INVITED PAPER

Next-Generation Multiple Access: From Basic Principles to Modern Architectures

By Eduard Axel Jorswieck, *Fellow IEEE*

ABSTRACT | The pressure to develop new network architectures and multiple access technologies is driven by increasing demands on network performance, number of devices, network traffic, and use cases. Recent advances in open radio access networks (RANs) with open interfaces and software-defined network functionalities allow adaptability in terms of medium access control and physical layer, but also flexibility in terms of network architectures. The aim of this tutorial is to provide a comprehensive overview of the current set of network architectures for wireless access together with next-generation multiple access technologies. It starts with the classical models for multiple access channel (MAC), broadcast channel (BC), and interference channel (IC) from network information theory and derives the fundamental results on capacity regions and their coding and signal processing schemes. Extensions to multicarrier, multiantenna, and multicell scenarios are discussed. The evolution from orthogonal to spatial-division multiple access (SDMA), nonorthogonal multiple access (NOMA), and rate splitting multiple access (RSMA) techniques and their performance guarantees are carefully explained. Recent advances toward multiconnectivity, cloud-RAN (C-RAN), and cell-free multiple access (CFMA) are explained. The data rate benefits of an anecdotal open RAN network are developed and the corresponding user data rates are calculated. Massive random and grant-free access schemes are also discussed. The tutorial concludes with a list of open research questions.

Manuscript received 29 February 2024; revised 16 May 2024; accepted 3 June 2024. This work was supported by the Federal Ministry of Education and Research (BMBF), Germany, through the Program of Souverän, Digital, and Vernetzt Joint Project 6G-RIC under Grant 16KISK031.

The author is with the Institute for Communications Technology, Technical University of Braunschweig, 38106 Braunschweig, Germany (e-mail: e.jorswieck@tu-braunschweig.de).

Digital Object Identifier 10.1109/JPROC.2024.3412423

KEYWORDS | Broadcast channel (BC); interference channel (IC); medium access control; multiple access channel (MAC); multiuser interference; network information theory; wireless communications.

NOMENCLATURE

AI	Artificial intelligence.
AOI	Age of information.
AWGN	Additive white Gaussian noise.
BBU	Baseband unit.
BC	Broadcast channel.
BPCU	Bits per channel use.
BS	Base station.
C-RAN	Cloud-radio access network.
CA	Collision avoidance.
CD	Collision detection.
CDMA	Code-division multiple access.
CFMA	Cell-free multiple access.
COMP	Cooperative multipoint.
CP	Cyclic prefix.
CPU	Central processing unit.
CSI	Channel state information.
CSMA	Carrier sense multiple access.
CU	Central unit.
DAB	Digital audio broadcast.
DAMA	Demand assigned multiple access.
DCC	Dynamic cooperation cluster.
DPC	Dirty paper coding.
DU	Distributed unit.
DVB	Digital video broadcast.
EMBB	Enhanced mobile broadband.
FDMA	Frequency-division multiple access.
FFT	Fast Fourier transform.
GFMA	Grant-free multiple access.

HSPA	High-speed packet access.
IC	Interference channel.
ICI	Intercell interference.
IID	Independent and identically distributed.
IoT	Internet of Things.
JT	Joint transmission.
LSA	Licensed shared access.
LTE	Long-term evolution.
MAC	Multiple access channel.
MEC	Mobile edge cloud.
MIMO	Multiple-input multiple-output.
MISO	Multiple-input single-output.
ML	Machine learning.
MMTC	Massive machine-type communication.
MRT	Maximum ratio transmission.
NFV	Network function virtualization.
NOMA	Nonorthogonal multiple access.
OFDM	Orthogonal frequency-division multiplexing.
OFDMA	Orthogonal frequency-division multiple access.
OMA	Orthogonal multiple access.
OTFS	Orthogonal time–frequency–space.
P2P	Point-to-point.
PAPR	Peak-to-average power ratio.
PDCP	Packet data convergence protocol.
PRB	Physical resource block.
PUPE	Per-user error probability.
RACH	Random access channel.
RAN	Radio access network.
RE	Resource element.
RIC	RAN intelligent controller.
RLC	Radio link control.
RRC	Radio resource control.
RS	Rate splitting.
RSMA	Rate splitting multiple access.
RU	Radio unit.
SC	Superposition coding.
SCMA	Sparse code multiple access.
SD	Successive decoding.
SDAP	Service data adaptation protocol.
SDMA	Spatial-division multiple access.
SDN	Software-defined networking.
SFN	Single-frequency network.
SIC	Successive interference cancellation.
SINR	Signal-to-interference and noise ratio.
SNR	Signal-to-noise ratio.
SVD	Singular value decomposition.
TDD	Time-division duplex.
TDMA	Time-division multiple access.
TIN	Treating interference as noise.
TS	Time sharing.
UAV	Unmanned aerial vehicle.
UMAC	Unsourced multiple access channel.
URLLC	Ultrareliable low-latency communications.
V2X	Vehicle-to-everything.
VNF	Virtual network function.
WLAN	Wireless local area network.
WSN	Wireless sensor network.
ZF	Zero forcing.

I. INTRODUCTION

Wireless communication has become the enabling technology for almost all digital technologies. The application domain has expanded beyond classical cellular communications to a wide range of applications in WSNs, IoTs, eHealth, V2X, UAVs, and body area networks, to name a few. As the resources for wireless communication are limited, i.e., spectrum, energy, time, and space, the challenge of multiple access has recently gained significant attention and momentum. A discussion on network architectures has also started [1].

The digital landscape continues to develop and evolve, with more services and use cases relying on the availability of fresh and reliable data. The current development of 6G and beyond wireless communication technologies is tailored to deliver increased speed, throughput, reliability, and ubiquitous connectivity. In [2], the limitations of 5G key performance indicator (KPI) are identified and targets for future research challenges are identified. Global coverage, including space–air–ground–sea, should be provided, with peak data rates up to terabits per second and user-experienced data rates above 1 Gb/s. Latencies below 1 ms are targeted. Scalability of connection density is up to 10^8 devices/km^2 and high positioning accuracy. The target for reliability (including security) is 99.99999%. The energy efficiency of the network should improve by a factor of 100 to 10^9 bits/J. In addition, computational capabilities, including ML support, are ubiquitously available at the mobile edge as well as on small IoT devices. The International Mobile Telecommunications 2030 (IMT-2030) Promotion Group has published a new recommendation [3] that identifies both enhanced capabilities for IMT-2030 and new capabilities for IMT-2030.

