_IDLE, there is no *centralized* information of the conditions of the UEs nor the channel, therefore most of the existing clustering methods originally intended for downlink transmission cannot be applied here. In addition, BS cannot do any central processing based on the instantaneous received information/measurements at the UEs at this stage, as the UEs are not attached to a cell yet. For example, the BS cannot gather all instantaneous measurements from the UEs and cluster the UEs accordingly. Last but not least, the UEs, including their channel conditions, can be very dynamic, especially when they are highly mobile. It is therefore extremely challenging to cluster the UEs in such a rapidly changing context.

The use of machine learning can help address the aforementioned challenges. For example, the UEs can be clustered based on their measurements at the initial access stage. Examples of such measurements could be Reference Signal Received Power (RSRP) and Reference Signal Received Quality (RSRQ), as indicators of SINRs between the UEs and the BS. In a conventional RACH procedure, collision is expected when the users are using the same preamble (randomly assigned to the user) and attempt to transmit at the same time/frequency slot, subject to a collision probability, denoted as P_c . Essentially, P_c is the probability where two users attempt to transmit at the same time/frequency slot, and are randomly allocated the same preamble. A classification machine learning algorithm employed at the UE can then be used to cluster the UEs, according to their measured power. For example, users whose measured power have significant difference can be clustered into one cluster, and they can use the same group of preambles, while another cluster of users will use a different group of preambles, therefore there will be no collision between two different clusters. On the other hand, users belong to the same cluster with sufficient difference in their received power facilitates their separation based on the NOMA concept, even when they attempt the transmission at the same time/frequency slot and with the same spreading code. Ideally if users grouped in the same cluster have sufficient differences in their power and can be separated via NOMA concept, there will be no collision happening in each cluster as well, and RACH will become collision free. In reality, however, the collision probability depends on the number of clusters, available preambles, and the number of users. Take, for example, a system with K clusters, P preambles, and N users. When $K=1$, the collision probability is P_c , the same as the conventional RACH, as if the users are not clustered. When $K=N=P$, the collision probability is zero, as it is equivalent to the case where every user is allocated a unique preamble. When $K < P < N$, the probability of collision is firstly reduced by a factor of K , due to the fact that different clusters will be using different groups of non-overlapping preambles, therefore collisions will not occur inter clusters. Let p_0 be the probability that the power difference is smaller than a threshold which makes them not separable at the receiver. Therefore, the collision probability in one cluster is $P_c p_0 / K$.

2.1.2 Enhanced grant-free access with advanced receivers

Grant-free uplink transmission is an essential enabler for 5G in order to reduce latency for URLLC and signalling overhead for short packets in mMTC scenarios. This is relevant for all use cases in Table 5-2 including URLLC and/or mMTC. Classical random-access approaches like the slotted ALOHA protocol are limited by packet collisions and only work for low system loads, defined as the average number of active users per slot (where a slot generally refers to a set of time-frequency resources). Therefore, 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 successfully decoded packets, as illustrated in Figure 2-2.

	Slot 1	Slot 2	Slot 3
User 1	1) Decode		2) Cancel
User 2		4) Cancel	3) Decode
User 3		5) Decode	

Figure 2-2 Illustration of coded random access with successive interference cancellation for three UEs (represented by rows) and three slots (represented by columns). The packets are decoded and cancelled in the given order.

The number of repetitions can be optimized for a given system load using tools from coding theory such as density evolution, and some further gains are possible by replacing the simple repetitions with more general packet erasure codes [PLC15]. However, a key assumption in most publications is that packets received without collision can always be correctly decoded and that the interference is ideally cancelled, which is in general not the case for short packet transmission over fading or fast changing channels. On the other hand, colliding packets are treated as erasures, which may be too pessimistic considering multi-user detection at the receiver, in particular in combination with multiple antennas. Some recent work showed the possible gains of multi-packet reception for a simplified channel model [SPL17].

In the course of this project, we will compare CRA schemes based on packet-level coding with NOMA on the physical layer. The latter allows for resolving collisions by UE-specific signatures using advanced receivers. Furthermore, channel codes with lower rates can be applied over multiple slots instead of encoding each slot individually, which is particularly important for short messages and fading channels. This approach would not change the total amount of used resources per message, but can increase the coding gain, since longer codewords suffer less from finite length losses compared to short codewords. We will evaluate the performance for different scenarios to obtain fundamental insights on the design of grant-free access for URLLC and mMTC services that have different requirements regarding reliability and power consumption.

2.1.3 Link-level comparison of NOCA and IDMA

In this section, let's focus on how Non-Orthogonal Code Access (NOCA) [ZMZ+16] and Interleave Division Multiple Access (IDMA) [PLW+06] use their dedicated signatures to separate the signals. For IDMA, user specific interleaving will be exploited for the users. For NOCA, the Zadoff-Chu (ZC) sequences are introduced to spread the user signals before superimposing them. Basically, the ZC sequences with spreading factor N_{zc} can generate $N_{zc} - 1$ root sequences. Each root index r can construct an orthogonal set with N_{zc} ZC sequences with cyclical shift. The cross-correlation of two arbitrary ZC sequences from different orthogonal sets (with different root index r) is $N_{zc}^{-1/2}$. Thus, $N_{zc} (N_{zc} - 1)$ non-orthogonal sequences can be generated with respect to a given spreading factor N_{zc} . In the NOCA system, every user selects randomly one ZC sequence from the complete non-orthogonal set. Obviously, a collision can happen. In [WZM+17], the NOCA receiver exploits hard Parallel Interference Cancellation (PIC) to recover the user signals. Throughout this section, for the first time Gaussian Approximation (GA) is

applied for NOCA to realize soft PIC, to balance the performance-complexity trade-off of NOCA system.

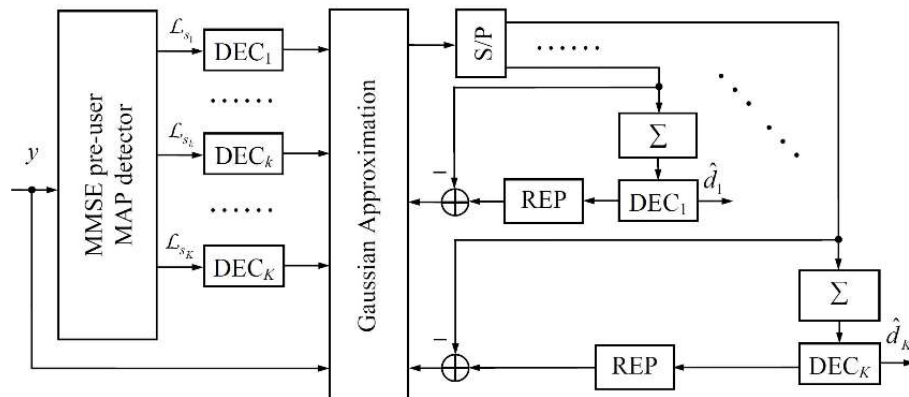


Figure 2-3 Generic NOCA receiver structure

Figure 2-3 presents the NOCA receiver with MMSE per-user MAP detector with GA. First of all, the MMSE criterion can provide a very rough estimation, even if the number of users K is bigger than the spreading factor N_{zc} . The soft-out information from the MMSE detector can serve as *a priori* information for the GA. The processing of the GA is similar to the Elementary Signal Estimator (ESE) of IDMA. Within the GA processor, the so-called soft PIC operation will estimate the first and second order statistics of the user signals. The soft-out values of GA processing will then be parallelized to K iterative loops simultaneously. Thus, the following expectations for NOCA and IDMA with respect to performance-complexity trade-off can be established:

- IDMA might exhibit better convergence capabilities by means of better randomization due to user-specific interleaving with relatively higher complexity.
- NOCA might outperform IDMA with very few iterations due to efficient MMSE filtering with relatively lower complexity.

The left part of Figure 2-4 presents the normalized complexity of NOCA and IDMA, respectively, by counting the number of numerical operations, such as addition, multiplication, division, exponential operation, and so on. IDMA is regarded as the benchmark performance. Full NOCA is a symbol-by-symbol algorithm. The symbol-based MMSE matrix operation makes it even more complex than IDMA. Thus, the NOCA-MMSE ‘5’ and NOCA-MMSE ‘25’ schemes are introduced, which reuse the MMSE weights every 5 symbols and 25 symbols, respectively. Both algorithms can reduce the computational complexity compared with IDMA, especially for 0th iteration. In the 1st iteration, since the computational requirement of GA becomes the dominant cost, the difference between NOCA and IDMA in sense of complexity becomes relatively smaller. It is observed that GA based NOCA-MMSE has lower complexity than IDMA, if no iteration or only one iteration is performed. Next, the link level performance in Figure 2-4 (right side) will be discussed. Under Wide-Sense Stationary Uncorrelated Scattering (WSSUS) channel with 60km/h user mobility, the NOCA-MMSE ‘25’ scheme does not work properly, due to the high fading rate. Notice that the NOCA-MMSE ‘5’ scheme can deliver proper performance without any iteration, and still outperforms IDMA with one iteration. From the 2nd iteration on, IDMA performs better due to the better convergence capability.

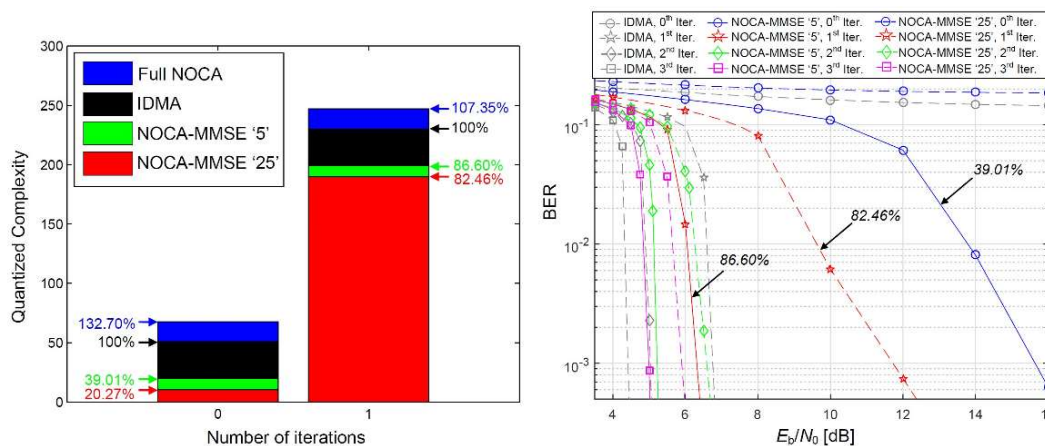


Figure 2-4 NOCA and IDMA comparison, 10 users, repetition rate $\frac{1}{4}$ (IDMA), spreading factor 4 (NOCA), WSSUS channel (60km/h), BPSK

GA based detection has been investigated applying a parallelized structure which makes it very appropriate for hardware implementation, while having strong latency constraints. Exploiting GA based detection for NOCA, the performance of NOCA is compared to IDMA under fair conditions. IDMA exhibits potential overloading capability for achieving high spectral efficiency and fast convergence with several iterations due to user-specific randomization. NOCA exhibits another kind of flexibility. With the pre-processing of MMSE per-user MAP detection, certain performance can be guaranteed even without iterations. MMSE is capable of delivering reliable soft-output to GA based detection. This allows NOCA to outperform IDMA for the 1st iteration, whose overall complexity is between 80% to 90% of that of IDMA. Once more than one iteration is affordable for GA based detection, IDMA turns out to be an effective option for quick convergence. For more details, see the ONE5G publication [C18].

2.2 Solutions for ultra-reliable and low latency communication

URLLC is a new service class introduced by 3GPP in 5G NR. 3GPP NR standards will need to support the URLLC requirements as follows [3GPP-38913]:

- Control plane latency of 10 ms, measured from a battery efficient state to start of continuous data transfer
- User plane latency down to 0.5 ms for both UL and DL. This is the time it takes to successfully deliver an application layer packet/message from the radio protocol layer 2/3 SDU ingress point to the radio protocol layer 2/3 SDU egress point via the radio interface in both uplink and downlink directions
- No mobility interruption time
- Up to 5-nines (99.999%) reliability, when targeting a 1 ms user plane latency.

The current discussions in 3GPP on NR URLLC is grouped into four different study items. The first deals with the support of separate channel quality indicator (CQI) and modulation and coding scheme (MCS) table(s) for URLLC. This includes the option of configuring two block error rate (BLER) targets for CQI reporting. The second agenda item studies the potential benefits of introducing a new Downlink Control Information (DCI) format with a smaller payload. Using a smaller DCI size permits lowering the DCI code rate, which in turn allows robust transmission on the user plane. The necessity of PDCCH repetition, which can be useful in achieving high reliability in certain scenarios such as one-shot transmission, is investigated next. The final item is a study on handling uplink (UL) multiplexing of transmission with different reliability requirements, which considers both intra-UE and inter-UE multiplexing.

This section presents various solution approaches for URLLC in the uplink. Though all layers of communication affect the reliability and the latency, it is important to eliminate competition among multiple users in the access layer and to improve the new packet design and receiver for reliable short packet transmission in the physical layer. As particular examples, different access solutions for URLLC are presented in Section 2.2.1, while Section 2.2.2 discusses optimized transmission techniques for short packets relevant for URLLC applications.

Several critical aspects are addressed when developing solutions for URLLC communication. Reducing the system overhead is an important issue, for example, by combining or even removing some of the procedures for channel training, user scheduling, and resource allocation. In this context, grant free (GF) access schemes emerge as a candidate solution, which have the potential of reducing the latency with respect to traditional grant-based (GB) approaches of the LTE radio standard. Other possible solutions covered here are based on hybrid schemes for radio access, which trade optimally quality of service (QoS), latency requirements and spectral efficiency.

Another aspect covered here is the reliable transmission in a multiuser scenario with potentially high number of system devices sporadically transmitting short packets. This scenario may be of interest in both mission-critical and mMTC applications. Therefore, fundamental importance is the study of the throughput-latency-reliability trade-off, given the number of system devices and the probabilities of user activation. In the context of short data transmissions, the design of frame structure conventionally consisting of preamble, metadata and data, is revisited.

2.2.1 Access Solutions for URLLC

Various challenges emerge for enabling URLLC for mission critical applications. Different technology components are needed to cope with: accurate channel state information, robust control channel design, reduced signaling overhead, efficient resource allocation for unpredictable traffic. The traditional dynamic resource scheduling procedure, as currently applied in LTE networks, is not suitable for machine type communication on mission critical applications. This is due to the stricter reliability requirements and overhead caused by potentially multiple messages exchanged between nodes to facilitate the data transfer (scheduling request followed by a resource grant, which in turn is followed by the actual data transmission). Grant-free schemes using semi-static configurations are an option to remove the signaling overhead caused by request/grant procedure, especially for predictable or periodic traffics. The configuration is done with radio resource control (RRC) signaling [3GPP-AH_NR2]. It includes, for instance, time and frequency resource allocation, modulation and coding scheme (MCS), power control settings and HARQ related parameters. In uplink, the configured devices should keep connected and synchronized, thus being ready for a URLLC transmission. In case of unpredictable traffic, configured resources can be shared by a number of users to reduce wastage. In this case, the transmissions are susceptible not only to inter-cell interference, but also to intra-cell interference.

Power control optimized settings 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, $1-10^{-5}$ transmission success probability within 1 ms. In this work the applicability of open loop power control (OLPC) for grant-free URLLC is investigated. The target is to optimize power control settings considering URLLC performance indicators. It is also investigated whether applying a power boosting for retransmissions can improve the success rate within the considered latency constraint. Detailed system-level simulations are used to assess the performance. The simulator includes modeling for many realistic effects, for instance, inter- and intra-cell interference, queuing, power saturation, time-frequency variant channel, HARQ combining, etc. Figure 2-5 illustrates the considered grant-free transmission scheme including HARQ and possibly power boosting. More details are available in [AJB+18].

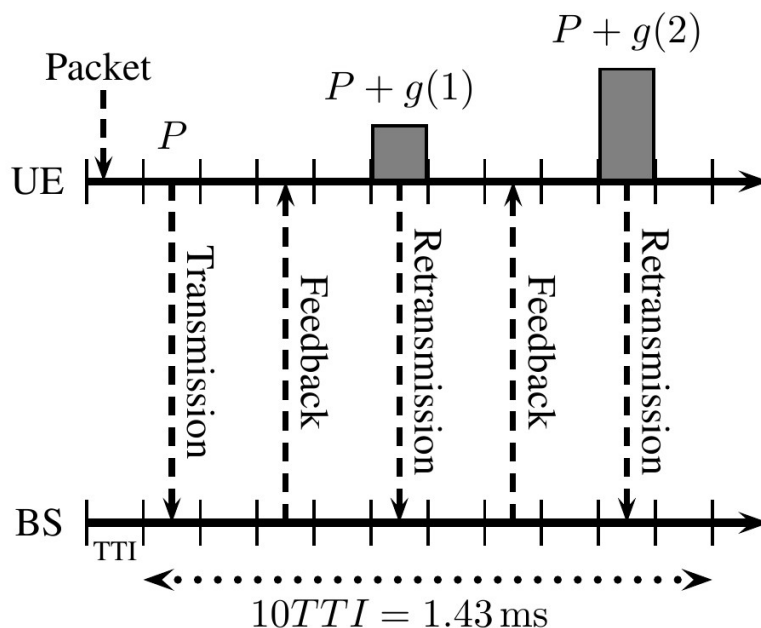


Figure 2-5 UMLLC Uplink Grant-Free Transmission with reactive HARQ and power boosting on the retransmissions. P is the transmit power without boost and $g()$ indicates the requested boost.

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

The simulation assumptions are based on the guidelines for system-level evaluation for UMLLC described in [3GPP-38802]. In summary, it is considered a multi-cell urban macro scenario (21 cells, 500 meters inter-site distance (ISD)). The receiver type is MMSE-IRC. The UE has a single antenna while the BS is equipped with two antennas. UEs are uniformly distributed outdoor and transmit small packets of 32 bytes generated according independent Poisson arrival processes. The transmission bandwidth is 10 MHz, 4GHz carrier frequency, 15 kHz of sub-carrier spacing and 2-symbol short-TTI. The MCS assumed is pre-configured to QPSK1/8. With that, it is assumed a maximum of 1 TTI for frame alignment, 1 TTI for processing in the base station or in the device side, 4 TTIs of round-trip-time.

Figure 2-6 shows the achieved outage probabilities at 1 ms as a function of the load for the different power control configuration. Firstly, it can be noticed that full path-loss compensation generally gives the best performance. With fractional path-loss compensation and without power boosting the achieved outage capacity was 800 packets per second (PPS)/cell. With full path-loss compensation, $P_0 = -104$ dBm and power boosting step of $PB_{step} = 10$ dB, a UMLLC load of 1200 PPS/cell can be achieved.

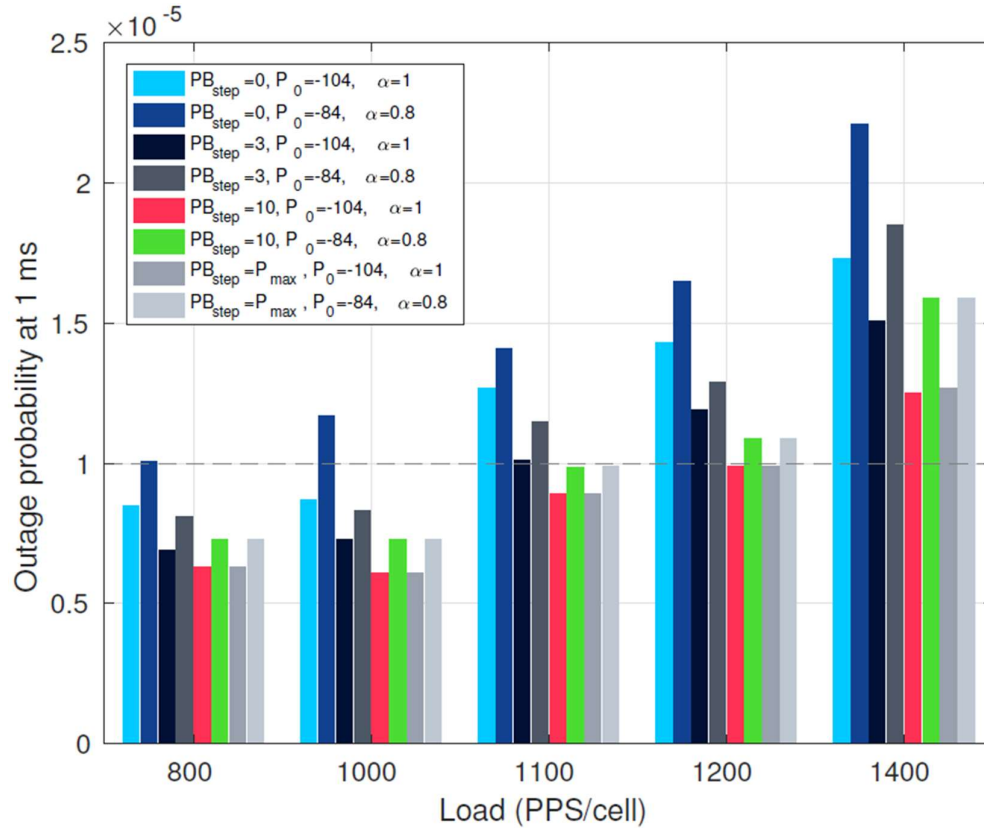


Figure 2-6 Outage at 1ms for different power control configurations and loads.

This study evaluates optimized power control settings to achieve better outage capacity while meeting the strict URLLC latency and reliability targets. Full path-loss compensation shows better performance and less sensitivity to P_0 setting than fractional path-loss compensation. This is due to the penalty applied to the cell edge UEs if $\alpha < 1$. The outage capacity gain when power boosting retransmissions are applied is approximately 20%. Higher gains could be observed in scenarios where the probability of having power limited UEs is lower. It is important to mention that the error rate of the first transmission should be low (such as, 10^{-3}), hence minimizing the excessive interference caused by retransmissions with boosting.

2.2.2 Short-Packet Transmission for URLLC Applications

Short Packet Structure for Ultra-Reliable Machine-type Communication

In classical information-theoretic analyses, the effect of the preamble and the metadata (e.g. control messages and synchronization signals) is usually not accounted for, since the resources they consume are negligible compared to the data. However, when the sizes of the preamble, the metadata, and the data are comparable (as it is for example in URLLC use cases), it is no longer obvious that the conventional frame or packet structure is a reasonable design. Machine-type communication requires rethinking of the structure of short packets due to the coding limitations and the significant role of the control information. In ultra-reliable low-latency communication (URLLC), it is crucial to optimally use the limited degrees of freedom (DoFs) to send data and control information. We consider the case of a point-to-point round-trip exchange consisting of the transmission of a short packet with acknowledgement (ACK). The model could be either UL or DL. We compare the detection/decoding performance of two short packet structures as shown in Figure 2-7(a): (1) time-multiplexed detection sequence and data (preamble structure); and (2) structure in which every packet carries both known pilot and unknown data together by split the transmit power (superimposed structure).

Figure 2-7(b) shows upper bounds/approximations of Packet Error Probability (PER) for the preamble case (solid lines) and superimposed case (dashed lines) for different SNR values. Black dots represent minima of PER values. We observe that the optimal overhead ratios, which is the length of preamble divided by the length of total superimposed signal, depend on the SNR and the target reliability, and that the superimposed structure achieves its minimum error probability at a higher overhead ratio. Furthermore, the PER achieved with superimposed detection sequences also experiences a degradation in terms of minimum PER because a larger fraction of resources is spent on detection overhead. The superimposed structure, however, offers enhanced adaptability as shown in Figure 2-7(c). Let us consider the case of a point-to-point round-trip exchange consisting of the transmission of a short packet of b bits and reception of a positive or negative ACK under stringent latency-reliability constraints. PER for three effective coding rates, R_{eff} , for optimal preamble size (dotted and dashed lines), optimal superimposed sequence (dashed lines) and pragmatic approach of adaptive coding rate for the preamble case (solid line). The coding rate for the data, defined as the length of short packet bits to the number of channel uses ratio, is changed every 1dB interval centered in the optimal computed rate for each integer SNR point, and labeled by the numbers next to it. For the system where the receiver and transmitter are aware of the SNR, the preamble and superimposed structures follow the optimal regimes. For the preamble case, this requires that the codewords are encoded with different rates. The superimposed structure is advantageous due to its flexibility to achieve optimal operation without the need to use multiple codebooks.

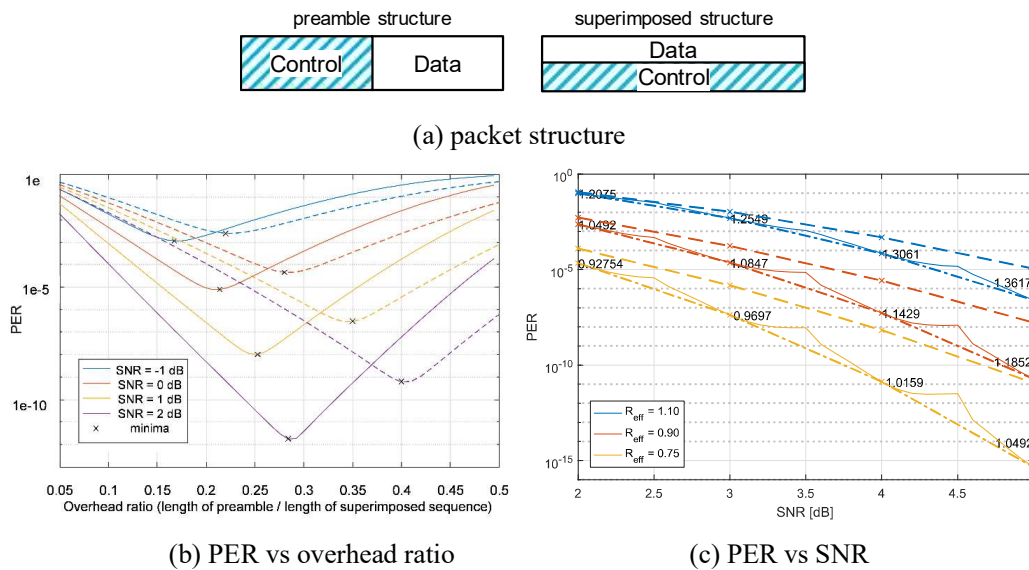


Figure 2-7 Packet structure and Packet error probability (PER)

Reliable schemes for short-packet transmission in massive MTC

While URLLC and mMTC use cases face different performance requirements in general, some aspects need to be addressed jointly. For example, in mMTC, where the main objective is to accommodate a large number of users sporadically transmitting short messages, the number of users that can be accommodated in the system is closely related to the latency incurred by the random-access procedure. In turn, in URLLC applications, the number of devices served by the system critically affects the targeted reliability of the system in the case where, due to latency or spectral constraints, the devices share the resources in a non-orthogonal fashion.

Against this background, we address a solution for reliable mMTC communication in the uplink, starting from the following assumptions: 1) 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; 2) we address the block fading scenario, where without a priori CSI

at neither the transmitter nor at the receiver, which is of particular relevance under the assumption of sporadic user transmissions; 3) the users (devices) are simple and can neither use precoding nor invest resources to enable channel estimation at the receiver, such that the communication is genuinely non-coherent. In addition, to account for latency-critical applications in the mMTC scenario and to decrease the control overhead, the users simultaneously perform initial access and communicate information to a joint receiver in a non-orthogonal fashion.

The transmission scheme is based on structured superposition coding and is motivated by the Sparse Regression Codes (SPARCs) recently introduced by Barron and Joseph [BJ12] for the additive (single user) Gaussian channel. In the multiuser scenario with L system users (UEs) in total, the n -th UE (when active) transmits information by linearly combining sequences from a predefined set $\mathcal{A}_n = \{\mathbf{a}_1, \dots, \mathbf{a}_{1M_n}\}$ of size B_n . The resulting transmit vector can be written in the form $\mathbf{x}^{(n)} = \mathbf{A}_n \mathbf{c}^{(n)}$, where \mathbf{A}_n is a $M \times B_n$ matrix whose columns contain the sequences of the set \mathcal{A}_n , and $\mathbf{c}^{(n)}$ is a binary vector. The information from UE n conveyed by $\mathbf{x}^{(n)}$ to the common receiver is contained in the support of $\mathbf{c}^{(n)}$ rather than the complex symbols in $\mathbf{s}^{(n)}$ itself. The inactive users are modelled as transmitting an all-zero sequence $\mathbf{x}^{(n)} = \mathbf{0}$, which is considered by the receiver as a separate codeword which appears with probability determined by the UEs activation probability. A special case of the proposed transmission scheme is when $\mathbf{c}^{(n)}$ is structured such that it selects k columns from \mathbf{A}_n (where $0 < k \leq B_n$), i.e. it linearly combines k sequences from the set \mathcal{A}_n . The effective transmission rate of UE n is then determined by the positions and the number of non-zero coefficients in $\mathbf{c}^{(n)}$. In a more general approach, $\mathbf{c}^{(n)}$ is selected from a code. When the applied code is linear, the transmitted codewords are obtained as $\mathbf{c}^{(n)} = \mathbf{G}_n \mathbf{b}^{(n)}$, where \mathbf{G}_n is a generator matrix, and $\mathbf{b}^{(n)}$ is the vector of information bits.

The joint active user and data detection problem is formulated from a Bayesian perspective. The enhanced receiver architecture leverages: 1) the design of the sets \mathcal{A}_n , $n = 1, \dots, L$ associated with the individual users; 2) the structure imposed on the transmit vectors $\mathbf{s}^{(n)}$ through the structure of the binary codewords $\mathbf{c}^{(n)}$; 3) signal sparsity induced by sporadic UE activity. For the high-dimensional inference problems with a potentially large number of system users, we rely on approximate inference (a form of hybrid generalized approximate message passing), which incorporates the structured prior information in a systematic way. The receiver assumes no a priori channel state information (CSI), i.e. it is agnostic to the instantaneous channel state realizations, which is of practical relevance in the scenario of massive MTC with sporadic transmission of short messages. An illustration for the graphical model representation of the Bayesian receiver architecture is provided in Figure 2-8.

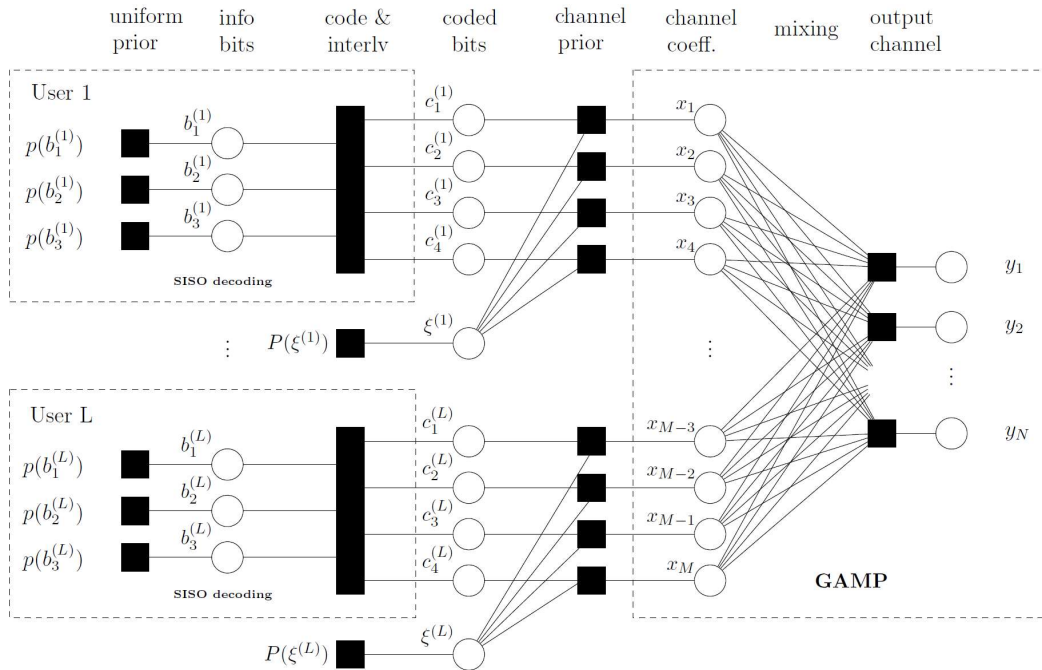


Figure 2-8 Graphical model representation for the Bayesian inference-based decoder.

Based on the observation vector \mathbf{y} , the decoder performs approximate Bayesian inference to decode the transmitted bit sequence \mathbf{b} . In each iteration, the inference is performed in two steps. In the first step (first half-iteration) the receiver performs Generalized Approximate Message Passing (GAMP) by treating the entries in the input vector \mathbf{x} as independent. In the second half-iteration an update is performed which accounts for the prior information for \mathbf{x} (probability of user activation, code structure, and channel statistics).

Selected preliminary results, which illustrate the potential of the scheme in an mMTC scenario are presented in Figure 2-9.

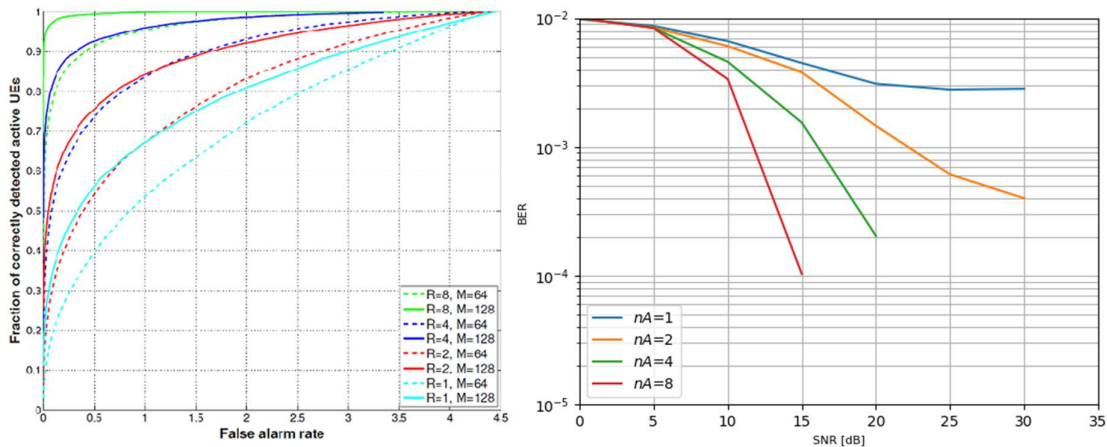


Figure 2-9. Left: Probability of correct user activity detection (for all active users) versus false alarm rate; Right: Single user performance: block error probability (BER) vs. total received SNR, for joint user activity detection and data decoding.

The scenario on the left in Figure 2-9 addresses $L = 256$ system users with activation probability $p = 0.1875$, jointly sharing $M = 64$ (respectively $M = 128$) resource elements for the random-access procedure (one step, grant free transmission without retransmissions or other mechanisms implemented at the higher layers). The system accommodates $R = 1 - 8$ RRHs in a C-RAN

scenario with fronthaul compression. The effective total received SNR (from all active users) is $SNR = 6 \text{ dB}$. The results are presented in terms of the probability of correct user activity detection (in total for all users) versus false alarm rate, where the false alarm rate is defined as the number of false alarms (inactive users detected as active) versus the number of active users. The results illustrate the effect of the number of RRHs on the reliability of user activity detection in a C-RAN scenario. The choice of the optional operational point (acceptable false alarm rate in the system and probability of correct detection) are not further discussed, as they depend on higher-layer considerations that are not taken into account here.

The scenario on the right in Figure 2-9 addresses a massive connectivity case with $L=1000$ system users and very short messages (each user has a codebook of size $B_n = B = 8$). The activation probability is $p = 0.1$ and the users jointly share $M = 250$ resource elements for the grant free transmission. The receiver accommodates $nA = 1$ – antennas, which, from modelling perspective, is the same as a C-RAN with $R = 1 - 8$ RRHs, but without fronthaul compression (i.e. quantization with infinite precision). The results are presented in terms of the block error probability BER (for joint user activity detection and data decoding) vs. the effective total received SNR (from all active users), for a single user (in a system with $L=1000$ users with probability of activation $p = 0.1$).

2.2.3 Interference mitigation for bi-directional URLLC

One of the key enablers of achieving ultra-reliable low latency communications (URLLC) is to use flexible numerology, which can be tuned to maximize the reliability of the wireless link [IX16]. Different numerologies may offer better resilience to challenging channel conditions. For example, large subcarrier spacing offer a better robustness to high Doppler shifts. In order to further optimize the system efficiency, different numerologies of different services may be packed adjacent to each other in the frequency domain as shown in Figure 2-10.

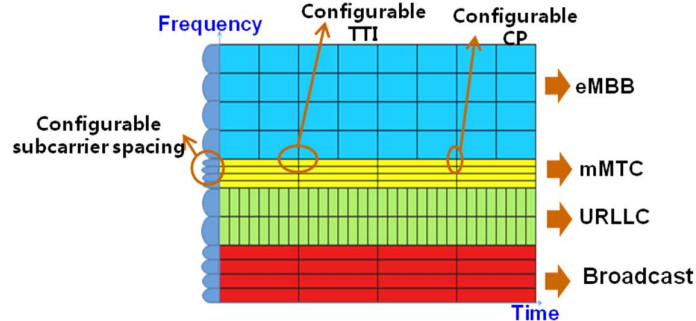


Figure 2-10 Adjacent services utilizing different numerologies

Traditionally, neighboring subbands are bordered by a guard band to minimize Out-Of-Band (OOB) noise, i.e., spectral leakage. In our work [IX17], we present a flexible radio frame structure where, under low latency constraints, large variations in the *bi-directional* data rates can be easily supported using time and frequency duplexing. As such, a fair subdivision of the radio resources which fit into the exact requested data rates becomes possible. To maintain high spectral efficiency, the bi-directional traffics are packed close to each other in frequency domain, which leads to considerable amount of crosstalk. Instead of using guard bands between neighboring subbands corresponding to different numerologies, we introduce a novel precoder to mitigate the crosstalk interference in baseband. A theoretical framework has been developed to derive the interference matrix which maps the transmit symbols from one numerology to the neighboring numerology as shown in Figure 2-11.

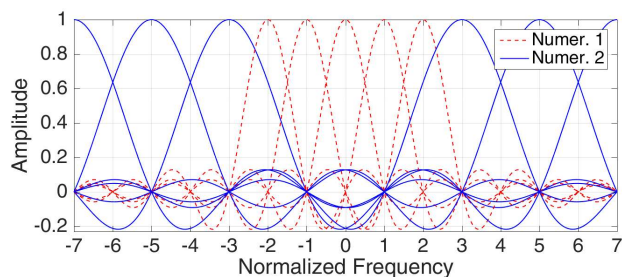


Figure 2-11 Spectra related to different subcarrier spacing of two numerologies (Numer.1 and Numer.2)

Knowing the interference matrix, the transmitter of the interfering signal is designed to precode the transmit signal so that it lies within the orthogonal space of the interference matrix. The orthogonal space can be derived from the Singular Value Decomposition (SVD) of the interference matrix. Naturally, this comes at the expense of a reduced data rate, since some subcarriers will be reserved for ensuring that the transmit signal is projected on the orthogonal space. There exist similar methods, e.g. by employing some cancellation carriers to reduce OOB. In [BCS06] the cancellation carriers are determined such that a transmitter occupying a band should have the lowest possible OOB emissions. Here, we seek for an interference mitigation solution tailored to the specific *non-orthogonality* of the subcarriers of DL and UL traffic. The subcarriers within the symbol containing *both DL and UL traffic* are precoded such that the Out Of Band (OOB) emissions are suppressed. It is further proposed to control the trade-off of data rate and OOB emission, by selecting precoding vectors which do not lie entirely in the orthogonal space, but contribute to a low power component in the OOB spectrum. Those vectors can be selected from the SVD decomposition by choosing pre/decoding vectors which correspond to low singular values.

Figure 2-12 shows the OOB emissions of a precoded transmit signal with different overheads [1.56%-2.73%]. ‘GB’ stands for guard band width. ‘PC’ stands for the proposed precoding scheme with different percentages of data reduction. $N=256$ is the FFT size. For comparison, we also plot the OOB emissions when a guard band is inserted which has a similar overhead as the precoded case for fair comparison. As shown, the smaller the overhead, the larger the OOB emissions due to the partially non-orthogonal precoding space. This trade-off can be tuned for different channel conditions to achieve a certain target SINR.

Depending on performance requirements, more or less subcarriers can be needed for the OOB emission suppression (instead of data transmission). For precoder determination, efficient matrix manipulation is usually required, meaning additional computational complexity. For details of this work, see [IX17]. Also, an IPR has been filed within ONE5G.

This work item is closed and the final results are published in [IX17]. Also, an IPR has been filed within ONE5G.