In terms of application scenarios, 5G started with the three main scenarios EMBB, MMTC, and URLLC. Driven by 6G operator and industry alliances, a number of new scenarios are emerging in different areas such as personal immersive applications, robotic automation, and remote data collection. 6G will continue to enhance and expand the above application scenarios to achieve further-embb (feMBB), ultra-mMTC (umMTC), and enhanced-uRLLC (euRLLC) [2]. Typical applications for each scenario are listed as follows—feMBB: cloud working, 3-D ultrahigh definition (UHD), extended or cross reality (XR), and holographic communication; umMTC: smart building, smart city, the IoT, and networked and autonomous systems; and euRLLC: self-driving car, teleoperations, missing critical applications, V2X, and tactile Internet. Clearly, there are tradeoffs between the conflicting KPIs. The more stringent the requirements, the more important it is to find appropriate tradeoff curves and corresponding Pareto bounds, e.g., for spectral versus energy efficiency [4].

To meet the enhanced and more stringent requirements envisaged for 6G and beyond, new technologies are being explored, including THz frequency bands, large smart surfaces, nonterrestrial networking (NTN), ML, visibile light communication (VLC), and quantum communications [5]. These new techniques provide the flexibility needed to address the conflicting KPIs and achieve efficient tradeoff limits. The more devices that are active, the more heterogeneous the service requirements, the more ways to schedule and allocate resources, and the more important multiple access schemes become. We need to operate communication systems close to or at the edge of the maximum achievable rate regions. Wireless access and fronthaul and midhaul links should be optimized together to avoid bottlenecks. This is why we are focusing on advanced next-generation multiple access techniques.

Modern MAC techniques will need to operate efficiently, flexibly, adaptively, and holistically, both from a cross-layer perspective and with respect to the underlying network architecture. Efficient operation includes resource efficiency with high spectral and energy efficiency as well as operational efficiency. Flexibility means that configuration parameters cover a wide operational range and support different use cases and scenarios. Adaptivity means that the network as a whole can sense and adapt to the environment. This implies robustness and resilience. For MAC, the adaptivity with respect to the information available at the transmitters and receivers is very important because it allows to achieve the high efficiency required. Holistic design means that models and approaches for system optimization must include all technological layers of the protocol stack. An example is the separation of source and channel coding in multiuser networks, which usually leads to suboptimal rates and results [6]. Holistic also refers to the joint design of the wireless access and the fronthaul and backhaul networks, as separation there also induces strong suboptimality [7].

The following trends for modern MAC networks are identified.

1) *Nonorthogonality:* Classical orthogonal resource allocation is not sufficient to support the increasing number of devices and their service requirements. SDMA, NOMA, and RSMA are techniques that establish scalable and efficient nonorthogonal resource allocations [8].
2) *Fundamental limits:* The solid foundation for system design is provided by network information theory. In [9], it is emphasized that the potential of (network) information theory needs to be unleashed to improve the efficiency and performance of communications, protocols, and hardware platforms.
3) *Synergies of multiple access techniques and network architectures:* The challenges of modern wireless communications networks will not be solved by a single multiple access technique but will require a combination of techniques carefully designed to exploit synergies [10].

The aim of this tutorial is to provide a comprehensive and accessible introduction to modern MAC techniques, taking into account the above observations and trends.

A. Scope of the Tutorial

The purpose of the tutorial is threefold: to provide background and concepts for physical layer coding and decoding schemes as elements of more complex network architectures. The MAC for modeling the uplink, the BC for modeling the downlink, and the IC for multicell or cell-less networks will be reviewed and their properties highlighted. Most multiple access scenarios will involve multiple carriers, multiple antennas, and multiple cells. Therefore, the peculiarities of multiuser network schemes are discussed in combination with spectral, spatial, and multicell interference scenarios. The corresponding resource allocation problems are briefly introduced and their solution structures explained. In particular, the MIMO BC and its capacity region require special attention since the capacity region is achieved by a combination of precoding, TS, beamforming, and power allocation. Models for multicell networks are discussed in more detail, and the underlying signal processing model and its resulting SINR expressions are highlighted. The framework described is general enough to be specialized to almost all MAC schemes considered in Sections IV and V. An overview of classical and more recent coordinated multiple access techniques shows how the underlying information-theoretic results are taken up. For the scenarios where capacity is known, the corresponding achievable rate regions can be compared with the ultimate capacity regions. A recent development in the research and standardization of cellular networks is the move toward grant-free access. It is interesting to see that WLANs and cellular networks are finally converging in the sense that WLAN is moving toward coordinated access schemes, while cellular is introducing grant-free access with uncoordinated initial access and corresponding collisions. The motivation behind both moves seems diametrically opposed; coordinated access can improve reliability by avoiding collisions, and grant-free or random access can improve latency by eliminating access request rounds. The interesting challenge is how to support both low latency and ultrahigh or even extreme reliability [11]. To improve the accessibility of the tutorial, we provide a running example of an open RAN architecture but avoid showing numerical simulation results of different network settings. This makes it easier for the reader to learn the communication theory concepts and methods, but we cannot explicitly show scalability with our small-scale anecdotal example.

This tutorial does not provide a complete summary of recent network information-theoretic results on capacity regions or the best-known inner and outer bounds for various network architectures. We limit our attention to random P2P codes because they are mature enough and achieve good performance. The exposition is based on the first-order capacity and rate expressions, i.e., Shannon rates, which require asymptotic block lengths. The finite block length rates [12] and their multiuser results [13] are not covered in this tutorial.

B. Related Tutorials

Exactly 30 years ago, the invited paper [14] on multiple access in wireless digital networks appeared in the PROCEEDINGS OF THE IEEE. At the time, preparations were underway for the third generation of mobile communications and the focus was on CDMA as the multiple access technique. Interestingly, the invited paper argues that the Advocates of Linux Open-source Hawaii Association (ALOHA) random access protocol and coordinated CDMA are different ways of looking at the same basic signals. In so-called DAMA architectures, a separate channel, called the request channel, is used by individual users to request capacity when needed. Since the protocol should scale with the number of users, random access was proposed in the DAMA. This early related tutorial shows that the discussions on grant-free and random access versus coordinated access are still relevant. In the current tutorial, we focus on coordinated access because it is the dominant type of access technique in cellular 3G-to-5G networks.

From a similar time but different perspective, Shamai and Wyner [15] and Somekh and Shamai [16] consider multiple access schemes in cellular setups and focus on wideband, TDMA, and CDMA. The approach is very similar to our tutorial in that it starts with achievable rates for specific cellular models, which complete the achievability proofs in the appendixes. The current tutorial covers more multiple access techniques, including SDMA, NOMA, RSMA, and CFMA.

A tutorial on multiple access technologies for beyond 3G mobile networks is provided in [17]. TDMA, CDMA, FDMA, combinations, and variants thereof are discussed. In particular, the combination of CDMA and FDMA, called multicarrier CDMA, is highlighted. However, the paper also includes SDMA with a list of potential capabilities but also requirements. It has a similar approach to our tutorial in that it provides a solid mathematical signal model as a basis for fair and sustainable comparisons and developments.

Jumping forward in time, the review article [18] provides a tutorial on NOMA for 5G and beyond. The main focus of the tutorial is the comparison between OMA and NOMA, which was important to highlight the benefits and gains of NOMA. Some combinations of NOMA with multiple-antenna systems are also illustrated. However, the scope of this article is limited to NOMA and its variants compared to classical OMA.

The review paper [19] considers the joint design of wireless resource allocation, edge computing, and storage capabilities of the MEC, and it provides a holistic view of MEC technology and its potential use cases and applications. A related survey paper [20] reviews multiple access schemes such as OMA, NOMA, OFDMA, and delta OMA (D-OMA). The focus is on variants of NOMA, including power and code domains. The schemes are compared using numerical simulations.

The tutorial [21] explains another multiple access scheme based on CDMA, called SCMA, where multiple users in SCMA are separated by assigning unique sparse codebooks. The basic principles of SCMA are highlighted, as well as several promising research directions.

The two tutorial and survey papers [22], [23] provide a comprehensive overview of RSMA, including comparisons with OMA, NOMA, and SDMA. The potential of RSMA for future mobile networks is highlighted.

Finally, the recent magazine paper [24] examines massive access and the fundamental challenges and gives an overview of the concept of massive access wireless communication and the current research on this important topic. For our tutorial, grant-free and massive random access are two side issues because we focus on coordinated multiple access.

C. Structure of the Tutorial

The tutorial is divided into three main parts, starting with the basics of network information theory for multiple access networks, followed by coordinated multiple access variants, and finally modern multiple access techniques.

Section III on the basic principles of network information theory highlights those techniques that are relevant and used in current and modern multiple access techniques. We start with the P2P channel, first without and then with interference and SIC. The main observations and results are highlighted and marked in the framed text. Then, the P2P channel with RS is explained. This may seem redundant and is not very common. However, for later applications within CFMA, this concept turns out to be very useful. Next, we show the capacity region of the classical MAC and the degraded BC and explain relevant results for the IC. These three basic network elements are the building blocks of the complex network architectures studied in the following sections. This section concludes with unifications and extensions to multicarrier, multiantenna, and multicell systems.

Section IV relates the basic information-theoretic results to current coordinated multiple access schemes. First, TDMA, OFDMA, CDMA, SDMA, and NOMA are described. The very recent extension of GFMA concludes this section. The main focus is on the newer multiple access variants, SDMA, NOMA, and GFMA. In SDMA, the concept of massive MIMO is introduced. In NOMA, different variants in multicarrier and multicell environments are discussed. While most historical references are omitted, certain survey papers are mentioned where further information and references can be found.

Section V on modern MAC techniques first introduces the open RAN architecture because it is universal and flexible to compare modern MAC schemes, including multiconnectivity, C-RAN, and CFMA. We conclude this section with a brief outlook on UMAC. An anecdotal example of an open RAN network with minimal configuration is used to illustrate the tradeoffs and achievable performance. Upper bounds as well as achievable rates with orthogonal, nonorthogonal, dual connectivity, and cell-free are computed and compared.

Please note that there is a very large body of work on modern MAC techniques and architectures, including NOMA and CFMA. It is not possible to mention all of this work and provide a complete bibliography. The selection of references in this article is subjective and does not claim to be complete. In order to keep the number of references reasonable, the classical information theory references on the elements of network information theory are only provided partly, but the book of the same name [25] from 2011 is cited instead. The book contains an extensive bibliography.

D. Notation

The notation of the tutorial paper is kept simple, and mathematical expressions and formulations are only mentioned if necessary to understand the underlying concepts. For the data rate expressions, we always use the logarithm with respect to base 2 in order to get the unit in bits/s. For notational convenience, we introduce the function $C(x) = \log(1+x)$. The variance of the noise is denoted by σ^2. The wireless channels are either described in terms of the channel coefficient $h \in \mathbb{C}$ with attenuation $|h|$ and phase $\arg h$ or in terms of the channel gain $|h|^2$. For the multiuser description, we need the set of users $\mathcal{K} = \{1,2,3,\ldots,K\}$ and subsets thereof $\mathcal{S} \subseteq \mathcal{K}$, where $\mathcal{S}^c$ denotes the complement of the set $\mathcal{S}$ given by $\mathcal{S}^c = \mathcal{K} \setminus \mathcal{S}$. The union of two sets is $\mathcal{A} \cup \mathcal{B}$. For TS, we need a linear combination of terms, where the weights are denoted by $0 \leq \lambda \leq 1$ with $\bar{\lambda} = 1 - \lambda$. The plus operation is $[a]^+ = \max(0,a)$. For the multiple-antenna sections, we need matrices denoted by capital boldface $\boldsymbol{H}$, vectors denoted by small boldface $\boldsymbol{h}$, determinant denoted by $\det(\boldsymbol{A})$, and diagonal matrix denoted by $\text{diag}(a_1,\ldots,a_n)$. The identity matrix is denoted by $\boldsymbol{I}$ and sometimes with index denoting the size $\boldsymbol{I}_r$. The matrix $\boldsymbol{A} \succeq 0$ is positive semidefinite, i.e., all eigenvalues are greater than or equal to zero. The conjugate transpose is denoted by $\boldsymbol{h}^H$. For different coding and decoding orders, the permutation π is a mapping from an index k to another index l, i.e., $l = \pi(k)$.