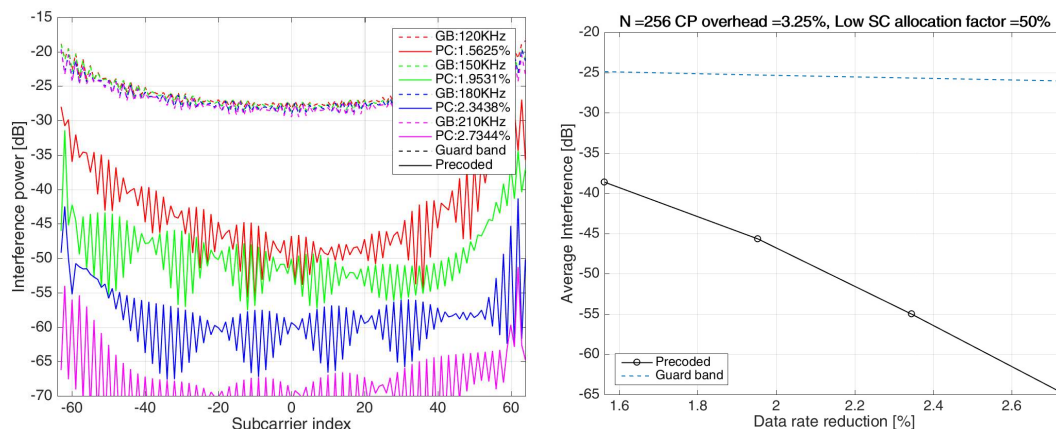


Figure 2-12 Left: Out-Of-Band emissions for precoded transmit signal and guard band inserted with different overhead. Right: average Out-Of-Band interference vs data rate reduction for precoded and guard band inserted.

3 Massive MIMO enablers towards practical implementation

Massive MIMO is a well-studied and investigated subject in the literature. Previous work includes references [LLS+14, MAM16-D54, FRZ+17, PDF17, BHS17], and notably the FP7 project MAMMOET (Massive MIMO for Efficient Transmission) [MAM16-D54], which has investigated fundamental limits and trade-offs. However, there are still many open challenges, in particular with respect to hardware complexity and CSI acquisition. ONE5G contributes contribution to the current state of knowledge by addressing implementation aspects of massive MIMO.

Flexible and scalable hardware architectures are needed to maximally utilize the potential gains of massive MIMO. This includes the array design (format and hybrid architectures), flexible and scalable hardware, i.e., reconfigurable baseband processing as enabler for multi-service (slicing). These, issues are addressed in Section 3.1.

CSI acquisition is still a crucial aspect of massive MIMO. In TDD mode, the challenges include pilot coordination and hardware imperfections. For FDD systems, we have to cope with constrained pilot signalling and CSI feedback [BLM16]. Especially the efficient estimation and compression of CSI at the user for downlink is an open challenge and addressed by most of the subsections in Section 3.2.

The most promising solution to reduce the hardware complexity and to overcome FDD limitations is hybrid precoding. The efficient analogue and digital precoder design for hybrid beamforming is addressed in Section 3.3. This also includes further applications of massive MIMO, like beamforming for backhaul in underserved areas and multi-cast transmission.

3.1 Flexible hardware architecture and multi-service support

The increase in hardware costs is a major concern for the deployment of massive MIMO. Therefore, solutions to reconfigure and tune hardware are a main enabler for massive MIMO, which is addressed in this subsection. Parts of the technology components will be implemented as proof-of-concept demos in WP5.

The impact of the array shape on system performance, cylindrical array structures are studied in Sections 3.1.1 and 3.1.2. The deployment of antennas for massive MIMO has a huge impact on the performance, however it is often neglected in most of the applications. Two technologies provide performance evaluations on this topic. First, depending on the user distribution, the deployment of antenna arrays, e.g. as $[8 \times 4]$ or $[1 \times 32]$ array can result to in sum-spectral efficiency gains of more than 100% compared to sub-optimal deployment. Second, using a cylindrical antenna deployment provides coverage gains for 50% of the users compared to sectorized planar array as shown in [KMT+18]. For usage in cellular deployments, ONE5G recommends to use joint precoding over all antennas instead of sectorizing the cylindrical array.

A flexible and fast configurable hardware architectures that considers a wide range of standards is proposed in Section 3.1.3. The reconfigurable hardware architecture proposed by ONE5G combines FEC features from 4G LTE, WiFi and 5G New Radio with respect to CRC, Coding, HARQ, Segmentation, and modulation without performance loss in the respective technology.

Finally, an analytic MIMO performance prediction that takes into account the channel hardening is provided in Section 3.1.4.

3.1.1 Impact of array shape in different deployments

In massive MIMO systems, for a given number of BS antennas, the performance strongly depends on the deployment scenario and on the UE distribution. In this study, we plan to understand how the antennas of a massive MIMO BS should be organized in order to maximize system performance. In these preliminary results, we consider an urban macro (UMa) hexagonal deployment with 7 sites, 3 macro BS per site, an inter-site distance (ISD) of 500 m and wrap-around. BSs are equipped with 64 antennas and transmit with a maximum power of 46 dBm at carrier frequency of 2 GHz on a system bandwidth of 10 MHz. We assume 10 single antenna UEs uniformly distributed per macro sector and consider the three-dimensional spatial channel model proposed by 3GPP [3GPP-36873]. We further assume the full buffer traffic model and perfect CSI at the BSs, each scheduling all its anchored UEs on the whole band, with equal power allocation and by using maximum ratio transmission (MRT) beamforming. We consider uniform planar arrays (UPA) with an antenna element spacing of 0.5λ , where λ is the wavelength associated to the carrier frequency. Moreover, we denote $R \times C \times P$ the structure of this array, where R is the number of rows, C the number of columns and P depends on the polarization, with $P=1$ referring to co-polarized antennas and $P=2$ to cross-polarized ones. In the results presented here, we assume cross-polarized antennas. Figure 3-1, shows the cumulative distribution function (CDF) of the UE signal-to-interference-plus-noise ratio (SINR) achieved by four different UPA shapes, all of them having 64 antennas. The reported SINR is computed for each UE as an average among all the resource blocks. Numerical results show that the best SINR is achieved by considering an array with just one row, i.e., $R=1$ and $C=32$, which is a very wide array, whereas the worst SINR is achieved by an array with eight rows, i.e., $R=8$ and $C=4$, which is a tall array. That happens because the UEs are uniformly distributed in the azimuth domain, with the BS-UE line-of-sight (LOS) link distributed over an azimuth range of 120° , whereas in the elevation domain UEs are distributed either outdoor or indoor in buildings with limited height, and the distribution of the BS-UE LOS link spans an elevation range of just 25° [3GPP-36873, Figure 8.2-6]. Therefore, in the UMa scenario, wide arrays must be used as they strongly outperform tall arrays. This is confirmed as well in Figure 3-2 where the cell spectral efficiency and cell border throughput achieved by these four UPA shapes are shown.

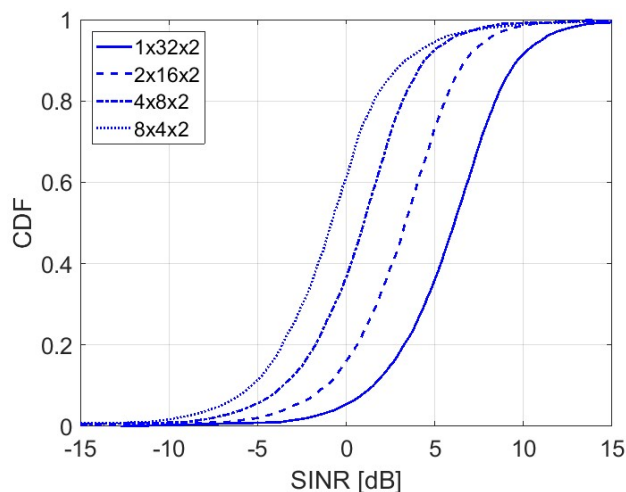


Figure 3-1 CDF of the UE SINR for different UPA shapes

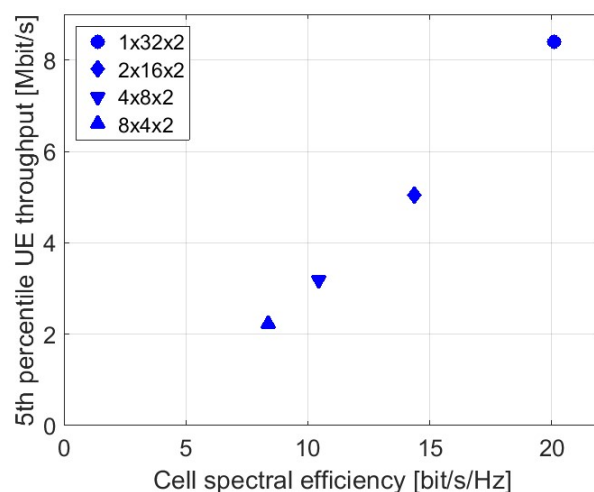


Figure 3-2 Cell border throughput versus cell spectral efficiency for different UPA shapes

In the future, we plan to assess different array shapes also in other scenarios, e.g., urban micro (UMi), evaluate the impact of imperfect CSI at the BS and compare different beamforming criteria.

3.1.2 Sector and Beam Management with Cylindrical Arrays

Most of massive MIMO research literature assumes uniform linear or planar array and standardization is focusing only on Uniform Planar Arrays (UPAs) [JKL+16], [3GPP36897]. One characteristic of UPAs is a direction dependent beamforming gain meaning that in some directions the beamforming gain is higher than in others, e.g. the array will point to areas of high traffic demands. However, these deployments are usually static resulting to the typical sectorization in the network. For example, in 3rd Generation Partnership Project (3GPP) system level assumptions [3GPP-36873] three UPAs are considered per Base Station (BS) location resulting to geometries similar as in Figure 3-3.

One disadvantage of these static deployments is that the array geometry cannot be dynamically adapted to changes in the environment. One example is illustrated in Figure 3-5 and Figure 3-6. During day time, the traffic demands are generated in the high beamforming gain direction of the sector, e.g. where cafes and restaurants are located. Figure 3-6 shows an evening time scenario, where the traffic is shifted to other directions, e.g. where a bar or late-night club is located. The

vision is that with Uniform Circular Arrays (UCAs) directions of sectors can be software -adapted with respect to changes in the user distributions [KMT+18].

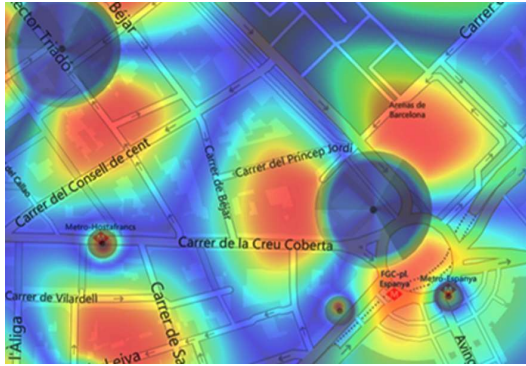


Figure 3-3: Wideband SINR due to sectorization with UPAs

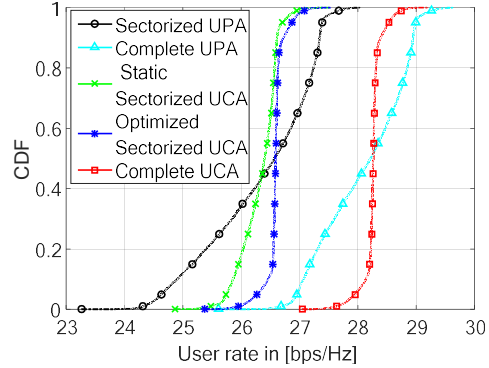


Figure 3-4: Single user spectral efficiency with MRT precoding without interference.

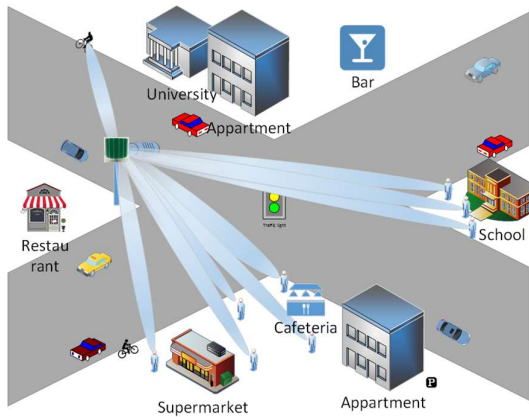


Figure 3-5: Daytime scenario.

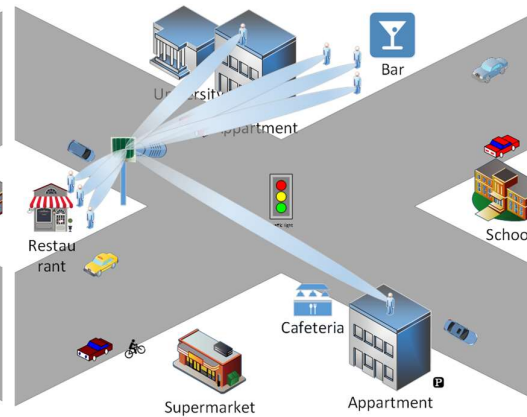


Figure 3-6: Evening time scenario.

Due to the novelty of this problem, we consider in a first step perfect Channel State Information (CSI) from all users to all UCA antennas to understand the performance tradeoffs and to evaluate potential gains as a motivation for future work.

For performance evaluation, a uniform $[8 \times 8]$ UPA is used, resulting into $3 \cdot 64$ total number of antennas for triple sectorization and compared with a $[8 \times 24]$ UCA, MRT/MMSE precoder with perfect CSI in single/multi user scenarios is assumed, respectively. Further details of the multiple-user multiple-sector OFDM system model are omitted here for the sake of brevity and can be found in [KMT+18]. The antenna geometries used for the numerical investigation are given in Figure 8-2 and Figure 8-3 and in appendix 8.1. The remaining simulations parameters are provided in Table 8-1 in Appendix 8.1.

In Figure 3-4, the basic trade-offs between the UPA and UCA are shown for the noise limited single user, single sector scenario with optimal Maximum Ratio Transmission (MRT) precoder. The spectral efficiency for the LoS scenario is given and it can be observed that the variance of the UCA is smaller than the UPA for both the complete and sectorized arrays. The crossing points at approximately 50% mean that in case of UPA higher peak rates for half the users are achieved by loss for the other half, compared to the UCA.

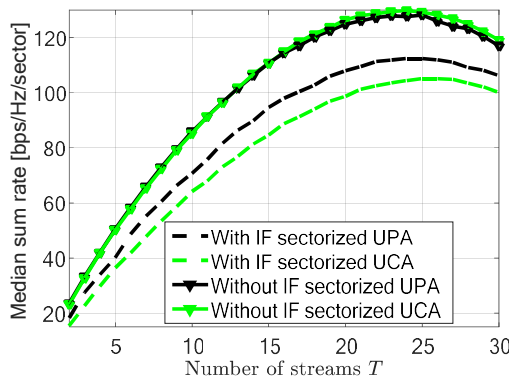


Figure 3-7: Impact from Inter-Sector Interference. IF in the legend stands for “Interference”

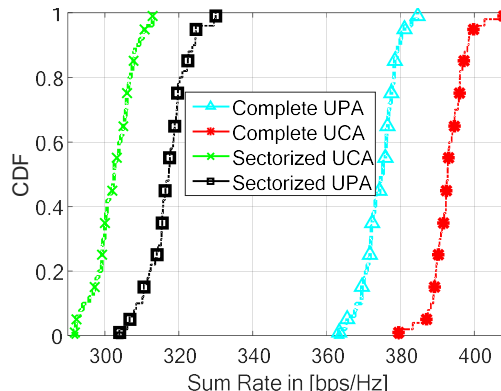


Figure 3-8: Sum rate over 3 sectors each serving 30 users in comparison with joint precoding over all antennas serving 90 users.

From the results in Figure 3-7 and Figure 3-8 on multiple-user multiple-sector it can be concluded that the complete UCA performs better than the complete UPA, however by dividing the UCA in three sectors, the sectorized UPA is better. In case of “complete” coherent transmission from all antennas is considered. This is due to the inter-sector interference and inherent coordination in the sectorized UPA. It is important to note that in the current investigation a uniform user distribution was assumed. More details can be found in [KMT+18]. As a next step, heterogeneous user distributions and trade-offs with different antenna element types will be investigated.

3.1.3 Flexible and fast reconfigurable HW architecture for multi-service transmission

The general context of flexible and fast reconfigurable hardware architecture is to evaluate the capability of communication systems to adapt to a wide range of services. This requirement amounts to make such an adaptation feasible for each of the modem’s digital processing module: e.g. filtering, synchronization, channel estimation, de-modulation, de-coding, etc. The following problems are to be solved.

- (a) Provided the analog/digital bandwidth compatibility, T , the adaptation has to be ensured for all services;
- (b) Given any mode, the complexity and/or consumption of the module has in the order of a dedicated circuit.

The second constraint above aims at satisfying technological feasibility by avoiding extensive use of processing unit based solutions.

As an illustration of these objectives, a derivation for Digital Front-End (DFE) blocks and coding/decoding schemes is carried out, and these principles are now illustrated on two module examples: Digital Front-End (DFE) and Forward Error Correction (FEC).

Digital Front-end

As a general overview, the DFE aims at compensating for Analogue Front-End limitations. It encompasses at least the following processes at both Tx and Rx sides:

1. Radio frequency (RF) impairment estimation and correction: e.g. linear effects as direct current (DC)-Offset, in phase/quadrature phase (IQ)-Imbalance;
2. Automatic Gain Control;
3. Filtering stages: blocker and adjacent channel rejection, in-band equalization;
4. Rate adaptation: adapt sampling rates from low flexible Analogue-to-Digital (Rx) / Digital-to-Analogue (Tx) Converters to arbitrary modulation rates.

- 5. Frequency de-multiplexing: convert frequency-multiplexed channels into time-multiplexed base-band channels and reciprocally.

In addition, the preamble detection followed by a frequency-time coarse synchronization step may be considered in the design. This is summarized in the case of a receiver in Figure 3-9:

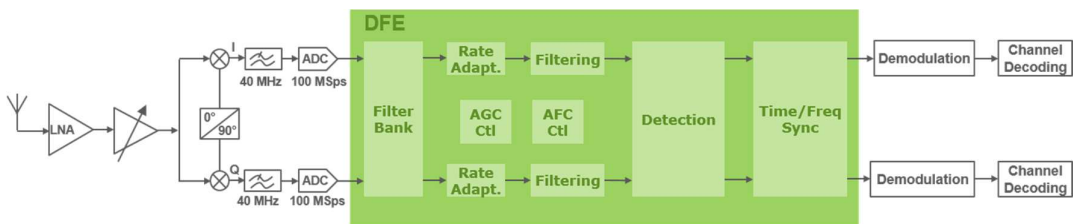


Figure 3-9 Rx Digital Front-End

Forward Error Correction

The Forward Error Correction (FEC) sublayer includes channel coding, with enhanced coding scheme like Low Density Parity Check Code (LDPC) or polar code, channel adaptation, as the rate and the mapping of the data stream, and the repetition mechanism (HARQ). The channel quality wraps the FEC configuration, so that the data rate and the reception performance maintain the best compromise.

New Radio [3GPP-38211] standard shows different constellation mappings (BPSK, QPSK, 16QAM, 64QAM, and 256QAM), different coding rates, and two main coding algorithms (LDPC/PUSCH-PDSCH, Polar code/PBCH - Control channels), with the diverse parametrization of Modulation and Coding Scheme gathered in [3GPP-38214]. The hardware architecture aggregates these functionalities, with low-latency reconfiguration and maintains the KPI performances. In addition, the system still keeps backward compatibility with already deployed networks such as LTE [3GPP-36212], aggregating Turbo and Convolutional Coding. This is summarized in Table 3-1.

Table 3-1. FEC parameters in LTE and NR systems.

Channel	NR PUSCH	LTE PUSCH	NR PUCCH	LTE PUCCH	NR PDSCH	LTE PDSCH	NR PDCCH	LTE PDCCH	NR PBCH	LTE PBCH
CRC	CRC16 /24	CRC24	CRC6 /11	CRC16	CRC16 /24	CRC24A	CRC24C	CRC16	CRC24C	CRC16
Coding	LDPC	Turbo	Polar	Conv.	LDPC	Turbo	Polar	Conv.	Polar	Conv.
HARQ	Yes	yes	No	No	Yes	Yes	No	No	No	No
Segmentation	Yes	Yes	No	No	Yes	Yes	No	No	No	No
Modulation	$\frac{\pi}{2}$ BPSK QPSK 16/64/ 256 QAM	QPSK 16/64 QAM	QPSK	QPSK	QPSK 16/64/ 256 QAM	QPSK 16/64 QAM	QPSK	QPSK	QPSK	QPSK

Figure 3-10 shows a merged FEC hardware architecture with several standardized physical layer designs (LTE, WiFi, NR).

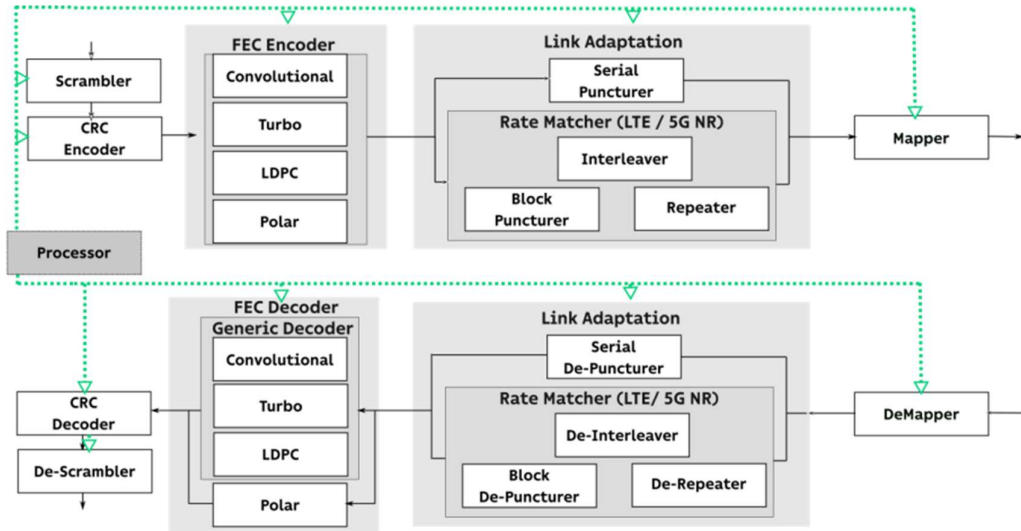


Figure 3-10 Forward Error Correction enabler

The FEC and DFE enablers are developed to address serially multiplexed services or several slices with hundreds of nanoseconds latency reconfiguration, through a small set of parameters following the data flow. It targets both Intel (former Altera) and Xilinx FPGA families, and will be integrated in the demonstration platforms proposed in One5G WP5. The DFE enabler is designed to address the “underserved area” scenario of PoC#4. The current version of the FEC enabler targets NR PDSCH and NR PUSCH configurations, and will be proposed in “Megacity scenario” by implementing it in WP5 PoC#2. The flexible hardware architecture supports the main constraints, in terms of data throughput, to be plugged on a larger Massive-MIMO demonstrator and background compatible with 4G LTE, not yet defined. Further details on the perspectives of the two TeC, their proposed demonstrations and some key performance results are provided in [ONE17-I51].

3.1.4 MIMO performance prediction

At link level, the main measure of performances is the capacity of the channel. However, most of the previous theoretical analyses of the massive MIMO channel were conducted using simple stochastic models. The goal of this contribution is to provide an efficient capacity prediction based on a limited number of parameters using a realistic ray-based channel model. Analysis of the stability of capacity is essential to practical design of MIMO systems, in particular considering scheduling, rate feedback, channel coding or modulation dimensioning. Massive MIMO offers two main gains: a more reliable signal (channel hardening) and an improved spectral efficiency. The reliability improvement of MIMO, called channel hardening is caused by two effects. The first one is an averaging phenomenon of the fading over the multiple links. The second one is based on the array gains (achievable at the emitter only in case of CSI at the Transmitter). The spatial multiplexing gain is obtained by using multiple data streams and is limited by the rank of the channel. We managed to highlight the two previously given gains of Massive MIMO in the problem of capacity analysis. This method reduces the complexity of the problem and gives a more explicit interpretation of the results.

The capacity of a MIMO point-to-point system is given by:

$$C = \log_2(\det(\mathbf{I}_{N_t} + \rho \bar{\mathbf{Q}} \mathbf{H}^H \mathbf{H})) \text{ bit/s/Hz}$$

Where $\rho = \frac{P_t}{N_0} \|\mathbf{H}\|^2$ is the received SNR assuming optimal precoding (whose stability is studied below), $\bar{\mathbf{Q}}$ is the normalized precoding matrix and \mathbf{H} is the normalized channel matrix. The channel influence separated into a spatial diversity term ρ and a spatial multiplexing term $\mathbf{H}^H \mathbf{H}$.

Channel Hardening.

Moving from single antenna systems to MIMO, the reliability of communication systems improves tremendously. On the one hand in single antenna systems, the signal is emitted from one single antenna and captured at the receive antenna as a sum of constructive or destructive echoes. This results in fading effects leading to a potentially very unstable SNR ρ (see above) depending on the richness of the scattering environment. On the other hand, in a MIMO system, with appropriate precoding, small-scale multipath fading is averaged over the multiple transmit and receive antennas. This yields a strong reduction of the received power fluctuations, hence the channel gain becomes locally deterministic essentially driven by its large-scale properties. This first main feature of MIMO systems is *channel hardening* [BHS17, NL17].

In the literature, the channel hardening effect is characterized by the coefficient of variation of the channel gain. It measures the variations of the diversity term ρ .

$$CV^2 = \frac{\text{Var}(\|\mathbf{H}\|^2)}{E\{\|\mathbf{H}\|^2\}^2}$$

It is observed in the literature that $CV^2 \xrightarrow[N_t, N_r \rightarrow +\infty]{} 0$, meaning that ρ behave deterministically for asymptotically wide antenna arrays. This measure has been previously investigated using simple Rayleigh fading models [BHS17]. We extended those studies using ray-based channel models and finite-size antenna arrays [RPM+18]. This measure of the stability of the SNR is essential to the practical design of MIMO systems, in particular on scheduling, rate feedback, channel coding and modulation dimensioning. Using this model, we managed to separate the influence of propagation conditions and antenna array topology. Moreover, the small-scale and large-scale contributions to channel variations are evidenced. Microscopic movements of the receiver will reduce large-scale fluctuation while small-scale variations will be bounded by $CV^2 < \frac{1}{N_r N_t}$.

The results of the contribution are given in the paper [RPM+18]. It provides a comprehensive study of channel hardening, taking into account antenna array topology and propagation conditions. It provides more precise insights than previous channel hardening studies using Gaussian based models. Figure 3-11 shows the influence of the number of antennas on channel hardening using a basic ray-based channel model (six rays with Gaussian distributed gains, direction of departure/arrival uniformly distributed). This model remains generic and is analytically tractable but the previously described results are applicable to a wider range of situations. Small-scale and large-scale variations are denoted by different colours. Small-scale fading is reduced by channel hardening whereas path loss variations and shadowing still remains.

Current studies focus on a scalable analysis of the stability of multiplexing. Further studies will include the influence of hybrid precoding and multiuser MIMO.

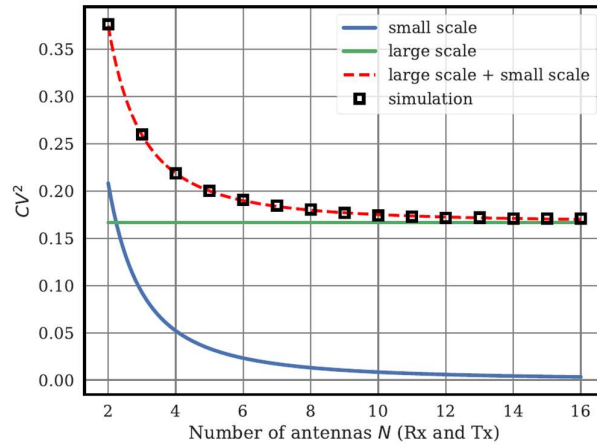


Figure 3-11 : Illustration of the channel hardening measure on a simple ray-based channel model. Small-scale and large-scale contributions are highlighted.

3.2 Efficient CSI acquisition in TDD/FDD and feedback compression in FDD

The gain of open loop or codebook based massive MIMO systems is limited by the codebook size or number of beam-directions, which in turn is limited by signalling/feedback overhead constraints. CSI at the transmitter and receiver of a MIMO link is of critical importance in order to achieve the diversity and capacity benefits offered by the use of multiple antenna elements. Acquiring CSI is typically performed at the initial transmission stage by use of training (pilot) signals. Even though channel estimation is a well investigated topic in the literature, the trend towards massive number of antenna elements introduces new challenges, namely, estimation of a very large number of variables, which, in turn, requires a proportional increase of the resources dedicated to training. The latter may be unacceptable in low coherence scenarios, e.g., due to mobility.

Since in a TDD system, the same frequency is used for transmission and reception, reciprocity between the UL and DL channel can be used. Calibration of the RF-frontend and a sufficiently long coherence interval are the prerequisites for channel reciprocity. This means in a TDD massive MIMO system, CSI is acquired by channel estimation based on UL pilots. For a FDD system, the situation is radically different. Here DL pilots are used to estimate the CSI at the user device. Afterwards, this information needs to be fed back to the access point. In this section we present efficient solutions for CSI estimation, pilot design and the feedback.

In contrast to LTE in 3GPP NR, the CSI acquisition for the purpose of feedback is exclusively based on dedicated reference signals in the DL and UL. In addition, there is also a procedure to align the beams for analogue/hybrid beamforming available. In addition to the periodic and aperiodic CSI reporting in LTE, NR also introduces a third scheme called semi-persistent CSI reporting. Currently, the CSI reporting is completely code-book based and can be flexibly restricted by the network (e.g. bandwidth, frequency granularity, etc.). In NR, the reference signals used for the purpose of demodulation are in contrast to LTE always precoded in the same way as the data transmission symbols. As in many parts of the NR standards, plentiful options to configure them are available (e.g. OFDM symbol distance, number of MIMO layers, etc.). All details regarding CSI feedback and demodulation can be found in [3GPP-38214].

As the channel estimation for a large number of antennas is complex the work in this section has shown that it is possible to achieve a sufficient channel estimation quality for a wide range of practical scenarios, while at the same time drastically reducing the computational complexity. We

also showed that the chosen schemes are in addition robust to additional non-linear impairments often encountered in consumer grade wireless systems.

For the CSI feedback in a TDD massive MIMO system we have shown how the systems performance can be improved by efficiently grouping the different users in a MU-MIMO setup. For the FDD CSI feedback we show how the codebook based CSI feedback present in 3GPP NR Rel. 15 could be extended to support a more explicit feedback of the channel and the interference at a moderate overhead. This does enable future enhancements of the system regarding multi point transmission and reception.

3.2.1 Parametric channel estimation for massive MIMO

Choosing an appropriate channel model is crucial for channel estimation. In particular, the objective is to use the model which allows to attain the highest data rate. In other words, we would like to find the model whose estimation causes the smallest capacity loss. To do so, we assume a physical model and allow the estimation model to deviate from it in order to optimize the estimation task in terms of capacity loss. The method presented here would be the same for any physical model, but we instantiate it for a static narrowband MISO/SIMO channel, because it leads to simpler mathematical expressions. The main idea is to use for estimation a model of the same form as the considered physical model, except for the number of paths. The estimation model is indeed made of p virtual paths, leading to estimates of the form

$$\hat{\mathbf{h}} = \sqrt{N} \sum_{k=1}^p d_k \mathbf{e}(\vec{v}_k),$$

Where N is the number of antennas, d_k is the complex gain of the k th path, \vec{v}_k its direction and $\mathbf{e}(\vec{v}_k)$ the corresponding steering vector. Such estimates belong to the set

$$\mathcal{M}_p \triangleq \left\{ \mathbf{x}, \mathbf{x} = \sum_{k=1}^p d_k \mathbf{e}(\vec{v}_k) \right\},$$

called model set, parameterized by the number of virtual paths p . The objective is then to find the optimal number of virtual paths with respect to the capacity loss. To do so, it is possible to use the relative mean squared error (rMSE):

$$\text{rMSE}(\hat{\mathbf{h}}) \triangleq \mathbb{E} \left[\frac{\|\mathbf{h} - \hat{\mathbf{h}}\|_2^2}{\|\mathbf{h}\|_2^2} \right].$$

This error measure can be decomposed into the sum of two terms:

$$\text{rMSE}(\hat{\mathbf{h}}) = \frac{\|\mathbf{h} - \mathbb{E}[\hat{\mathbf{h}}]\|_2^2}{\|\mathbf{h}\|_2^2} + \frac{\text{Tr}[\text{cov}(\hat{\mathbf{h}})]}{\|\mathbf{h}\|_2^2},$$

The first one being called the *bias* and the second one the *variance*. Intuitively, the bias should be reduced when p increases. On the contrary, the variance should be increased when p increases. This is confirmed by our previous study of the variance term [LP17], showing that it is at least (for efficient estimators)

$$\frac{\text{Tr}[\text{cov}(\hat{\mathbf{h}})]}{\|\mathbf{h}\|_2^2} = \frac{3p}{\text{SNR}}.$$

Regarding the bias term, defining the appropriate oracle estimator leads to

$$\frac{\|\mathbf{h} - \mathbb{E}[\hat{\mathbf{h}}]\|_2^2}{\|\mathbf{h}\|_2^2} = \frac{\|\mathbf{h} - \text{proj}_{\mathcal{M}_p}(\mathbf{h})\|_2^2}{\|\mathbf{h}\|_2^2}.$$

Unfortunately, this quantity is difficult to compute exactly (it is NP-hard). We thus proposed to compute it approximately using a variant of the QR matrix factorization. Moreover, an analytical study is currently ongoing to bound this bias term.

Preliminary results are shown in Figure 3-12 for 16 and 256 antennas, respectively (certain curves for the LS estimator at low SNR do not appear because the relative error was greater than one). To obtain these results, channels have been generated according to the considered physical model using the NYUSIM [SMR17] channel simulator. For these channels, the average number of physical paths is fifty. Then the rMSE has been computed for various values of p and compared to the classical least-squares (LS) estimator, exhibiting a bias/variance tradeoff and an optimal number of virtual paths to consider. It is interesting to note that this optimal number is much smaller than the number of physical paths for any SNR (around a dozen against fifty). Moreover, although the LS estimator performs almost as well as the physics-based estimator for 16 antennas (at reasonably high SNR), the physics-based is much better with 256 antennas, confirming the benefit of using a precise parametric model for massive MIMO.

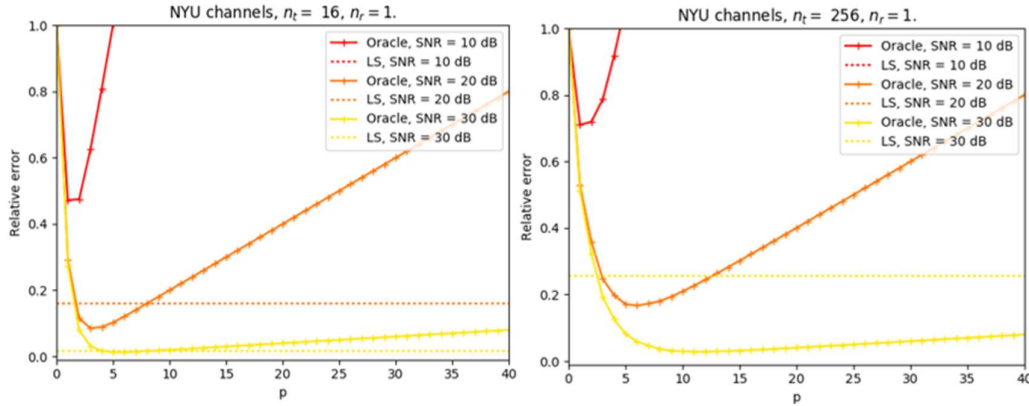


Figure 3-12: Performance of the physical channel estimator for 16 and 256 antennas with respect to the number of virtual paths.

In order to get access to the data rate loss caused by channel estimation, the rMSE should be linked to the capacity loss. We managed to derive a bound on this capacity loss with respect to the rMSE and the SNR (mathematical developments are not shown here for brevity reasons, they will be reported in an upcoming paper), which allows to quantify the worst-case data rate loss caused by the imperfect channel estimation procedure. The bound can be used to adjust the number of virtual paths to estimate, in order to respect a certain data rate constraint. For example, estimating only 5 virtual paths in a millimeter wave massive MIMO channel is sufficient to attain 90% of the optimal capacity (attained only if the channel is perfectly known). The bound is tight for a large rMSE (around 1), and it can be made smaller if we constrain the rMSE.

Future work should compare thoroughly the physical channel estimators to other estimators used in practice such as the LMMSE, as well as studying in-depth the assumptions on which the physical models rely on. Moreover, the study should be generalized to take into account frequency selectivity and user mobility.

3.2.2 Hierarchical sparse channel estimation for multiuser massive MIMO with reduced training overhead

Application of the compressive sensing (CS) framework [FR13] to the problem of mMIMO CSI acquisition is one of the most promising approaches towards highly efficient, accurate, and low-overhead channel estimation procedures [HG17a]. The motivation for exploiting CS techniques comes from the fact that the wireless channel is inherently sparse, due to the finite and typically small number of (resolvable) propagation paths [S02], [BHS+10]. Therefore, identification of the paths' properties (namely, delay, angle of arrival, and complex gain) is sufficient to obtain CSI. Most of the previously proposed CS-inspired CSI algorithms for the massive MIMO (mMIMO) channel (a) consider the narrowband signalling scenario and (b) require multiple observations in time in order to first obtain an accurate estimate of a certain covariance matrix, which is later to be processed for extracting CSI information (see [HG17a] for an overview). However, this

approach is non-applicable in the case of wideband signalling as typically used in the eMBB use cases 5 and 6 in Table 5-2, and it may also require an excessive number of pilot overhead which is not acceptable for mMTC applications such as the use case 3 in Table 5-2.

CS-inspired channel estimation schemes were recently proposed also for the wideband and low-latency scenarios (see, e.g., [CY16], [YGS+16], [HG17b]). These works clearly demonstrate the performance potential of CS estimation algorithms via means of numerical simulations. However, they fail to provide rigorous analytical results in terms of performance.

Towards understanding the performance and training overhead limits of CS channel estimation in wideband mMIMO, we propose a novel CS-inspired channel estimation algorithm for the uplink of a single cell serving multiple UEs. The UEs are assumed to have a single antenna, whereas the BS is equipped with a Uniform Linear Array (ULA) of $M \gg 1$ elements. The key idea is that the channel of a single UE can be represented as a sparse matrix in the, so called, delay/angular domain. However, this matrix is not simply sparse, but its sparsity pattern, although random, can be shown to possess a well-defined structure, referred to as *hierarchical sparsity*.

Based on this observation, we proposed a CS-inspired channel estimation algorithm that explicitly takes into account the hierarchical sparsity property. In addition, by introducing a rigorous mathematical framework regarding hierarchically sparse vectors, we provide scaling laws for the pilot overhead required that is sufficient to obtain reliable (in the sense of bounded error norm) channel estimates. Please refer to [WRF+18] for details. It is noted that analogous analytical results are not available in the literature since ignoring the hierarchical property of sparsity leads to an intractable analysis.

Figure 3-13a demonstrates the performance of the proposed algorithm (“HiIHT” label in the figure) for the single user channel estimation case. For this example, a ULA of $M = 256$ elements and OFDM transmissions of $N = 1024$ subcarriers were considered. The channel impulse response consists of L distinct paths, with a maximum delay spread equal to $\frac{1}{4}$ of the OFDM symbol period. The, so called, on grid case was considered (see [WRF+18] for details), where the delays and angle of arrivals for each path were randomly and uniformly distributed on a two-dimensional grid. The path gains were modelled as independent, complex Gaussian random variables. The system SNR was set to 10 dB. The figure shows the channel estimate mean squared error (MSE) versus the, so called, pilot overhead ν_τ , which equals the ratio of number of subcarriers used for channel estimation purposes to N . It can be seen that the MSE decreases with increasing pilot overhead, as expected. However, excellent performance can be achieved for very small overhead (in the order of 0.01), even for large values of L . The MSE performance of a conventional CS algorithm, which ignores the hierarchical sparsity property, is also shown (“IHT” label in the figure). It can be seen that the algorithm requires a significantly large minimum pilot overhead to achieve a reasonable performance.

Figure 3-13b demonstrates the performance of the proposed algorithm for the multiuser case, where all users send their pilots on the same subcarriers (see Appendix 8.2 for details). The system parameters are the same as above with the difference that there are now $U = 4$ UEs out of which only V are active, with $L = 3$ paths for each user channel. It can be seen that the algorithm performs very well in this setting as well, with a moderate cost in minimum required overhead compared to the single user case. The minimum overhead scales proportionally to the number of active users V as the algorithm efficiently exploits the sparsity in the user activity domain.

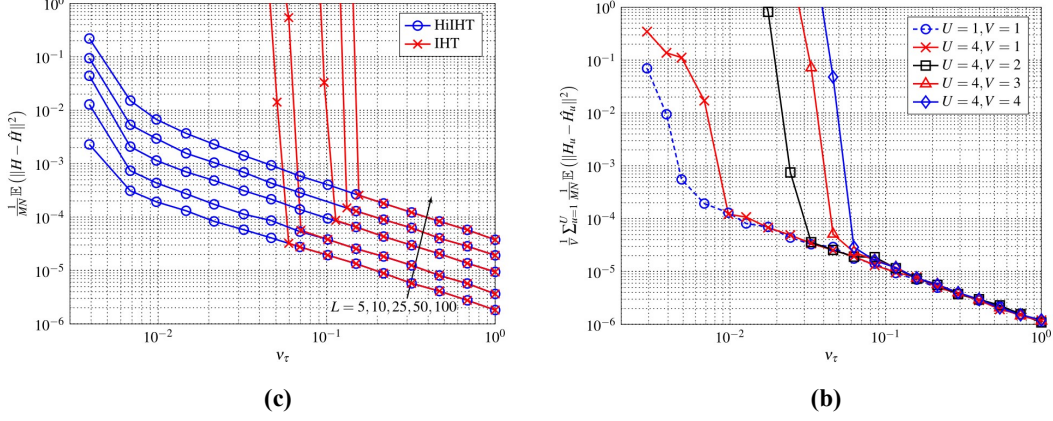


Figure 3-13 a) Mean squared error (MSE) performance of proposed algorithm (HiHTP) and conventional algorithm (IHT) for the single user case. (b) MSE performance of proposed algorithm for the multiuser case.