II. BASIC PRINCIPLES FROM INFORMATION THEORY

Many of the basic results in this section date back to the first 50 years of information theory. The excellent survey in [26] includes some of these fundamental results. The book [25] on network information theory covers most of the multiuser channel results listed in this section. The selection of results and their presentation are tailored to the specific needs of this tutorial. The channel codes considered in this section are limited to the class of P2P channel codes [27], as they are sufficient to describe the achievable rate regions results [28].

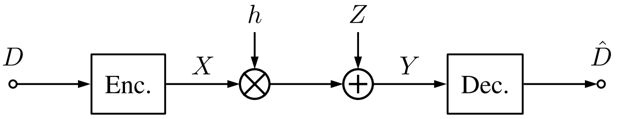

Fig. 1. *P2P channel model with encoder, decoder, fading, and AWGN.*

A. Point-to-Point Channel

As a warm-up, we start the tutorial with the P2P single-input–single-output channel, as shown in Fig. 1. The message D is encoded into codewords X and sent into the channel. The input signal is multiplied by a channel coefficient $h \in \mathbb{C}$, whose absolute value corresponds to the attenuation and phase corresponds to the delay. We assume that the channel coefficient stays the same for the whole duration of a codeword. It is a quasi-static channel model. Then, AWGN Z is added to the received signal, corresponding to the thermal noise of the low-noise power amplifiers in the receiver. Finally, the decoder tries to obtain an estimate $\hat{D}$ of the message D.

In this tutorial, we will always assume that the receiver has perfect CSI. This is motivated by the fact that the CSI can be obtained relatively easily using pilot signals and channel estimation. In this case, the capacity of the P2P channel is simply described by the well-known Shannon capacity formula

$$C_{\text{p2p}} = \log\left(1 + \frac{|h|^2 \cdot P}{\sigma^2}\right) = C\left(\frac{|h|^2 \cdot P}{\sigma^2}\right) \tag{1}$$

where $C(x) = \log(1+x)$ is the capacity function, the upper term in the fraction $|h|^2 \cdot P$ corresponds to the useful received signal power, while the lower part corresponds to the noise variance σ^2 of the AWGN. The capacity in (1) is achieved by a P2P Gaussian codebook at the transmitter with transmit power P [25].

1) Point-to-Point Channel With Interference: In order to prepare for the upcoming multiuser scenarios, let us add another additive interference term corresponding to interference to Fig. 1 (as shown in Fig. 2).

The additional interference I is a random variable, similar to the AWGN Z, which models one or multiple interfering signals. With the additional interference I, the following rate is achievable:

$$R_{\text{p2p}}^{\text{TIN}} \leq C\left(\frac{|h|^2 \cdot P}{\sigma^2 + \sigma_I^2}\right) \tag{2}$$

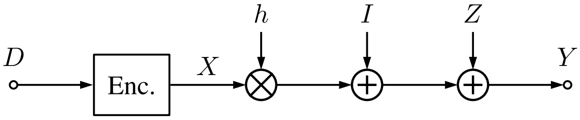

Fig. 2. *P2P channel model with encoder, fading, interference, and AWGN. The decoder is omitted as two variants are discussed.*

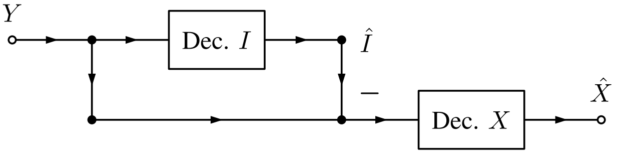

Fig. 3. *SIC performed at the receiver in the P2P with interference. First interference I is decoded and subtracted, and then, X is decoded. Each decoding operation leads to one rate constraint.*

where σ_I^2 is the interference power. The rate in (2) is achieved by a P2P Gaussian codebook at the transmitter with transmit power P and realized for the worst case interference distribution, which is Gaussian with variance σ_I^2 [29]. In this setup, the receiver simply treats the interference as additional noise and labels the scheme TIN. This is the optimal strategy when the structure of the interference at the receiver is not known or when the receiver is unable to decode the unknown interference.

Another option at the receiver is to decode the interference before decoding the signal of interest. This scheme is called SIC or SD and is an important technique in multiuser scenarios. The operation at the decoder is illustrated in Fig. 3.

If the decoding of the interference is successful, then the achievable rate increases and it is given by

$$R_{\text{p2p}}^{\text{sic}} \le C\left(\frac{|h|^2 \cdot P}{\sigma^2}\right) = C_{\text{p2p}} \qquad (3)$$

which corresponds to the channel capacity from (1). The interference term I in the denominator has simply disappeared. However, this only works if the decoder can successfully decode the interference. This means that the data rate of the transmission of the interference signal is less than the achievable rate at the receiver, which treats the intended signal as noise. If we assume that the Gaussian codebook for the interference has a rate of R_I, then SIC is feasible if the rate condition below is met [30]

$$R_I \le C\left(\frac{\sigma_I^2}{\sigma^2 + |h|^2 \cdot P}\right). \qquad (4)$$

Note that in the fraction within the logarithm in (4), the roles of the intended signal and the interference have been reversed because, in SIC, the task is to decode the interference first while treating the signal of interest as additional noise.

The important observation in (4) is that there is a condition for a receiver to decode the interference. There are several cases where the condition in (4) can be simplified, e.g., if the interference is caused by a common transmitter (as in BC) and the interference contains the data for a weaker receiver [which has a received SNR smaller than the receiver considered in (3)], then the condition in (4) is automatically met. This sometimes leads to the myth that *interference can be canceled if it is stronger than the signal*.

However, this is not the case if the interference is caused by another transmitter, e.g., in the IC or a multicell system. Then, the rate R_I of the interference is important and must be considered. In some cases, it is possible to choose the rate R_I because the transmitters know that SIC will be performed at some receivers. Then, the rate R_I can be reduced to meet the condition in (4). This is the price of SIC, which is usually another additional rate constraint (4) that must be met.

SIC applied at the receiver can improve the achievable data rate. However, this is only possible if an additional condition on the rate of the interferers is fulfilled.

This is the first basic result and building block we need to develop the MAC, BC, and IC results.

B. Point-to-Point Channel With Rate Splitting

The second important observation is obtained when we consider the so-called RS approach for a single P2P channel. Equipped with the result of SIC, we could come up with the following idea illustrated in Fig. 4.

The data stream D with rate R is split into two data streams D_1 and D_2 with lower rates R_1 and R_2. Obviously, $R_1 + R_2 = R$. Each data stream is encoded separately into codewords X_1 and X_2. These are superimposed into the codeword $X = (p)_1^{1/2} X_1 + (p)_2^{1/2} X_2$ and sent into the channel. The transmit power is split between the two codewords $p_1 + p_2 = P$.

At the receiver side, the two superimposed codewords are decoded using SIC or SD as shown in Fig. 3 (replacing I with X_1 and X with X_2). As we have emphasized, this SIC operation leads to *two* rate constraints. The rate constraints depend on the decoding order. On the one hand, if X_1 is decoded first, treating X_2 as noise, the following rate constraints are obtained:

$$R_1^{1\to 2} \le C\left(\frac{|h|^2 p_1}{\sigma^2 + |h|^2 p_2}\right)$$
$$R_2^{1\to 2} \le C\left(\frac{|h|^2 p_2}{\sigma^2}\right). \qquad (5)$$

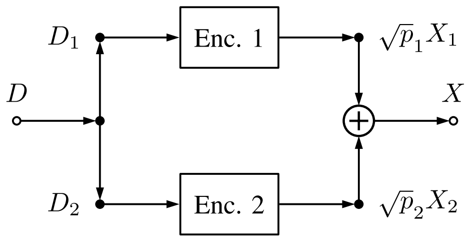

Fig. 4. *RS for the P2P channel. The data are split into two data streams with rates R_1 and R_2. Both are encoded separately in Enc. 1 and Enc. 2, and superimposed and sent into the channel.*

On the other hand, if X_2 is decoded first, treating X_1 as noise, the following rate constraints are obtained:

$$\begin{aligned} R_2^{2\to1} &\le C\left(\frac{|h|^2 p_2}{\sigma^2 + |h|^2 p_1}\right) \\ R_1^{2\to1} &\le C\left(\frac{|h|^2 p_1}{\sigma^2}\right). \end{aligned} \tag{6}$$

The total achievable rate is simply the sum of the two rates. It is easy to check that for both decoding orders, the total achievable rate is equal to the P2P channel capacity, i.e., for decoding order 1 → 2

$$\begin{aligned} R^{1\to2} &= R_1^{1\to2} + R_2^{1\to2} \\ &= \log\left(\sigma^2 + |h|^2 p_1 + |h|^2 p_2\right) - \log\left(\sigma^2 + |h|^2 p_2\right) \\ &\quad + \log\left(\sigma^2 + |h|^2 p_2\right) - \log\left(\sigma^2\right) \\ &= C\left(\frac{|h|^2 (p_1 + p_2)}{\sigma^2}\right) = C\left(\frac{|h|^2 P}{\sigma^2}\right). \end{aligned} \tag{7}$$

This scheme is called RS. It is also known as the broadcast approach in communication networks [31]. Broadcasting targets multiple receivers with different receive powers and SNR. The idea is that receivers with better receive power can decode more layers of the superimposed codewords, while weaker receivers only decode a few.

> For a single-user P2P channel, RS applied at the transmitter and SIC performed at the receiver achieve the capacity of the P2P channel.

In other words, instead of using a single Gaussian codebook, the transmitter could split the message into any number of messages encoded at different rates and perform RS, and at the receiver, SIC is applied to achieve the capacity of the P2P channel. In a simple single-user setup, this scheme only increases complexity with the number of encoders and decoders. In addition, the SIC can lead to error propagation in practice. In multiuser scenarios, this approach can easily improve the achievable rates, as explained in the next sections.

C. Multiple Access Channel

The system model for the two-user MAC is shown in Fig. 5. Each user has an independent message D_1 and D_2, which are separately encoded into two codewords X_1 and X_2. The senders choose their transmit power p_1 and p_2. In the MAC, there are usually no total power constraints. Both codewords undergo fading h_1 and h_2 corresponding to the channels from the two transmitters to the receiver. Both signals and AWGN are summed and enter the decoder. The decoder's task is to estimate the two messages $\hat{D}_1$ and $\hat{D}_2$.

Comparing Fig. 5 with the P2P channel with RS, some similarities can be observed. The main difference is that the two codewords have different fading h_1 and h_2. Another difference is that the two messages belong to two different users, and therefore, their rates R_1 and R_2 are conflicting in the sense that an increase in one user's rate could lead to a decrease in the other user's rate. Therefore, the performance metric is not a single rate R, but a so-called achievable rate region consisting of tuples $(R_1, R_2) \in \mathbb{R}_+^2$.

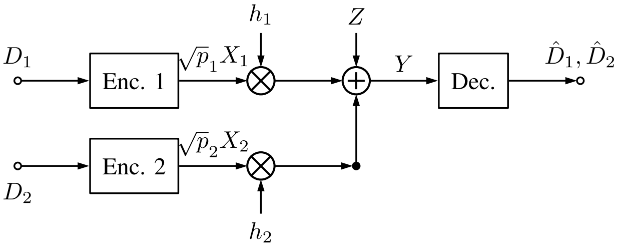

Fig. 5. *Two-user fading MAC with AWGN.*

Since the received signal is a superposition of two faded codewords and AWGN, one approach to decoding is to use SIC. The receiver can choose the decoding order 1 → 2 or 2 → 1. The achievable rates are calculated using the approach described for the P2P channel as follows. First, for the decoding order 1 → 2

$$\begin{aligned} R_1^{1\to2} &\le C\left(\frac{|h_1|^2 p_1}{\sigma^2 + |h_2|^2 p_2}\right) \\ R_2^{1\to2} &\le C\left(\frac{|h_2|^2 p_2}{\sigma^2}\right) \end{aligned} \tag{8}$$

and next for the decoding order 2 → 1

$$\begin{aligned} R_1^{2\to1} &\le C\left(\frac{|h_2|^2 p_2}{\sigma^2 + |h_1|^2 p_1}\right) \\ R_2^{2\to1} &\le C\left(\frac{|h_1|^2 p_1}{\sigma^2}\right). \end{aligned} \tag{9}$$