3.2.3 On the amount of downlink training for FDD correlated massive MIMO scenarios

When simple least squares (LS) estimators are used at the UEs, the DL training overheads are proportional to the number of BS antennas M . Another disadvantage of LS estimators is that they do not capture the channel spatial structure, as Minimum Mean Square Error (MMSE) estimators do. It was shown in previous works (see, e.g., [BX17], [JMC+15]) that optimized reduced rank training (i.e. training durations $T < M$ or $T \ll M$) combined with MMSE estimators can result in accurate channel estimates. Optimized training sequences are ones that are chosen according to the users' spatial covariance matrices.

In this contribution, we address the open problem of choosing a suitable training duration to ensure a bounded achievable rate loss to the one achieved with perfect CSI at the BS. To that end, we need to define some quantities. Namely, let $h_k \in \mathbb{C}^M$ and $C_k \in \mathbb{C}^{M \times M}$ be the channel vector and spatial covariance matrix of user k , $1 \leq k \leq K$. Further, let $S \in \mathbb{C}^{M \times T}$ be the training sequence matrix used by the BS with the T different sequences as its columns. Note that the training duration T is not fixed in advance; rather, it will be optimized according to the users' channel covariance matrix, as elaborated later. If user k performs MMSE estimation and obtains an estimate \hat{h}_k , then the channel estimation error covariance matrix as a function of S reads

$$\Phi_k(S) = E \left[\|h_k - \hat{h}_k\|_2^2 \right] = C_k - C_k S (S^H C_k S + \sigma_t^2 I_T)^{-1} S^H C_k \quad (3-1)$$

where σ_t^2 is the additive noise power in the training phase. For analytical tractability, we focus on zero-forcing precoders in the following. Denote $R_{per,k}$ to be the achievable rate of user k with perfect CSI at the BS, R_k to be the rate achieved with estimates $\hat{h}_1, \dots, \hat{h}_K$, and $\Delta_k = R_{per,k} - R_k$ to be the rate loss due to estimation errors. If P is the total transmit power at the BS, and σ_d^2 is the additive noise power in the data phase that is experienced by each user, the SNR in the data phase is defined as $SNR = P/\sigma_d^2$. Then, we have the following key result.

Theorem

Let $\lambda_1(\Phi_k(S))$ be the largest eigenvalue of $\Phi_k(S)$ and $c = O(1)$ be a constant. Then the condition $\lambda_1(\Phi_k(S)) \leq c SNR^{-1}$ is sufficient to keep a bounded ergodic rate loss, i.e., $E[\Delta_k] < \log_2(1 + c)$ as $SNR \rightarrow \infty$.

The proof is presented in [BX18]. Though the relation $E[\Delta_k] < \log_2(1 + c)$ can only be proven asymptotically, it holds for finite SNRs as observed by extensive numerical results. The quantity c is a design parameter that can be used to control that maximum allowed rate loss. This theorem is very important because it specifies sufficient conditions of the DL estimation errors that keep

the rate loss finite in the high SNR regime. The next task would be to explicitly find training sequences that satisfy the conditions $\lambda_1(\Phi_k(S)) \leq c \text{SNR}^{-1}$ for $k = 1, \dots, K$ as explained next.

Sequence construction

Simultaneously finding T sequences that satisfy the derived conditions for all users is intractable. Therefore, we proceed in a different manner, namely, by looking for a subset of sequences that satisfy the conditions separately for each user. For user k , this corresponds to a given number of eigenvectors corresponding to the largest T_k eigenvalues of C_k . Then, the range of the found subsets is combined to find the desired T sequences. Guidelines on how to find T_k are given in [BX18]. Note that the condition $\lambda_1(\Phi_k(S)) \leq c \text{SNR}^{-1}$ implies a higher channel estimation accuracy as SNR increases; thus, T_k and T increase with the SNR.

Numerical results

We investigate the optimized training duration when 1) the parameter c is varied, and 2) when the channel correlations are varied. We consider $K = 8$ users with azimuth mean angle of arrivals uniformly distributed between $\{-52.5^\circ, \dots, 52.5^\circ\}$. The BS is equipped with an 8×16 uniform rectangular array ($M = 128$) with half a wavelength antenna spacing in both horizontal and vertical directions. The channel correlations follow a Laplacian angular spectrum model with horizontal and vertical standard deviations β_H and β_V , respectively. Larger values of β_H and β_V imply that the channel entries are less correlated and that the channel power comes from more directions (i.e. larger effective rank of the spatial covariance matrix) than the case with small β_H and β_V .

For the case ($\beta_H = 10^\circ, \beta_V = 2^\circ$), Figure 3-14 (left) shows that increasing c reduces T . Increasing c reduces the channel estimation accuracy, and therefore the training duration is reduced. Figure 3-14 (right) shows that an increasing correlation (i.e. decreasing β_H and β_V) reduces T . This is due to the fact that decreased values β_H and β_V results in the channel power being concentrated in fewer directions, which decreases the needed T_k and therefore T .

Future work will focus on sequence codebook design, and implicit covariance feedback methods that facilitate the implementation of the proposed method.

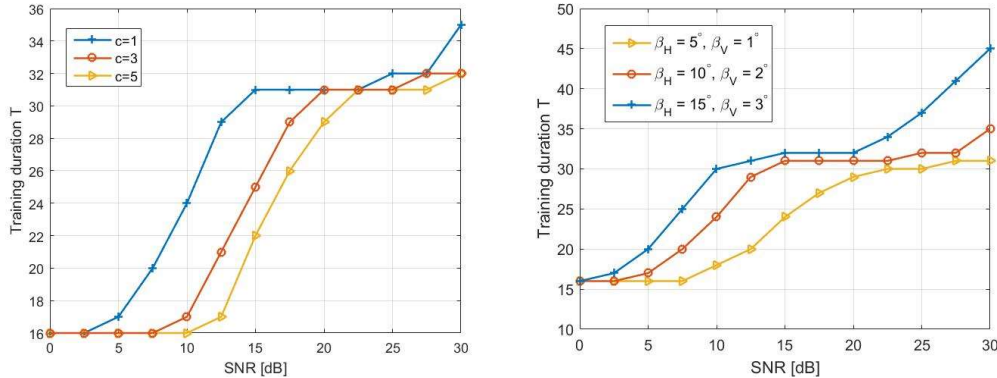


Figure 3-14 (left) Effect of c on the training duration (right) Effect of correlations on the training duration

3.2.4 Improving CSI acquisition through spatial multiplexing – TDD/FDD

In order to achieve the performance gain of Massive MIMO, accurate CSI estimates are required at the transmitter. Acquiring this information comes with an increase in signalling, feedback and processing overhead. As a matter of fact, CSI estimates can be acquired using uplink training in TDD mode. Consequently, the number of scheduled users for uplink training is restricted due to the limited coherence slot length and training overhead. In FDD systems, CSI feedback overhead

grows very rapidly with the number of system antennas, which also limits the number of scheduled users. In this section, we will address the CSI acquisition in both TDD and FDD and develop opportunistic solutions for pilot allocation in TDD systems and CSI feedback in FDD systems.

The scenario considered here is aligned with the description of scenario “Outdoor hotspot and Indoor offices” described in D2.1 [ONE17-D21].

CSI feedback for FDD massive MIMO systems

In FDD, massive MIMO channel estimation overhead scales linearly with the number of system antennas. This can be very limiting, since it results in limiting the number of scheduled users. Consequently, new feedback schemes should be developed in order to reduce the CSI feedback overhead. Recently, joint Spatial Division and Multiplexing (JSDM) was proposed to address this issue [NAA+14]. It works by partitioning users into groups, where users within each group have, ideally, the same channel covariance eigenspace. It proceeds by splitting the downlink beamforming into two stages: an outer precoder that depends on the channel statistics, and an inner precoder that depends on the instantaneous effective channel realizations. The role of the precoders is to suppress inter-group and intra-group interference respectively. The dimensions of the effective channel are significantly less than the number of antennas, thanks to the outer precoder projection. Even with this reduction in CSI at Transmitter (CSIT) feedback, the same sum capacity can be achieved for the corresponding MU-MIMO downlink channel if the eigenspaces of groups are mutually orthogonal.

In realistic scenarios, users might have similar but not necessarily identical second order downlink channel statistics. This dictates the incorporation of a clustering algorithm to partition users into groups with sufficiently similar covariance eigenspaces. On top of that, with a high number of users uniformly distributed across the cell, the eigenspaces of the groups are far from meeting the orthogonality condition and a reduction of the number of simultaneously served groups is required. Consequently, user grouping is of paramount importance. Previous works on this matter include the proposed K-means, hierarchical and K-medoids clustering algorithms in order to group users based on their channel second order statistics [NAA+14] [Sun17]. These approaches are limited by a major shortcoming as they require a prior estimation of the number of clusters, which is unknown a priori. Hierarchical clustering suffers from a large computational complexity, which can become very problematic in dense connectivity scenarios. This inspired us to propose a novel similarity measure along with a new clustering scheme where the number of clusters is not required to be known. We also developed, by using graph theory tools, a low complexity scheduling scheme that outperforms all currently proposed methods in both sum-rate and throughput fairness. In the sequel, we will provide the main lines of the solution and one can refer to [MHA+17] for more details. The first step consists in clustering the users without passing the target number of clusters as a parameter. To do so, we construct a complete graph $G_c = (V_c, E_c)$ where V_c is the set of vertices and E_c represents the edges. Each vertex represents an user and an edge would have a $\langle +1 \rangle$ label to signal that these two users are preferred to be in the same cluster while $\langle -1 \rangle$ label refers to the opposite case. The labels are determined using a similarity metric defined in [MHA+17]. Our goal now is therefore to produce a partition of the graph's vertices in a way that agrees as much as possible with the edge labels. To do so, we propose a cost function as the total disagreements of our resulting partitioned graph. The total disagreements cost is defined as the overall negative weights inside a cluster added to the positive weights between clusters. Consequently, the graph partitioning problem can be formulated as a combinatorial problem. We refer to [MHA+17] for more details on the mathematical formulation.

The obtained graph partitioning problem turns out to be NP-hard. Nevertheless, one can deal with it through efficient approximation algorithms (one can refer to [MHA+17]). Once the clustering is performed, the next step is naturally to schedule the clusters or the groups for CSI feedback. The previous studies in this area have not considered fairness between users. Consequently, we have developed an optimization problem that takes into account the fairness among the clusters

(and hence the users) and that schedules the clusters for CSI feedback. A description of the problem can be found in Appendix 8.3. For more details, one can refer to [MHA+17].

We have then analyzed the obtained scheduling problem and have proposed a low complexity solution. We proceeded by building a directed interference graph and applying appropriate transformations on the resulting graph in order to convert the problem to a vertex coloring problem. This transformation is one of the main contributions in this work and requires several steps. We then use a simple yet effective maximal independent set based vertex coloring algorithm that achieves a $O\left(\frac{n}{\log(n)}\right)$ -approximation of the optimal solution, where n is the number of vertices of the graph. Groups that are assigned the same color represent a subset of groups that are allowed to transmit simultaneously. In the end, at the start of each coherence time, the schedule (i.e. the color) that leads to the largest utility is selected. Preliminary results show promising gains in terms of both throughput and rate fairness as one can see in the next two figures where our scheme is compared to the JSDM and SLNR schemes developed respectively in [NAA+14] and [Sun17]. It is worth mentioning that the fairness is defined using the known Jain's index. Simulations also show a CSI feedback reduction of approximately 6 to 7 folds, compared with the full CSIT case in [NAA+14].

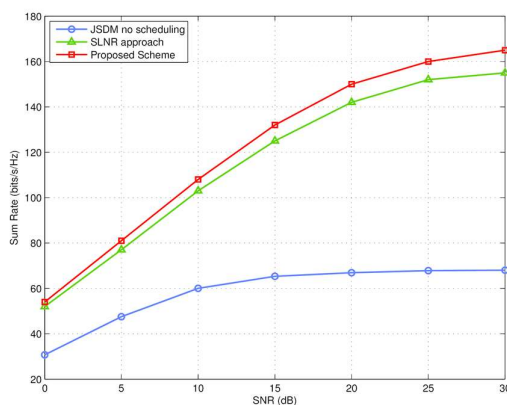


Figure 3-15 Comparison of sum spectral efficiency vs. SNR

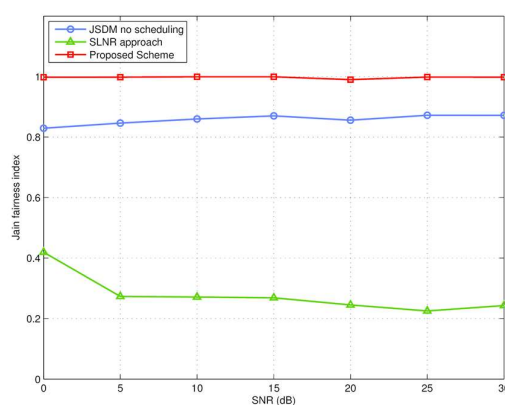


Figure 3-16 Comparison of Jain's fairness index vs. SNR

Uplink training in TDD massive MIMO systems

In TDD systems, spatial multiplexing can also be exploited in order to mitigate co-pilot interference. This allows achieving higher spectral efficiency per user or scheduling more users with the same training overhead. In both cases, the result will be an increase in the system performance. We aim at adapting second order statistics based user grouping to a TDD mMIMO system where CSI is obtained with uplink training and channel reciprocity. In fact, allocating pilot sequences as a function of the second order statistics of each UE channel can lead to an efficient mitigation of pilot contamination and, consequently, results in higher spectral efficiency. Here, covariance based grouping is also utilized. Practically, this leads to group the users having similar channel covariance matrices into the same cluster. The users in the same cluster are allocated different pilots while the same pilot sequences are reused in all clusters. This method suffers from the following main shortcoming. Users might have similar but not necessarily identical second order channel statistics, which leads to a substantial overlapping between clusters. We therefore propose an alternative approach for user clustering in order to address this issue. We consider a multicell TDD mMIMO system, in which, the diversity of the covariance eigenspaces is exploited in order to allow for a more aggressive pilot reuse, within each cell, while mitigating copilot interference. This allows for an increase in the number of scheduled users for the same training overhead while providing a gain in the achievable spectral efficiency. Since a multicell system is considered, both intra and inter-cell copilot contamination should be addressed.

We choose to decouple the two problems and address them successively. In order to deal with intra-cell copilot interference, we adopt a novel approach that aims at constructing copilot user groups. In each cell, any given copilot group is formed such that it contains UEs with minimum overlapping in their signal's eigenspaces. The idea is to associate each user with a set of beams that span its main covariance eigenspace. We consider uniform linear arrays (ULAs) for which, the eigenvectors of the channels covariance matrices can be approximated by the columns of a unitary discrete Fourier transform (DFT) matrix. After obtaining the UE's specific decoding matrices, the BSs derive copilot user groups. In opposition to prior covariance based clustering schemes, where each user signal is processed with a group specific matrix, the proposed approach enables to take into consideration the actual covariance eigenspace of each user providing a larger gain in spectral efficiency. Once copilot user groups are formed, we address the issue of inter-cell copilot interference through an efficient cross-cell training sequence allocation scheme. This allocation is based on the second order statistics of the interference channels. Using this information, the network is able to allocate specific uplink training sequences to copilot UE's groups in different cells, such that the resulting interference can be managed efficiently using the previously defined eigenspace based receivers. Preliminary results show promising gains in terms of spectral efficiency as one can see in the next figure where we compare our scheme with the classical scheme where the same pilots are used in all cells with no clustering.

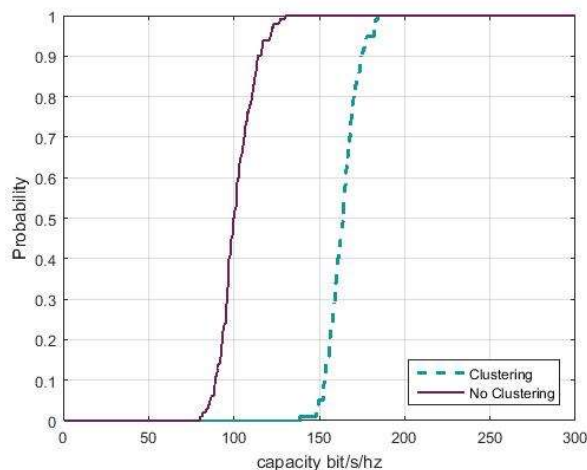


Figure 3-17 Spectral Efficiency of the proposed clustering scheme

3.2.5 Efficient feedback schemes for more accurate CSI and advanced precoding

In FDD, channel knowledge is conveyed to BS via uplink feedback from UE. So far in LTE, only **implicit** feedback is supported, i.e. UE feeds back the best Precoding Matrix Index (PMI) in a predetermined codebook known at BS and UE side [3GPP-38802]. In order to support advanced radio concepts such as multi-TRP coordination, envisioned in NR phase II, explicit CSI knowledge at the BS side is highly beneficial.

In this study, we consider the feedback system shown in Figure 3-18, where long term and wideband matrix \mathbf{W}_1 is used to build the DFT grid of beam (GoB) reducing the dimension of the channel from M antennas to $2L$ beams (L beams per polarization). At the UE side, the DL CSI H_{DL} is compressed (using the compression function \mathbf{C}) and fed back to the gNB, where it will be decompressed and used to build the short term and frequency selective precoding \mathbf{W}_f .

In order to exploit the underlying time domain sparsity of the channel, we consider explicit CSI feedback based on time domain channel impulse response a.k.a. explicit time domain feedback

scheme (ETF). We feedback the most significant taps of the sparse time domain channel impulse response (CIR) back to the gNB.

As shown in Figure 3-19, the limited channel frequency response (CFR) knowledge at the UE side, due to the presence of the guard bands, leads to a loss of sparsity in the CIR. As explained in [AVW18], this can be considered as a case of underdetermined system of equations. However, since the sought CIR vector is originally sparse, compressive sensing techniques can be used to recover the channel sparsity.

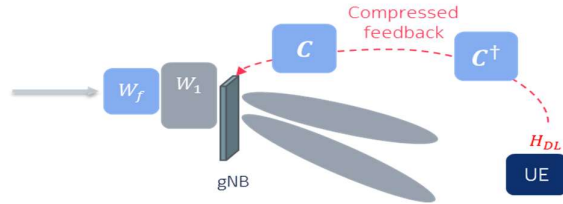


Figure 3-18 Explicit CSI Feedback System Model

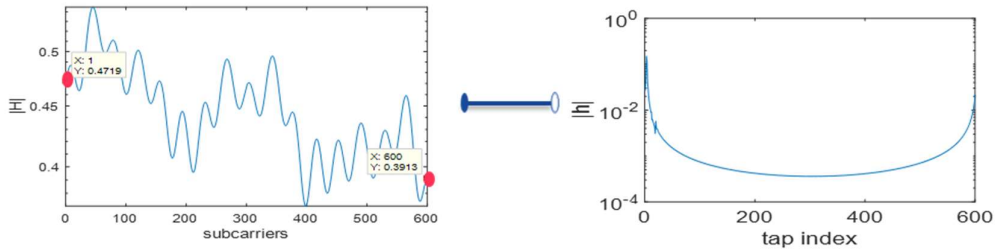


Figure 3-19 Limited CSI Knowledge Causes Loss of Sparsity

Preliminary results for explicit CSI feedback based ETF, show superior performance against the recently standardized NR Type II CSI feedback, in an UMi 64x2 MU-MIMO setting with rank 1 transmission. In each sector, 10 UEs were randomly dropped. UE speed of 3 km/h is used. We assumed a bandwidth of 10MHz with 50 physical resource blocks (PRBs), at a carrier frequency of 2GHz. The feedback overhead was fixed to 264 bits for both schemes (NR type II CSI and ETF).

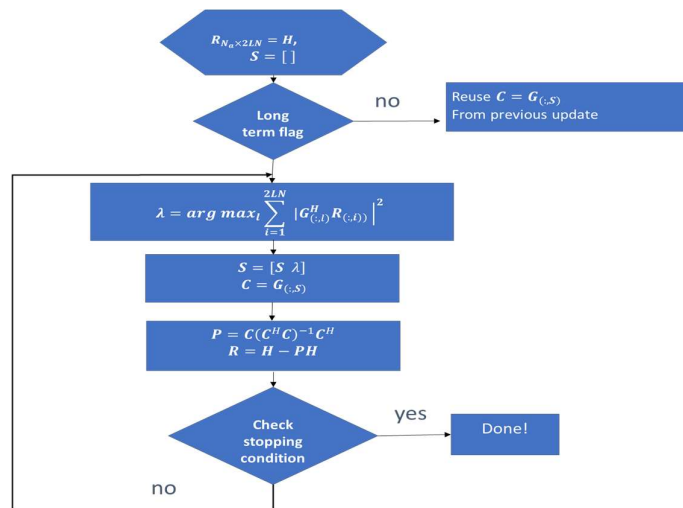


Figure 3-20 OMP Flowchart

A sparse recovery algorithm is applied at the UE side, namely orthogonal matched pursuit (OMP) [CW11], so as to compress the CSI feedback. Figure 3-20 depicts the flow chart of the OMP, which is a greedy heuristic iterative algorithm working on identifying the channel support, i.e., the positions of the most significant taps. More details can be found in [AVW18]. In every iteration step, the algorithm seeks to find the most significant tap, implicitly exploiting the common channel support assumption. After that the contribution of that significant tap is subtracted from the residual CRF. The algorithm can stop after a fixed number of steps (what we assume here in this work) or until the power of the residual matrix goes below a predetermined value. The final output of the OMP is the channel support vector S .

As shown Figure 3-21, with proper CSI compression, an overall gain of around 12.7% is achieved using the ETF combined with OMP compared to state-of-the-art implicit feedback scheme. In the future, we plan to investigate other feedback compression schemes with lower UE complexity.

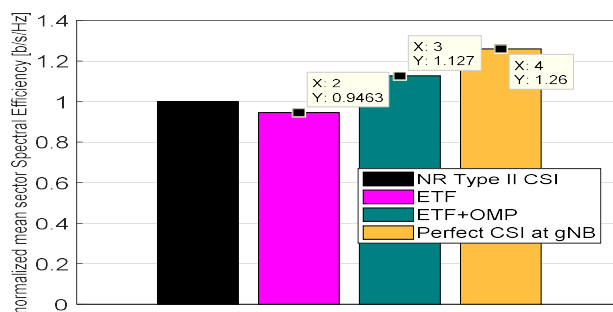


Figure 3-21 Normalized spectral efficiency for the different feedback schemes

3.2.6 Joint investigation of UL channel estimation and massive MIMO detection regarding robustness

The increased number of antennas for massive MIMO systems in combination with the larger bandwidth envisioned for future mobile broadband systems will lead to a considerable increase in computational complexity. Since massive MIMO systems have been proven to approximately achieve channel capacity with linear MU-MIMO detection methods [YH15], we concentrate on these methods. Traditionally, MIMO detection is investigated independently of other receiver effects of the system. In this work we investigate channel estimation and MU-MIMO detection in combination with other impairments of the system regarding performance and complexity. This is especially interesting for practical systems, where many idealistic assumptions are not satisfied. Therefore, we especially wanted to test if the developed methods are robust to modelling mismatch and channel estimation errors. The details of this work are given in [RN18].

The channel estimation is based on 2x1-D Wiener Filter based interpolation as described in [HKR97]. We use the 3GPP NR Type 1 OFDM pilot structure lately agreed in 3GPP [3GPP-38211]. This pilot structure provides orthogonal reference signals for up to 8 users. The interpolation filter across subcarriers is developed using a model for the Power Delay Profile (PDP) of the channel, and additional estimation of the modelling parameters.

The MU-MIMO detection algorithm we investigate is based on coordinated descent with binary step-size called Dichotomous Coordinate Descent (DCD). In general coordinate descent algorithms successively optimize an objective function one direction at a time. In [RN18] we show how this algorithm is derived from relaxing the optimization problem defined by ML MIMO detection. As the step-size is binary, all multiplications can be implemented by bit-shift operations and the complexity is reduced. Compared to the classical MMSE approach, there is also no assumption about the noise covariance matrix.

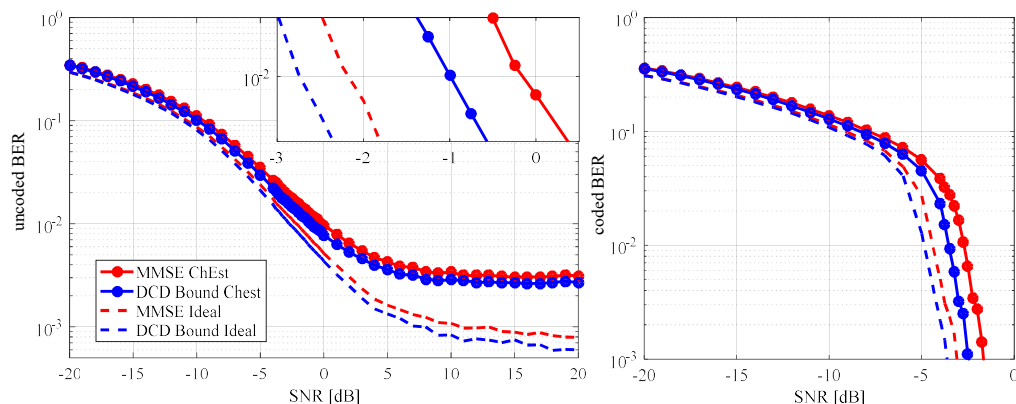


Figure 3-22 Simulations results for un-coded (left) and coded (right) BER for MMSE and Sequential DCD MU-MIMO detection with full channel knowledge (Ideal) and real channel estimation (ChEst).

The performance results are shown in Figure 3-22. The simulations parameters and the corresponding complexity results are also shown in [RN18]. In the comparison of the BER performances, we can observe that the performance of the DCD algorithm is always better than the one of MMSE. Since the DCD algorithm is more robust towards channel estimation errors, the performance gap increases if channel estimation is used instead of perfect channel knowledge. As we can see from the Figure 3-22, these results carry over to the case considering channel coding.

The results for the computational complexity show a 16% improvement of DCD over MMSE in the case that MMSE equalizer cannot be reused for multiple OFDM symbols. In the case that this matrix can be reused 14 times the complexity advantage is reduced to 10%.

3.3 Analog and/or Digital Beamforming/Precoding

To support high-throughput and reliable wireless links, beamforming techniques are essential. There are extensive digital precoders for multiuser scenarios reported in the literature such as zero forcing (ZF) and minimum mean squared error (MMSE) beamforming. In contrast to digital precoders, an analog precoder is comprised of a set of phase shifters, which are not able to adjust signal amplitudes. This constant modulus feature makes the design of analog precoder difficult as the optimization problem will become nonconvex. As a result, iterative procedure is usually used to compute analog precoders. Recently, a hybrid beamforming system, which consists of a digital precoder and an analog precoder in a transceiver, has widely attracted attentions from both academia and industry. This hybrid structure provides a balanced solution to a massive MIMO system with a limited number of radio frequency (RF) chains. Although there is not a specific antenna structure or implementation algorithm standardised in the 3GPP NR, a highly flexible channel state information (CSI) framework has been established to provide sufficient support for different settings. This includes both codebook-based and non-codebook-based precoding, CSI feedback for useful and interference channels. In this subsection, a number of beamforming techniques will be discussed to satisfy different ONE5G use cases.

Hardware complexity reduction is a main enabler for future massive MIMO systems. In this section, ONE5G achieves complexity reduction by using 1-bit phase shifter with a hybrid SLNR precoding scheme with less than 5% performance loss compared to full digital precoding. Complementing this, complexity reduction can also be achieved with full digital precoding using low resolution ADCs. It is shown in ONE5G that in some scenarios digital precoding with low resolution ADCs is more energy efficient and achieves a higher sum-rate than hybrid precoding systems [RPS+18].

Furthermore, functionality split is one of the enablers to deploy massive MIMO RRHs, where the sum-rate performance is constraint by the fronthaul capacity. For limited fronthaul links ONE5G

shows that trace-weighted analog beamformer always outperforms state of the art equal combining [PKC+17].

The flexibility of massive MIMO facilitates the use of in-band backhaul in underserved areas. In such a scenario, it is shown that the degrees of freedom can be used for efficient interference reduction to achieve reliable backhaul links up to 5 km. As a complementary technique, ONE5G proposes to use Probabilistically Shaped Coded Modulation (PSCM) [PX17] for additional SNR gain.

3.3.1 Genetic algorithm assisted hybrid beamforming for wireless fronthaul

Cloud radio access network (CRAN) forms a network by centralizing BBUs in a data center while distributing RRHs across the whole network layout. The common public radio interface (CPRI) has been used to connect the BBU and RRUs. However, the fibre implementation of CPRI introduces high cost in fronthaul and limits flexibility of the fronthaul network. Nowadays, the wide spectrum in millimeter wave (mmWave) frequency bands provides the potential of fronthaul links with quasi-fibre throughput and enhanced flexibilities. A working group has been established in IEEE to standardize the next generation fronthaul interface (NGFI) [NGFI] which includes the discussion of wireless fronthaul. The application of wireless fronthaul is able to provide sufficient flexibilities to support the ‘Live Event Experience’ use case in WP2 (ONE17-D21].

To further reduce the cost of deploying wireless fronthaul, a hybrid beamforming technique with a limited number of RF chains can be used. Details of the genetic algorithm assisted hybrid beamforming for wireless fronthaul can be found in Appendix 8.4.

Mean remote node sum rates of hybrid Signal-to-Leakage-plus-Noise Ratio (SLNR) beamforming, which maximizes the received signal power while minimizing power leakage to other cells, in an independently and identically distributed (i.i.d.) Rayleigh fading channel were depicted in Figure 3-23. Digital SLNR, zero forcing (ZF), and hybrid ZF were shown as reference. In general, SLNR slightly outperformed ZF in the low SNR regime. Performance of hybrid SLNR with 1 or 2 bits phase resolution analog precoder was approximately 15% to 7% worse than digital SLNR in the SNR range between -12 dB to 0 dB. This performance gap reduced to 5% in the moderate SNR regime. Meanwhile, the complexity of the hybrid SLNR beamforming will be N_{iter} times the complexity of the digital SLNR beamforming, where N_{iter} is the number of iterations. However, the overall complexity and cost will be offset by analog precoders with only 1 or 2-bit phase resolution.

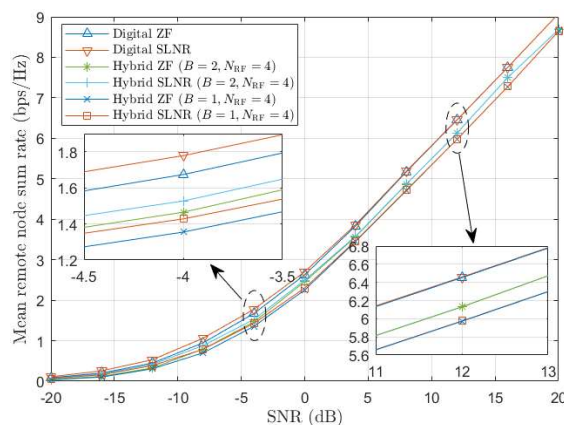


Figure 3-23 Mean remote node sum rate comparison of digital/hybrid SLNR/ZF beamforming schemes ($N_T = 8$, $K = 3$, $N_{R,l} = 1 \forall l$)

Beam patterns of digital and hybrid SLNR beamforming are illustrated in Figure 3-24. It was assumed in this figure that remote nodes were in line-of-sight positions. It can be seen in Figure

3-24 (left) that with digital SLNR beamforming, the shaped beams were sharp and with clear spatial boundaries. The hybrid SLNR beamforming with 1-bit phase shifters in Figure 3-24 (right) had reasonably well shaped beams. However, remote nodes received weaker energy due to the limited capability of the analog precoder. More importantly, a strong beam was observed in hybrid SLNR, which can cause inter-cell interference in a multicell network. This implied that current focus in the literature of hybrid beamforming design approximating the performance of fully digital beamforming is not sufficient. Impact of hybrid beamforming on multicell networks is essential as well.

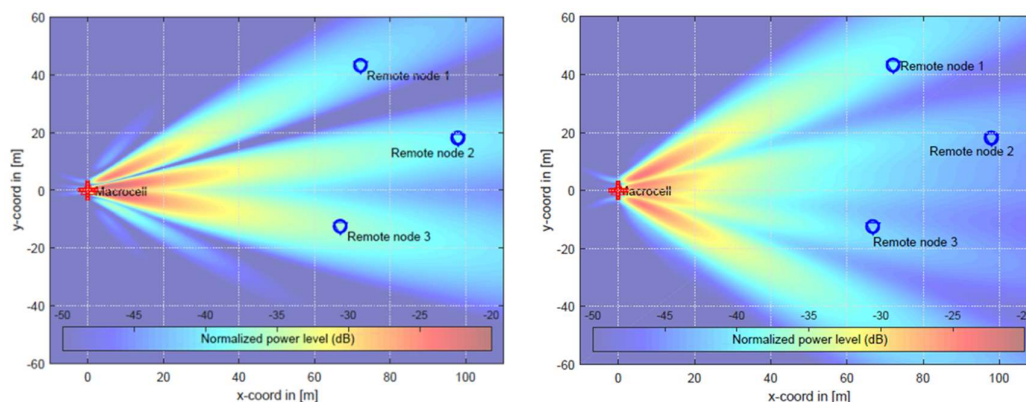


Figure 3-24 Beam patterns of digital SLNR (left) and hybrid SLNR beamforming (right).

Considering that the implementation of hybrid SLNR was significantly simpler than digital SLNR, the hybrid SLNR structure was able to provide better tradeoff between complexity and performance in implementing wireless fronthaul. For future work, joint considerations in multicell hybrid SLNR structure will be considered.

3.3.2 Hybrid array architectures covering different deployment scenarios

Massive MIMO using large antenna arrays with a high number of antenna elements has the capability to establish orthogonal spatial channels free of intra cell interference, allowing spatial multiplexing to a high number of users distributed within the coverage area. However, this high spectral efficiency can be fully exploited only if a high number of spatially separated UEs are served simultaneously. In practical deployments the number of UEs or simultaneous MIMO streams is varying and can be low during a substantial amount of time. The present work therefore aims at evaluating the impact of array size and architecture on the spectral efficiency and user throughput under different deployment conditions, and identifying most efficient array architectures.

Hybrid arrays have the potential for efficient hardware implementation, because they map the high number of antenna elements to a lower number of antenna ports [GRM+16]. The number of antenna ports provides the freedom to assign the MIMO layers to the UEs, and therefore needs to be dimensioned according to the number of simultaneously served UEs. The obvious advantage is that the hardware effort and power consumption for Digital-to-Analog Conversion (DAC) and the RF chains scale with the number of UEs and not with the number of antenna elements. For verification of the benefits the array architecture according to Figure 3-25 has been used and different array sizes and port assignments have been compared by system simulations.

In Figure 3-25 a number of simultaneous MIMO streams K is precoded to feed P antenna ports, with each port mapped to a subarray with M antenna elements to form an array with $N = P \cdot M$ elements.

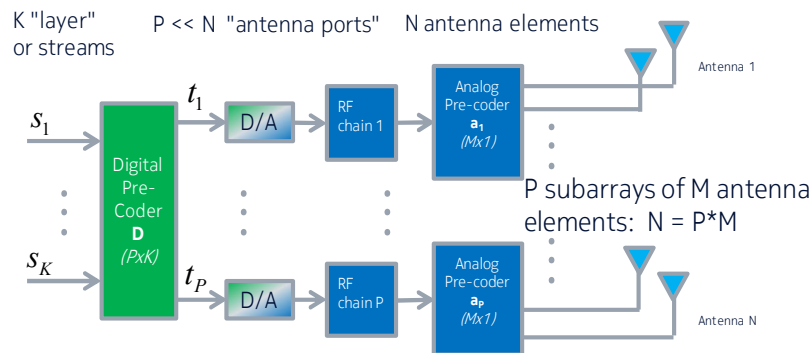


Figure 3-25: Hybrid array architecture

System simulations with different array architectures and sizes have been conducted. Spectral efficiency for UEs with a cross-polarized antenna have been evaluated for the 3GPP urban macro (UMa 3D) channel [3GPP-36873] and with zero-forcing (ZF) and eigen-beamforming (EBF) precoder. For details see Appendix 8.5.

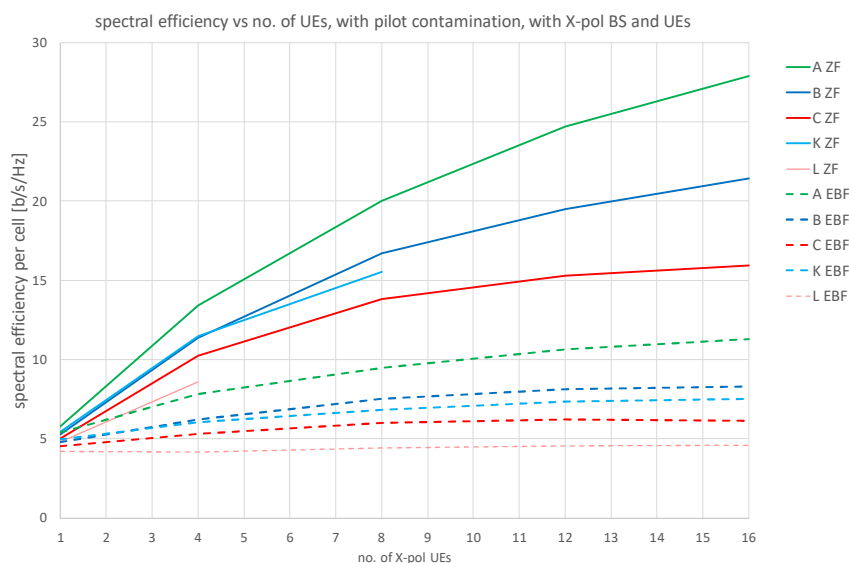


Figure 3-26: Spectral efficiency versus the number of simultaneously served UEs, each served with 2 x-polarized MIMO layers, with different array sizes

The simulation results in Figure 3-26 reveal that for zero forcing (ZF) precoding the spectral efficiency shows high variation versus the number of simultaneously served UEs. For low numbers of simultaneously served UEs the spectral efficiency is almost the same for all array types. Especially array type B and L, both with same aperture size, but with 64 and 32 antenna ports, show almost identical performance. Only for large number of UEs the increased array size (size A (256 elements) > size B (128 elements) > size C (64 elements)) has benefits in terms of cell spectral efficiency, whereas the array types K (128 elements) and L (64 elements), which have only 32 and 16 ports, respectively, are not able to serve more than $P/4$ UEs with ZF algorithm due to the lack of degrees of freedom. The eigen-beamformer (EBF) shows much less increase of spectral efficiency versus number of simultaneously served UEs, but has also similar performance for low number of UEs. So, for a low number of simultaneous UEs the spectral efficiency is not differing much for all array types, whereas the hardware effort for the lower number of antenna ports and antenna size of the smaller arrays is significantly reduced. Only for a large number of UEs the increased array sizes in combination with a high number of antenna ports have benefits in terms of spectral efficiency.

The next steps of the work will focus on a more detailed analysis on optimum shape of the subarrays and design criteria for different deployments, to find the optimum array architectures for specific deployments.

3.3.3 Multicast Beamforming

The need for MIMO multicast (aka groupcast) was identified in the 3GPP Rel-14 study item [3GPP-38913], where it was 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 evolved multimedia broadcast multicast services (eMBMS) and single cell point to multipoint (SC-PTM) do not exploit multi-antenna transmission such as transmit diversity and spatial multiplexing. However, it is expected that future systems will make extensive use of massive MIMO antennas. This is motivated by new automotive and industrial use cases, including safety applications, traffic flow control, and MTC services [ACSM17]. While today’s eMBMS/SC-PTM standards are mainly targeting streaming and file delivery, 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-38913].

The availability of CSI at the base station enables the application of beamforming techniques, which is especially important for higher frequencies. However, multicast beams must be jointly adapted to different UEs channels, which poses additional challenges compared to conventional beamforming. This is illustrated in Figure 3-27, which shows the array power gain vs. angles of departure. Multicast (or groupcast) means that the same message is to be transmitted to a group of UEs. This can be done either jointly, by using a single beam, which is formed as a function of all users’ CSI, based on the max-min fair strategy introduced in [SDL06]. We assume partial CSI obtained by beamformed reference signals on the basis of a DFT grid of beams.

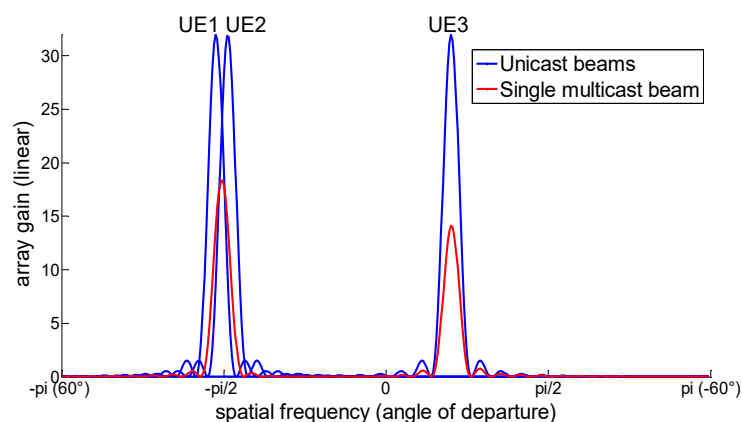


Figure 3-27 Multicast beamforming. The same signal is transmitted over each beam

Here we study the impact of UE grouping and UE distribution on the performance of multicast beamforming. We assume that the total transmission power is constant. The power is proportional to the area under the array gain curve (see Figure 3-27). If the UEs have different angles of departure, then this means a reduction of 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. A fundamental and important question to be clarified is: which of these options is better? We have carried out numerical simulations for different UE distributions, numbers of UE, and grouping strategies. The results are shown in Figure 3-28.