One interesting observation is that the sum rate for both decoding orders is the same

$$\begin{aligned} R_1^{1\to2} + R_2^{1\to2} &= R_1^{2\to1} + R_2^{2\to1} \\ &\le C\left(\frac{|h_1|^2 p_1 + |h_2|^2 p_2}{\sigma^2}\right). \end{aligned} \tag{10}$$

From (8) and (9), it is clear that the two decoding strategies lead to two different rate pairs $(R_1^{1\to2}, R_2^{1\to2})$ and $(R_1^{2\to1}, R_2^{2\to1})$. The use of TS allows to achieve linear combinations of any two achievable rate tuples, in particular the two from (8) and (9). The corresponding achievable rate region is illustrated in Fig. 6.

The achievable rate region in Fig. 6 can be characterized by the following three inequalities obtained from (8) to (10). They correspond to the three half-planes in Fig. 6, the intersection of which gives the achievable rate region

$$R_1 \le C\left(\frac{|h_1|^2 p_1}{\sigma^2}\right)$$

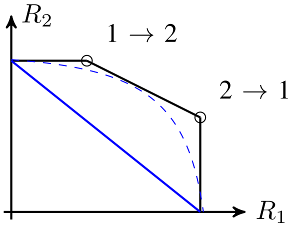

Fig. 6. ***Achievable rate region for the two-user MAC with AWGN. The two circles indicate the achievable rate pairs with the two decoding orders. The line between the two circles is achieved by TS.***

$$\begin{aligned} R_2 &\le C\left(\frac{|h_2|^2 p_2}{\sigma^2}\right) \\ R_1 + R_2 &\le C\left(\frac{|h_1|^2 p_1 + |h_2|^2 p_2}{\sigma^2}\right). \end{aligned} \tag{11}$$

In fact, the achievable rate region characterized in (11) cannot be improved. There is a converse proof showing that this is indeed the capacity region of the MAC.

The extension to any number of users K is straightforward and the capacity region is described by the following inequalities for all subsets of users $\mathcal{S} \subseteq \mathcal{K} = \{1, 2, \ldots, K\}$:

$$\sum_{k\in\mathcal{S}} R_k \le C\left(\frac{\sum_{k\in\mathcal{S}} |h_k|^2 p_k}{\sigma^2 + \sum_{l\in\mathcal{S}^c} |h_l|^2 p_l}\right) \tag{12}$$

where $\mathcal{S}^c = \mathcal{K} \setminus \mathcal{S}$ is the complement of $\mathcal{S}$.

The AWGN MAC is very well understood from information theory and the capacity region is fully characterized [32]. As we will see later, the extension to multiple antennas is straightforward.

The AWGN MAC capacity region is achieved by P2P codes, SIC with both decoding orders, and TS. The sum capacity is achieved by any decoding order.

D. Broadcast Channel

The system model for the two-user BC is illustrated in Fig. 7. The transmitter encodes two messages D_1 and D_2 into one codeword X transmitted over two separate channels with fading h_1 and h_2 and AWGN to the two receivers. Each receiver is only interested in one message, either D_1 or D_2.

A simple way is to first encode the two messages into two codewords X_1 and X_2, weight them with the power allocation $(p_1)^{1/2}$ and $(p_2)^{1/2}$ and send the superposition

$$X = \sqrt{p_1} X_1 + \sqrt{p_2} X_2. \tag{13}$$

This coding scheme is called SC. The sender has a total power constraint $p_1 + p_2 \le P$. It reminds us of the P2P RS coding scheme, where a user's data are split into two messages, which are coded separately and then superimposed into a codeword, as discussed in Section II-B.

Each receiver can then decide whether it first decodes the superimposed message of the other user or decodes its own message with TIN. Let us assume that the first channel is better,[1] i.e.,

$$|h_1|^2 \ge |h_2|^2 \tag{14}$$

and the first receiver performs SIC and decodes the codeword for the second user first. This is only successful if

$$R_2' \le C\left(\frac{|h_1|^2 p_2}{\sigma^2 + |h_1|^2 p_1}\right). \tag{15}$$

After decoding the interference, the first user can then decode its own codeword if

$$R_1 \le C\left(\frac{|h_1|^2 p_1}{\sigma^2}\right). \tag{16}$$

The weaker receiver could also try to perform SIC. According to (4), this would lead to the rate constraints for the data rate of the first user as

$$R_1' \le C\left(\frac{|h_2|^2 p_1}{\sigma^2 + |h_2|^2 p_2}\right). \tag{17}$$

Clearly, the rate constraint in (17) is stricter than (16). Therefore, the second receiver would constrain the rate of the first user too much, if SIC is applied. Therefore, the second receiver decodes its own codeword directly with TIN. The rate is

$$R_2 \le C\left(\frac{|h_2|^2 p_2}{\sigma^2 + |h_2|^2 p_1}\right). \tag{18}$$

From two constraints (16) and (18), the second one is stricter because of (14). Therefore, the constraint in (16) on R_2' is automatically satisfied and can be omitted. This results in the following achievable rate region for all

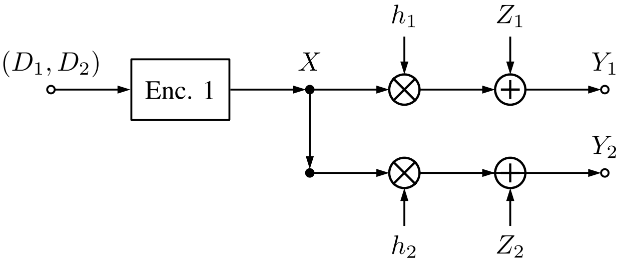

Fig. 7. ***Two-user BC with AWGN.***

[1] Note that such an order of receivers is always possible for scalar channels. This means that one receiver is the weaker, observing a more noisy observation than the stronger receiver. This is called degraded BC.