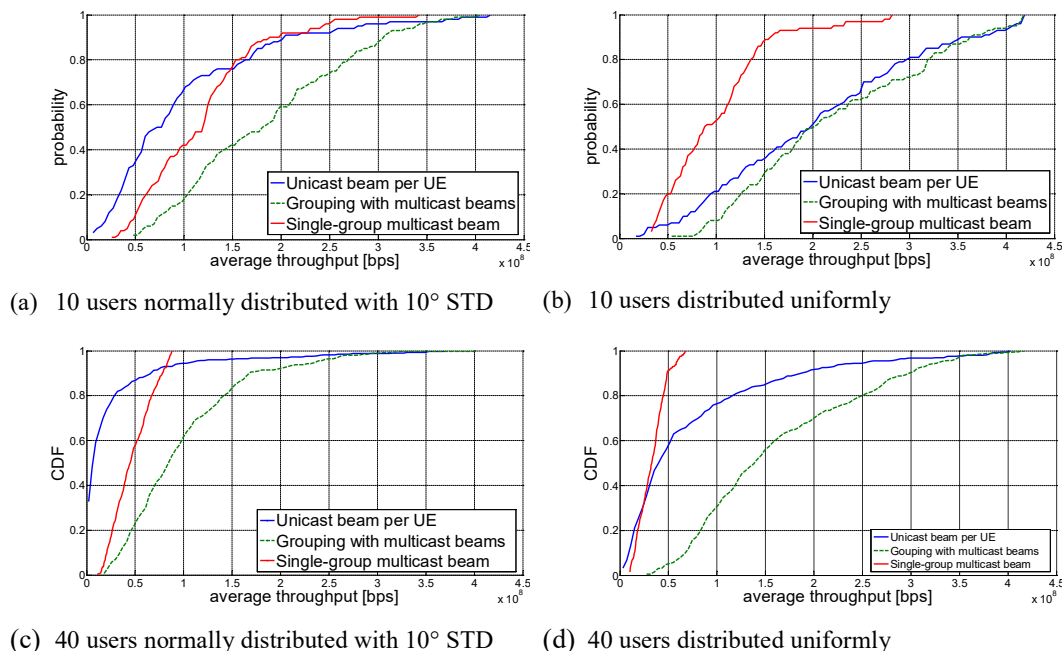


Figure 3-28 Comparison of unicast and multicast MIMO transmission

The simulation is configured as follows: $\lambda/2$ uniform linear array with 32 antennas, azimuth only. 40 UEs are distributed over a 120° sector. Urban macro channel UMa is assumed. The resulting SNRs are calibrated against [3GPP-36814]. UEs that are close in angles are grouped together. Three different curves correspond to different ways of mapping UEs to beams. We consider three strategies: 1) unicast beam for every UE (blue), 2) one beam per group (green), 3) single multicast beam for all UE (red). The UEs that belong to the same group and to different beams are scheduled to orthogonal frequency or time resources (100 MHz). This determines the number of resource blocks for each UE. For each beam, the MCS is selected according to the worst-case CQI. The resulting throughput is computed for the typical 3GPP numerology.

The results from Figure 3-28 show that the single beam strategy does not perform well for the extreme case of uniform distributions. In this case it is better to use unicast beams with scheduling and spatial multiplexing. The outcome is different if all UEs are spatially close to each other, in which case the single beam strategy is better. The results are explained by the tradeoff between the achievable array gain and the multicasting gain (sharing of resources).

The best performance is achieved by the heuristic grouping strategy, which combines the advantages of the other two (extreme) cases. In our simulations, a simple heuristic grouping strategy was used, which groups together UEs transmitting over the same DFT beams. The gain is becoming larger as the number of UEs grows.

3.3.4 Decentralized beamforming algorithms

In TDD, users send pilots in the uplink for the BSs to estimate the channels in the uplink and deduce the downlink channels making use of reciprocity. However, this estimation process is likely to be accompanied by errors denoted as the channel estimation is affected by errors. Hence in general the CSI at the BS side is not perfect. This imperfection in CSI at the BS side will heavily decrease the performance of classical precoders (designed to maximize sum rate for perfect CSI) and in turn decrease the achievable sum rate. We study optimal precoders in the case of imperfect CSI at the BSs by solving the expected sum rate stochastic maximization problem using the difference of convex functions approach (see Appendix 8.6). 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.

The performance of our proposed scheme, denoted by Combined Channel Estimate and Covariance CSIT, is evaluated through numerical simulations. We compare it to the EWSMSE approach in [NGC12] and to a naive approach where the sum rate maximization problem is solved but perfect CSI is assumed and the obtained precoder is used for the imperfect CSI case. The latter approach is denoted by Channel Estimate Only. Figure 3-29 and Figure 3-30 shows the achievable sum rate versus SNR for a MISO (or a MIMO) system composed of 2 cells, with 8 single-antenna (or two-antennas) users in total and 8 antennas per BS, the transmit covariance matrices considered are low rank matrices (i.e. rank equal to 2). The low-rank property of the correlation matrices is motivated by the work in [YGC14]. We use Monte Carlo simulation with 100 channel realizations; the channel estimation error has 3 times less gain than the channel estimate. As seen in both figures, the proposed approach provides higher sum-rate than the other approaches. For future work, performance of MIMO in multicell scenarios will be studied. Details of the proposed algorithm are provided in Appendix 8.6.

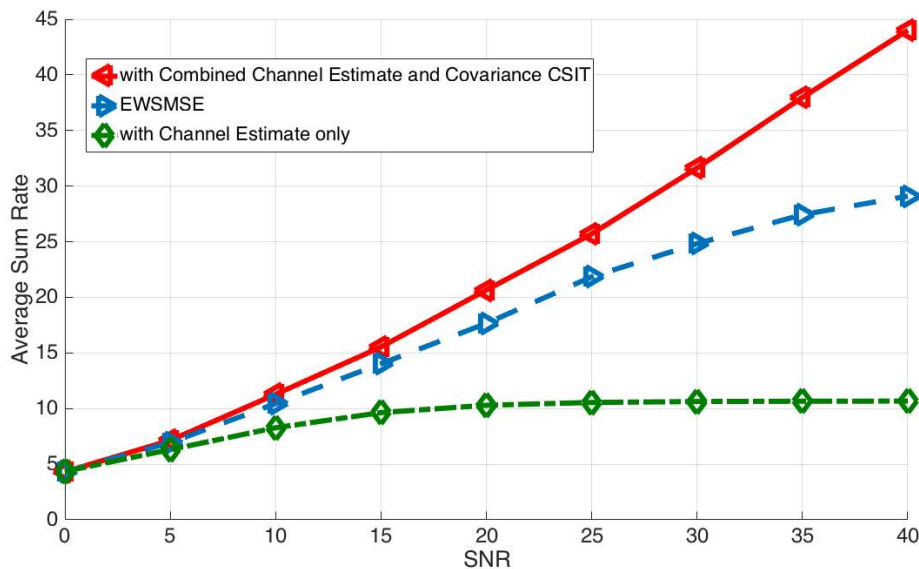


Figure 3-29 Sum rate comparison: Correlated low rank matrices, 1 antenna/user,4 users/cell, 8 antennas per base station, 2 cells, 25% estimation error

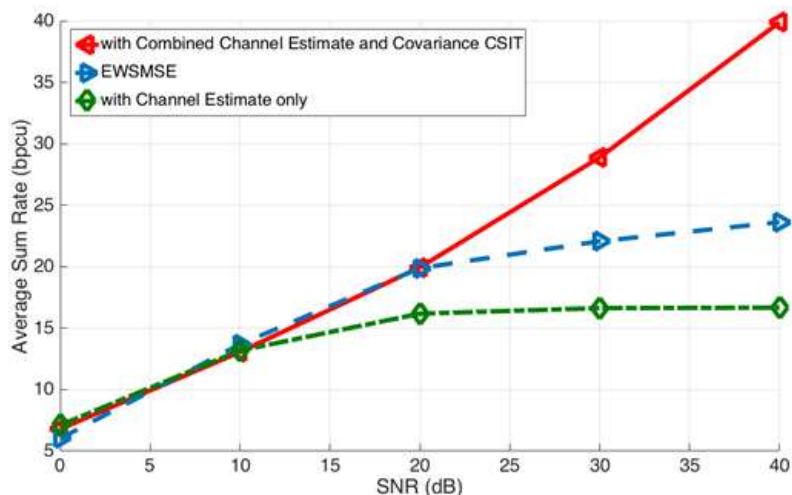


Figure 3-30 Sum rate comparison: Correlated low rank matrices, 2 antennas/user, 4 users/cell, 8 antennas per base station, 2 cells, 25% estimation error

3.3.5 Massive MIMO with Hybrid Analog-Digital Precoding in a CRAN Architecture

Cloud RAN (CRAN) architecture [CM-WP14] allows for splitting the functionalities of massive MIMO systems [Mar10]. In this CRAN architecture, remote radio head (RRHs) are equipped with a large number of antennas, connected to a baseband unit (BBU) via wired fronthaul links. Each RRH only performs basic functions, such as beamforming in the radio-frequency (RF) domain, while the BBU manages all the digital functions, such as channel estimation and beamforming in the digital domain [MRH+17]. This functionality split is thereby envisaged to enable the massive deployment of massive MIMO RRHs.

One major system bottleneck of the functionality split comes from the fact that the BBU can observe the channel between RRHs and user equipments (UEs) only through the analog beamformers. Estimating the exact channel state information (CSI) thus requires several sequential test measurements [AAL+14], which entails latency with the risk of outdated CSI. To avoid this problem, we propose an analog beamformer design that relies only on second-order channel statistics, i.e. spatial channel covariance matrices. Since spatial covariance matrices change less frequently than instantaneous CSI, this approach reduces channel estimation complexity as well as the fronthaul consumption for exchanging CSI.

Another technical bottleneck of the aforementioned functionality split is the data traffic volume transported through the fronthaul links. Due to the limited fronthaul capacity, it is not always desirable to transport all data streams from the BBU to the RRHs. Especially when fronthaul capacity is extremely limited, this may induce too much fronthaul quantization noise, degrading the overall data rate. In this respect, we suggest transporting only part of the data streams by adjusting each RRH's number of active RF chains, which affects the number of analog beams.

Given the analog beamformer design, we assume the BBU perfectly knows the instantaneous observed channels and constructs a regularized zero-forcing (RZF) digital beamformer. The effectiveness of the proposed hybrid precoding design under the functionality split is validated by simulation under different numbers of active RF chains and fronthaul capacities.

It is noted that our analog beamformer design stems from [PPY+17], where the analog beams are determined as the strong singular vectors of the sum of the spatial covariance matrices of all UEs. Our analog beamformer design is based on a weighted sum of the covariance matrices, allowing

to account for UE channels with different energy. Detailed procedure of the proposed algorithm can be found in Appendix 8.7 and further explanation can be found in [PKC+17].

We evaluate the performance of the proposed algorithm by means of Monte Carlo simulations. In the simulated system, we consider each of the 2 RRHs is equipped with a uniform linear array with 64 antenna elements with inter-element distance of half wavelength, whereas each of the 3 users has a single antenna. We consider the users are apart from RRHs at distances 1000m, 500m, and 100m.

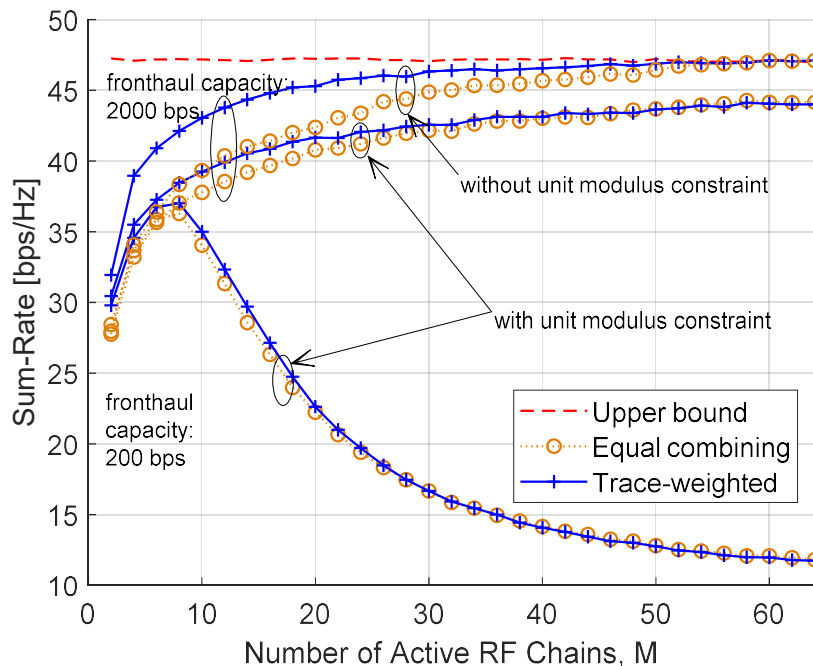


Figure 3-31: Sum-rate vs. the number of active RF chains

We evaluate the performance of the proposed method for design of the analog beamformer as a function of the number M of active RF chains per RRH. In Figure 3-31, the sum rate obtained by our proposed method (trace-weighted) is compared to that of a similar method which applies equal weighting to the covariance of all users (equal combining) for two different fronthaul capacities 200 and 2000bps. As an upper bound, the sum-rate performance of a fully-digital RZF precoder based on instantaneous CSI and operating over fronthauls with unlimited capacity is also presented. Our proposed design outperforms the design with equal combining in terms of sum-rate, especially when the number of active RF chains is low. In addition, the figure illustrates the impact of RF chain activation on the sum-rate. For sufficient fronthaul capacity (2000bps), sum-rate is a monotone increasing function of the active number of RF chains, even after exceeding the number of UEs. In this case, activating the whole set of RF chains always provides the highest sum rate. For insufficient fronthaul capacity (200bps), on the contrary, the quantization noise variance after fronthaul compression severely grows as M increases, and thus activating only a subset of the available RF chains better improves sum rate. We also evaluate the impact of enforcing the unit modulus constraint, which reflects that the analog beamformer changes phase only. As can be observed, the unconstrained algorithm reaches the performance of the bound for sufficiently large M , while the unit-modulus design has a small but significant loss in comparison.

3.3.6 Enhanced Backhauling

Wireless backhaul for coverage enhancement in underserved areas

In a low Average Revenue Per User (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 backhaul (BH) link between the BSs and the core network, when using a classical optical fiber 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. This wireless link should be a long-range link (several kilometres) and an adaptive link to compensate for signal impairments caused by weather conditions. Moreover, a sufficient data rate should be provided. 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 thereby to improve the spectral efficiency of the system. The spatial interference on the UE side is managed by a mMIMO precoding technique. The proposed solution is illustrated by Figure 3-32, where (a) a wired BH and (b) the proposed in-band wireless BH with mMIMO are illustrated. We focus on the downlink of a TDD OFDM system.

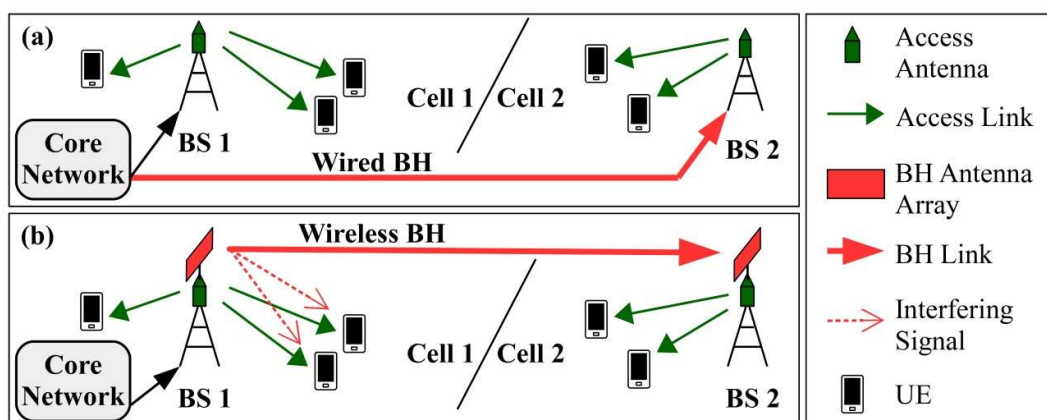


Figure 3-32 Illustration of (a) a wired BH and (b) the proposed in-band wireless BH with mMIMO.

In [LZL15], the authors either do not take into account the interference on the UE side or completely suppress it. For this project, a new mMIMO precoding technique, called RZF-CI, is proposed in order to maximize the received power on the receiving BS, while limiting the interference power level on the user side to a predetermined value Δ_{max} . This precoder is detailed in Appendix 8.8. Thereafter, the performance of the RZF-CI precoder is compared to the performance of the MRT precoder, which does not take into account the interference on the UE side, and to the ZF precoder, which completely suppress the interference on the UE side. Simulations are thus carried out using the 3GPP 38.901 RMa LOS spatial channel model [3GPP-38901]. For the simulations, the BH antenna arrays are identical uniform rectangular arrays facing each other with 16 antennas per column, 4 antennas per line and 10 cm inter-antenna spacing. The height of these BH antenna arrays is of 35 m and they are separated by 10 km. Moreover, the FDMA method is used on the access antennas to serve multiple UEs and the simulations focus on one UE. The considered UE is equipped with an omni-directional antenna and is located on the straight line joining the BS 1 and the BS 2. Figure 3-33 gives (a) the BER on the UE side as a function of the SNR (with 2 km between the BS 1 and the UE) and (b) the normalized received power on the BS 2 as a function of the distance between the BS 1 and the UE. The MRT, the ZF and the RZF-CI precoders (with $\Delta_{max} = 0.5$ dB and $\Delta_{max} = 1$ dB) are used. In terms of BER on the UE side, the degradation is controlled by the value of Δ_{max} with the RZF-CI precoder, while the performance with the MRT precoder is highly impacted by the in-band communication. In terms of received power on the receiving BS, the RZF-CI precoder allows an important gain compared to the received power with the ZF precoder (up to 8 dB), while the loss is rather limited compared to the received power with the MRT precoder (up to 1.5 dB).

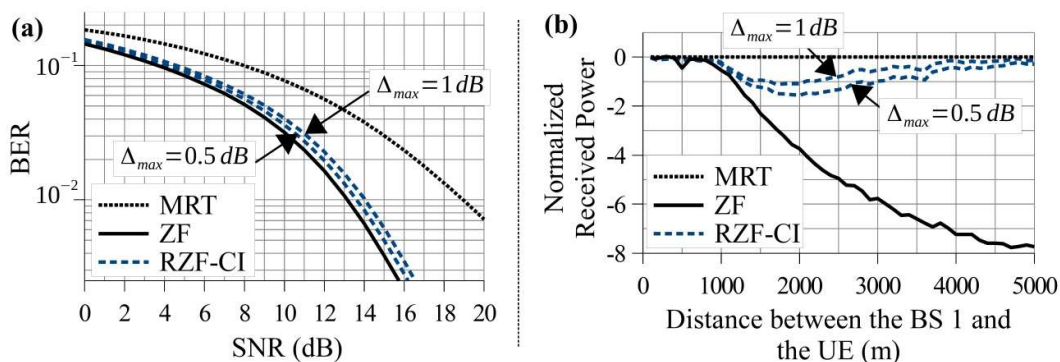


Figure 3-33 (a) BER on the UE side as a function of the SNR and (b) normalized received power on the BS 2 as a function of the distance between the BS 1 and the UE.

For future works, the impact of a desynchronization between the access and BH links will be studied. Moreover, we will try to reduce the complexity of the system thanks to hybrid beamforming.

Signal shaping for backhaul channels

Our main focus is on the ‘megacities’ scenario. Though, we will also contribute to the use cases related to ‘underserved areas’, e.g., long range connectivity in remote areas with smart farming application. Probabilistically Shaped Coded Modulation (PSCM) allows for flexible and efficient transmission on a wireless links [R1-1700076]. Therefore, the initial study of the PSCM complexity and throughput is necessary.

PSCM has been proposed as a modulation scheme for NR (see [R1-1700076]), therefore the investigation of distribution matching schemes is relevant for the 3GPP standardization. More details can be found in [PX17].

PSCM [BSS15], is proposed as a complement of the previous techniques. PSCM improves the performance of flexible backhauling. We have evaluated two architectures for fixed length distribution matching [BR13]. Distribution matching is essential for PSCM, especially the probabilistic amplitude shaping (PAS) [BSS15], a modulation scheme recently proposed for next generation wireless communications. PAS can be seen as a combined source-channel coding scheme where a source decoder, called a distribution matcher (DM) or shaping encoder (ShEnc) [R1-1700076], is introduced before a systematic channel encoder (ChEnc) at the transmitter [BSS15]. The receiver performs the inverse operations by introducing a source encoder, called the inverse DM or shaping decoder (ShDec), after a channel decoder (ChDec), see Figure 3-34. The parallel DM architecture, called the bit-level distribution matcher (BL-DM) [PX17b], was compared with the symbol-level constant composition distribution matcher (CCDM) [SB16]. Simulation over AWGN results confirmed higher throughput of the BL-DM without performance loss compared to the symbol-level CCDM, when parallel processing, e.g. with multiple processors, can be used (Figure 3-35, Figure 3-36). For sequential processing, e.g. single core systems, the symbol-level CCDM achieves higher throughput, i.e., lower number of operations per symbol. We also proved that when the output distributions of the BL-DM and the symbol-level CCDM are the same, i.e., they can be represented as a product of binary distributions, the BL-DM achieves lower or equal rate-loss, i.e. BL-DM can achieve higher transmission rate with the same parameters as SNR, FER, and signal distribution than symbol-level CCDM.

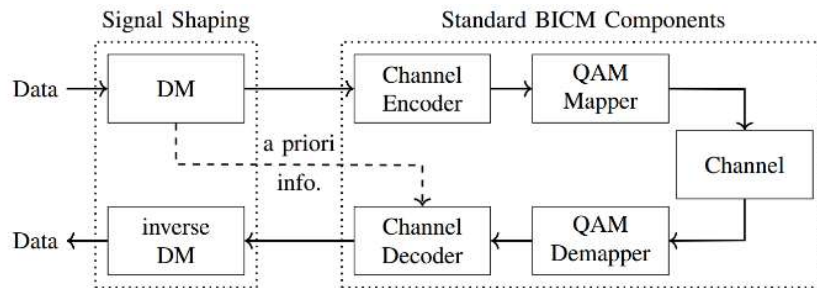


Figure 3-34 The PAS system investigated

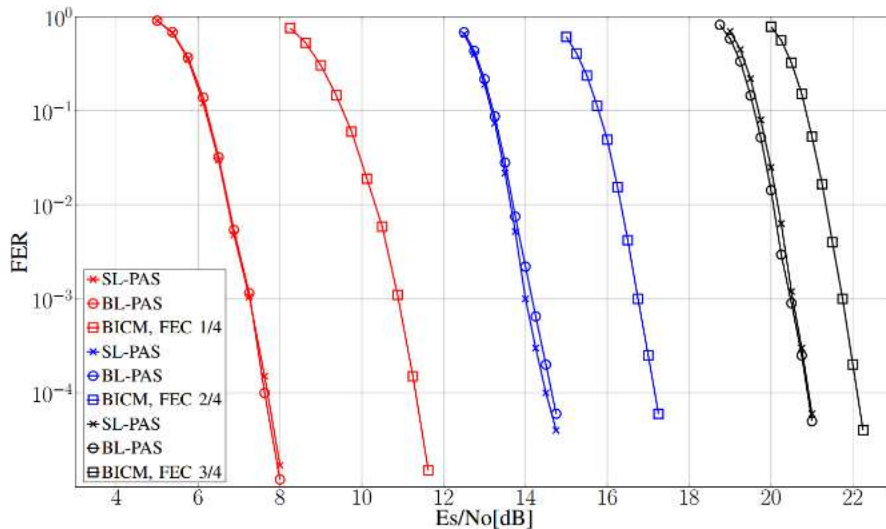


Figure 3-35 Frame error rate (FER) for two DM architectures: BL-PAS = BL-DM+PAS, SL-PAS = symbol-level CCDM+PAS.

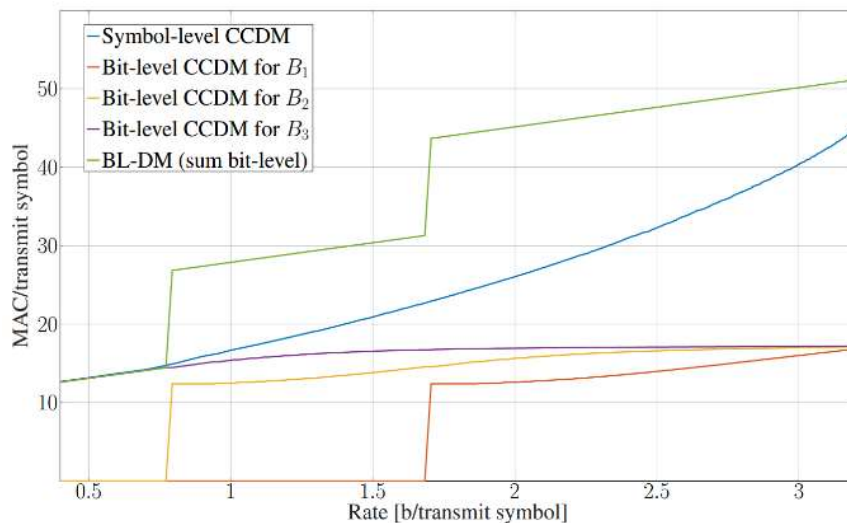


Figure 3-36 Number of operations in terms of MAC (multiply-accumulate) operations for the CCDM (blue) and the BL-DM with parallel processing (violet). The processing of B_1 , B_2 , B_3 , can be done in parallel, thus resulting in a lower number of MAC/symbol than symbol-level CCDM. However, the sum of MAC/symbol is higher for B_1 , B_2 , B_3 .

3.3.7 A Comparison of Hybrid Beamforming and Digital Beamforming with Low-Resolution ADCs for Multiple Users and Imperfect CSIR

The use of a large antenna array combined with a large bandwidth is a huge challenge for the hardware implementation; essentially the power consumption will limit the design space. At the moment, analogue/hybrid beamforming is considered as a possible solution to reduce the power consumption. Analog or hybrid beamforming systems strongly depend on the calibration of the analogue components. Another major disadvantage is the large overhead associated with the alignment of the Tx and Rx beams of the base and mobile stations. Specifically, if high gain is needed, the beamwidth is small and thus the acquisition and constant alignment of the optimal beams in a dynamic environment is very challenging [BHR+15], [MH15]. Another option to reduce the energy usage while keeping the number of antennas constant is to reduce the resolution of the ADCs. This can also be combined with hybrid beamforming.

The contribution of this work can be summarized as follows:

- Achievable rate analysis for hybrid beamforming systems and digital systems with low resolution ADCs in a multiuser, multipath, wideband scenario. In addition, the effects of transmitter impairments, channel estimation errors are considered.
- Analysing the channel estimation error considering the reference signal patterns already agreed upon for 3GPP NR (5G).
- Illustrating the energy efficiency - spectral efficiency trade-off considering the power consumption of the receiver RF front-end for hybrid and digital beamforming.

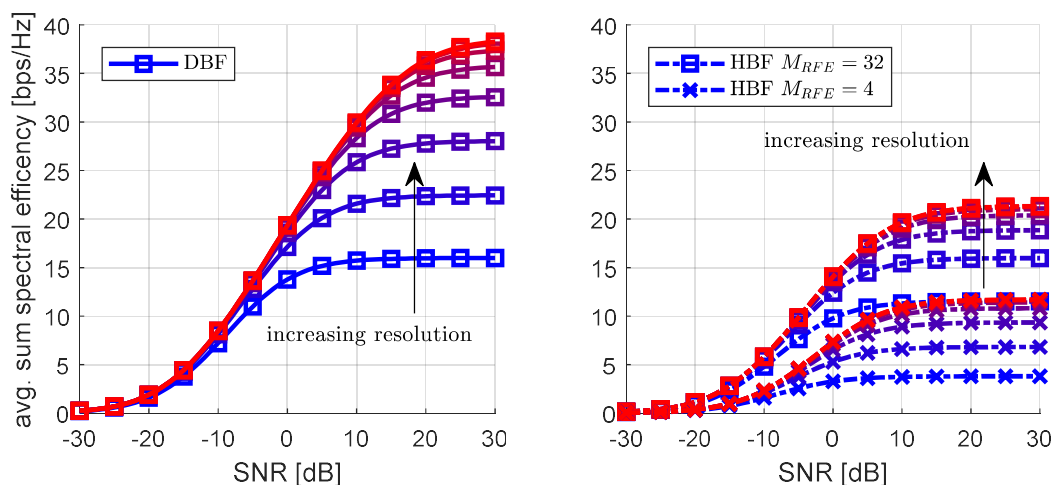


Figure 3-37 Avg. sum-spectral efficiency for different receiver configurations and ADC resolution 1-8 bit different number of RF chains M_{RFE} .

Figure 3-37 shows the average achievable rate over multiple channel realizations. The resolution in bits increases from the bottom to top for each group of curves. From the Digital Beamforming (DBF) results in Figure 3-37 we see that at high SNR the rate saturates and there is only minor improvement above a resolution of 5 bits. The reason for this saturation is the transmitter impairments.

Figure 3-38 shows the achievable rate and energy efficiency at different SNR values. For each curve the ADC resolution increases from the leftmost point of the curve. This point represents 1 bit resolution for all ADCs, or 1 bit resolution for the ones with lower resolution ADCs in the case of mixed-ADC DBF. For all cases we see that the DBF system is more energy efficient compared to Hybrid Beamforming (HBF). The major reason for this is that the digital system

retains all available degrees of freedom. We can see that as the SNR increases (Figure 3-38) the smaller the improvement of additional RF chains. The explanation for this is that even though we gain more degrees of freedom we still need to divide them among the users. In Figure 3-38 we see that there is little difference between having 8 or 16 RF chains. As the SNR increases the optimal resolution in terms of energy efficiency also increases. As predicted from the achievable rate curves, above a resolution of 5 bits the energy efficiency decreases for all cases.

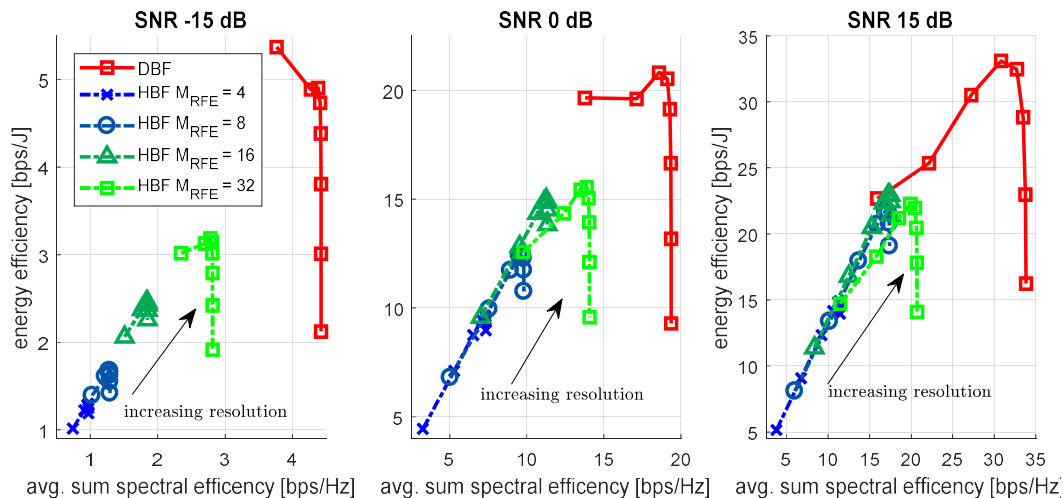


Figure 3-38 Spectral and energy efficiency of digital beamforming hybrid beamforming with 64 receive antennas and 4 users at different SNR with ADC resolution 1-8.

The evaluations in this work showed that low resolution ADC digital beamforming systems are more energy efficient and achieve a higher rate than hybrid beamforming systems for multiuser scenarios. The reason is that the sub-arrays of hybrid beamforming must focus on a single user. Evaluations with mixed ADC configurations showed that such systems can achieve different achievable rates and energy efficiency values around the ones achieved by a uniform ADC configuration. Future extensions should consider the following points. For the hybrid beamforming case, the evaluation only shows the result if the beams are already aligned. As shown in [BHR+15], beam alignment can require a large overhead. In addition, considering what degree of power disparity among the users is possible for different ADC resolutions also provides an interesting scenario to evaluate. Additional details of the evaluation can be found in [RPS+18].

4 Advanced link coordination based on CRAN/DRAN and massive MIMO

The ever-increasing network infrastructure density increases the total network throughput due to the corresponding increase of spatial frequency reuse. However, this reuse gain comes with the cost of increased interference levels, which may harm user throughput and, therefore, must be taken into account. Towards this end, the paradigm of CRAN is one of the most promising approaches that are considered extensively by 3GPP. In particular, as described in detail in [ONE17-D31], the radio functionalities of a node B (base station) are split in a way such that RAN functions that can be centralized (typically corresponding to higher layer operations and some, but not all, lower layer operations) are performed by a central unit (CU). The CU is connected via high speed links to multiple distributed units (DUs) and/or remote radio heads (RRHs). The natural benefit of this approach is to enable coordinated transmissions at the signal (PHY) level and/or scheduling/resource-allocation (RRC, MAC) level. What is equally important is the flexibility this architecture provides towards efficient implementation of the network slicing

concept for diverse type services such as eMBB and mMTC. This flexibility allows the system to operate in any state between fully centralized (CRAN) and fully distributed (DRAN).

However, the CRAN/DRAN architecture also comes with its own, unique challenges and issues that are still open in 3GPP and need to be addressed. The goal of this section is to provide advanced solutions/enablers for CRAN/DRAN that are grouped and presented into the following two major categories: (a) Physical layer techniques and procedures for CRAN/DRAN and (b) Resource allocation and traffic management in CRAN/DRAN. The CRAN architecture described in [ONE17-D31] is considered throughout this section.

The first category identifies challenges and provides solutions related to CRAN-specific physical layer techniques/procedures such as CSI acquisition/feedback, and beamforming. Even though these procedures are mature in current cellular networking, they require a complete re-design in a CRAN setting, since direct application of current approaches are facing severe performance as well as scalability issues. Particular focus is given on the design of efficient (low overhead) training and feedback signalling for CSI acquisition and reporting, respectively.

The second category identifies and provides solutions related to resource allocation, scheduling, and traffic management. The studies considered in this category cover a wide range of topics, including mapping virtual networks to a given physical CRAN topology (e.g., for implementing network slicing), comparison of performance among different coordination schemes with various levels of complexity and/or CSI requirements, as well as relay-assisted coordinated uplink transmission with limited CSI.

4.1 Physical layer techniques and procedures for CRAN/DRAN

CRAN considerations in 3GPP NR are in a very preliminary stage and currently only limited towards providing a flexible architecture that can enable efficient network slicing/virtualization [ONE17-D31]. There is currently no provision by the standard for sophisticated cooperative and/or coordinated transmissions that are inherently available in CRAN. This is due to CRAN-specific physical layer challenges for which no mature solutions are currently available. This is the reason that the specification of the “lower layer split” is still pending [ONE17-D31].

The major challenge from the perspective of physical layer techniques and procedures is clearly related to the acquisition and reporting of *global* CSI (assuming the issues related to backhaul/fronthaul are resolved). This is because, optimal network decisions in terms of, e.g., transmission mode selection, resource allocation, user scheduling, and power control, require, ideally, knowledge of the link quality of *all possible pairs* of RRHs and UEs, even if not all of the links (if not most of them) are strong. This means acquiring and reporting CSI for tens (or more) of links per UE, for a total of hundreds (or more) of links in the CRAN/DRAN system. Incorporating the conventional CSI acquisition and feedback approaches in the CRAN/DRAN setup will clearly incur a tremendous cost in signalling overhead both for acquisition (by means of training signals) as well as reporting (by means of feedback signals) of CSI.

The major focus of this subsection is with respect to issues related to CSI acquisition and feedback/reporting. It provides an overview of the signalling framework currently considered in 3GPP NR (Rel-15) and identifies related limitations and challenges, towards solutions that achieve, with reasonable signalling overhead, (close to) optimal performance in terms of, for example, MCS selection and UE-to-RRH association. In addition, a new paradigm for global CSI acquisition with reduced training overhead, based on the recently emerged field of compressive sensing is proposed. Finally, this section proposes a machine-learning-aided signal detection approach towards optimal operation that is robust to the inherent interference appearing in CRAN with multiple concurrent transmissions (e.g., mMTC) as well as fronthaul capacity limitations. The solutions provided here are highly relevant to CRAN related work/study items for future releases of NR (Rel-16 and beyond), such as NR CoMP, (flexible) duplexing, and mMTC.

This section considers some of the physical layer challenges towards achieving the full premise of sophisticated cooperative/coordinated CRAN transmissions. As the availability of accurate and global CSI available at both transmitter(s) and receiver(s) is critical in CRAN, the challenge of low-overhead signalling for acquiring and reporting CSI was extensively investigated. A summary of the current RS framework in 3GPP was provided and its limitations regarding scalability of CSI acquisition and reporting were identified. For CSI acquisition, the potential of compressive sensing techniques towards obtaining global CSI was investigated, where it was shown that low-overhead training is only possible under sufficiently large path losses (e.g., dense urban scenarios and/or high carrier frequencies). In addition, a novel interference suppression technique for CRAN, based on concepts from machine learning and non-linear filtering was proposed.

These studies are highly relevant to eMBB and mMTC use cases, as the CRAN, under cooperative transmissions, can naturally accommodate a huge number of devices with maximum data rates.

4.1.1 Reference signal framework in new radio for cooperative transmission

In 3GPP 5G NR system, advanced MIMO transmission techniques are very important technology component to fulfil the high spectrum efficiency requirement. Different MIMO schemes such as SU(MU)-MIMO and CoMP are further enhanced in NR compared to LTE system. Moreover, due to the essential support of the transmission in high frequency spectrum, e.g., mmWave frequency band, beam-centric transmission schemes are extensively developed.

In 3GPP NR [3GPP-38802], beam management is defined as a set of layer 1/layer 2 (L1/L2)² procedures to acquire and maintain a set of transmit-receive points (TRPs) and/or UE beams that can be used for DL and UL transmission/reception. Specifically, the following DL L1/L2 beam management procedures are supported within one or multiple TRPs:

- DL beam alignment: Used to enable UE measurement on different TRP Tx beams to support selection of TRP Tx beams/UE Rx beam(s).
- DL beam refinement: Used to enable UE measurement on different TRP Tx beams to possibly change inter/intra-TRP Tx beam(s).
- UE receive beam refinement: Used to enable UE measurement on the same TRP Tx beam to change UE Rx beam in the case UE uses beamforming.

In addition to the above beam management procedure, other advanced CoMP transmission techniques such as dynamic point selection (DPS), (non-)coherent joint transmission (JT) are also supported in NR. To support all these advanced transmission schemes, the following reference signals are supported in NR.

Synchronization signal block

In NR, synchronization burst set is comprised of a number of synchronization signal blocks (SSB). Each SSB shall be transmitted in a specific beam direction. Regardless of the periodicity of synchronization burst set, in each 5ms time window where SSBs are transmitted, different maximum number of SSBs are supported for different carrier frequencies. Within this 5 ms window, number of possible candidate SS block locations is L. The maximum value of L, for different frequency ranges are

- For frequency range up to 3 GHz, L is 4
- For frequency range from 3 GHz to 6 GHz, L is 8
- For frequency range from 6 GHz to 52.6 GHz, L is 64

² L1 refers to PHY sublayer, while L2 refers to MAC/RLC/PDCP sublayers.

The detected SSBs shall be used for time/frequency synchronization source for UE at least during initial access stage. Even after UE is RRC connected with the network, in addition to time/frequency tracking, SSBs shall be also used for beam tracking.

CSI-RS

Similar to LTE, channel state information reference signal (CSI-RS) is used in NR for both beam management and CSI acquisition and feedback. Three types of CSI-RS are supported as follows:

- Periodic CSI-RS (p-CSI-RS): is configured by RRC signaling, transmitted periodically, and can be used for time-frequency tracking and CSI acquisition and feedback.
- Semi-persistent CSI-RS (sp-CSI-RS): is configured by RRC signaling, and activated/deactivated dynamically by physical downlink channel, and can be used for time-frequency-spatial tracking, and CSI acquisition and feedback.
- Aperiodic CSI-RS (a-CSI-RS): is scheduled by physical downlink channel, and used for spatial tracking and CSI acquisition and feedback.