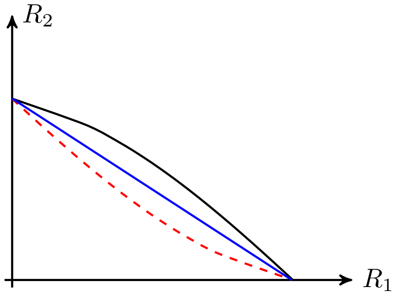

Fig. 8. *Capacity region for the two-user BC with AWGN. The blue line corresponds to TS, the black line corresponds to the capacity region, and the red dashed line corresponds to the achievable region with DPC and the opposite coding order.*

$0 \le p_1, p_2 : p_1 + p_2 \le P$:

$$\begin{aligned} R_1 &\le C\left(\frac{|h_1|^2 p_1}{\sigma^2}\right) \\ R_2 &\le C\left(\frac{|h_2|^2 p_2}{\sigma^2 + |h_2|^2 p_1}\right). \end{aligned} \tag{19}$$

It turns out that there does not exist any better coding and decoding scheme and the region in (19) is the capacity region of the degraded BC. In contrast to the MAC region, no TS is required, varying the power allocation is sufficient to obtain the complete boundary of the capacity region. It is shown in Fig. 8.

The capacity region of the degraded AWGN BC is well understood [33], [34]. We observe the following.

> For a degraded AWGN BC channel, the capacity region is achieved by power control, SC at the encoder, SIC at the stronger receiver, and TIN at the weaker receiver.

There exists also an alternative to SC and SIC, which is based on Gelfand–Pinsker coding, called DPC for the AWGN channel. The idea of DPC is to perform the opposite operation to SIC in the MAC in the downlink direction. Since the interference that is created by superimposing codewords is known at the encoder, it can be removed with smart encoding for the upcoming codeword. The encoding order is opposite of the corresponding SIC order from the MAC. With using DPC, the precoding order can be selected freely because it does not depend on the received signal power. Using DPC, the rate pair in (19) as well as the other order shown in the following can be achieved [35]:

$$\begin{aligned} R_1 &\le C\left(\frac{|h_1|^2 p_1}{\sigma^2 + |h_1|^2 p_2}\right) \\ R_2 &\le C\left(\frac{|h_2|^2 p_2}{\sigma^2}\right). \end{aligned} \tag{20}$$

In the scalar AWGN BC, the region in (19) is larger than the region obtained with the other coding order in (20). In Fig. 8, the SC SIC region is shown in black solid line, while the region in (20) is shown in red dashed line. In the degraded BC, neither DPC nor TS is required to achieve the capacity region.

E. Interference Channel

The combination of the MAC and BC results in the IC. The IC with AWGN is illustrated in Fig. 9. There are two separate encoders and two separate receivers. Each encoder has one message for its intended receiver. Both encoders use the channel at the same time on the same frequency. This leads to interference in addition to AWGN. The direct channel coefficients are denoted by h_{11} and h_{22}, while the ICs are denoted by h_{12} and h_{21}.

In order to simplify the exposition, we will choose the standard interference model in which the direct links are normalized to $h_{11} = 1 = h_{22}$. The ICs are symmetric and set to $h_{12} = (a)^{1/2} = h_{21}$.

In the IC, the coding and decoding techniques from MAC and BC can be applied and combined. It depends on the strength of the interference, i.e., on the value of a, which choice of coding and decoding techniques is most appropriate. There are a few cases in which the capacity region of the IC is known. For strong interference $a > 1$, the capacity region is achieved by using P2P codes and SIC at both receivers. The capacity region is the intersection of the two MAC capacity regions

$$\begin{aligned} R_1^s &\le C\left(\frac{p_1}{\sigma^2}\right) \\ R_2^s &\le C\left(\frac{p_2}{\sigma^2}\right) \\ R_1^s + R_2^s &\le C\left(\frac{\min\left(p_1 + ap_2, p_2 + ap_1\right)}{\sigma^2}\right). \end{aligned} \tag{21}$$

Furthermore, for weak interference $a \approx 0$, the sum capacity is achieved by P2P codes and TIN at both receivers [36], [37], [38]. The individual achievable rates are

$$\begin{aligned} R_1^w &\le C\left(\frac{p_1}{\sigma^2 + ap_2}\right) \\ R_2^w &\le C\left(\frac{p_2}{\sigma^2 + ap_1}\right). \end{aligned} \tag{22}$$

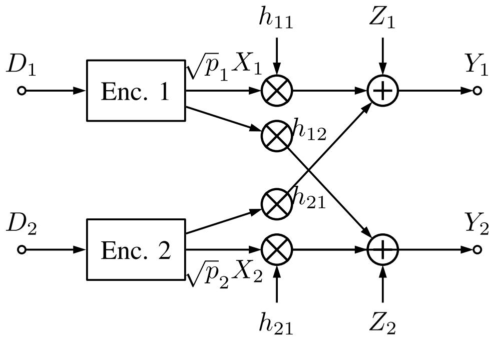

Fig. 9. *Two-user fading IC with AWGN.*

Finally, for moderate interference $0 < \epsilon < a < 1$, the best-known strategy, which achieves the largest rate region, is a combination of TS and RS [39]. If we restrict ourselves to use random code ensembles, RS, SC, and TS, then it is the largest rate region achievable [28]. The simplified RS achievable rate region without TS is described by $R_1 = R_{11} + R_{12}$ and $R_2 = R_{22} + R_{21}$ with

$$\begin{aligned}
&R_{11}^r \le C\left(\frac{\lambda_1 p_1}{\sigma^2 + \lambda_2 a p_2}\right), \; R_{22}^r \le C\left(\frac{\lambda_2 p_2}{\sigma^2 + \lambda_1 a p_1}\right)\\
&R_{12}^r \le C\left(\frac{\bar{\lambda}_1 p_1}{\sigma^2 + \lambda_1 p_1 + \lambda_2 a p_2}\right)\\
&R_{12}^r \le C\left(\frac{\bar{\lambda}_1 a p_1}{\sigma^2 + p_2 + \lambda_1 a p_1}\right)\\
&R_{21}^r \le C\left(\frac{\bar{\lambda}_2 a p_2}{\sigma^2 + \lambda_2 p_2 + \lambda_1 a p_1}\right)\\
&R_{21}^r \le C\left(\frac{\bar{\lambda}_2 a p_2}{\sigma^2 + p_1 + \lambda_2 a p_2}\right)
\end{aligned} \tag{23}$$

with RS parameters $0 \le \lambda_1$ and $\lambda_2 \le 1$.