Transmission configuration information

In NR, concept called transmission configuration information (TCI) is used for UE to support advanced transmission schemes such as CoMP. UE can be configured by RRC with one or more TCI states. Each configured TCI state includes reference signal IDs used for time/frequency synchronization reference and spatial quasi-collocated (QCLed) reference signal IDs. The time/frequency synchronization reference signal can be SSB, or p-/sp-CSI-RS. The spatial QCLed RS can be SSB, p-/sp-/a-CSI-RS. For each scheduled data transmission, the respective downlink control information (DCI) includes the index of configured TCI defining the time/frequency synchronization reference and spatial QCLed RS for the demodulated reference signal (DMRS) of the scheduled data channel. By virtue of signaled TCI, UE can determine the proper receive beamforming filter for the scheduled data channel with certain transmit beam direction.

Amplitude quantization for Type-2 codebook based CSI feedback

In NR system, two types of codebook, namely Type-1 and Type-2 codebook, have been standardized for CSI feedback in the support of advanced MIMO operation. Both types of codebook are constructed from 2-D DFT based grid of beams, and enable the CSI feedback of beam selection as well as PSK based co-phase combining between two polarizations. Moreover, Type-2 codebook based CSI feedback can report both the wideband and subband amplitude information of the selected beams. As a result, it is envisioned that more accurate CSI shall be obtained from the Type-2 codebook based CSI feedback so that better precoded MIMO transmission can be employed by the network. To reduce the CSI feedback signalling, 1 bit based subband amplitude with only two quantization levels is supported in combination to 3 bits based wideband amplitude feedback. Typically, wideband amplitude shall be calculated as the linear average amplitude of the beam over all subbands. However, due to the coarse subband amplitude quantization, it has been observed in case of joint wideband and subband amplitude feedback, the linear average based wideband amplitude can lead to a large amplitude quantization errors. In this project, we develop two methods for joint wideband and subband amplitude calculations. Specifically, both optimal and sub-optimal methods are proposed. The optimal method can achieve the minimum amplitude quantization errors at the cost of a relatively large computation complexity. And by virtue of a derived scaling factor, the sub-optimal method exhibits clearly smaller quantization error than the conventional linear average based method especially for the channel with large frequency selectivity. The details of the developed work can be referred to [MMF18].

4.1.2 Efficient signalings in massive MIMO multi-node networks

Both centralized and distributed radio access networks (CRAN/DRAN) are being considered in 5G. In particular, a number of network coordination schemes have been provided and discussed in the new radio coordinated multipoint (NR-CoMP). The JT scheme, where data for one UE can

be transmitted from multiple transmission/reception points (TRPs) in the same time-frequency resource, includes two types of transmissions, the so called, *non-coherent JT* (NCJT) and *coherent JT*.

In non-coherent JT, multiple data layers are independently transmitted to a UE by different transmission points in a single-user MIMO transmission. While in coherent JT, multiple layers are jointly transmitted to a UE by a set of transmission points in a single-user MIMO transmission.

There are a number of problems related to signalling in coordinated networks, where the design of efficient signalling is critical in order to reduce signalling overhead and/or enhance feedback quality from UE. For example, there are different network coordination methods, including coordinated scheduling/coordinated beamforming (CSCB), JT and DPS. Typically, the UE needs to measure and generate multiple CSI reports based on multiple interference hypotheses. UE can, in principle, report all the different CSI measurements corresponding to all hypotheses, or only report a subset of CSI measurements when the network dynamically triggers such a feedback from the UE. In either case, multiple CSI reports are needed. As a result, massive signalling overhead can be created, which is not desirable in a mMTC scenario.

In other occasions, enhancement of UE feedback with respect to the existing NR CSI framework may be needed in coordinated networks. For example, the numerous possible resource allocation schemes with NCJT result in different interference levels on given PRBs that should, in principle, be reported by the UE.

Figure 4-1 shows an example of the different interference scenarios under different resource allocation schemes, where schemes 1-3 correspond to having PRBs (1) fully overlapped, (2) partially overlapped, and (3) no overlapped resource allocations, respectively. In such a case, it is desired that modulation and coding schemes (MCS) should also be selected differently according to the different interference types. Figure 4-2 shows an example of PRB allocation for two UEs and the resulting interference. The left panel shows the MCS assignment for the non-overlapped PRB of UE 1 (where two TRPs serve one UE each), and the right panel shows the MCS assignment for the overlapped PRB (where both TRPs serve UE1). Note that the interference power is different in these two cases, implying that, in principle, these two MCSs should be chosen differently. However, in the current framework, multiple MCSs for one codeword is not supported yet.

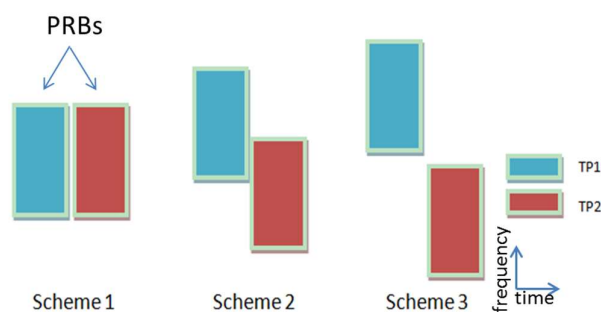


Figure 4-1 Different resource allocation schemes result in different interference scenarios.

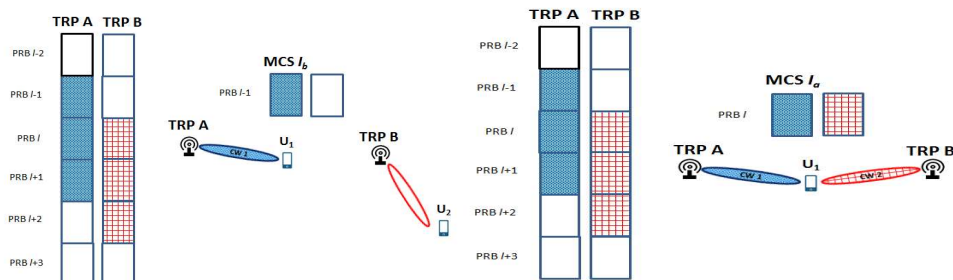


Figure 4-2 Illustration of the interference scenarios on different PRBs.

One of the work under this topic that is currently ongoing in WP4, studies flexible Radio Resource Management (RRM) for Coordinated Multi-Point (CoMP), considering three different network coordination scenarios: no coordination, partial coordination, and full coordination. RRM schemes are proposed. The proposed scheme uses offset RRM with UE feedback mechanism, allowing flexible RRM solutions to optimize PRB allocation and multiple MCS configuration for three scenarios with different level of coordination between TRPs involved in NCJT, depending on different UE interference measurement capabilities and backhaul conditions. Another aspect of signalling that should be considered is a new reference signal in NR, i.e., the phase tracking reference signal (PTRS), which can be used common phase error (CPE) compensation in CoMP. The proposed scheme is expected to provide flexibility to deal with different coordination levels among multiple BSs, improve the radio resource management efficiency, enhance the robustness against different backhaul conditions and interference measurement hypotheses, and reduce the complexity. Ongoing work is being carried out to finalise the details of the work, and the results will be provided in a future deliverable.

4.1.3 Compressive channel estimation in CRAN

In CRAN, multiple low-complexity, low-cost RRHs are distributed over the network coverage area and are all connected to a central, cloud-based baseband unit whose task is to jointly process the signals received from or sent to the UEs in the system. This centralized-by-design network architecture is, in principle, able to realize the vision of large-scale, multi-cell cooperative networks [SWJ15]. However, a major challenge towards this goal is the need for *global* CSI [PSL+16], [SWJ15], i.e., estimation of the quality of all links among the RRHs and the UEs. With a large number of RRHs expected at least for eMBB use cases, the standard training procedure based on orthogonal pilot sequences that is also currently employed by 3GPP, results in an unacceptably large overhead [SZL14], rendering the feasibility of sophisticated cooperative/coordinated transmission techniques questionable.

Recently, various research efforts have been made towards reducing the training overhead in CRAN (e.g., [SZL14], [ZYZ17], [XRL15], [HQC+17]). All these works are based on the premise that, in a large-scale CRAN deployment, only a small subset of the RRHs will have significant contribution to the downlink received energy by any UE (similar consideration holds for the uplink as well). Therefore, a virtual *channel sparsification* [WBS+15], [FZY16] is assumed at the UE side, which considers only the strongest RRH-to-UE links for channel estimation purposes and ignores the remaining links. This approach effectively allows for reduced training overhead using small-length pilot sequences. The effectiveness of this approach has been verified in the literature numerically, providing valuable insights on the effectiveness of small-length training sequences. However, the *operational conditions*, e.g., propagation losses and small-scale fading statistics, under which the channel sparsification assumption is applicable are not clear.

In this work, the performance potential of applying compressive sensing (CS) techniques towards reducing downlink training overhead in large-scale CRAN deployments is investigated. In

particular, for a CS-motivated training sequence design that is independent of the number and positions of RRHs and UEs, the performance of the, so called, oracle estimator is analytically obtained. The oracle estimator has *a priori* knowledge of the set of RRHs with the strongest links to a certain UE, but not their channel values, and its performance serves as an optimistic estimate for the performance of many practical CS estimation algorithms (that have no such *a priori* information) [CRW12], [CRM14]. By employing tools from stochastic geometry, a closed-form upper bound of the oracle estimator mean squared error performance is obtained, that is averaged over the distribution of RRH positions and channel statistics and is *independent* of the UE position within the CRAN deployment. The bound expression is tight for large scale, dense CRAN deployments and provides insights on how estimation performance is affected by (a) system *design* parameters, e.g., training sequences length and number of (strongest) channels to be estimated by each UE, and (b) *operational conditions*, e.g. RRH density and small-scale fading statistics. Detailed derivation of this bound as well as numerous related insights can be found in [SW18].

Figure 4-3 demonstrates the mean squared error (MSE) of the estimate of the global downlink CSI, obtained by a UE located at the centre of a square area where $N_{RRH} = 500$ RRHs are independently and uniformly distributed. The channel between an RRH and the UE is considered to be flat fading (i.e., a single PRB is considered) and takes into account small- and large-scale fading. The latter is characterized by the value of the path loss exponent α , assuming the standard model where the transmitted signal power decreases proportionally to $d^{-\alpha}$, where $d > 0$ is the propagation distance. The system utilizes non-orthogonal training sequences, each of duration (length) $N_p < 500$ symbols and randomly assigned among RRHs, with the training symbols independently generated as complex Gaussian variables. During the channel estimation stage, all RRHs transmit their sequences simultaneously and the UE is called to obtain the channel estimates from the superimposed received sequences. Note that the standard, orthogonal-training approach would require training sequences of length greater than or equal to 500 symbols (channel uses), which can be considered as an excessive overhead under certain conditions (e.g., high mobility).

The figure shows the MSE obtained by applying the standard basis pursuit (BP) algorithm of CS theory (e.g., see [FR13]) for the estimation of the global CSI. In addition, the theoretical MSE achieved by the oracle estimator is depicted. It can be seen that improved performance is obtained by an increased value of the path loss exponent α . In particular for $\alpha \geq 4$, very good performance can be achieved with a training overhead reduction much greater than 50% (compared to the standard, orthogonal training approach). However, performance is poor for small values of α which suggests that CS-based estimation of the global CSI is a viable approach in dense CRAN deployments, however, *only under sufficiently large path-loss factor* as in e.g., dense urban scenarios and/or high carrier frequencies.

Future work will consider incorporating the insights of the oracle estimator analysis to the design of practical CS-based algorithms.

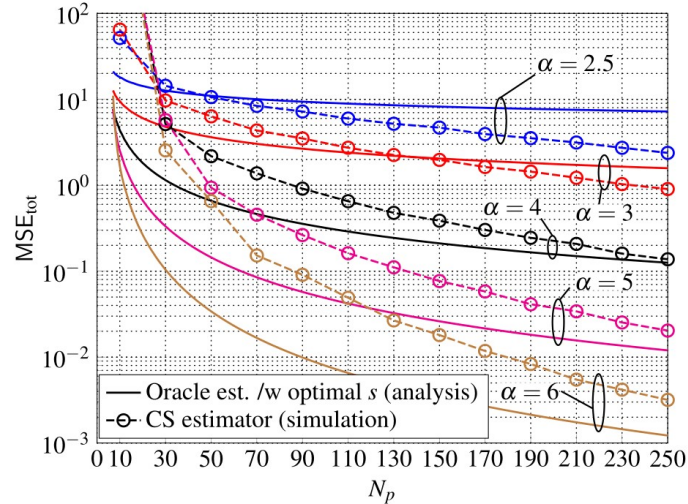


Figure 4-3 MSE of the global CSI estimate for the CS-based channel estimator as a function of the training sequence length N_p . The analytical expression for the MSE of the oracle estimator is also shown.

4.1.4 Nonlinear mechanisms in cell-less systems

Cell-less systems are a promising approach for enabling massive connectivity in mMTC/IoT systems. In contrast to (conventional) cellular-systems based systems, in the cell-less uplink, devices can broadcast to several radio access points (RAPs) without associating with any of them. As mentioned in previous sections, C-RAN is an important candidate for implementing cell-less systems. Under the assumption of an ideal high-capacity fronthaul, most of the complex baseband processing is migrated to the CU. However, the traditional “functional split” (i.e., to perform radio access at RRH and baseband processing at CU) in C-RAN systems with a capacity-limited fronthaul may not be appropriate. Especially, in a massive connectivity scenario, the data rate requirement on fronthaul links might exceed the fronthaul capacity. Therefore, in recent studies, the virtue of performing some “local” baseband processing at RRHs followed by data fusion at the CU has been investigated (see, e.g., [USD17]). It has been shown that this local processing strategy (termed Detect&Forward (D&F) in the following) outperforms the standard Quantize&Forward (Q&F) C-RAN processing for a low-capacity fronthaul.

Against this background, we propose a machine learning-based D&F symbol detection method for cell-less C-RAN systems. In contrast to information theoretic compression approaches, our aim is to achieve a low (Gray-coded) bit error rate (BER) which is of a practical importance from the point of view of reliability at a given device transmission rate. We consider a limited-capacity fronthaul and only a few antennas at the RRH for reasons to do with cost and complexity. Intuitively, a distributed D&F strategy can only compete well with the centralized joint processing in a Q&F system if the local detection/processing at RRHs exhibits a good performance. Due to the availability of a small number of antennas at the RRHs, we employ non-linear or partially-linear detection of [ACY+18], which has been shown to outperform the conventional MMSE-SIC systems (with SNR disparity and symbol-level SIC) when only a few antennas are available. This method is summarized with performance comparison with MMSE-SIC in Appendix 8.9. The soft-information (in the form of likelihood values) about the local symbol detection at the RRHs is then fused/combined at the CU. To this end, we develop a low-complexity set-theoretic method to estimate likelihood (pdf) functions at the RRHs. There are many methods that can accomplish the fusion of likelihood values including consensus and optimal log-likelihood quantization. For the purposes of simulation and comparison with traditional Q&F, we perform simple scalar quantization of the likelihood values. Please see Appendix 8.9 for more details about forwarding strategies and the algorithms used to obtain the results below.

We assume a Rayleigh block fading channel, in which the channel remains constant for a certain number T_c of channel symbols in a coherence block. During the first $T_t < T_c$ channel symbols in this block, training of the nonlinear detection framework (a filter function f) is performed and for $t > T_t$ detection/data communication takes place. These assumptions are typical in learning-based communication systems. The learning and detection is performed using Algorithm 1 in [ACY+18].

Figure 4-4 (left) demonstrates performance comparison between the set-theoretic D&F (solid lines) strategy in comparison with the conventional centralized joint processing of Q&F for uncoded QPSK modulation. Device SNRs (from the set $\{-3 \text{ dB}, -2 \text{ dB}, \dots, 9 \text{ dB}, 10 \text{ dB}\}$) and channels (having Rayleigh distribution) are chosen independently at random and results are averaged over 10000 experiments in total. The D&F strategy with likelihood fusion described above outperforms Q&F for low fronthaul capacity. The results justify the use of the nonlinear detection [ACY+18] for a small number of antennas is $M = 3$, number of RRH $R \in \{2, 3, 4\}$, and a device cluster size $K = 6$. Note that device clusters are allocated disjoint orthogonal frequency blocks (RBs) (i.e., there is no inter-cluster interference that we need to consider). This enables us to perform simulations for a single cluster. But the processing can be applied to multiple clusters in parallel. The results justify the use of the nonlinear detection [ACY+18], and the D&F strategy outperforms Q&F for low fronthaul capacity. Figure 4-4 (right) demonstrates the performance gains that can be achieved by exploiting the inherent spatial diversity offered by cell-less systems. In comparison with the centralized nonlinear detection using a single base station (BS) [ACY+18], cell-less systems achieve a better performance over the range of the training time. This result shows that the training time can be significantly reduced by exploiting the spatial diversity offered by a cell-less system.

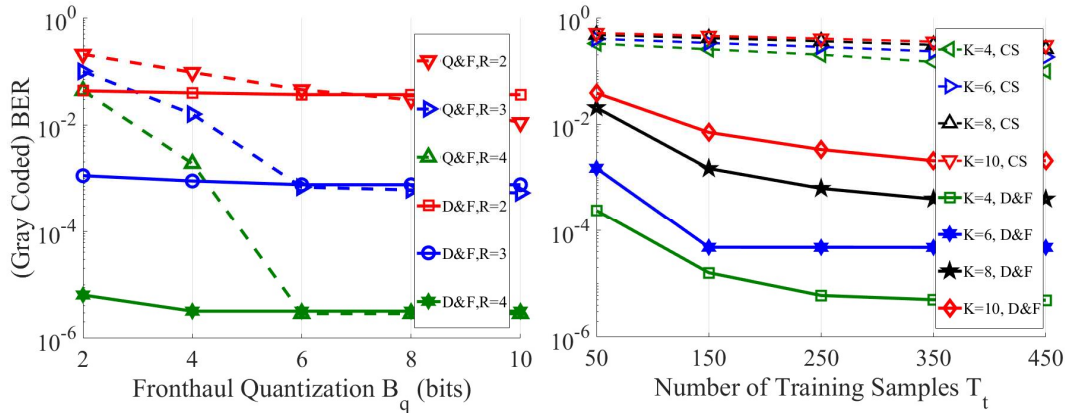


Figure 4-4 (left) Performance of D&F Vs. Q&F with $M = 3$ antennas at each RRH and $K = 6$ devices transmitting in the uplink. The results are shown for an increasing number of RRHs denoted by $R \in \{2, 3, 4\}$ and 100 training samples. (right) Performance of D&F (with $R = 3, B_q = 4$) Vs. Single BS (CS) [ACY+18]. The results are shown for an increasing cluster size $K \in \{4, 6, 8, 10\}$.

4.2 Resource allocation and traffic management in CRAN/DRAN

In order to offer connectivity and services to ubiquitous users, the evolution of network architectures can benefit from Cloud technologies. CRAN and DRAN that are attracting increasing interest in 3GPP emerged as two viable architectures that permit to achieve remarkable improvement in terms of latency reduction, higher data throughput or better energy efficiency for 5G networks. However, several challenges arise in CRAN/DRAN, such as wireless interferences,

backhaul and functionality issues, which need to be addressed through judicious resource allocation and prudent scheduling paradigms. In this subsection, the issue of resource allocation and traffic management in CRAN/DRAN is addressed from a lower layer operations perspective (see [ONE17-D31] for higher layer solutions). Aspects such as favorable propagation in cell-free massive MIMO, the problem of flexible functionality assignment, taking into account the required QoS as well as the capacity constraints of the links are addressed. Moreover, the performance of different network coordination schemes is studied, with a particular emphasis on centralized and distributed non-coherent joint transmission of the TRPs. The optimization of user's long-term throughput for slow-fading Multiple Access Multiple relay Channel (MAMRC) is also investigated.

In Section 4.2, several aspects of resource allocation and traffic management in CRAN/DRAN were investigated. Firstly, the problem of user scheduling in cell-free massive MIMO was investigated. After showing that this problem is NP-hard, a polynomial time solvable sub-optimal solution has been developed using semidefinite relaxation technique. Secondly, the problem of flexible functionality assignment that takes into account the functional architecture as well as the traffic load was studied. Based on an optimization problem formulation, the system can adjust to varying traffic requirements, optimally exploiting the functional split flexibility offered by the CRAN architecture. Then, a coherent joint-transmission that optimizes the beamforming in order to maximize the users' utility function was considered. Cross-link interference (CLI) resulting from duplexing in CRAN was investigated. Numerical analyses revealed an improvement in the downlink throughput. This performance enhancement was obtained by means of proper CLI management, which suppresses the interfering channel that is partly available at the target UE. Finally, the problem of optimizing the long-term throughput for cooperative incremental redundancy retransmissions in the slow-fading MAMRC was investigated. Three different strategies that are based on a centralized perspective were adopted and their performance was validated through simulations.

The above studies represent novel approaches on resource allocation and traffic management, towards optimal utilization of the CRAN architecture and its associated flexibility. The solutions and corresponding performance results provide guidelines for the upcoming standardization of CRAN-specific aspects by 3GPP, such as lower layer function split and operation under dynamic TDD.

4.2.1 Architecture optimization for Cell-less mMIMO systems

In order to tackle the ever-increasing amount of data traffic traversing wireless networks, a transformation of the radio access network (RAN) is necessary. One of the most promising architectures for future RANs is Cloud radio access network (CRAN). Leveraging multiple low-complexity, low-cost and distributed access points (APs) with a central, cloud-based baseband unit that handles signal processing, CRAN enables to reduce the cost of deployment and operating networks while providing considerable coverage and performance gains. Coupling cloud architecture with a great number of APs results in what is known as Cell-Free Massive MIMO (CF mMIMO) [NAY+17], or equivalently, Cell-less Massive MIMO (Cl mMIMO). Cl mMIMO systems aim at harvesting the considerable spatial multiplexing gain of conventional Massive MIMO together with the aforementioned benefits of cloud architecture. The published works on Cl mMIMO systems proved the energy efficiency and spectral efficiency gains in addition to the uniform QoS that these systems can provide [NAY+17], [NAM+17]. Therein, favorable propagation and channel hardening were leveraged in order to assess Cl mMIMO performance. Favorable propagation refers to the mutual orthogonality among the vector-valued channels of the user terminals. It is one of the determining properties of the radio channel that is exploited in massive antenna systems. One of the major takeaways is that the spatial correlation that results from the distributed deployment of APs may have a detrimental impact on the favorable propagation. More specifically, users that are relatively close to each other will incur high spatial correlation, which will jeopardize the mutual orthogonality of the users' channel [CB17].

In our work, we analyze how spatial correlation between users' channels vector influences favorable propagation, in CI mMIMO systems. We then explore how favorable propagation can be improved through resource allocation by taking into account solely the large-scale fading and the number of available APs. We establish a design optimization problem based on the perspective of user's scheduling. The proposed design advocates how spatial user grouping can improve favorable propagation. However, the resulting optimization problem is difficult to solve in general. We demonstrate that the formulated problem is NP-hard and we invoke semidefinite relaxation (SDR) approach to design a polynomial time solvable randomized procedure to find a sub-optimal solution to the NP-hard problem. In addition to that, to increase users' throughput, we investigate the problem of bandwidth allocation for the resulting scheduling design strategy. More details on this work can be found in the Appendix 8.10.

We explore the impact of the proposed design optimization approach on the performance of Cell-Free Massive MIMO. Two metrics are considered, namely, Large-scale fading correlation and average achievable downlink throughput. We consider a circular region having an area of 1 Km^2 where M APs and K users are randomly located according to a uniform distribution. We consider C spatial groups with a maximum of τ users per group.

The large-scale fading coefficients include the impact of the path loss, which is computed using a three-slope path loss mode [NAY+15], and lognormal shadow fading with standard deviation σ_{sh} . All other simulation parameters are summarized in the following table:

Table 4-1 Simulation parameters

Carrier frequency	1.9 GHz	Bandwidth	20 MHz
Noise figure	9 dB	AP antenna height	15 m
σ_{sh}	8 dB	User antenna height	1.7 m
Coherence slot	200 samples	Downlink power, pilot power	200, 100 mW

We present the performance of the proposed resource allocation approach and compare it with a conventional CI mMIMO system. The conventional system allows for non-orthogonal access for all active users to the entire system bandwidth.

Figure 4-5 shows the impact of the proposed spatial grouping (SG) on the normalized large-scale fading correlation for different values of M . To do so, we investigate the CDFs of the normalized large-scale fading correlation for $K = 20$ and $C = 4$, averaged over all users. The latter is given by

$$E_{k,j}(\mathbf{P}) \left\{ \frac{\sum_{m=1}^M \beta_{mk} \beta_{mj}}{\sqrt{\sum_{m=1}^M \beta_{mk}^2} \sqrt{\sum_{m=1}^M \beta_{mj}^2}} < l \right\}, \text{ for all } k, j$$

We can see that appropriate spatial user grouping substantially reduces spatial correlation within each group, which results in more favorable propagation environment (more orthogonal user channels). From Figure 4-5, we observe that SG enables to achieve, for $M = 100$, spatial correlation levels comparable to conventional CF mMIMO with $M = 200$ [NAY+17]. This means that SG can be a very practical alternative to network densification. This means the performance of CF mMIMO can be improved without requiring the deployment of more APs, which further improves the cost efficiency of CI mMIMO systems, and CRAN in general.

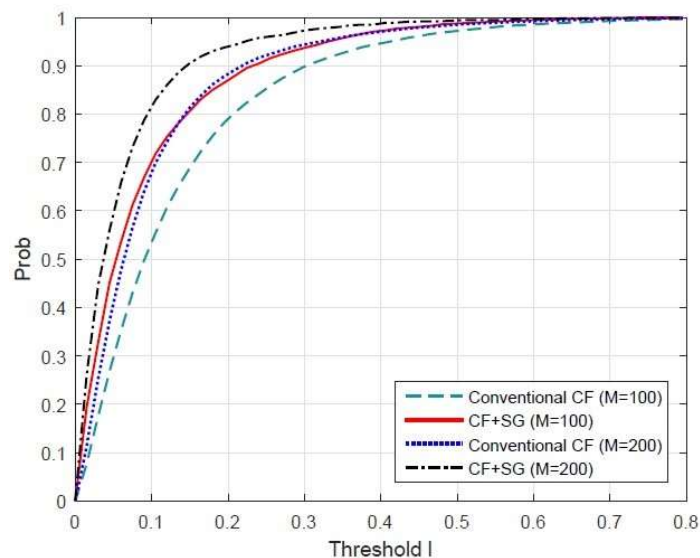


Figure 4-5 Comparison of CDFs of normalized large-scale correlation for $K = 20$, $C = 4$

Figure 4-6 exemplifies the average downlink net throughput versus the number of APs (M) for K equal to 10 and 15, respectively. Figure 4-6 demonstrates that the performance of both approaches improves as M increases. This is a direct consequence of array gain increase. In addition, Figure 4-6 also shows that the proposed SG with bandwidth allocation outperforms the conventional CI system. This is due to the fact that SG improves favorable propagation within each group. Indeed, SG enables gains of 18.5 % and 17.05 %, for $(M = 100, K = 10)$ and $(M = 100, K = 15)$, respectively. Consequently, appropriate user grouping and selection can be a more practical and cost-efficient alternative to increasing the number of APs. Indeed, for $K = 15$ SG achieves approximately the same downlink throughput (14.956 *Mbits/s*) at $M = 85$ as the conventional CF system with $M = 100$. Again, this means that improving the performance of CI mMIMO and CRAN can be achieved through architecture optimization without requiring a costly deployment of more APs.

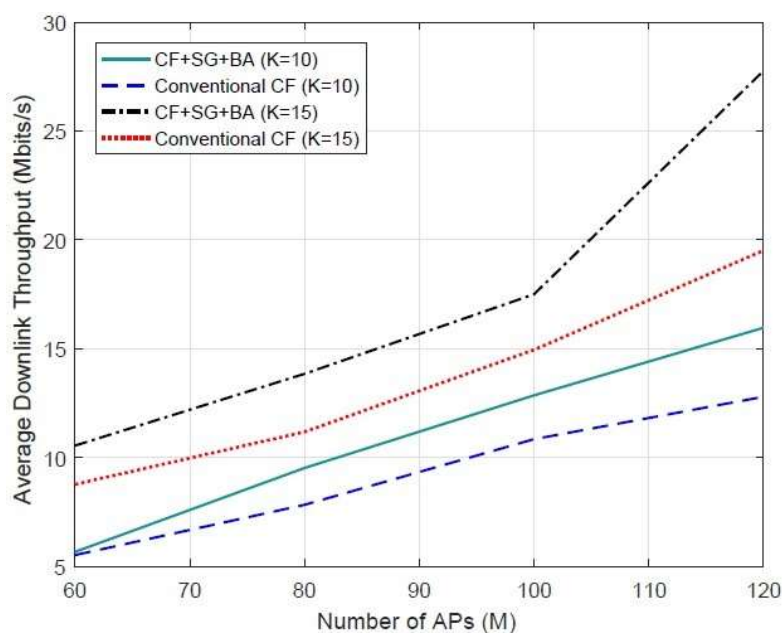


Figure 4-6 Average Downlink Throughput versus the number of APs and different K values

4.2.2 Optimised functionality placement and resource allocation in a CRAN/DRAN context

Motivation and problem description

As a part of a study item for New Radio (NR), 3GPP started studying different functional splits between central and distributed units. They have proposed about 8 possible options [5GPPP-ARCH] shown in Figure 4-7 and also discussed in Sec. 2 of [ONE17-D31]. Based on this functional split, we study the functions of the LTE protocol stack, which can be partitioned in distinct elements and assigned to different network units.

We consider a network architecture that contains a Central Unit (CU) and several Distributed Units (DUs) (explained also in [ONE17-D31]), integrating the DU and RRH in the same node. A CU (e.g. BBU) is a logical node that includes the gNB functions, except those functions allocated exclusively to the DU. It controls the operation of DUs over front-haul interface. A DU (also referred to as RRH) is a logical node that includes a subset of the gNB functions, depending on the functional split option. Its operation is controlled by the CU.

In contrast to the fixed functional split provided by the CRAN and DRAN architectures, a flexible split is proposed in order to choose an optimal operating point between full centralization and local execution, adapting to network characteristics as well as current service requirements. This approach is relevant mainly for the Underserved Areas scenario and “Long range connectivity in remote areas with smart farming application” and “Smart grid, connected lighting and energy infrastructure” use cases from WP2 and can also be employed in the context of network slicing.

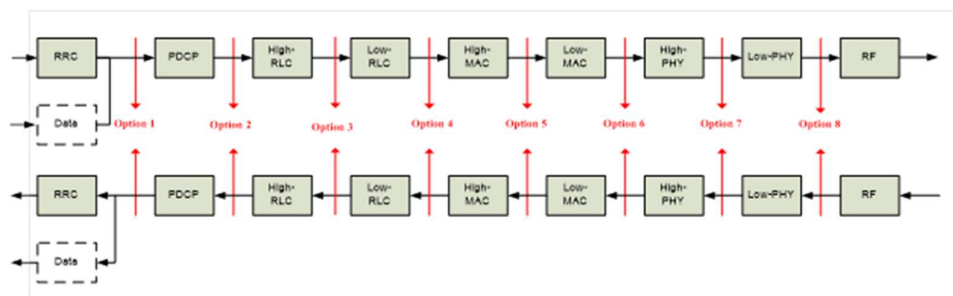


Figure 4-7 Function Split between central and distributed unit [3GPP-38801]

Problem statement

We consider a functional graph $G(F, K)$, where each node corresponds to a functional entity (FE) and each edge connects interacting FEs, and a system layout graph $G(S, L)$ consisting of the available server entities (SEs) and the communicational channels (wired or wireless backhaul links) among them (detailed description can be found in Appendix 8.11). Our objective is to assign the FEs to SEs.

The factors contributing to cost are:

- The number of SEs that will need to be activated, assuming the cost (related mostly to energy consumption) associated with the activation of an SE to be given by $B = \{b_j \mid \forall j \in S\}$
- The cost (computational latency) of running an FE on a specific SE, depending on the computational intensity of the FE and the computational power of the SE
- The cost (data latency) imposed by the communication among FEs that are running on different SEs, calculated by the amount of data to be transferred and the associated data rate
- The traffic rate served by each DU (RRH), acting as a factor to its performance metrics.

The constraints of our problem address the following aspects:

- Each FE should be allocated to a single SE: $A_j \cap A_{j'} = \emptyset, j, j' \in S$
- The capacity constraints of each SE should be respected.
- The capacity constraints of each communicational link should be respected.
- The required QoS is provided by every DU (RRH) to its associated traffic. For this purpose, we consider the chain as a queuing system, assume Poisson traffic arrival and constrain the traffic intensity.

Solution approaches

An optimal solution to the previously discussed optimization problem can be based on Mixed Integer Programming. However, the inherent computational intensity of the algorithm is prohibitive. In order to provide a solution in a quick and efficient manner, a simulated annealing algorithm is developed. In the next steps, heuristic algorithms, including consolidation and load balancing approaches will be evaluated.

The cost function used to evaluate each solution is given below. More details about the function and constraints, as well as the Simulated Annealing algorithm, can be found in Appendix 8.11.

$$f(A, B, P, N) = w_1 * \sum_{j \in S} b_j y_j + w_2 * \sum_{\forall i \in F, \forall j \in S} [x_{ij} * p_{ij}] + w_3 * \sum_{\forall i, i' \in F} [(1 - z_{ii'}) * n_{ii'}]$$

Results

For our tests we assume a CU and 10 DUs (a subset of which are enabled) serving stable total network traffic (traffic load is represented by a relative quantity and Poisson traffic arrival is assumed), while QoS requirements are imposed by constraining the traffic intensity of each queuing system (with an upper bound varying between 0.85 and 0.95). In Figure 4-8 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.

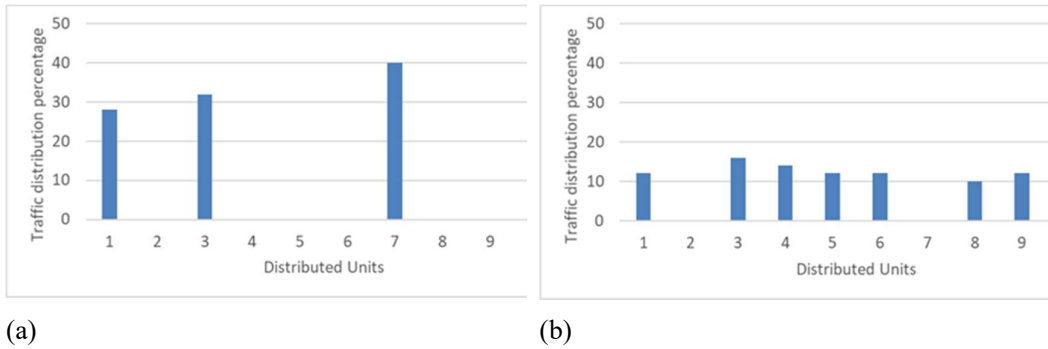
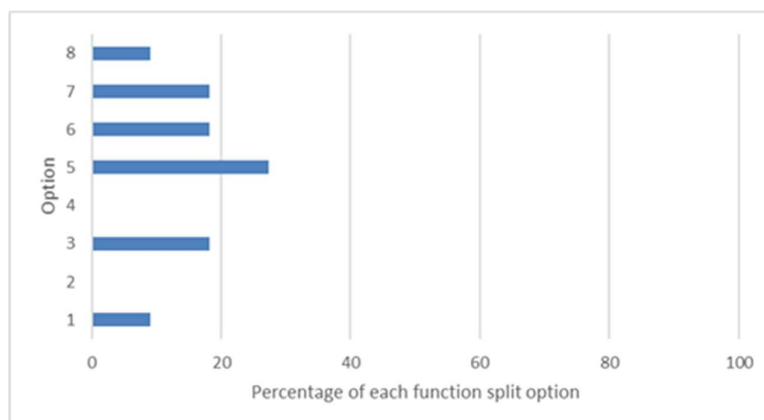


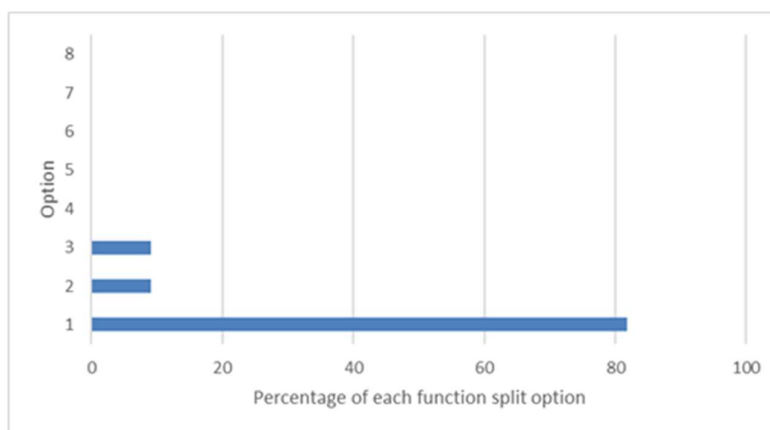
Figure 4-8 Traffic distribution percentage to distributed units for different QoS requirements

Using the function split options 1-8 as shown in Figure 4-7, with Option 1 and 8 being the most decentralized and centralized alternatives respectively, assuming 40 DUs and altering the total network traffic, we get Figure 4-9 showing the percentage of active DU branches that use each functional split option in the final solution to which the algorithm converges. In (a), a lot of the DUs select a more centralized option, keeping the cost function to a lower level. An increase in total traffic leads to (b), where the network utilizes a more distributed approach in order to provide

an acceptable QoS. Further results, mainly regarding the Underserved Areas scenario are expected.



(a)



(b)

Figure 4-9 Percentage of each function split option

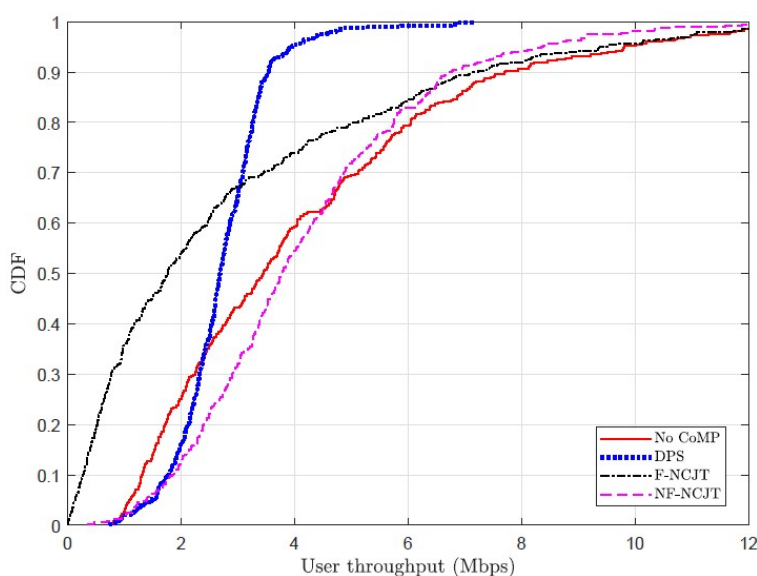
4.2.3 Centralized and distributed multi-node schedulers for non-coherent joint transmission

In the ‘Outdoor hotspots and smart offices with AR/VR and media applications’ use case in WP2, it is required that robust and high data rate connections are established for UEs. Both the CRAN and distributed RAN (DRAN) are the main network architectures to support these applications via different network coordination schemes. In 3GPP, several downlink coordinated multipoint (CoMP) schemes such as joint transmission (JT) and dynamic point selection (DPS) were investigated and evaluated. The JT category can be further divided into two groups, i.e., coherent JT (CJT) and non-coherent JT (NCJT). CJT performs joint beamforming from all coordinated TRPs, which can be regarded as a distributed multiple-input multiple-output (MIMO) system. Contrarily, NCJT allows coordinated TRPs to transmit independent layers to the target UE. NCJT is often further categorized into fully overlapped NCJT (F-NCJT) and non-fully overlapped NCJT (NF-NCJT), respectively. F-NCJT requires that resources are allocated on equivalent physical resource blocks (PRBs) on each coordinated TRP, whereas NF-NCJT sees further decoupling of the TRPs, allowing for flexible allocation across the available PRBs of each. A comparison between different network coordination schemes is shown in Table 4-2.

Table 4-2 Comparison between different network coordination schemes.

	CJT	DPS	F-NCJT	NF-NCJT
Centralized scheduler	yes	yes	yes	no
Joint precoder	yes	no	no	no
Fully overlapped PRB	yes	no	yes	no
User data sharing	yes	yes	yes	no

System level simulations have been performed to evaluate DPS/F-NCJT/NF-NCJT in an indoor scenario. Detailed descriptions and simulation settings can be found in Appendix 8.12. It can be observed in Figure 4-10 that F-NCJT has the worst cell-edge and median user throughputs due to limited scheduling options. DPS has the best performance in cell edge (5% user throughput) because of minimum inter-cell interference. It also has the lowest 95% user throughput and the median user throughput. NF-NCJT improves the median user throughput, although it has a worse cell-edge user throughput than DPS and a worse cell center user throughput than the standard scheme where CoMP is not used.

**Figure 4-10 CDFs of user throughputs of different network coordination schemes (5 users per TRP)**

These results imply that, in order to maximize the throughputs of different users, dynamic switching between different network coordination schemes should be explored. In this regard, efficient signalling should be used to support switching. Also, the performance of CJT can be included as well. These will be further investigated during the project.

4.2.4 NR duplexing with CRAN and network coordination

Cloud radio access network (CRAN) forms a network by centralizing baseband units (BBUs) while distributing radio remote heads (RRHs) across the whole network. This network architecture enables transception points (TRPs) to exchange information in a more efficient manner and thus is able to provide much tighter network coordination. In addition to the technique introduced in Section 4.2.3 where TRPs are coordinating in the spatial and frequency domains, duplexing (dynamic TDD) can also utilize tight coordination between TRPs in the time domain to provide more flexibility in the ‘Outdoor hotspots and smart offices with AR/VR and media applications’ use case. However, duplexing introduces a limiting factor known as cross link interference (CLI) which occurs when two TRPs are operating in different subframe directions, i.e., one is uplink and the other is downlink. As a result, the downlink TRP will interfere the uplink signal in the neighbour TRP (TRP-to-TRP CLI) and the uplink UE will interfere the downlink signal of the UE in the neighbour cell (UE-to-UE CLI). As the channels between TRPs

are much more stable than channels between UEs, the UE-to-UE CLI is more difficult to handle and thus UE-to-UE CLI management is required.

A UE-to-UE CLI management procedure is proposed here in ONE5G that CLI is handled within the NR network coordination framework. The uplink UE is regarded as another interfering TRP from the downlink UE’s perspective. Then, the uplink TRP provides demodulation RS (DMRS) and scheduling information of its own cell to the downlink TRP. As a result, the downlink TRP can assign zero-power CSI RS (ZP CSI-RS) overlapping DMRS locations of the interfering UE to the downlink UE, such that the downlink UE can perform channel measurement on the interfering channel and perform advanced interference rejection algorithms to suppress UE-to-UE CLI. Figure 4-11 shows the uplink DMRS locations of the interfering UE and the assigned ZP CSI-RS to the desired UE. It can be observed that the ZP CSI-RS overlaps with the uplink DMRS. The detailed procedure of this UE-to-UE CLI management can be found in Appendix 8.13.

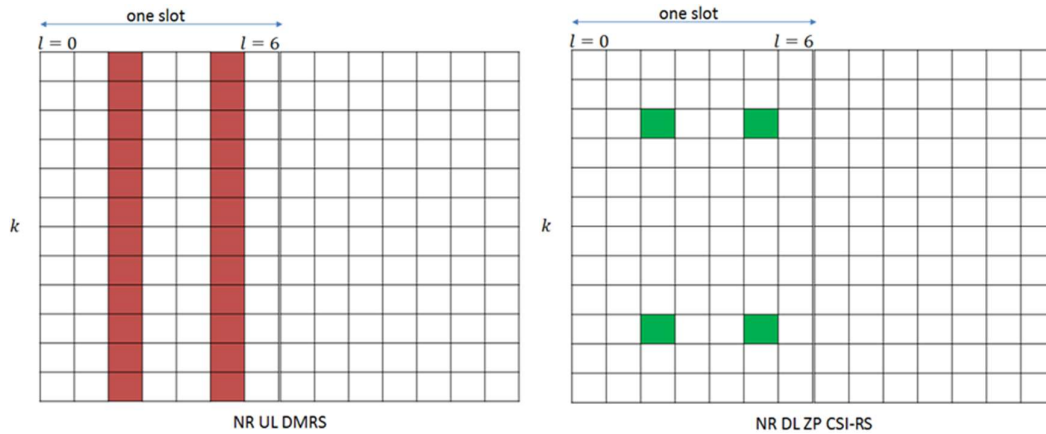


Figure 4-11 UL DMRS locations of the interfering UE (left) and DL ZP CSI-RS (right) for the desired UE

Figure 4-12 illustrates the impact of the proposed CLI management method on both downlink and uplink user throughputs in an indoor environment. Simulation settings can be found in Section 8.13. It can be seen that with CLI management, the downlink throughput improves by 30%. This gain is due to the fact that the target UE has certain knowledge of the interfering channel and therefore can suppress interference. On the contrary, the gain of CLI management is negligible in the uplink because TRP-to-TRP is less significant.

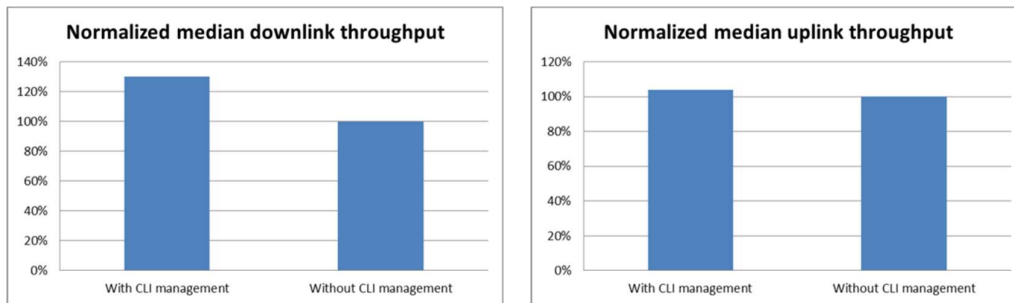


Figure 4-12 Median DL throughput comparison (left) and median UL throughput comparison (right)

For future work, in addition to the time domain coordination, a joint space-time-frequency coordination for a CRAN will be investigated.

4.2.5 Centralized scheduling on the uplink of the Multiple Access Multiple Relay Channel (MAMRC)

Centralized scheduling strategies for cooperative incremental redundancy retransmissions in the slow-fading Multiple Access Multiple Relay Channel (MAMRC) are investigated.

$(M,L,1)$ -MAMRC is considered, where $M \geq 2$ independent users (sources) communicate with a single destination using a help of L half-duplex dedicated relays, who apply Selective Decode-and-Forward (SDF) protocol (Figure 4-13).

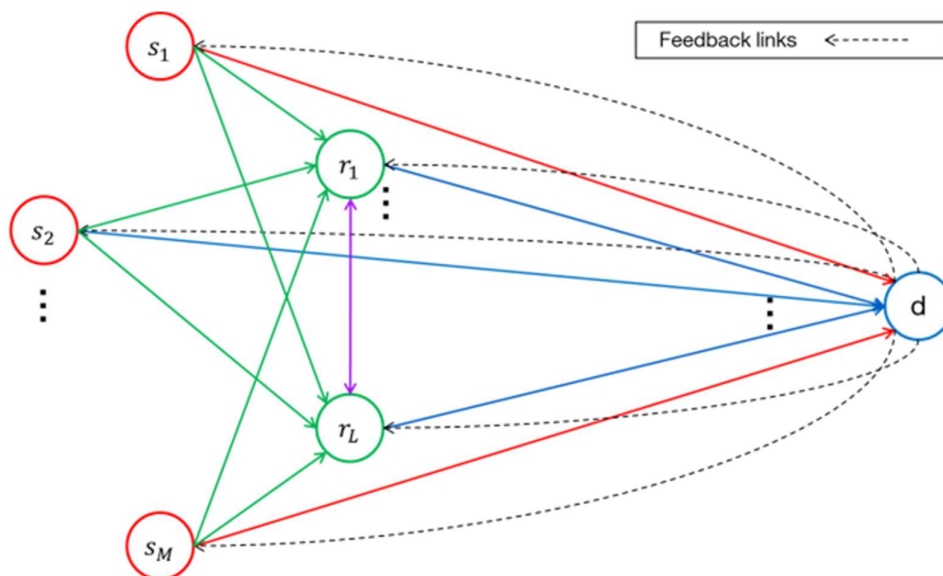


Figure 4-13 Orthogonal Multiple Access Multiple Relay Channel (OMAMRC) with feedback.

Multiple Access is orthogonal in time (OMAMRC), where transmissions occur in consecutive time-slots. A Cyclic Redundancy Check (CRC) is appended to each source message to allow relays and the destination to detect any decoding error. Sources transmit successively for the first phase. The second phase consists of a limited number of time slots for retransmissions. In each time slot of the second phase, the destination schedules a node (being a relay or a source) to retransmit, conditional on the knowledge of the correctly decoded source sets of each node (which is itself for a source) and a partial knowledge of the global CSI (CSI is available only at the receiver of each link). Scheduling decision is conveyed over perfect limited feedback broadcast control channel, while the information about decoded source sets of relays are conveyed over forward coordination control channels. A scheduled relay performs a cooperative retransmission using Incremental Redundancy (IR) Hybrid Automatic Repeat Request (HARQ) mechanism on its correctly decoded source messages. Joint Network-Channel Coding (JNCC) framework is used.

The goal is to maximize the long-term aggregate throughput by applying the proper centralized scheduling strategy of the sources, under the fairness constraint that the node selection should not depend on the initial rates of the sources. So, this work is suitable for eMBB services. Three different node selection strategies have been proposed [CVM+18]:

1. *Strategy 1*: by exhaustively going through all different node selection alternatives, the destination chooses the node which brings the highest number of newly decoded sources (equivalent to the maximization of the cardinality of the set of the successfully decoded sources of the destination). If there are multiple nodes that bring the same number of decoded sources at the destination, the selected node is the one having the highest mutual information between itself and the destination.

2. *Strategy 2*: the selected node is the one having the highest mutual information between itself and the destination, where all the nodes that were able to decode at least one source from the set of unsuccessfully decoded sources at the destination are candidates.
3. *Strategy 3*: a node with the highest product of the mutual information between itself and the destination and the cardinality of its set of successfully decoded sources is selected.

The performance of three proposed selection strategies is evaluated in terms of long-term aggregate throughput by using Monte-Carlo simulations. Two different selection strategies are used as a benchmark. The first one, referred to as “Reference 1” in the figure legends, is the strategy 1 from [MVB16], which is based on the minimization of the probability of the common outage event after each round. The other one, referred to as “Upper-bound” in the figure legends, is based on the exhaustive search approach for the optimal activation sequence of nodes with respect to normalized long-term aggregate throughput. The latter approach has no interest in practice, since the required knowledge of the CSI of all links in the network would incur extremely large feedback overhead.

Figure 4-14 shows the long-term aggregate throughput as a function of the average SNR of each link in the network (symmetric link scenario). The symmetric initial rates are chosen from a discrete Modulation and coding scheme (MCS) family whose rates belong to $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5\}$ [bits per channel use], such that they maximize the long-term aggregate throughput with respect to the average SNR (slow link adaptation). Here, the slow link adaptation is very simple since the rates of the sources and the SNR are equal. The slow link adaptation for “Reference 1” strategy is illustrated by the black dotted line that corresponds each to a given initial rate (the same for each source) ranging from 0.5 to 3.5 [bits per channel use]. It simply takes the envelope of these curves. The gain of the proposed strategies (within the framework of slow link adaptation) compared with “Reference 1” is approximately 1 dB.

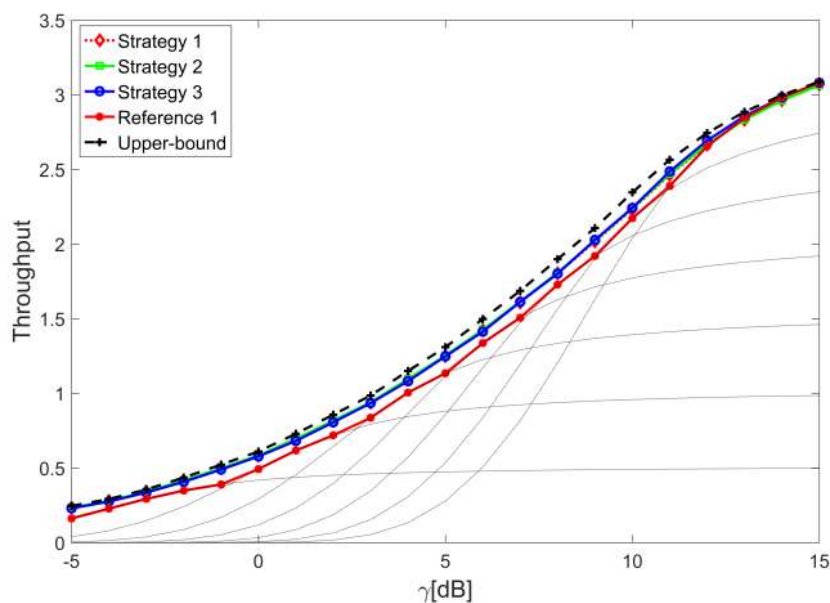


Figure 4-14 Long-term aggregate throughput of different strategies with slow link adaptation and symmetric link scenario.

As the next step, efficient slow-link adaptation algorithms will be investigated in both symmetric and asymmetric link scenarios with respect to different quality of services.

5 Use cases, system-level evaluation and proof of concept

5.1 Proof of Concept (PoC)

Seven of the most promising ONE5G technical components (TeCs) described in the previous sections have been selected as candidates for PoC implementation (see also the WP5 internal report [ONE5G-I51]). These TeCs capture advances from the three major research topics considered in WP4, namely, (a) future proof multi-service access, (b) massive MIMO enablers, and (c) advanced link coordination in CRAN/DRAN. The following table provides a summary description of the selected TeCs.

Table 5-1. Candidate technologies for implementation in WP5

Topic	TeC	High-level description	Details	WP5 PoC
URLLC	Power control optimized settings for grant-free URLLC	User Scheduling and RRM for URLLC services	Sec. 2.2.1	PoC#1
Beamforming	decentralized beamforming algorithms for sum rate maximization via large system analysis.	Optimal precoder design taking into account imperfect CSI	Sec. 3.3.4	PoC#2
Beamforming	Sector and beam management with cylindrical antennas	Investigate cylindrical 3D antenna arrays for improved coverage and capacity	Sec. 3.1.2	PoC#3
Flexible HW implementation	Flexible and fast reconfigurable HW architecture for multi-service transmission	FEC and Digital Front-end architecture design and efficient implementation	Sec. 3.1.3	PoC#2, PoC#4
Network slicing	Optimised functionality placement and resource allocation in CRAN/DRAN context	Optimal allocation of functional to physical network entities for diverse multi-service support by network slicing	Sec. 4.2.2	PoC#2
Signalling overhead reduction in CRAN	Compressive channel estimation in CRAN	Minimizing training overhead required to obtain global CSI in a massive CRAN setting	Sec. 4.1.3	PoC#1
Multi-cell interference coordination/cancellation	Nonlinear mechanisms in cell-less systems	Design of non-linear and low-complexity detection filters for reliable detection of UE signals by combining soft information from multiple RRHs.	Sec. 4.1.4	PoC#3

5.2 Connection to WP2 use cases

Table 5-2 provides an overview on how the WP4 technologies contribute to the KPIs and use cases defined by WP2, as described in Deliverable D2.1 [ONE17-D21].

Table 5-2. Overview on WP4 technologies and their contribution to WP2 use cases

WP4 Technology	Primary KPIs	Contribution	WP2 use cases

Grant Free Access	cost/energy efficient mMTC	Reduces signalling load, and hence can be more energy efficient	3,4,7,8
	latency	Reduces latency as UEs do not need to wait for access grant	1,2
NoMA	cell throughput	Increased capacity region, more efficient usage of multi-user channels	1,6
	connection density (mMTC)	Enables "overloading" due to overlapping radio resources in the time and frequency domain	1,6
HARQ	Latency, signalling overhead	Low-complexity protocols	1,2,4
Beamforming	Cost and energy efficiency	Hybrid designs	1,3,4,5,6
	coverage	Array gain for backhauling solutions	4,9
	throughput	Beam management, integrated multicast/unicast, multicell coordination	1,3,5,6
grouping and scheduling for CSI acquisition	robustness	Avoiding pilot interference. Better support of massive MIMO techniques	1,3,5,6
Enhanced modulation and coding	coverage	~1dB SNR improvement by probabilistically shaped coded modulation (PSCM) for the backhaul	4,9
CSI estimation and feedback	Reduced complexity	Exploiting channel properties	1,3,4,5,6
	Robustness and impairments	Reduced training overhead and CSI feedback designs (exploiting structural channel properties, e.g., sparsity) for TDD and FDD	1,3,4,5,6
CRAN, network MIMO	Maximum/average number of active UEs, area traffic capacity, experienced UE data rates	Efficient global CSI acquisition with small signalling overhead, non-linear filtering for joint processing of UE transmissions from multiple APs	3,5,6
Impact of array formats	Throughput	Better matches the array geometry to given user distributions, thus improved array gains	1,3,4,5,6
Inter-numerology interference mitigation	Future-proof design	Slicing support, coexistence of services	1,2,3,4,7,8,9

Reconfigurable baseband processing	Future-proof design	Slicing support, coexistence of services	1,2,3,4,7,8,9
Resource allocation and traffic management in CRAN/DRAN	Maximum/average number of active UEs, area traffic capacity, experienced UE data rates	Scheduling solutions suitable for both current 3GPP specs as well as future, sophisticated CRANs with network slicing support	3,5,6

5.3 Contribution to ONE5G scenarios

The ONE5G project aims at optimizing the KPI mainly defined for the "Underserved Areas" and "Megacities" scenarios, which are described in Deliverable D2.1 [ONE17-D21]. The technical solutions provided by WP4 contribute to the scenarios 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. This is included in Objective 5 of the project proposal ("...to propose adaptations to allow sustainable provision of wireless services in underserved areas under constrained circumstances"). 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 shaping, which adds additional SNR gain. Furthermore, non-orthogonal multiple access (NoMA) is a promising technology for massive 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 is developing 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, CSI acquisition, and hardware optimization. For massive access and URLLC, we obtain gains from NoMA and enhanced HARQ. These two 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.

5.4 Link-to-System Model

The evaluation of large networks, many devices, or a long time, is hardly feasible for MIMO technologies. On the other hand, massive MIMO is a key-technology for 5G and evaluation on system or network level is essential. Therefore, an abstraction between the link to system level is necessary to be able to speed up the simulations without simulate all the process at the PHY layer. This abstraction consists in defining metrics that allow simulating the whole system by doing some approximations.

In [WCT+14] a PHY-layer abstraction model is proposed and adapted to fit in the 5G system-level platform. The information flow chart and principle components of the PHY-layer abstraction model are shown in Figure 5-1 with a numerical example. First, the PHY layer MIMO simulation

is performed for technology component A with parameter configuration according to the use case or scenario requirements. From these simulations, the following two outputs are required for the system level abstraction:

1. The number of spatially multiplexed users per time-frequency resource
2. The achieved user spectral efficiency over the geometry or SNR

In Figure 5-1, the cumulative distribution function (CDF) of the number of multiplexed user is given for output one. The user spectral efficiency over the geometry is given for output two. Note, that as an alternative to the geometry also the SNR can be used, e.g. if inter-cell interference in the PHY layer simulation is not explicitly simulated and considered as noise. The geometry here means the un-precoded sum power ratio of the serving BS antennas over all other (interfering) BS antennas for a user. The geometry can be interpreted as a wideband “unprecoded” SINR, see Figure 5-1.

In the second step, the system level or network layer simulation is performed assuming MIMO technology component A . Therein, (multiple) users on a time-frequency resource have to be selected according to the mapping table or curve from PHY layer simulation, where simplest case is to use always the median value. Note that the active user selection in the system level simulation may depend on traffic and mobility models. After the user selection, the SNR or geometry of these users is determined and used as input to the mapping from SNR or geometry to user spectral efficiency. Note, that the user spectral efficiency from the PHY layer MIMO simulation can also include the case that users have multiple antennas and maybe receive multiple data streams. With spectral efficiency of the active users, the whole spectral efficiency for the given time-frequency resource is obtained in the system level simulation for the given MIMO technology component A . Thus, different technology components can be evaluated and compared in the same simulation environment, however it maybe not possible to combine different technology components.

Note, that the SINR values from the mapping table can directly be used as input to Mutual Information based exponential SNR Mapping (MIESM) and/or Mean Mutual Information per Bit (MMIB) models as commonly used in 3GPP NR simulations.

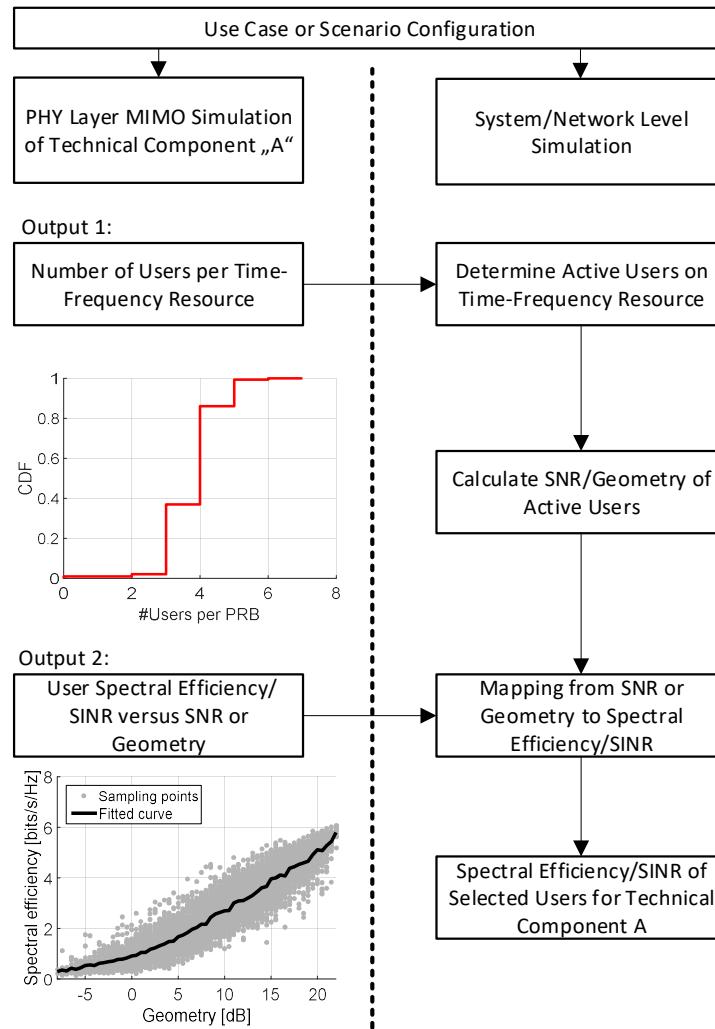


Figure 5-1 Spectral Efficiency of the proposed clustering scheme

6 Conclusions and future work

Report D4.1 presents intermediate results of WP4. The research topics and solutions are closely aligned with the WP2 use cases, and with upcoming 3GPP WIs/SIs, which aim to provide massive access with improved link quality to various service types.

Non-orthogonal multiple access (NOMA) is a promising technique to enable efficient massive access. We have proposed to incorporate the use of grant-free uplink transmission and NOMA for massive machine-type communications (mMTC) uplink. Also, improved design of user signatures and NOMA with spreading and interleaving will be included in the general WP4 NOMA framework. In this report, preliminary results are presented, which will be further refined within the project time frame. Another WP4 topic is related to ultra-reliable low-latency communications (URLLC). WP4 has developed means to achieve an optimal balance between reliability and latency by using grant-free transmissions for short packets.

Massive multiple-input multiple-output (mMIMO) will be used in 5G to boost throughput and improve link qualities. We are investigating practical massive MIMO array design issues. It can be observed that array layouts impact the performance in different deployment scenarios. Although planar arrays are the main focus in 3GPP at the moment, we have investigated alternative arrangements of planar arrays to provide optimal performance. Also, we have found that a complete uniform circular array (UCA) layout can provide promising gain. A technology related to a vertical use case is massive beamforming for multicast (groupcast), based on channel state information (CSI) at the transmitter. However, acquisition of CSI in massive MIMO can introduce a large overhead and induce high complexity. We have proposed different schemes to reduce the overhead such as opportunistic feedback and compressed explicit feedback. Complexity of massive MIMO can be reduced by efficient detection with dichotomous coordinate descent (DCD) and by applying hybrid beamforming structures with one-bit phase resolution analog precoders. We have also demonstrated how limited-capacity fronthaul/backhaul suppresses the performance gains of massive MIMO. In this case, hybrid beamforming can be used to reduce the required fronthaul load and probabilistically shaped coded modulation (PSCM) can be used to improve backhaul capability.

To further leverage the gain from massive MIMO, coordinated transmissions are studied. Both centralized radio access network (CRAN) and distributed RAN (DRAN) are essential deployment variants for 5G. We focus on optimizing physical layer procedures and resource allocation techniques for CRAN/DRAN. We explore a CSI framework for cooperative transmission and efficient signalling schemes. By utilizing large-scale sparsity, CSI overhead can be reduced using compressed sensing (CS). Also, cell-less network structures are beneficial for cell edge users in the case of CRAN. Regarding resource allocations, we have shown that different network coordination schemes have their respective strengths to handle different scenarios. In a network as complex as 5G, it is important to flexibly switch between different network coordination schemes. We have also discovered that allowing terminals to exchange CSI via sidelinks can improve the performance of the system within a CRAN framework.

For each group of technologies, their requirements and applicability with respect to the upcoming 3GPP NR standard has been assessed. NOMA and URLLC solutions are highly important topics and subject to current discussions in 3GPP. The solutions for flexible hardware architectures, CSI acquisition and hybrid array architectures are features which partially target efficient operation within the constraints of 3GPP. Some of them also go far beyond current 3GPP concepts with high potential for future standard extensions.

For some of the presented technologies, a proof of concept will be demonstrated in WP5. Further, the connection to the use cases of WP2 and the specific contribution to the ONE5G scenarios “Underserved Areas” and “Megacities” have been addressed, including the specific use case categories eMBB, long-range and cost/energy efficient mMTC, and reliability and low latency

communications. Although this report focuses on “preliminary results”, we already can state that the various presented technologies can well address the objectives and goals of ONE5G.

7 References

- [3GPP-36212] 3GPP T.S. 36.212, “Multiplexing and channel coding,” V12.6.0, Sep. 2015.
- [3GPP-36814] 3GPP. Further advancements for E-UTRA physical layer aspects, V9.2.0, March 2017
- [3GPP-36873] 3GPP T.R. 36.873, “Study on 3D channel model for LTE,” V12.2.0, Jun. 2015.
- [3GPP36897] 3GPP T.R. 36.897, “Study on Elevation Beamforming/Full-Dimension (FD) MIMO for LTE,” V13.0.0, Jun. 2015.
- [3GPP-38211] 3GPP T.S. 38.211, “Physical channels and modulation,” V1.0.0, Sep. 2017.
- [3GPP-38211] 3GPP TS38.211 NR; Physical channels and modulation
- [3GPP-38214] 3GPP T.S. 38.214, “Physical layer procedures for data,” V15.0.0, Dec. 2017.
- [3GPP-38801] 3GPP T.R. 38.801, “Study on new radio access technology: Radio access architecture and interfaces,” V14.0.0, Mar. 2017.
- [3GPP-38802] 3GPP T.R. 38.802, “Study on New Radio Access Technology–Physical Layer Aspects,” V14.2.0, Sep. 2017.
- [3GPP-38901] 3GPP T.R. 38.901, “Study on channel model for frequencies from 0.5 to 100 GHz,” V14.0.0, Mar. 2017.
- [3GPP-38913] 3GPP T.R. 38.913, “Study on Scenarios and Requirements for Next Generation Access Technologies,” V14.2.0, Mar. 2017
- [3GPP-AH_NR2] Final Report of 3GPP TSG RAN WG1 #AH_NR2 v1.0.0, Qingdao, China, June, 2017
- [5GPPP-ARCH] 5GPPP, View on 5G Architecture, July 2016, available at <https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-5G-Architecture-WP-July-2016.pdf>
- [AAL+14] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, “Channel estimation and hybrid precoding for millimeter wave cellular systems,” *IEEE J. Sel. Topics Signal Process.*, vol. 8, pp. 831–846, Oct. 2014.
- [ACSMI17] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, A. Iera. “Multicasting over Emerging 5G Networks: Challenges and Perspectives”, *IEEE Network*, vol.: 31, no. 2, pp. 80-89. Mar. 2017.
- [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
- [AJB+18] R. Abreu, T. Jacobsen, G. Berardinelli, K. Pedersen, I. Z. Kovács, and P. Mogensen, "Power Control Optimization for Uplink Grant Free URLLC", Accepted in 2018 IEEE WCNC, Barcelona, Apr. 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
- [BCS06] S. Brandes, I. Cosovic, and M. Schnell, “Reduction of out-of-band radiation in OFDM systems by insertion of cancellation carriers,” *IEEE Communications Letters*, vol. 10, no. 6, pp. 420–422, 2006
- [BHR+15] C. N. Barati et al., "Directional Cell Discovery in Millimeter Wave Cellular Networks," in *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 6664-6678, Dec. 2015. doi: 10.1109/TWC.2015.2457921

- [BHS+10] W. U. Bajwa, J. Haupt, A. M. Sayeed and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proceedings of the IEEE*, 98 (6), pp. 1058-1076, 2010
- [BHS17] E. Björnson, J. Hoydis, L. Sanguinetti, "Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency", *Foundations and Trends® in Signal Processing*, 2017
- [BJ12] A. Joseph and A. R. Barron, "Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 2541–2557, May 2012.
- [BLM16] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten Myths and One Critical Question," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 114–123, Feb. 2016
- [BR13] Georg Böcherer, Rana Ali Amjad, "Block-to-Block Distribution Matching", arxiv.org/abs/1302.1020
- [BSS15] G. Bocherer, 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.
- [BX17] S. Bazzi and W. Xu, "Downlink training sequence design for FDD multiuser massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 18, pp. 4732-4744, Sep. 2017.
- [BX18] 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.
- [C18] Chen, Yejian; "Exploiting Gaussian Approximation for Non-Orthogonal Coded Access," accepted by 2018 IEEE 87th Veh. Technol. Conf. (VTC'18 Spring), Porto, Portugal, June 2018.
- [CB17] Z. Chen and E. Bjoernson, "Can We Rely on Channel Hardening in Cell-Free Massive MIMO?" in 2017 IEEE Globecom Workshops, Dec 2017, pp. 1–6.
- [CF14] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Comm. Pur. Appl. Math.*, vol. 67, no. 6, pp. 906–956, 2014.
- [CM-WP14] China Mobile White Paper ver. 3.0, "C-RAN: The road towards green RAN," Jun. 2014.
- [CRM14] G. Coluccia, A. Roumy, and E. Magli, "Exact performance analysis of the oracle receiver for compressed sensing reconstruction," In *Acoustics, Speech and Signal Processing (ICASSP)*, 2014 IEEE International Conference on, May 2014.
- [CRW12] W. Chen, M. Rodrigues, and I. Wassell, "On the use of unit-norm tight frames to improve the average mse performance in compressive sensing applications," *IEEE Signal Process. Lett.*, vol. 19, no. 1, pp. 8–11, Jan. 2012.
- [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," accepted to IEEE ICC'18, Kansas City, USA, May 2018.
- [CW11] T. T. Cai and L. Wang, "Orthogonal Matching Pursuit for Sparse Signal Recovery With Noise," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4680–4688, July 2011.

- [CY16] Z. Chen and C. Yang, "Pilot decontamination in wideband massive mimo systems by exploiting channel sparsity," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5087–5100, July 2016.
- [DLC+2017] Ding, Zhiguo, Yuanwei Liu, Jinho Choi, Qi Sun, Maged ElKashlan, Chih-Lin I and H. Vincent Poor. "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks." *IEEE Communications Magazine* 55 (2017): 185-191.
- [FR13] S. Foucart and H. Rauhut, *A mathematical introduction to Compressed Sensing*. Birkhäuser, 2013
- [FRZ+17] G. Fodor et al., "An Overview of Massive MIMO Technology Components in METIS," *IEEE Communications Magazine*, vol. 55, no. 3, Mar. 2017
- [FZY16] C. Fan, Y. J. Zhang, and X. Yuan, "Dynamic nested clustering for parallel PHY-layer processing in cloud-RANs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1881–1894, Mar. 2016.
- [GRM+16] S. Gimenez, S. Roger, D. Martin-Sacristan, J. Monserrat, P. Baracca, V. Braun, H. Halbauer, "Performance of Hybrid Beamforming for mmW Multi-antenna Systems in Dense Urban Scenarios", *IEEE PIMRC*, Sept. 4-7, 2016, Valencia, Spain
- [HG17a] S. Haghighatshoar and G. Caire, "Massive MIMO channel subspace estimation from low-dimensional projections," *IEEE Trans. Signal Process.*, vol. 65, no. 2, pp. 303–318, Jan. 2017.
- [HG17b] S. Haghighatshoar and G. Caire, "Massive MIMO Pilot Decontamination and Channel Interpolation via Wideband Sparse Channel Estimation," *IEEE Trans. Wireless Commun.*, Januar 2017, submitted. [Online]. Available: arXiv preprint arXiv:1702.07207
- [HJA18] S. Hajri, Juwendo Denis and M. Assaad, "Enhancing Favorable Propagation in Cell-Free Massive MIMO Through Spatial User Grouping", submitted to SPAWC 2018.
- [HKR97] P. Hoeher, S. Kaiser and P. Robertson, "Two-dimensional pilot-symbol-aided channel estimation by Wiener filtering," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP) 1997.*, vol. 3, Munich, Bavaria, Germany, Apr. 1997, pp. 1845–1848.
- [HMS14] R. Heckel, V. I. Morgenshtern, and M. Soltanolkotabi, "Super-resolution radar," *CoRR*, vol. abs/1411.6272, 2014. [Online]. Available: <http://arxiv.org/abs/1411.6272>
- [HQC+17] Q. He, T. Q. S. Quek, Z. Chen, and S. Li, "Compressive channel estimation and multi-user detection in CRAN," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.
- [IX16] M. Ibrahim and W. Xu, "On Numerology and Capacity of the Self-Contained Frame Structure," 2016 *IEEE Globecom Workshops (GC Wkshps)*, Washington, DC, 2016, pp. 1-6.
- [IX17] Mohamed Ibrahim, Wen Xu A Frame Structure with Precoding for Bidirectional Low Latency Applications *Globecom URLLC Workshop Dec 2017*
- [JKL+16] H. Ji, Y. Kim, J. Lee, E. Onggosanusi, Y. Nam, J. Zhang, B. Lee, and B. Shim, "Overview of full-dimension MIMO in LTE-advanced pro," *IEEE Comm. Mag.*, vol. PP, no. 99, pp. 2–11, October 2016.
- [JM9] S. Jokar and V. Mehrmann, "Sparse solutions to underdetermined kronecker product systems." *Lin. Alg. Appl.*, vol. 431, no. 12, pp. 2437–2447, 2009.

- [JMC+15] Z. Jiang, A. F. Molisch, G. Caire, and Z. Niu, "Achievable rates of FDD massive MIMO systems with spatial channel correlation," *IEEE Trans. Wir. Commun.*, vol. 14, no. 5, pp. 2868-2881, May 2015.
- [KMT+18] M. Kurras et al., "On the Application of Cylindrical Arrays for Massive MIMO in Cellular Systems," in *WSA 2018; 22th International ITG Workshop on Smart Antennas*, 2018, pp. 1-7
- [LLS+14] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," *Selected Topics in Signal Processing*, *IEEE Journal of*, vol. 8, no. 5, pp. 742-758, Oct. 2014
- [Lo99] T. Lo, "Maximum ratio transmission," *IEEE Trans. Commun.*, vol. 47, no. 10, pp. 1458-1461, Oct. 1999.
- [LP17] L. Le Magoarou and S. Paquelet, Parametric channel estimation for massive MIMO, arXiv:1710.08214
- [Lue97] D. G. Luenberger, *Optimization by Vector Space Methods*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1997.
- [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.
- [MAM16-D54] MAMMOET Deliverable 5.4, "Publishable Summary", FP7-ICT-2013-11, Project reference: 619086
- [Mel99] M. Melanie, *An Introduction to Genetic Algorithms.*, 5th ed., MIT Press, Massachusetts, 1999.
- [MH15] J. Mo and R. W. Heath Jr., "Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5498-5512, Oct. 2015.
- [MHA+17] A. Maatouk, S. Hajri, M. Assaad, H. Sari, S. Sezginer, "Graph Theory Based to Users Grouping and Downlink Scheduling in FDD massive MIMO", submitted, Oct. 2017.
- [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*.
- [MRH+17] A. F. Molish, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO – a survey," *IEEE Commun. Mag.*, vol. 55, pp. 134-141, Sep. 2017.
- [MVB16] A. Mohamad, R. Visoz and A. O. Berthet, "Cooperative Incremental Redundancy Hybrid Automatic Repeat Request Strategies for Multi-Source Multi-Relay Wireless Networks," *IEEE Commun. Lett.*, vol. 20, no. 9, pp. 1808-1811, Sept. 2016.
- [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.
- [NAM+17] E. Nayebe, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and Power Optimization in Cell-Free Massive MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4445-4459, July 2017.
- [NAY+17] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-Free Massive MIMO Versus Small Cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834-1850, March 2017.

- [NGC12] F. Negro, I. Ghauri, and D.T.M. Slock, "Sum Rate Maximization in the Noisy MIMO Interfering Broadcast Channel with Partial CSIT via the Expected Weighted MSE," in Proc. IEEE ISWCS12, Paris, France, Aug. 2012.
- [NGFI] IEEE Standard Association, "NGFI - Next Generation Fronthaul Interface", Available: <https://standards.ieee.org/develop/wg/NGFI.html>
- [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.
- [ONE17-D21] ONE5G Deliverable. Scenarios, KPIs, use cases and baseline system evaluation, Dec 2017
- [ONE17-D31] ONE5G Deliverable 3.1, "Preliminary Multi-Service Performance Optimization Solutions for Improved E2E Performance," May 2018.
- [ONE17-I51] ONE5G Internal Report IR5.1, "Preliminary report on implementation and integration of PoC components" May 2018.
- [PDF17] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "NR: The New 5G Radio Access Technology," IEEE Communications Standards Magazine, vol. 1, no. 4, pp. 24–30, Dec. 2017
- [PHS05] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: channel inversion and regularization," Communications, IEEE Transactions on, vol. 53, no. 1, pp. 195–202, 2005.
- [PKC+17] J. Park, D. M. Kim, E. de Carvalho, and C. N. Manchón, "Hybrid Precoding for Massive MIMO Systems in Cloud RAN Architecture with Capacity-Limited Fronthauls," Sept. 2017. Available online: <https://arxiv.org/abs/1709.07963>
- [PLC15] E. Paolini, G. Liva and M. Chiani, "Coded Slotted ALOHA: A Graph-Based Method for Uncoordinated Multiple Access," IEEE Trans. Inf. 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.
- [PPY+17] S. Park, J. Park, A. Yazdan, and R. Heath, "Exploiting spatial channel covariance for hybrid precoding in massive mimo systems," IEEE Trans. Signal Processing, vol. 65, pp. 3818–3832, Jul. 2017.
- [PSL+15] E. Paolini, C. Stefanovic, G. Liva and P. Popovski, "Coded random access: applying codes on graphs to design random access protocols," IEEE Commun. Mag., vol. 53, no. 6, pp. 144-150, June 2015.
- [PSL+16] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," IEEE Commun. Surveys Tuts., vol. 18, no. 3, pp. 2282– 2308, Third Quarter 2016.
- [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.
- [PX17b] M. Pikus and W. Xu, "Bit-level probabilistically shaped coded modulation," IEEE Commun., Lett., vol. 21, no. 9, pp. 1929–1932, Sept. 2017.
- [R1-1700076] 3GPP R1-1700076, "Signal shaping for QAM constellations," Huawei, HiSilicon, RAN1 NR Ad Hoc, Jan. 2017.
- [R1-1715576] 3GPP R1-1715576, "Discussion on NoMA study for Rel-15 SI", Huawei, HiSilicon, RAN1#NRAH3, Sep. 2017.

- [R1-1801414] 3GPP R1-1801414, "Work Plan for Rel-15 SI on NR NOMA", ZTE, RAN1 #92, Feb. 2018.
- [R1-1802005] 3GPP R1-1802005, "Transmitter side signal processing schemes for NoMA", Samsung, RAN1 #92, Feb. 2018
- [RN18] K. Roth, J. A. Nossek, "Robust massive MIMO Equilization for mmWave systems with low resolution ADCs", IEEE Wireless Communications and Networking Conference (WCNC 2018) Workshops, Apr. 2018.
- [RPM+18] M. Roy, S. Paquelet, L. L. Magoarou, and M. Crussiere, "MIMO Channel Hardening: A Physical Model based Analysis", 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. PP, no. 99, pp. 1-1. doi: 10.1109/JSTSP.2018.2813973
- [S02] A. M. Sayeed, "Deconstructing multiantenna fading channels," IEEE Trans. Signal Process., vol. 50, no. 10, pp. 2563-2579, Oct. 2002.
- [SB16] P. Schulte and G. G. Bocherer, "Constant composition distribution matching," IEEE Trans. Inf. Theory, vol. 62, pp. 430-434, Jan. 2016
- [SDL06] N.D. Sidiropoulos, T.N. Davidson, Zhi-Quan (Tom) Luo. Transmit Beamforming for Physical-Layer Multicasting. IEEE Trans. Sig. Proc., vol. 54, no. 6, June 2006.
- [SK06] Suman, B. & Kumar, P. J Oper Res Soc (2006) 57: 1143. <https://doi.org/10.1057/palgrave.jors.2602068>
- [SMR17] S. Sun, G. R. MacCartney Jr., and T. S. Rappaport, "A Novel Millimeter-Wave Channel Simulator and Applications for 5G Wireless Communications," 2017 IEEE International Conference on Communications (ICC), May 2017
- [SPL17] C. Stefanovic, E. Paolini and G. Liva, "Asymptotic Performance of Coded Slotted ALOHA with Multi Packet Reception", IEEE Commun. Lett., vol. PP, no. 99, pp. 1-1.
- [STS07] M. Sadek, A. Tarighat, and A. H. Sayed, "A leakage-based precoding scheme for downlink multi-user MIMO channels," IEEE Trans. Wireless Commun., vol. 6, no. 5, pp. 1711-1721, May 2007.
- [STY09] K. Slavakis, S. Theodoridis and I. Yamada, "Adaptive Constrained Learning in Reproducing Kernel Hilbert Spaces: The Robust Beamforming Case," IEEE Trans. Signal Process., vol. 57, no. 12, pp. 4744-4764, Dec. 2009
- [Sun17] X. Sun et. al, "Agglomerative user clustering and downlink group scheduling for FDD massive MIMO systems", in Proc. IEEE ICC'17, May 2017, France
- [SW18] S. Stefanatos and G. Wunder, "Performance Limits of Compressive Sensing Channel Estimation in Dense Cloud RAN," submitted to IEEE International Conference on Communications (ICC'18), 2018. Available: <https://arxiv.org/abs/1710.10796>
- [SWJ15] J. Schreck, G. Wunder, and P. Jung, "Robust iterative interference alignment for cellular networks with limited feedback," IEEE Trans. Wireless Commun., vol. 14, no. 2, pp. 882-894, Feb. 2015.
- [SYY98] H. Stark, Y. Yang, and Y. Yang, Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics. New York, NY, USA: John Wiley & Sons, Inc., 1998.

- [SZL14] Y. Shi, J. Zhang, and K. Letaief, "CSI overhead reduction with stochastic beamforming for cloud radio access networks," in Proc. of IEEE Int. Conf. on Commun. (ICC), Sydney, Australia, Jun. 2014.
- [TAG+12] K. Tsagkaris, M. Akeziidou, A. Galani, P. Demestichas. (2012), Evaluation of signalling loads in a cognitive network management architecture. Int. J. Network Mgmt, 22: 235-260. doi:10.1002/nem.803
- [WBS+15] G. Wunder, H. Boche, T. Strohmer, and P. Jung, "Sparse signal processing concepts for efficient 5G system design," IEEE Access, vol. 3, pp. 195–208, 2015.
- [WCT+14] Green transmission technologies for balancing the energy efficiency and spectrum efficiency trade-off
- [WRF+18] 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. Available: <https://arxiv.org/abs/1803.10994>.
- [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.
- [XRL15] X. Xu, X. Rao, and V. K. Lau, "Active user detection and channel estimation in uplink CRAN systems," in Proc. IEEE ICC'15, London, UK, Jun. 2015, pp. 2727–2732.
- [YGC14] H. Yin, D. Gesbert, and L. Cottatellucci, "Dealing With Interference in Distributed Large-Scale MIMO Systems: A Statistical Approach," IEEE J. Sel. Topics Signal Process., vol. 8, no. 5, pp. 942--953, Oct. 2014.
- [YGS+16] L. You, X. Gao, A. L. Swindlehurst, and W. Zhong, "Channel acquisition for massive mimo-ofdm with adjustable phase shift pilots," IEEE Trans. Signal Process., vol. 64, no. 6, pp. 1461–1476, March 2016.
- [YH15] S. Yang and L. Hanzo, "Fifty years of mimo detection: The road to large-scale mimos," IEEE Commun. Surveys Tuts., vol. 17, no. 4, pp. 1941–1988, Fourthquarter 2015.
- [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.
- [ZYZ17] J. Zhang, X. Yuan, and Y. J. Zhang, "Locally orthogonal training design for cloud-RANs based on graph coloring," IEEE Trans. Wireless Commun., vol. 16, no. 10, pp. 6426–6237, Oct. 2017.

8 Appendix

8.1 Sector and Beam Management with Cylindrical Arrays

System Model

In this work, a multi-user multi-sector Orthogonal Frequency Division Multiplexing (OFDM) downlink system with K Mobile Stations (MSs) and L sectors at the same location is considered, where \mathcal{L} is the set of cells $\{1, \dots, L\}$ and \mathcal{K} the set of MSs $\{1, \dots, K\}$. Each MS is equipped with M receive antennas and each sector with N transmit antennas. With this, the receive signal \mathbf{y}_k at MS k is given by

$$\mathbf{y}_k = \sum_{l \in \mathcal{L}} \mathbf{H}_{k,l} \mathbf{V}_l \sqrt{\hat{\mathbf{P}}_l} \mathbf{x}_l + \mathbf{n}_k,$$

where $\mathbf{H}_{k,l} \in \mathbb{C}^{M \times N}$ is the downlink channel matrix from sector l to MS k and $\mathbf{V}_l \in \mathbb{C}^{N \times T_l}$ is the precoding matrix applied by sector l with $T_l = |\mathcal{T}_l|$ being cardinality of the set $\mathcal{T}_l = \{1, \dots, T_l\}$ of spatial layer ids transmitted by sector l . Furthermore, $\hat{\mathbf{P}}_l \in \mathbb{R}_+^{T_l \times T_l}$ is the diagonal power allocation matrix at sector l with the power constraint $P_{\text{RB}} \geq \text{trace}(\hat{\mathbf{P}}_l)$, $\mathbf{x}_l \in \mathbb{C}^{T_l}$ is the vector of transmit data symbols at sector l to be send to the MSs, and $\mathbf{n}_k \in \mathbb{C}^M$ denotes the Additive White Gaussian Noise (AWGN) vector with covariance $\mathbb{E}[\mathbf{n}_k \mathbf{n}_k^H] = \mathbf{I}_M \sigma_k^2$. $\mathbb{E}[\cdot]$ is the expectation operator, σ_k^2 is the noise power comprising the receiver and thermal noise, and \mathbf{I}_M is an identity matrix of size M . Each column of the precoding matrix $\mathbf{V}_l = [\mathbf{V}_{l,1}, \dots, \mathbf{V}_{l,T_l}]$ is a spatial multiplexed stream or layer and assigned to a certain MS. Due to the complexity in signal processing caused by the large number of antennas N considered in massive MIMO this work is focused on linear precoding such that the number of spatial layers is constraint by the number of sector antennas $T_l \leq N$. Assuming that MS k is served by sector l on stream $t \in \mathcal{T}_l$ the receive signal from above can be divided into three parts according to

$$\begin{aligned} \mathbf{y}_{k,t} &= \underbrace{\mathbf{H}_{k,l} \mathbf{v}_{t,l} \sqrt{[\hat{\mathbf{P}}_l]_{t,t}}}_{\tilde{\mathbf{h}}_{k,t}} x_{t,l} + \underbrace{\sum_{\substack{j \in \mathcal{T}_l \\ j \neq t}} \mathbf{H}_{k,l} \mathbf{v}_{j,l} \sqrt{[\hat{\mathbf{P}}_l]_{j,j}}}_{\boldsymbol{\vartheta}_{k,t}} x_{j,l} \\ &+ \underbrace{\sum_{\substack{m \in \mathcal{L} \\ m \neq l}} \mathbf{H}_{k,m} \mathbf{V}_m \sqrt{\hat{\mathbf{P}}_m} \mathbf{x}_m}_{\mathbf{z}_k} + \mathbf{n}_k, \end{aligned}$$

where the intra-sector interference caused by streams $j \neq t$ are aggregated in $\boldsymbol{\vartheta}_{k,t} \in \mathbb{C}^M$ and streams from other BSs are denoted as inter-sector interference in $\mathbf{z}_k \in \mathbb{C}^M$. The subscript l of stream t is omitted for notational brevity and not required, since t corresponds to the stream with symbols for MS k and sector l is the serving sector of MS k . The linear receive filter at MS k for stream t is denoted by $\mathbf{w}_{k,t} \in \mathbb{C}^M$, e.g. Minimum Mean Square Error (MMSE) as in [3GPP-36829]. Assuming Gaussian distributed normalized receive symbols the Signal to Interference and Noise Ratio (SINR) of stream t at MS k including post-processing at the receiver is obtained by

$$\gamma_{k,t} = \frac{\mathbf{w}_{k,t}^H \tilde{\mathbf{h}}_{k,t} \tilde{\mathbf{h}}_{k,t}^H \mathbf{w}_{k,t}}{\mathbf{w}_{k,t}^H \tilde{\mathbf{Z}}_{k,t} \mathbf{w}_{k,t}},$$

where $\tilde{\mathbf{Z}}_{k,t} \in \mathbb{C}^{M \times M}$ is the receive interference covariance plus noise matrix obtained by

$$\tilde{\mathbf{Z}}_{k,t} = \boldsymbol{\vartheta}_{k,t} \boldsymbol{\vartheta}_{k,t}^H + \mathbf{z}_k \mathbf{z}_k^H + \mathbf{I}_M \sigma_k^2.$$

Finally, the capacity of the effective channel-link between sector l serving user k on stream t in the presence of AWGN is characterized by the Shannon-Hartley theorem as

$$R_{k,t} = B \log_2(1 + \gamma_{k,t}).$$

where B is the bandwidth of the link and Rk, t is given in [Bit/s]. Each user k is assigned to the sector with the largest sum power over all channel elements.

User Distribution

The single antennas users are distributed uniformly on a circle with a 50m radius as seen in Figure 8-1.

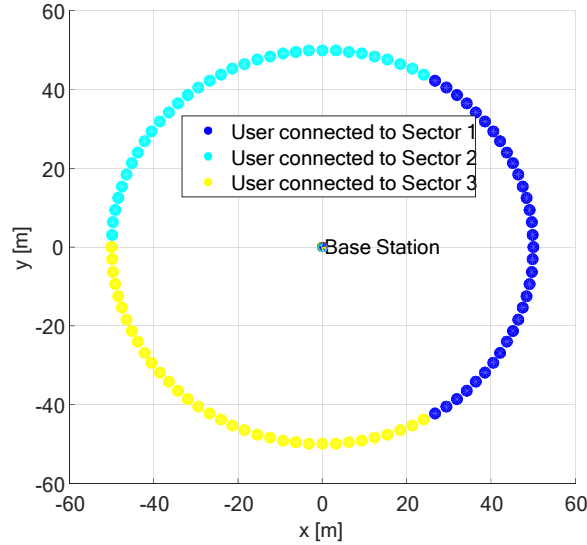


Figure 8-1: User distribution

Antenna Array Geometries

For UPAs used in the baseline scenario we consider M_{UPA} rows of antennas, N_{UPA} columns of antennas and L identical arrays at the same location oriented in directions as given in Figure 8-2 for $M_{\text{UPA}} = N_{\text{UPA}} = 8$ and $L = 3$. In the proposed UCAs we consider M_{UCA} vertical antennas and N_{UCA} equidistant distributed columns on a circle, see the example with $M_{\text{UCA}} = 8$ and $N_{\text{UCA}} = 24$ in Figure 8-3. For fair comparison of both array geometries the constraint of equal number of antennas is considered such that $LM_{\text{UPA}}N_{\text{UPA}} = M_{\text{UCA}}N_{\text{UCA}}$.

UCA Antenna-to-Sector Mapping

In the numerical evaluation two antenna mapping modes are considered, denoted as “adaptive” and “static” selection and defined as follows. Consider the set of UCA antennas denoted as \mathcal{A} , then the UCA is divided into L disjunctive sets $\mathcal{A}_l \subset \mathcal{A}$ of antennas mapping the antenna elements to independent sector as in the sectorized UPA case. Furthermore, we assume the ordered list of antenna elements per user \mathcal{A}_k such that $|h_{a_1}| > |h_{a_2}| > \dots > |h_{a_{Ll}}|$, where a_i denotes the i -th element of \mathcal{A}_k .

In adaptive sectorization each user selects the first $M_{\text{UPA}}N_{\text{UPA}}$ antenna elements from the ordered list \mathcal{A}_k . Note, that the “adaptive” mode can only be considered for single layer transmission such that $T_l = 1$, since each user selects individually his antennas with the largest receive power.

In static sectorization the set of antennas \mathcal{A}_l for sector l is the same for each user by selecting N_{UPA} consecutive columns from the UCA. Note, that due to the uniform user distribution and UCA symmetry the performance is independent from the starting column.

Table 8-1: Simulation assumptions

Parameter	Value
Channel model	QUADRIGA http://quadriga-channel-model.de/ version 2.0.0-664 and setting to match 3GPP calibration
Scenario	Urban micro, line of sight and non-line of sight
Single antenna element type	Patch element
Gain of main lobe	≈ 9.8 dB
Horizontal, vertical half power beam width	64° , 64°
Antenna geometry	UCA and UPA, see Figure 8-2 and Figure 8-3
UPA	$M_{UPA} = N_{UPA} = 8, L = 3$
UCA	$M_{UCA} = 8$ and $N_{UCA} = 24$
Transmit power per sector	40 dBm
Bandwidth	18 MHz, LTE resource grid
User distribution	100 user uniform on a circle with radius 50 m
User selection	Random
Precoder	Single user – MRT [Lo99]. Multi-user – MMSE [PHS05]

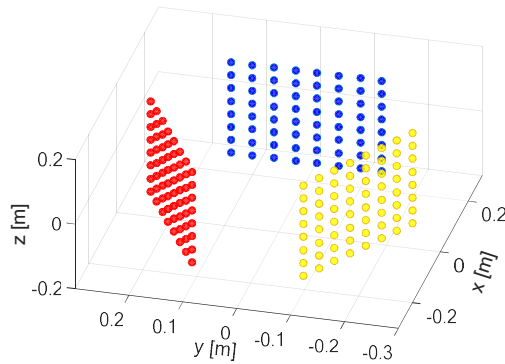


Figure 8-2: Sectorized UPA with colour coded sectors.

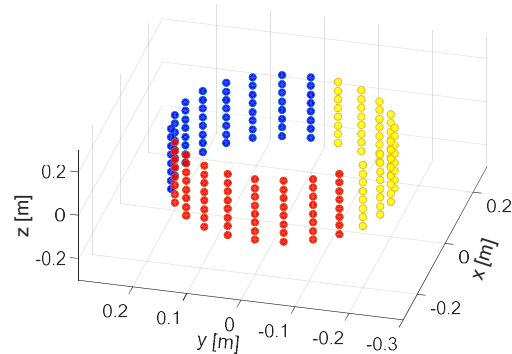


Figure 8-3: Sectorized UCA with colour coded sectors.

8.2 Hierarchical sparse channel estimation for multiuser massive MIMO with reduced training overhead

This section considers the extension of the system model and approach presented in [WRF+18] to the multiuser setting. Please refer to [WRF+18] for the single user setting and system model as well as the notations used in the following.

Let U denote the number of users assigned the exact same set of O_τ subcarriers to transmit their training (signature) sequences for channel estimation purposes. Let $c_u \in \mathbb{C}^{O_\tau}$ denote the signature sequence of user u . Then, considering for simplicity only a single OFDM symbol at time t , the BS observes the superposition of all U training sequences as

$$Y(t) = \sum_{u=0}^{U-1} \text{diag}(c_u(t))H_u(t) + Z(t),$$

where $H_u(t)$ denotes the channel matrix of user u .

A critical issue for identifying the channels $\{H_u(t)\}_{u=0}^{U-1}$, or, equivalently, their delay/angular representations $\{W_u(t)\}_{u=0}^{U-1}$, is the selection of the signatures $\{c_u(t)\}_{u=0}^{U-1}$. We propose the following sequences for the case where $U \leq N/D$, where D is the cyclic prefix length: $[c_u(t)]_l = e^{-i2\pi D^{-1}l} [c_0(t)]_l, l \in [0, \tau]$, that is the signatures of the users are frequency shifted versions of a baseline sequence (the sequence of user $u = 0$), which, without loss of generality, we set it as a sequence of ones. By exploiting the delay-angular representation described in [WRF+18], it is easy to see that, with the proposed sequence assignment, signal $Y(t)$ can be written as

$$Y(t) = A_\tau \tilde{W}(t) A_0^H + Z(t),$$

where $\tilde{W}(t) = [W_0^T(t), \dots, W_{U-1}^T(t)]^T$, consisting of the sparse delay-angular representations of the user channels. Therefore, the multiuser channel estimation problem has transformed to the single user channel estimation problem investigated in [WRF+18], with the difference that we are estimating the delay/angular representation matrix $\tilde{W}(t)$ that contains information for all user channels. An example of $\tilde{W}(t)$, shown in Figure 8-4, for a case with $U = 3$ and only the first and third users transmitting (the second is inactive). It is clearly seen that $\tilde{W}(t)$ possesses a rich hierarchical structure of multiple levels: the level of users, within which there are the levels of angle and delay (for each individual user channel). Therefore, its estimation can be naturally accommodated by the proposed single user channel estimation algorithm, by simply taking now into account the additional hierarchy level of users.

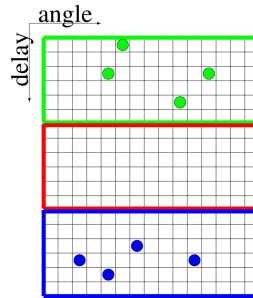


Figure 8-4 Example support of matrix $\tilde{W}(t)$ for the case of three users in total with only the first and third active.

8.3 CSI feedback for FDD massive MIMO

In this section, we will provide more details about our opportunistic feedback scheme for FDD systems. Our solution consists of providing a novel similarity measure along with a new clustering scheme where the number of clusters is not required to be known. We also develop, by using graph theory tools, a low complexity scheduling scheme that outperforms all currently proposed methods in both sum-rate and throughput fairness. We will describe briefly here the developed clustering and scheduling scheme. One can refer to [MHA+17] for more details.

1) Correlation Clustering

- **Similarity Measure**

In the previous literature, the similarity between user 1 and user 2 is solely taken based on their covariance's eigenstructures $(U_1 U_1^H, U_2 U_2^H)$ without taking into account the energy of the modes. In our case, we will be applying our similarity measure on the whole covariance matrices (R_1, R_2) . The motivation behind this is that differences in the eigenstructures of weak modes should contribute less than the one's of strong modes. We define the new similarity measure as follows:

$$d(R_1, R_2) = \frac{\text{Tr}(R_1^H R_2)}{\|R_1\|_F \|R_2\|_F}$$

This similarity measure is lower bounded by 0 and upperbounded by 1. A value of 0 corresponds to the case where R_1 and R_2 are orthogonal while a value of 1 takes place when R_1 and R_2 are collinear.

- **Clustering Algorithm**

Unlike the previously proposed approaches, we seek to use a clustering algorithm without passing the target number of clusters as a parameter. To do so, we take advantage of the ease of threshold design presented by our proposed similarity metric. An interesting way to do so is by choosing a threshold, denoted by Degree of OverLap, high enough such as if $d(R_1, R_2) \geq DOL$ then users k and k' can be thought to be laying in the same correlation space. Unlike other similarity metrics, this threshold is easily determined. One can simply say if the degree of overlap between the two spaces is above 0.95 then consider them as highly similar and are preferred to be assigned to the same cluster. Based on this, we can construct what we will call a complete graph $G_c = (V_c, E_c)$ where each vertex represents a user and an edge would have a $\langle +1 \rangle$ label to signal that these two users are preferred to be in the same cluster while $\langle -1 \rangle$ label refers to the opposite case..

Our goal now is therefore to produce a partition of the graph's vertices in a way that agrees as much as possible with the edge labels. To do so, we propose a cost function as the total disagreements of our resulting partitioned graph. The total disagreements cost is defined as the overall negative weights inside a cluster added to the positive weights between clusters. Our partitioning problem can be hence formulated as follows:

$$\begin{aligned} \text{minimize} \quad & J = \sum_{(u,v) \in E_c^+} x_{uv} + \sum_{(u,v) \in E_c^-} (1 - x_{uv}) \\ \text{s.t.} \quad & x_{uv} + x_{vw} \geq x_{uw} \quad \forall u, v, w \in V_c \\ & x_{uv} = x_{vu} \quad \forall u, v \in V_c \\ & x_{uv} = \begin{cases} 0 & \text{if } u \text{ and } v \text{ are in the same cluster} \\ 1 & \text{Otherwise} \end{cases} \end{aligned}$$

The constraints take into account the symmetry of x_{uv} and the triangular inequality satisfied by these variables. This inequality ensures that the optimization formulation is well posed. For instance if three users u, v and w are such that u and v are in the same group and v and w in the same group then u and w are in the same group. In other words, v cannot belong to two groups. What makes this clustering formulation interesting is that there is no need to specify the target number of clusters. Instead, the resulting optimal number of clusters could be any value from 1 to K depending on what fits our graph the most. The detailed clustering algorithm can be found in [MHA+17].

2) Downlink Scheduling

- **Problem Formulation**

After grouping users with the aforementioned clustering framework, we can now deal with the orthogonality aspect. In realistic scenarios, groups do not lay in mutual orthogonal channel covariance spaces and inter-group interference can therefore limit the overall performance. One can seek to reduce this interference by applying appropriate outer precoding techniques but it is insufficient. Therefore, adopting a scheduling scheme is of paramount importance for the overall performance. The main idea is to schedule the clusters for CSI feedback such that the selected clusters have almost orthogonal channels, which reduces the interference. The previous studies in

this area have not considered fairness between users. Consequently, we introduce weights W_{g_k} in order to incorporate fairness in our scheduling scheme.

We formulate a scheduling framework which maximizes the total weighted sum-rate such that the SIR is higher than a minimum threshold, which allows the users to communicate at a minimum rate. Let x_g be a boolean variable that indicates if the group g is scheduled for CSI feedback. We can formulate our scheduling problem:

$$\begin{aligned} & \text{maximize} && \sum_{g=1}^G x_g \sum_{k=1}^{K_g} W_{g_k} R_{g_k} \\ & \text{subject to} && \text{SIR}_g \geq \alpha_g \quad g=1, \dots, G \\ & && x_g \in \{0,1\} \quad g=1, \dots, G \end{aligned}$$

- **Scheduling Scheme**

In order to solve our scheduling problem, we propose a 2-step scheme based on graph theory. The first step deals with both the SIR constraint and a combinatorial difficulty faced in our problem. The second step aims to find the appropriate combination of groups to be able to solve the problem.

a) Elimination: We can picture each vertex in the graph as a sink of interference that undergoes successive iterations. In the first iteration, all groups are considered to be active. The outer precoder of each group is calculated. For each vertex $g \in V$, the SIR condition is tested. If it is violated, the edge $e(g', g)$ with the highest weight is eliminated. In other words, the group that interfere most with g is chosen to be eliminated. At the next iteration, we have a new graph due to the edges removal from the previous iteration and therefore the outer precoders are to be recalculated. This time however, the outer precoder of each vertex $g \in V$ is calculated based on the eigenspace of neighboring vertices only. We repeat the same procedures of the first step and the weight of the edges of neighboring vertices only are recalculated. The stopping criteria would be if an iteration resulted in no new deleted edges. In other words, if simultaneous scheduling of neighboring vertices in the resulting graph will not violate the SIR condition of each of them.

b) Grouping: After proceeding with the elimination step, our SIR constraint can be replaced by making sure that two simultaneously scheduled groups should have an edge between them in G_u . Therefore, our optimization problem is turned into:

$$\begin{aligned} & \text{maximize} && \sum_{g=1}^G x_g \sum_{k=1}^{K_g} W_{g_k} R_{g_k} \\ & \text{subject to} && x_g + x_{g'} \leq 1 \quad \forall (g, g') \in E_u \\ & && x_g \in \{0,1\} \quad g=1, \dots, G \end{aligned}$$

To solve this new problem, we recall that for a well chosen $\alpha_g, \forall g$, an edge exists between two vertices in $G_u = (V, E_u)$ only if they barely interfere and hence scheduling them together would normally increase their sum utility. Our goal is therefore to find combinations of groups that are adjacent one to the other in G_u while covering the whole vertex set V . We seek to find the smallest number of cliques that cover V , where we emphasize "smallest" to ensure that each clique have the largest number of groups possible inside. Essentially, we are trying to solve the minimal clique vertex cover problem. The minimal clique vertex cover problem is known to be equivalent to vertex coloring on the complement graph \bar{G}_u , a well known NP-Complete problem. Knowing that vertex coloring seeks to partition the set of vertices into the smallest number of independent sets,

one can see the connection between the two problems since a subset of vertices is a clique in G_u if and only if it is an independent set in \bar{G}_u . We will use a simple yet effective maximal independent set based vertex coloring algorithm that achieves a $O\left(\frac{n}{\log(n)}\right)$ -approximation of the optimal solution and apply it on \bar{G}_u . Each group g will be then assigned a color. Groups that are assigned the same color represent a subset of groups that are allowed to transmit simultaneously. In the end, at the start of each coherence time, the schedule (i.e. the color) that leads to the largest utility is selected. More details can be found in [MHA+17].

8.4 Genetic Algorithm Assisted Hybrid Beamforming for Wireless Fronthaul

Let us consider a wireless fronthaul system in time division duplex (TDD) mode using hybrid beamforming system with a macrocell (transmitter) and K remote nodes (receivers). The transmitter is equipped with N_T transmit antennas and N_{RF} radio frequency (RF) chains. The l th remote node is equipped with $N_{R,l}$ receive antennas. The received signal vector of the l th remote node is given by

$$\mathbf{y}_l = \mathbf{H}_l \mathbf{A} \sum_{k=1}^K \mathbf{D}_k s_k + \mathbf{n}_l = \mathbf{H}_l^E \sum_{k=1}^K \mathbf{D}_k s_k + \mathbf{n}_l \quad (8-1)$$

where \mathbf{H}_l is the $N_{R,l} \times N_T$ channel matrix between the l th remote node and the transmitter, \mathbf{D}_k is the $N_{RF} \times 1$ digital precoding vector, \mathbf{n}_l is Gaussian noise vector, \mathbf{A} is the $N_T \times N_{RF}$ analog precoding matrix, and $\mathbf{H}_l^E = \mathbf{H}_l \mathbf{A}$ is the effective (distorted) channel matrix. An analog precoding matrix consists of phase shifters only. Also, the value of the phase of a phase shifter is restricted by B bits and B equals to 1 or 2 in this design to minimize cost.

Given the analog precoder \mathbf{A} , The digital precoder is designed to maximize the SLNR [STS07]. Let $\tilde{\mathbf{H}}_l^E = [\mathbf{H}_1^E, \mathbf{H}_2^E, \dots, \mathbf{H}_{l-1}^E, \mathbf{H}_{l-1}^E \dots \mathbf{H}_K^E]$, the digital precoder for the l th remote node can be expressed as $\mathbf{D}_l \propto \mathbf{v}_l^1$ [STS07], where \mathbf{v}_l^1 is the eigen vector corresponding to the maximum eigen value $\lambda_{\max}^{(l)}$ of

$$\left(M_l \sigma^2 \mathbf{I} + (\tilde{\mathbf{H}}_l^E)^H \tilde{\mathbf{H}}_l^E \right)^{-1} (\mathbf{H}_l^E)^H \mathbf{H}_l^E. \quad (8-2)$$

Also, the digital precoder needs to be normalized such that the aggregate precoding matrix has norm one, i.e., $\|\mathbf{A} \mathbf{D}_l\| = 1$.

The design of analog precoder is to maximize the maximum eigen value of Eq. (8-2) for each remote node. This optimization is done via the genetic algorithm. In the context of genetic algorithm, a chromosome is a solution to the optimization problem. Population represents the collection of all chromosomes, which is initialized randomly. The fitness function represents the survival probability of a chromosome. To maximize the maximum eigen value, the fitness function is defined as

$$f(\mathbf{A}) = R \approx \sum_{k=1}^K \log_2 \left(1 + \lambda_{\max}^{(k)} \right). \quad (8-3)$$

The genetic algorithm based analog precoder optimization procedure involves three basic operators, i.e., selection, crossover, and mutation. The selection process is based on the roulette wheel selection [Mel99]. The selection probability of \mathbf{A} is $f(\mathbf{A}) / (\sum_{\mathbf{A}'} f(\mathbf{A}'))$. Each selection operation randomly picks two chromosomes in the population according to their selection probabilities. The picked chromosomes are called parents. Next, Crossover randomly interchanges bits between the selected pair of chromosomes (parents) according to a predefined crossover probability p_c . A typical value of p_c is 0.7 [Mel99]. The resulting chromosomes are known as children. Next, mutation is performed on the two children by mutating each bit with mutation probability p_m . A typical value of p_m is 0.001 [Mel99]. Then, the mutated children will be placed into the new population. The overall procedure for analog precoder design can be summarized as:

```

1: Initialize population and set generation number  $g=1$ ;
2: while  $g <$  maximum generation number
3:     Remote nodes transmit reference signals to the macrocell;
4:     for each chromosome  $\beta$  in the population
5:         Adjust analog precoder  $\mathbf{A}$  according to  $\beta$ ;
6:         Measure distorted effective channel  $\mathbf{H}_k\mathbf{A}$  for all  $k$ ;
7:         Calculate fitness  $f(\mathbf{A})$  of  $\beta$ ;
8:     end for
9:     Perform selection, crossover, mutation, and update the population;
10:     $g = g+1$ ;
11: end while
12: Output the chromosome with the highest fitness in the population as the analog precoder;

```

8.5 Hybrid array architectures covering different deployment scenarios

The hybrid array architecture allows different degrees of freedom for selection of the number of MIMO streams K , the number of antenna ports P and the number of array elements N . This results in various possible subarray sizes and shapes, which form the entire array and impact the achievable spectral efficiency in different deployment scenarios. The used array configurations (number of elements and ports) are given in Table 8-2. Each port is represented by a subarray with 4, 2 and 1 x-polarized antenna element, as indicated in Figure 8-5. The simulation parameters are shown in Table 8-3.

Table 8-2: Array configurations used for simulations

Array type	No. of elements	No. of ports
A	256 (128 x-pol elem.)	64
B	128 (64 x-pol elem.)	64
C	64 (32 x-pol elem.)	64
K	128 (64 x-pol elem.)	32
L	64 (32 x-pol elem.)	16

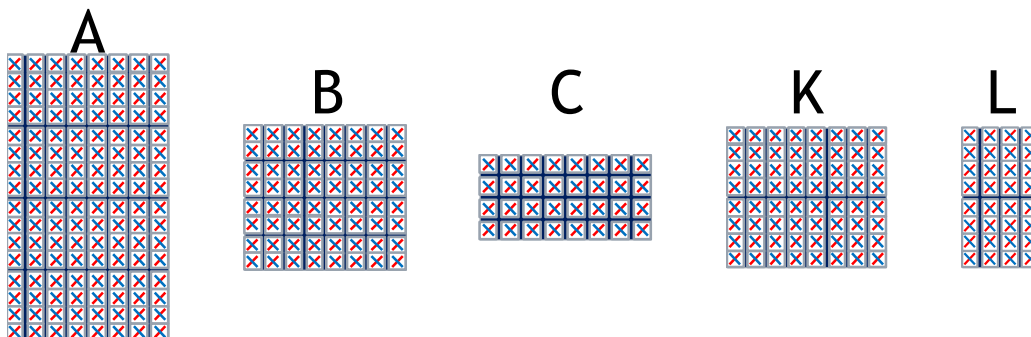


Figure 8-5 Array types: x-pol elements arranged in subpanels with different size

Table 8-3: System simulation parameters

Parameter	Value
Channel Model	UMa 3D according to [36.873]
ISD	500m
Nr. Cells	7 Cells, i.e. 21 Sectors with Wrap Around
Scheduler	All mobiles in a sector scheduled in parallel
BS Antenna Configuration	256 / 128 / 64 elements, 64 / 32 / 16 ports (see Table 8-2)

BS TX Power	46 dBm for 256 elements, 43 dBm for 128 elements, 40 dBm for 64 elements
UE Antenna Configuration	1x1x2
UE TX Power	23 dBm
User Distribution	1, ..., <Nr. Ports/4> Users per Sector
Channel Measurements	Channel Reciprocity assumed: UL Measurements, 1 Sample per Coherence Block
Beamforming	Zero Forcing (ZF), Eigenbeamforming (EBF)
Channel Estimation	UL Pilot Contamination considered. <Nr. Users> orthogonal Pilots per Sector. Pilot Reuse 1 between Sectors.
Link Adaptation	Perfect Link Adaptation
Traffic Model	Full Buffer
Bandwidth	10 MHz
Carrier Frequency	2 GHz
BS height	25m

8.6 Decentralized beamforming algorithms

The weighted sum rate maximization problem under imperfect CSI consideration can be written as follows:

$$\mathbf{Q} = \begin{cases} \max_{\mathbf{H}|\bar{\mathbf{H}}} \sum_k u_k \ln \det(\mathbf{I}_M + \mathbf{H}_{k,b_k}^H \mathbf{R}_k^{-1} \mathbf{H}_{k,b_k} \mathbf{Q}_k) \\ \sum_{k:b_k=j} \text{tr}\{\mathbf{Q}_k\} \leq P_j^{\text{BS}} \text{ for each } j \end{cases}$$

where M is the number of antennas at the BS, \mathbf{Q}_k is the transmit covariance matrix of user k , \mathbf{H}_{k,b_k} is the channel from user k 's BS towards user k . The constraints are used to limit the power transmitter per BS. The total and interference plus noise receive covariance matrices are given by, respectively,

$$\begin{aligned} \mathbf{R}_k &= \mathbf{H}_{k,b_k} \mathbf{Q}_k \mathbf{H}_{k,b_k}^H + \mathbf{R}_{\bar{k}}, \\ \mathbf{R}_{\bar{k}} &= \sum_{i \neq k} \mathbf{H}_{k,b_i} \mathbf{Q}_i \mathbf{H}_{k,b_i}^H + \sigma^2 \mathbf{I}_N \end{aligned}$$

where \mathbf{C}_p and \mathbf{C}_r are the channel estimation error covariances at the transmit and receive sides respectively.

If the number of transmit antennas M becomes very large, we get a convergence of any term of the following:

$$\mathbf{H}\mathbf{Q}\mathbf{H}^H \rightarrow E_{\mathbf{H}}\mathbf{H}\mathbf{Q}\mathbf{H}^H = \bar{\mathbf{H}}\mathbf{Q}\bar{\mathbf{H}}^H + \text{tr}\{\mathbf{Q}\mathbf{C}_p\}\mathbf{C}_r$$

The objective function can be rewritten as:

$$\begin{aligned} \text{EWSR} &= u_k \ln \det(\mathbf{I}_N + \check{\mathbf{R}}_{\bar{k}}^{-1} (\mathbf{H}_{k,b_k} \mathbf{Q}_k \mathbf{H}_{k,b_k}^H + \text{tr}\{\mathbf{Q}_k \mathbf{C}_{p,k,b_k}\} \mathbf{C}_{r,k})) \\ &+ \sum_{i=1, i \neq k}^K u_i \ln \det(\mathbf{I}_N + \check{\mathbf{R}}_{\bar{i}}^{-1} (\mathbf{H}_{i,b_i} \mathbf{Q}_i \mathbf{H}_{i,b_i}^H \\ &+ \text{tr}\{\mathbf{Q}_i \mathbf{C}_{p,i,b_i}\} \mathbf{C}_{r,i})) \end{aligned}$$

With

$$\begin{aligned}\check{\mathbf{R}}_k &= E_{\mathbf{H}|\bar{\mathbf{H}}}\mathbf{R}_k = \sigma^2\mathbf{I}_N + \sum_i \mathbf{H}_{k,b_i} \mathbf{Q}_i \mathbf{H}_{k,b_i}^H + \text{tr}\{\mathbf{Q}\mathbf{C}_{p,k,b_i}\}\mathbf{C}_{r,k}, \\ \check{\mathbf{R}}_{\bar{k}} &= E_{\mathbf{H}|\bar{\mathbf{H}}}\mathbf{R}_{\bar{k}} = \sigma^2\mathbf{I}_N + \sum_{i \neq k} \mathbf{H}_{k,b_i} \mathbf{Q}_i \mathbf{H}_{k,b_i}^H + \text{tr}\{\mathbf{Q}\mathbf{C}_{p,k,b_i}\}\mathbf{C}_{r,k}.\end{aligned}$$

We isolate the EWSR of the links other than the k^{th} link $EWSR_{\bar{k}}$, we perform this operation for each user k , after linearization and augmenting the EWSR function with constraints, we get a set of K per-user problems given by:

$$\max_{\mathbf{G}_k} \ln \det(\mathbf{I}_{d_k} + \mathbf{G}_k^H \check{\mathbf{B}}_k \mathbf{G}_k) - \text{tr}\{\mathbf{G}_k^H (\check{\mathbf{A}}_k + \lambda_{b_k} \mathbf{I}_M) \mathbf{G}_k\}$$

where d_k is the number of data streams received at user k , G_k is the beamforming matrix ($\mathbf{Q}_k = \mathbf{G}_k \mathbf{G}_k^H$) and

$$\begin{aligned}\check{\mathbf{B}}_k &= \mathbf{H}_{k,b_k}^H \check{\mathbf{R}}_{\bar{k}}^{-1} \mathbf{H}_{k,b_k} + \text{tr}\{\mathbf{C}_{r,k} \check{\mathbf{R}}_{\bar{k}}^{-1}\} \mathbf{C}_{p,k,b_k} \\ \check{\mathbf{A}}_k &= -\frac{\partial EWSR_{\bar{k}}}{\partial \mathbf{Q}_k} \\ &= \sum_{\substack{i=1 \\ i \neq k}}^K u_i [\mathbf{H}_{i,b_k}^H (\check{\mathbf{R}}_i^{-1} - \check{\mathbf{R}}_i^{-1}) \mathbf{H}_{i,b_k} \\ &\quad + \text{tr}\{(\check{\mathbf{R}}_i^{-1} - \check{\mathbf{R}}_i^{-1}) \mathbf{C}_{r,i}\} \mathbf{C}_{p,i,b_k}]\end{aligned}$$

With respect to \mathbf{G}_k , we get:

$$\mathbf{G}_k = P_k^{1/2} \text{eigenmatrix}(\check{\mathbf{A}}_k + \lambda_{b_k} \mathbf{I}_M, \check{\mathbf{B}}_k)$$

Where P_k is determined by waterfilling and λ_{b_k} is determined by bisection.

8.7 Massive MIMO with Hybrid Analog-Digital Precoding in a CRAN Architecture

In CRAN architecture, all the analog beamformers of RRHs are determined at the BBU. Such decisions rely on jointly optimizing the size of analog beamforming matrix of RRH l , $\mathbf{F}_{RF,l}$ and the entities of $\mathbf{F}_{RF,l}$, i.e. the number of active RF chains per RRH M_l and $[\mathbf{F}_{RF,l}]_{ij}$, revealing the following technical challenges.

One major difficulty comes from the optimal decision of M_l , which affects not only the size of the analog beamformer $\mathbf{F}_{RF,l} \in \mathbb{C}^{N \times M_l}$, where N is the number antenna per RRH but also the size of the digital beamformer $\mathbf{F}_{BB} \in \mathbb{C}^{\bar{M} \times K}$, where $\bar{M} = \sum_{l=1}^L M_l$. Therefore, adjusting M_l affects the quantization noise variance \mathbf{Q}_l since fronthaul compression is performed after applying \mathbf{F}_{BB} . Because of per-RRH transmit power constraints, \mathbf{F}_{BB} is normalized by a single quantity α at the BBU, it requires a search for the \hat{l} -th RRH producing the largest transmit power for a given channel and quantization noise.

We propose a sub-optimal algorithm that sequentially optimizes $[\mathbf{F}_{RF,l}]_{ij}$ and M_l , while exploiting spatial covariance matrices and the large-scale approximated SINR.

1) Construction of $[\mathbf{F}_{RF,l}^*]_{ij}$

Suppose a given number of active RF chains $M_l \leq \hat{M}$, where \hat{M} is the number of RF chains per RRH. For this given value, $[\mathbf{F}_{RF,l}]_{ij}$ is constructed based on the trace-weighted approach.

2) Selection of M_l^*

After the construction of $[\mathbf{F}_{RF,l}]_{ij}$ for each $M_l \in \{1, 2, \dots, \widehat{M}\}$, we find the optimal M_l^* that maximizes the sum-rate.

3) Unit modulus constraint on $[\mathbf{F}_{RF,l}]_{ij}$

Up to this point, the unit modulus constraint has been neglected. This constraint is re-introduced after the selection of M_l^* by simply setting the entries of the analog beamformers to have unit modulus and the same phases as the unconstrained beamformer.

The aforementioned procedures are detailed in [PKC+17].

8.8 Description of the RZF-CI precoder

System model

The downlink of a TDD OFDM system is considered and the FDMA method is used on the access antenna to serve the multiple UEs. The BH antenna array on the BS 1 is equipped with N_t antennas, while the BH antenna array on the BS 2 is equipped with N_r antennas. However, because of the long distance and the strong LOS component between the two BH antenna arrays, the N_r antennas of the second BH antenna array can be seen as an equivalent single antenna with a high directivity and pointing towards the first BH antenna array. Additionally, the UEs are assumed to be equipped with a single antenna. The system model is described in Figure 8-6 focusing on one sub-carrier. As the FDMA method is used for the access links, at most one UE is impacted by the BH link on the considered sub-carrier.

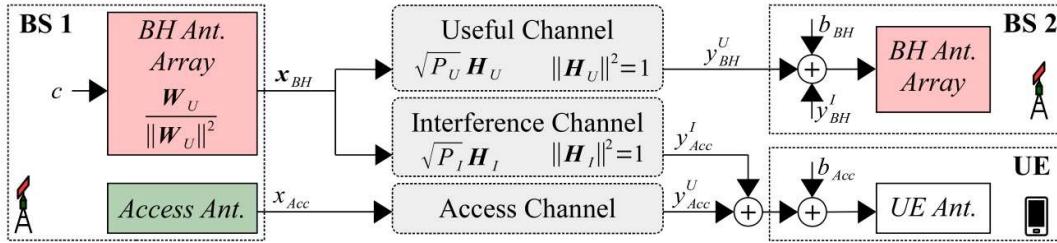


Figure 8-6 System model.

Useful received power on the BS 2

On the considered subcarrier, the BH antenna array of the BS 1 is used to send the data c of variance σ_c^2 to the second BH antenna array. This transmission is realized thanks to the precoder described by the vector \mathbf{W}_U of size $(N_t \times 1)$. Therefore, the vector of precoded data is defined by $\mathbf{x}_{BH} = \mathbf{W}_U / \|\mathbf{W}_U\| \cdot c$. The useful channel between the two BH antenna arrays is defined by the vector $\sqrt{P_U} \mathbf{H}_U$ of size $(1 \times N_t)$, with P_U being the power of this useful channel and \mathbf{H}_U being the normalized useful channel vector ($\|\mathbf{H}_U\|^2 = 1$). Therefore, the received useful signal on the BH antenna array of the BS 2 is defined by $y_{BH}^U = \sqrt{P_U} \mathbf{H}_U (\mathbf{W}_U / \|\mathbf{W}_U\|) \cdot c$ and the received useful power on the BH antenna array of the BS 2 is thus:

$$|y_{BH}^U|^2 = P_U \sigma_c^2 \frac{|\mathbf{H}_U \mathbf{W}_U|^2}{\|\mathbf{W}_U\|^2}. \quad (8-4)$$

Damage caused by the BH link on the UE side

On the UE side, the received signal y_{Acc} is defined by $y_{Acc} = y_{Acc}^U + y_{Acc}^I + b_{Acc}$, where y_{Acc}^U is the useful data coming from the access antenna, y_{Acc}^I is the interfering data coming from the BH antenna array of the BS 1 and b_{Acc} is the noise component of variance σ_b^2 . The interference channel between the BH antenna array of the BS 1 and the UE is defined by the vector $\sqrt{P_I} \mathbf{H}_I$ of size $(1 \times N_t)$, with P_I being the power of the interference channel and \mathbf{H}_I being the normalized

interference channel vector ($\|\mathbf{H}_I\|^2 = 1$). Therefore, the interfering data is defined by $y_{Acc}^I = \sqrt{P_I} \mathbf{H}_I (\mathbf{W}_U / \|\mathbf{W}_U\|)$. c and the received interference power on the UE side is thus:

$$|y_{Acc}^I|^2 = P_I \sigma_c^2 \frac{|\mathbf{H}_I \mathbf{W}_U|^2}{\|\mathbf{W}_U\|^2}. \quad (8-5)$$

In order to evaluate the damage caused by the BH link on the UE side, the metric Δ is defined as follows:

$$\Delta = \frac{\rho}{\gamma} = \frac{P_I \sigma_c^2 \frac{|\mathbf{H}_I \mathbf{W}_U|^2}{\|\mathbf{W}_U\|^2} + \sigma_b^2}{\sigma_b^2} = 1 + \frac{|\mathbf{H}_I \mathbf{W}_U|^2 P_I \sigma_c^2}{\|\mathbf{W}_U\|^2 \sigma_b^2} \quad (8-6)$$

with $\rho = |y_{Acc}^U|^2 / \sigma_b^2$ and $\gamma = |y_{Acc}^U|^2 / (|y_{Acc}^I|^2 + \sigma_b^2)$ being respectively the SNR and the SINR on the UE side. Therefore, the higher the value of Δ , the bigger the impact of the BH link on the UE side.

RZF-CI precoder

The RZF-CI precoder described thereafter is designed to increase the value of $|y_{BH}^U|^2$ compared to the classical ZF precoder, while keeping the value of Δ under the predetermined value Δ_{max} . It needs the following inputs:

- the normalized channel vectors \mathbf{H}_U and \mathbf{H}_I , estimated on the BS 1 side during the uplink,
- the SNR ρ , estimated on the UE side in a preliminary step, when the BH link is not active or when a ZF precoder is used ($y_{Acc}^I = 0$),
- the SINR γ_{MRT} , estimated on the UE side in a preliminary step, when a MRT precoder is used.

The RZF-CI precoder is then computed as follows:

$$\mathbf{W}_U = \begin{cases} \mathbf{H}_U^H & \text{if } \Delta_\infty \leq \Delta_{max} \\ (1 + \alpha_0) \mathbf{H}_U^H - \Gamma_H \mathbf{H}_I^H & \text{if } \Delta_\infty > \Delta_{max} \end{cases}, \quad (8-7)$$

with $\Delta_\infty = \rho / \gamma_{MRT}$, $\Gamma_H = \mathbf{H}_I \mathbf{H}_U^H$ and:

$$\alpha_0 = \frac{(1 - |\Gamma_H|^2) + \sqrt{(1 - |\Gamma_H|^2) \left(\frac{\Delta_\infty - 1}{\Delta_{max} - 1} - |\Gamma_H|^2 \right)}}{\frac{\Delta_\infty - 1}{\Delta_{max} - 1} - 1}. \quad (8-8)$$

8.9 Nonlinear Mechanisms in Cell-less Systems

We denote by M the number of receive antennas at the BS, and by K the number of transmit devices whose signals need to be detected at the BS. The authors in [STY09] show that nonlinear filtering design in the powerful framework of Gaussian reproducing kernel Hilbert space (RKHS) outperforms linear filtering especially when $K > M$. The authors also demonstrate the superior angular resolution of the nonlinear adaptive filter. The computational advantage of this approach is that the nonlinear filtering in the original space becomes a linear filtering task in an infinite dimensional RKHS. Precise details can be found in [ACY+18, STY09].

Nonlinear filters, however, suffer from lack of sufficient robustness against environmental changes. Their performance, in dynamic scenarios, e.g., in which devices transmit sporadically, may deteriorate. To exploit the benefits of both linear and nonlinear filters, we propose a partially linear filter in the sum space of the linear and the Gaussian RKHSs. In more detail, we denote by H_L and H_G the real RKHSs associated with the linear and the Gaussian kernel, respectively. The partially linear filter is defined as an element of the RKHS $H: H_L + H_G: \{\omega_L f_L + \omega_G f_G: f_L \in H_L, f_G \in H_G\}$, where $\omega_L, \omega_G > 0$ are some weights. The filter is trained using training samples by

utilizing the *adaptive projected subgradient method* (APSM) which consists of an online sequence of projections onto convex sets defined by the training samples:

$$f_{t+1} = f_t + \left(\sum_{j \in J_t} q_j^n P_{C_j}(f_t) - f_t \right)$$

where P_{C_j} is the projection on the set C_j defined by the j th sample weighted by $w_j > 0$. The index set J_t is a subset (e.g., comprising the latest 10 samples) of the training samples up to time t . It is well known that under certain assumptions we can obtain a point arbitrarily close to the intersection of all but a finite number of sets C_j . For more details of ASPSM please see [ACY+18].

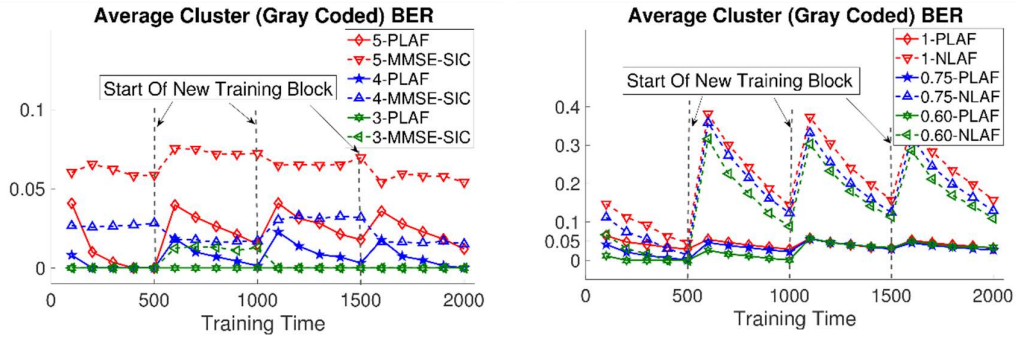


Figure 8-7 (left) Bit Error Rate for PLAF and MMSE-SIC for $M=3$ and $K=5,4,3$. (right) Bit Error Rate for PLAF and NLAF for $M=3$, $K=5$, and active device $\%=100, 75$ and 60 .

Figure 8-7 shows the comparison between the partially linear filter (PLAF), the linear filter followed by successive interference cancellation (MMSE-SIC), as well as the purely nonlinear filter (NLAF) of [STY09]. Note that the performance is shown for the case of un-coded QAM modulation for a single BS. We present the average performance of a cluster of devices where it is assumed that other clusters connected to the BS are assigned orthogonal resources. To simulate a dynamic environment, such as the one expected in massive machine-type communication (mMTC), we switch the channel and user distribution (in terms of active users) after every 500 samples. We validate the model after every 100 samples such that the points in the graph stand for the BER expected if training is stopped at that point. The PLAF outperforms the MMSE-SIC scheme when the number of devices exceeds the number of antennas (fixed at $M = 3$). In contrast to the NLAF, the PLAF shows much more robustness against channel and user distribution switching. In this experiment we perform simulation with different percentages of active users in the system.

In the following we present how the partially linear beamforming described above can be performed in a distributed fashion in a cell-less scenario.

Distributed Detection in C-RAN Cell-less Systems

We now describe the two forwarding strategies. For simplicity we assume real signalling, but the complex case is a straightforward extension using the bijection between real and complex numbers described in [ACY+18]. The result of using this bijection is that the real received signal has now dimension $2M$. Moreover, the procedure is shown for a single device because device symbols for each device are detected in parallel. We also assume BPSK signalling for clarity. The simulation results are, however, presented for QPSK in Section 4.1.4. The detection of the modulation symbol is performed using a filter $f: \mathbb{R}^{2M} \rightarrow \mathbb{R}$.

- **Quantize and Forward (Q&F)**

The received signal $\mathbf{r}^l \in \mathbb{R}^{2M}$, $l \in \{1, 2, \dots, R\}$, at each RRH is forwarded to the CU after quantization using B_q bits per component. The quantized signals received from each RRH are stacked to obtain a signal $\mathbf{r}^{cu} \in \mathbb{R}^{2M}$. In this way, RRHs act like a distributed antennas system. The training and detection for the filter $f: \mathbb{R}^{2M} \rightarrow \mathbb{R}$ is performed at the CU using Algorithm 1 in [ACY+18].

- **Detect and Forward (L&F)**

The training and detection with $f_l: \mathbb{R}^{2M} \rightarrow \mathbb{R}$ is performed at each RRH using the Algorithm 1 in [ACY+18]. During detection, likelihood values from each RRH are fused at the CU to estimate the BPSK symbol $b(t)$ broadcasted by the device at time t :

$$\hat{b}(t) = \text{sgn} \left(\sum_{l=1}^R \log \frac{\mathcal{L}_\ell(+1, \mathbf{r}^l(t))}{\mathcal{L}_\ell(-1, \mathbf{r}^l(t))} \right)$$

where $\text{sgn}(x)=1$ if $x \geq 0$ otherwise $\text{sgn}(x)=-1$. The likelihood functions $\varphi^l(f_l(\mathbf{r}^l(t))|+1) := \mathcal{L}_\ell(+1, \mathbf{r}^l(t))$ and $\varphi^l(f_l(\mathbf{r}^l(t))|-1) := \mathcal{L}_\ell(-1, \mathbf{r}^l(t))$ are estimated using a projection-based set theoretic method outlined below after the training period. During detection, the likelihood values can be computed and fused at the CU using many methods including simple scalar quantization and consensus approaches. These approaches are not the focus of this work.

Set-Theoretic Estimation of Likelihood Functions

In this section, we present a general low-complexity technique for obtaining a reliable approximation of a pdf given a sample set of independent and identically distributed (i.i.d) samples. We denote a random source by \mathbf{X} and perform a ‘‘set theoretic approximation’’ of the pdf $\varphi_{\mathbf{X}}$ by utilizing available prior knowledge. The prior knowledge includes general properties of pdfs and knowledge derived from a given sample set $\mathcal{D}_{\mathbf{X}} := \{x_1, x_2, \dots, x_N\}$ which we assume to consist of i.i.d observations of \mathbf{X} . These sample sets can be easily constructed given the trained filter and its response to the training data that was used for training during time T_t .

We start by assuming that $\varphi_{\mathbf{X}} \in L^2(\mathbb{R})$, where $L^2(\mathbb{R})$ is the Hilbert space of square (Lebesgue) integrable functions equipped with the inner product $(\forall f, g \in L^2) \langle g, f \rangle_{L^2} := \int_{\mathbb{R}} g(x)f(x)dx$ and norm $\|f\|_{L^2}^2 = \langle f, f \rangle_{L^2} < \infty$. We follow a similar approach to [SYY98], but in contrast to [SYY98], we restrict $\varphi_{\mathbf{X}}$ to the closed subspace $= \{\varphi \in L^2 | \varphi = \sum_{i=1}^N w_i \kappa(\cdot, x_i)\}$, $N \in \mathbb{Z}_{\geq 0}$, $(\forall i \in \overline{1, N}) w_i \in \mathbb{R}$; $(\forall i \in \overline{1, N}) \kappa(x, x_i) := (1/\sqrt{2\pi\sigma^2}) \exp\left(\frac{-|x-x_i|^2}{2\sigma^2}\right)$, $\sigma > 0$, $x, x_i \in \mathbb{R}$. We equip \mathcal{G} with an inner product $\langle h, p \rangle_{\mathcal{G}} = \langle h, p \rangle_{L^2}$ and the norm $\|f\|_{\mathcal{G}}^2 = \langle f, f \rangle_{\mathcal{G}}$ such that \mathcal{G} is a Hilbert space. The reason for working in this subspace is that the algorithms we employ consist of inner product operations which are very convenient to compute in this subspace with well-known closed-form solutions. In the following we fix the centres x_i by using the values in $\mathcal{D}_{\mathbf{X}}$ as centres for the Gaussian kernel above. We observed a good performance for this heuristic.

In the light of the above, the objective now becomes to find a $\varphi^* \in \mathcal{G}$ that agrees with all the available information we have about $\varphi_{\mathbf{X}}$. More precisely, suppose that the prior information amounts to the fact that $\varphi_{\mathbf{X}}$ is a member of Q closed-convex sets, i.e., $(\forall q \in \overline{1, Q}) \varphi_{\mathbf{X}} \in C_q \subset \mathcal{G}$. Then an approximation of $\varphi_{\mathbf{X}}$ is a solution to the ‘‘set feasibility problem’’: find $\varphi^* \in \mathcal{G}$ such that $\varphi^* \in \bigcap_{q=1}^Q C_q$. Set feasibility problems can be solved by using a plethora of projection algorithms which are well-known for their simplicity. Moreover, projection operations often involve inner-

products that have convenient closed-forms in \mathcal{G} which results in a low complexity of the algorithms. We now present the construction of the sets C_q and details of our algorithm.

Consider the event $\{a_q \leq \mathbf{X} \leq b_q\}$ (\mathbf{X} is the random variable defined above) where the probability of this event $\Pr[a_q \leq \mathbf{X} \leq b_q] = \bar{p}_q$ is unknown. Given a sample set $\mathcal{D}_X := \{x_1, x_2, \dots, x_N\}$, we can divide the range of values in \mathcal{D}_X in intervals $(q \in \overline{1, Q}) [a_q, b_q]$, where Q is a design parameter. Since the intervals $(q \in \overline{1, Q}) [a_q, b_q]$ are calculated from \mathcal{D}_X , \bar{p}_q is a random variable. We follow the approach in [SYY98] to calculate the 95% confidence intervals $\mathcal{P}_q := [P_q^L, P_q^H]$ such that $(q \in \overline{1, Q}) \Pr[P_q^L \leq \bar{p}_q \leq P_q^H] \approx 0.95$. These calculations are computationally inexpensive. We omit the details here due to space limitation. Since φ_X is a pdf, it must be a member of each set C_q given as $C_q := \{\varphi \in \mathcal{G} \mid \Pr[a_q \leq \mathbf{X} \leq b_q] = \int_{a_q}^{b_q} \varphi(x) dx \in \mathcal{P}_q\}$, which also implies that $\varphi_X \in \bigcap_{q \in \overline{1, Q}} C_q \subset \mathcal{G}$. Furthermore, φ_X must also satisfy “necessary” conditions for a pdf: $\int_S \varphi_X(x) dx = 1$ and $\varphi_X(x) \geq 0$, where $S \subset \mathbb{R}$ is the support which we assume is bounded. We denote by C_{Q+1} and C_{Q+2} , the sets of functions that satisfy the two necessary conditions, respectively, and note that we now must have that $\varphi_X \in \bigcap_{q \in \overline{1, Q+2}} C_q \subset \mathcal{G}$.

We assert without proof that all the sets defined above are closed and convex [SYY98]. Therefore, we may employ standard sequential *projection onto convex sets* (POCS) algorithm to find a point in the intersection $\varphi_X \in \bigcap_{q \in \overline{1, Q+2}} C_q$. However, in some rare cases this intersection may be an empty set, in which case the approximation might not be sufficiently close to sets C_{Q+1} and C_{Q+2} . To overcome this problem, we instead employ the parallel projection algorithm given below which guarantees convergence to a point that minimizes the weighted sum of minimum distances (in the sense of least squares) from each set. It is important to note that if $\bigcap_{q \in \overline{1, Q+2}} C_q \neq \emptyset$, both methods converge weakly to a point in the intersection.

- **Parallel Projection Algorithm**

Consider the distance $\phi(\varphi) := \sum_{q=1}^{Q+2} \beta_q \|\varphi - P_{C_q}(\varphi)\|_{\mathcal{G}}^2$. For every choice of $\varphi_{(0)} \in \mathcal{G}$ and every choice of $(q \in \overline{1, Q+2}) \beta_q > 0, \sum_{q=1}^{Q+2} \beta_q = 1$, the sequence $\varphi_{(n)}$ generated by

$$\varphi_{(n+1)} = \sum_{q=1}^{Q+2} \beta_q P_{C_q}(\varphi_{(n)})$$

converges weakly to a $\varphi^* \in \operatorname{argmin} \phi(\varphi) \in \mathcal{G} \subset L^2$ [SYY98].

We provide the details of how to calculate projections P_{C_q} in the next section. Before we proceed further, we describe some basic results pertaining to the subspace \mathcal{G} to be utilized in the next section. Denote by $G \in \mathbb{R}_+^{N \times N}$ the positive semidefinite Gram matrix with entries $(\forall i, j \in \overline{1, N}) [G]_{i,j} := \langle \kappa(\cdot, x_i), \kappa(\cdot, x_j) \rangle_{\mathcal{G}}$. For a function $h \in L^2$, define a vector-valued mapping $\xi: h \mapsto [\langle h, \kappa(\cdot, x_1) \rangle_{\mathcal{G}}, \dots, \langle h, \kappa(\cdot, x_N) \rangle_{\mathcal{G}}]^T \in \mathbb{R}^N$. The projection of h onto \mathcal{G} denoted by $P_{\mathcal{G}}(h)$ (also called the “closest point” to h in \mathcal{G}) is given as $P_{\mathcal{G}}(h) = \sum_{i=1}^N \zeta_i(h) \kappa(\cdot, x_i)$, $(\forall i \in \overline{1, N})$; $\zeta_i(h) \in \mathbb{R}$ is the i th component of $\zeta(h)$. The vector $\zeta(h)$ is the solution to $G\zeta(h) = \xi(h)$ [Lue97].

- **Calculation of Projections for Parallel Projection Algorithm**

As previously mentioned, all the required inner products (in the space \mathcal{G}) in the following have well-known closed-form solutions that are omitted due to space limitation. Let $\varphi_{(0)} := \sum_{i=1}^N w_{(i,(0))} \kappa(\cdot, x_i)$, with $(\forall i \in \overline{1, N}) w_{(i,(0))} \in \mathbb{R}$ arbitrary, in the parallel projection algorithm

above. In the following, we show how to calculate each $\mathbf{P}_{\mathcal{C}_q}(\varphi_{(n)})$ in the parallel projection algorithm.

- **Projection on Sample Sets**

As discussed above, we must have that $(\forall q \in \overline{1, Q}) \int_{a_q}^{b_q} \varphi(x) dx = \int 1^q(x) \varphi(x) dx \in \mathcal{P}_q = [\mathbf{P}_q^L, \mathbf{P}_q^H]$, where $1^q(x) = 1$ if $x \in [a_q, b_q]$, otherwise $1^q(x) = 0$. The projection $\mathbf{P}_{\mathcal{C}_q}(\varphi_{(n)})$ onto the closed-convex set $\mathcal{C}_q = \{\varphi \in \mathcal{G} \mid \langle \mathbf{P}_{\mathcal{G}}(1^q), \varphi \rangle_{\mathcal{G}} = \int 1^q(x) \varphi(x) dx \in \mathcal{P}_q\}$ is given as

$$\mathbf{P}_{\mathcal{C}_q}(\varphi_{(n)}) = \begin{cases} \varphi_{(n)} - \frac{q^q - \mathbf{P}_q^H}{\|\mathbf{P}_{\mathcal{G}}(1^q)\|_{\mathcal{G}}^2} \mathbf{P}_{\mathcal{G}}(1^q), & \text{if } q^q - \mathbf{P}_q^H > 0 \\ \varphi_{(n)} - \frac{q^q - \mathbf{P}_q^L}{\|\mathbf{P}_{\mathcal{G}}(1^q)\|_{\mathcal{G}}^2} \mathbf{P}_{\mathcal{G}}(1^q), & \text{if } q^q - \mathbf{P}_q^L < 0 \\ \varphi_{(n)}, & \text{otherwise} \end{cases}$$

where $q^q := \langle \mathbf{P}_{\mathcal{G}}(1^q), \varphi_{(n)} \rangle_{\mathcal{G}}$.

- **Normalization**

As discussed above, we must have that $\int_{-\infty}^{\infty} 1^S(x) \varphi(x) dx = 1$, where $1^S(x) = 1$ if $x \in S$, otherwise $1^S(x) = 0$. The projection $\mathbf{P}_{\mathcal{C}_{Q+1}}(\varphi_{(n)})$ onto the closed-convex set $\mathcal{C}_{Q+1} = \{\varphi \in \mathcal{G} \mid \langle \mathbf{P}_{\mathcal{G}}(1^S), \varphi \rangle_{\mathcal{G}} = \int_{-\infty}^{\infty} 1^S(x) \varphi(x) dx = 1\}$ is given as

$$\mathbf{P}_{\mathcal{C}_{Q+1}}(\varphi_{(n)}) = \begin{cases} \varphi_{(n)} - \frac{\langle \mathbf{P}_{\mathcal{G}}(1^S), \varphi_{(n)} \rangle_{\mathcal{G}} - 1}{\|\mathbf{P}_{\mathcal{G}}(1^S)\|_{\mathcal{G}}^2} \mathbf{P}_{\mathcal{G}}(1^S). \end{cases}$$

- **Non-negativity**

As discussed above, we must have that $\varphi_{\mathbf{x}}(x) \geq 0$. Let $\varphi_{(n)} = \sum_{i=1}^N v_i \kappa(\cdot, x_i)$ and $\mathbf{v} = [v_1, v_2, \dots, v_N]^T$. We omit the proof for the assertion that the projection $\mathbf{P}_{\mathcal{C}_{Q+2}}(\varphi_{(n)})$ is given as $\mathbf{P}_{\mathcal{C}_{Q+2}}(\varphi_{(n)}) = \sum_{i=1}^N w_i \kappa(\cdot, x_i)$, where $(i \in \overline{1, N}) w_i$ is the i th component of

$$\mathbf{w}^* \in \operatorname{argmin}_{\mathbf{w} \geq 0} \frac{\mathbf{w}^T \mathbf{G} \mathbf{w}}{2} - \mathbf{w}^T \mathbf{G} \mathbf{v},$$

which is a quadratic program (QP) and can be solved by any standard convex solver.

- **Complexity**

The complexity of the learning algorithm is discussed in [ACY+18]. Regarding the complexity of Parallel Projection Algorithm, the complexity is dominated by the projection required for the non-negativity. This projection is a QP which can be solved in polynomial time or better since \mathbf{G} is positive semi-definite, and the problem is convex.

8.10 Architecture optimization for Cell-less mMIMO systems

System model

We consider a CI massive MIMO system that consists of K single omni-directional antenna users that are served simultaneously by M single antenna APs. In this work, it is assumed that $K \ll M$ and that the APs are using the same time-frequency resources. APs are randomly located within a given coverage area and are managed by a central processing unit (CPU) to which they are connected through perfect back-haul links. The CPU handles part of the physical layer processing such as data coding and decoding. Let $g_k \in \mathbb{C}^{M \times 1}$, denote the complex channel vector

between user k and all the APs. Specifically, the m -th element, g_{mk} is the channel coefficient between the k -th user and the m -th AP and is modeled as follows

$$g_{mk} = \sqrt{\beta_{mk}} h_{mk}$$

Where $h_{mk} \sim CN(0,1)$, $m = 1, \dots, M$, $k = 1, \dots, K$ denote the small-scale fading coefficients while β_{mk} , $m = 1, \dots, M$, $k = 1, \dots, K$ denote the large scale fading coefficients which include the impact of path-loss and shadowing.

We also consider that the system is operating in TDD mode where the channel estimates are obtained using uplink training with orthogonal pilot sequences [NAY+17].

Improving Favorable propagation: Which users can be active simultaneously?

Favorable propagation represents an important property in large antenna systems. It refers to the mutual orthogonality between users' vector wireless channel. With favorable propagation, the overall system's performance is guaranteed to be very appealing with simple linear processing since the effect of interferences is considerably attenuated. Practically, favorable propagation cannot be exactly met, but it can be approximately achieved. This is the case when the number of APs grows large and the channels are said to provide asymptotically favorable propagation. In this case the condition of asymptotically favorable propagation can be formulated as follows

$$\frac{g_k^* g_j}{M} \rightarrow 0, \quad M \rightarrow \infty \quad \text{for } k \neq j$$

Nevertheless, a very large number of APs is neither practical nor cost-efficient. Consequently, we consider a different perspective to obtain asymptotically favorable propagation with an acceptable number of APs.

The intuition here comes from a bound on the complementary cumulative distribution function of the inner product between two given users' channel which can be stated as follows

$$P \left\{ \frac{g_k^* g_j}{M} > \theta \right\} \leq \frac{1}{1 + \frac{M^2 \theta^2}{\sum_{m=1}^M \beta_{mk} \beta_{mj}}}$$

Concretely, to improve favorable propagation, the bound above should be made as low as possible for users that are active on the same time-frequency resources (for each value of θ, M). This goal can be achieved by scheduling users on the same resources if the inner product of their large-scale fading coefficients is low. Consequently, one plausible way to reduce spatial correlation is to resort to appropriate selection of active users, which can be achieved by spatial user grouping.

Graphical Modeling and Spatial User Grouping

In the considered setting, the first step for architecture optimization is to construct a spatial correlation graph that captures the level of favorable propagation for a set of users, which are active simultaneously. More specifically, we design an undirected favorable propagation graph $G(V, E)$. The set of vertices V represents the users in the coverage area and each edge $e_{k,j} \in E$ is associated with a weight $\omega_{k,j} = \sum_{m=1}^M \beta_{mk} \beta_{mj}$, that quantifies large scale fading correlation between the two users' channels. In this case, user grouping reduces to minimizing the spatial correlation between the channels of users that belong to the same group.

The user grouping problem is formulated as the following combinatorial optimization problem

$$\begin{aligned} \max_{x_{k,c}} \quad & \sum_{c=1}^C \sum_{k \in V} \sum_{j \in V, j \neq k} \omega_{k,j} (1 - x_{k,c}) \\ \text{s. t.} \quad & \sum_{c=1}^C x_{k,c} \leq \alpha, \quad k \in V \end{aligned}$$

$$\sum_{k \in V} x_{k,c} \leq \tau, \quad c = 1, \dots, C$$

where C denotes the total number of groups and α , the maximum number of groups to which a user can belong at the same time. The above problem is NP-hard (refer to [HJA18] for proof). Since the problem is NP-hard, its global optimal solution cannot be found by mean of polynomial time solvable algorithms. Our goal is to design a low-complexity algorithm to find a local optimal solution. To do so, we will resort to semidefinite programming. We define the following change of variables

$$x_c = (x_{1,c}, \dots, x_{K,c})^T, \quad y_c = 2x_c - \mathbf{1}_K$$

$$W = \begin{bmatrix} 0 & \omega_{2,1} & \cdots & \omega_{K,1} \\ \omega_{1,2} & 0 & \cdots & \omega_{K,2} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{1,K} & \omega_{2,K} & \cdots & 0 \end{bmatrix}$$

Where $\mathbf{1}_K$ is a column vector with 1 entry. After applying this change of variables the problem above can be reformulated as follows

$$\max \quad \frac{1}{4} \sum_{c=1}^C (\varepsilon - y_c^T W y_c)$$

$$S. t. \quad \sum_{c=1}^C y_c \leq \alpha_{eq}, \quad k \in V$$

$$Tr(diag(y_c)) \leq \tau_{eq}, \quad c = 1, \dots, C$$

$$y_c \in \{-1, 1\}^K, \quad c = 1, \dots, C$$

where $\tau_{eq} = 2\tau - K$ and $\alpha_{eq} = 2\alpha - C$. We then propose to combine the semidefinite relaxation method with the *Schur complement in order to solve the above problem*. The detailed proposed algorithm can be found in [HJA18].

Bandwidth Allocation Problem

Once spatial user grouping is performed, we proceed to optimizing Bandwidth allocation to the different groups. Indeed, as favorable propagation is optimized within each group, the users from different groups need to be scheduled on orthogonal resources in order to harvest the advantages of the improved channel orthogonality. The resulting Bandwidth allocation problem can be formulated as follows

$$\max_{0 \leq \gamma_c \leq 1} \quad \sum_{c,k \in \Delta(c)} \gamma_c R_{k,c}$$

$$S. t. \quad R_k^{th} \leq \sum_{c=1}^C \gamma_c R_{k,c}, \quad k = 1, \dots, K$$

$$\sum_{c=1}^C \gamma_c \leq 1$$

Where $R_{k,c} = B \frac{T_c - |\Delta(c)|}{T_c} \log_2 \left(1 + \frac{\rho_d (\sum_{m=1}^M \vartheta_{mk})^2}{\rho_d \sum_{m=1}^M \sum_{k' \in \Delta(c), k' \neq k} \beta_{mk} \vartheta_{mk'+1}} \right)$ denotes the rate of user k within group c . R_k^{th} denotes the minimum rate requirement of user k .

The above problem is a convex linear optimization problem and the optimal solution can be found using the interior point method.

8.11 Optimised functionality placement and resource allocation in a CRAN/DRAN context

Problem statement

We consider the previously explained set F of functional entities (FEs), defined as the minimal components to which functionality can be fractionated. In addition, the corresponding graph $G(F, K)$, where each node corresponds to an FE and each edge connects interacting FEs, is provided. The computational load corresponding to each FE i is φ_i . Edges are weighted according to the amount of data transferred between the FEs $k_{ii'}$, where $i, i' \in F$.

Furthermore, we are given a system layout graph $G(S, L)$ consisting of the available server entities (SEs, logical units consisting of the CU and DUs), represented by the set S , and the communicational channels (wired or wireless backhaul links) among them, represented by the set L [TAG+12]. As a future extension, the set S could also contain central, regional and edge clouds. The computational capacity of each server entity is given as c_j , where $j \in S$. Also, elements of the set L are characterised by their average data rate $r_{jj'}$, where $j, j' \in S$.

Our objective is to assign the FEs to SEs. Let A_j , where $j \in S$, denote the set of FEs that will be assigned to SE j . We are looking for the minimum cost allocation $A = \{A_j | j \in S\}$, $A_j \subseteq F$ that satisfies a set of capacity and performance constraints, as well as the distribution of network traffic to each DU.

Problem formulation

We introduce the set of variables y_j , where $j \in S$, that take the value 1 (0) depending on whether the candidate SE j is (is not) activated. Furthermore, in order to describe the allocation of FEs to SEs, we introduce the variables x_{ij} , where $i \in F, j \in S$, that take the value 1 (0) depending on whether FE i is (is not) assigned to SE j . We also define the set of variables $z_{ii'}$, where $i, i' \in F$ that take the value 1 (0) depending on whether FEs j and j' are (are not) assigned to the same SE, calculated as:

$$z_{ii'} = \sum_{\forall j \in S} [x_{ij} * x_{i'j}] \quad (8-9)$$

The problem of obtaining A may be reduced to the following problem:

Minimize:

$$f(A, B, P, N) = w_1 * \sum_{\forall j \in S} b_j y_j + w_2 * \sum_{\forall i \in F, \forall j \in S} [x_{ij} * p_{ij}] + w_3 * \sum_{\forall i, i' \in F} [(1 - z_{ii'}) * n_{ii'}] \quad (8-10)$$

Subject to:

$$\sum_{\forall j \in S} x_{ij} = 1, \forall i \in F \quad (8-11)$$

$$\sum_{\forall i \in F} [x_{ij} * \varphi_i] \leq c_j, \forall j \in S \quad (8-12)$$

$$\sum_{\forall i, i' \in F} [x_{ij} * x_{i'j'} * k_{ii'}] \leq r_{jj'}, \forall j, j' \in S \quad (8-13)$$

The cost function used to evaluate each solution is given in (8-10). Each of its terms represent one of the cost factors mentioned in 4.2.2. In order to enhance the algorithm performance, they are all normalized [SK06]. The corresponding weights, as well as the QoS requirements are dependent on the use case.

Constraint (8-11) ensures that each functionality is allocated to exactly one resource, while (8-12) and (8-13) guarantee that the capacity of resources and links respectively is respected. This capacity is calculated in order to ensure the given QoS requirements.

Solution approach: Simulated Annealing

The Metropolis criterion is used for efficient candidate generation. In each iteration, the potential next steps are given by a set of candidate moves. Three types of moves are considered:

- A step towards centralization of one branch (path from the CU to one DU).
- A step towards decentralization of one branch.
- Offloading a portion of traffic from one branch to another.

The initial temperature, cooling rate and number of iteration to reach equilibrium in each temperature are fine tuned for each configuration, namely number of DUs, traffic intensity, computational and data transfer load and power etc.). As an ending criterion, the number of temperatures for which the best solution has not changed must exceed a threshold and the candidate next step acceptance ratio must be below 2%.

8.12 Centralized and distributed multi-node schedulers for non-coherent joint transmission

The main difference between centralized/distributed multi-node schedulers in the context of NR-CoMP is the level of information exchange between TRPs, e.g., scheduling information, CSI measurements, control signalling, etc. This is shown in Figure 8-8 (left). Ideally, centralized scheduler can be optimum. However, in practice, overhead introduced by centralized scheduler is the main factor limited its performance. As a result, distributed schedulers are also widely used. Figure 8-8 (right) shows an example of PRB allocations of DPS/F-NCJT/NF-NCJT. DPS chooses one TRP within the CoMP set for the target UE, blanking others. For F-NCJT, multiple TRPs transmit on the same PRBs to the target UE without data exchange among the CoMP set. For NF-NCJT, multiple TRPs individually transmit to the target UE without data exchange among the CoMP set.

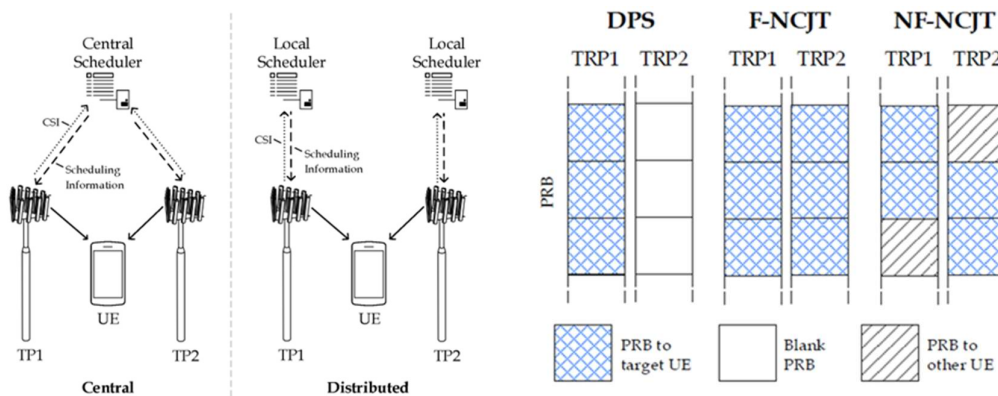
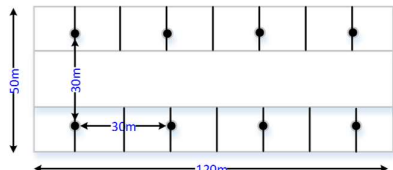


Figure 8-8 Centralized and distributed schedulers (left) and examples of PRB allocations of DPS/F-NCJT/NF-NCJT (right).

An investigation into the performance difference of interlayer interference mitigation (F-NCJT) versus the possible gain achievable from flexible scheduling (NF-NCJT) is performed using a

system level simulator. The simulation is performed in an indoor hotspot scenario with assumptions listed in Table 8-4. Results suggest that even distributed schedulers can perform well comparing to centralized schedulers.

Table 8-4. Key evaluation assumptions.

Attributes	Values or assumptions
Layout	<p>Indoor TP: Number of TPs: N=8, per 120m x 50m</p> 
Carrier Frequency	3.5 GHz
Bandwidth	10MHz (50RBs)
Subcarrier Spacing	15kHz
Channel model	TR 36.814 Indoor Hotspot with 3D distance
TP antenna configuration (M,N,P)	ULA with M=1, N=1, P = 2 with polarization Model -2 from TR 36.873
TP Tx power	24dBm
TP antenna pattern	2D omni with 5dBi gain (According to TR 36.814)
TP antenna height	6m
UE antenna height/UE dropping	1.5m, uniform
Maximum CoMP measurement set size	Baseline 3TPs.
UE antenna gain	According to TR 36.873
UE receiver noise figure	9 dB
Traffic model	Non- full buffer FTP traffic model 1, S = 0.5Mbytes
UE receiver	MMSE-IRC
UE antenna	4Rx, 0°/90° polarization slants, 0.5 wavelength spacing with polarization Model -2 from TR 36.873
Feedback assumption	CQI, PMI and RI reporting delay is 5 ms
Transmission mode	TM10 based
Channel estimation	Realistic, $\tilde{\mathbf{H}} = \mathbf{H} + \alpha\mathbf{N}$ where α is the power of interference plus noise

8.13 NR duplexing with CRAN and network coordination

The philosophy of the proposed UE-to-UE CLI management technique is that the downlink UE regards the uplink UE as another interfering TRP. This requires certain signalling exchange between the downlink and UL TRPs. The general procedure is depicted in Figure 8-9. A message is sent from the UL TRP to the DL TRP, informing the DL TRP about the configurations of UL DMRS of the interfering UE. The configurations of the UL DMRS may include the locations of the UL DMRS, cyclic shift of the UL DMRS sequence, orthogonal sequence, etc. The DL TRP can then assign ZP CSI resources for the desired UE to measure the channel information of the interfering UE. An optional message from the DL TRP to the desired UE can be sent, for the purpose of estimating the complex channel coefficients at the desired UE side. Then, the interfering UE will send UL signals to the UL TRP and corrupt the DL signal of the desired UE. The desired UE may then measure interference at the ZP CSI resources and perform advanced receiver algorithms such as interference rejection combining receiver to decode its DL signal. The simulation settings of the UE-to-UE CLI management are listed in Table 8-5.

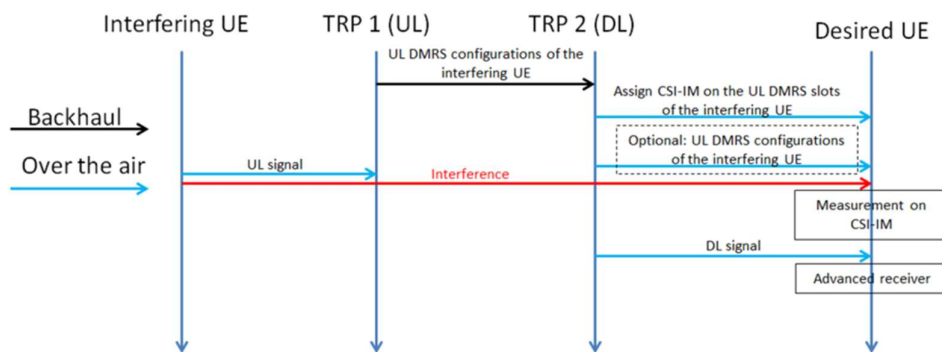


Figure 8-9 Procedure of the proposed UE-to-UE CLI management.

Table 8-5. Key evaluation assumptions for NR duplexing.

Attributes	Values or assumptions
	Indoor TP: Number of TPs: N=8, per 120m x 50m
Layout	
Carrier Frequency	3.5 GHz
Bandwidth	10MHz (50RBs)
Subcarrier Spacing	15kHz
Channel model	TR 36.814 Indoor Hotspot with 3D distance
TP antenna configuration (M,N,P)	ULA with M=1, N=1, P = 2 with polarization Model -2 from TR 36.873; Ntx=4
TP Tx power	24dBm
TP antenna pattern	2D omni with 5dBi gain (According to TR 36.814)
TP antenna height	6m
UE antenna height/UE dropping	1.5m, uniform
UE antenna gain	According to TR 36.873
UE receiver noise figure	9 dB
Traffic model	Non- full buffer FTP traffic model 1, S = 0.5Mbytes, $\Lambda_{DL}=1/s$, $\Lambda_{UL}=1/s$
UE receiver	MMSE-IRC
UE antenna	2Rx, 0°/90° polarization slants, 0.5 wavelength spacing with polarization Model -2 from TR 36.873
Feedback assumption	CQI, PMI and RI reporting delay is 5 ms
Transmission mode	TM10 based
Channel estimation	Realistic, $\tilde{\mathbf{H}} = \mathbf{H} + \alpha\mathbf{N}$ where α is the power of interference plus noise