In Fig. 10, the different schemes and their achievable rate regions for the AWGN IC are illustrated. In Fig. 10, the capacity region for strong interference $a = 1.3$ is illustrated (red solid line shows the capacity region achieved with SIC). In addition, the achievable rate regions by TIN for weak interference $a = 0.1$ are shown as blue dashed line. Finally, the achievable rate regions for moderate interference $a = 0.8$ are illustrated for TIN (black dashed) and for simplified RS (black dashed-dotted).

In the IC with AWGN, the coding and decoding strategy depends on the strength of the interference. The P2P codes, RS, TIN, SIC, and combinations of these using TS achieve either the capacity region or the best-known achievable rate region.

Interestingly, despite many attempts, the moderate IC still has an unsolved capacity region. One approach is to approximate the gap between the inner and outer boundaries of the IC [40]. More recently, the corner points of the capacity region of the weak IC are characterized in [41]. A special class of fading binary ICs is considered in [42].

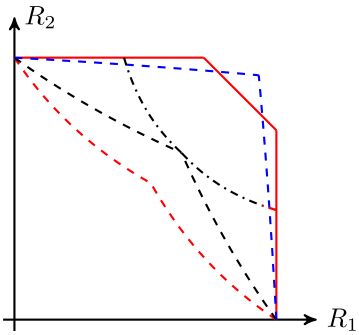

Fig. 10. *Capacity and achievable rate regions for the standard two-user IC with AWGN and different interference strengths a for $\sigma^2 = 1$. Red solid line shows the capacity region achieved with SIC for strong interference, blue dashed lines show the capacity region achieved with SIC for weak interference, and black dashed TIN, simplified dashed-dotted RS, shows the capacity region achieved with SIC for moderate interference.*

III. UNIFICATIONS AND EXTENSIONS OF BASIC MODELS

This section discusses the extensions to the simple channel models used in the last section to derive the capacity and achievable rate region results for the basic network elements. In particular, we summarize the extensions to multiantenna transceivers, and the extension to multiple carriers and the multicell network scenarios. The main focus is to validate whether the main observations regarding the optimal coding and decoding schemes hold. This provides a unified view of the fundamental limitations of multiple access technologies.

A. Multiple-Carrier Systems

Applying a multicarrier modulation scheme at the transceivers results in a number of parallel channels. For any multiuser system, this opens up more degrees of freedom in scheduling users on different carriers and allocating power across the spectral domain.

The first important observation is that separating the coding into parallel P2P encoders and decoders achieves the capacity of the parallel P2P links. Furthermore, the optimal spectral power allocation for a P2P link is known as the *waterfilling* solution. It is classical for AWGN P2P channels [43]. For other scenarios, including different performance metrics [44] or other types of CSI, there are recent views and algorithms [45]. The general formulation of the problem for K parallel channels is

$$\begin{aligned}
&\max_{p_1,\ldots,p_K \ge 0} \sum_{k=1}^{K} f_k\left(p_k\right)\\
&\text{s.t.} \sum_{k=1}^{K} p_k \le P
\end{aligned} \tag{24}$$

where f_k are real, increasing, smooth, and strictly concave functions. $P > 0$ is the maximum transmit power constraint. The programming problem in (24) is a convex programming problem [46]. Therefore, the Karush–Kuhn–Tucker (KKT) optimality conditions are sufficient and necessary. The solution can be characterized or efficiently found numerically. The analytical characterization leads to the waterfilling-type solution with water level $\nu > 0$

$$p_k^* = \left[\tilde{f}_k'\left(\nu\right)\right]^+, \quad \sum_{k=1}^{K} p_k^* = P \tag{25}$$

with $[x]^+ = \max(0, x)$. ν is chosen to fulfill the power constraint. $\tilde{f}'$ is the inverse function of the first derivative of f. Although the formulation in (25) looks simple, its implementation may not be since the $[]^+$ operation is not differentiable. The multiuser spectrum optimization problem is studied in [47], where it is shown that the duality gap of the optimization problem is always zero under TS or if the number of carriers K approaches infinity, regardless of the convexity of the objective function.

For parallel MACs, the coding problem becomes separable, too, i.e., each individual parallel MAC can be treated separately [48]. The power control problem becomes coupled by the sum power constraint. Overall, the power control problem is very similar to the P2P case. In particular, the sum-rate maximization problem reads as

$$\begin{aligned}&\max\nolimits_{p_{11},p_{12},\ldots,p_{KM}\geq 0} \sum_{k=1}^{K} f_k\left(\sum_{m=1}^{M} h_{km}p_{km}\right)\\ &\text{s.t.} \sum_{k=1}^{K} p_{km} \leq P \quad \forall 1 \leq m \leq M \end{aligned} \tag{26}$$

where h_{km} is the channel gain of user $m \in \{1,\ldots,M\}$ on parallel channel k. The problem in (26) is also a convex programming problem. The extension of P2P waterfilling to the problem in (26) is an iterative procedure called *iterative waterfilling*. Fix the power allocation of all users, consider one user m, and perform P2P waterfilling. Iterate over the users until the sum rate does not change. This algorithm efficiently finds the global optimal power allocation. The generalization to vector channels, discussed in Section III-B, is also possible [49].

For parallel BCs, the capacity region is characterized in [50] and it turns out that the parallel BCs are also separable. The power control problem also has a long history [51] but also very recent results [52]. A common approach to solving power control problems for the BC, i.e., downlink, is to solve the corresponding reciprocal MAC, i.e., uplink, and problem and apply the uplink–downlink duality [53]. This leads to a very similar problem for maximizing the sum rate in BC as in (26), and the same iterative waterfilling algorithm can be used to solve the power control problem efficiently.

Finally, parallel ICs are not always separable [54], i.e., separate encoding and decoding of the parallel ICs is suboptimal. The smart and simple counterexample in [54] shows that it is possible to achieve a strictly larger rate region with joint encoding over the parallel channels. The highest possible rate using separate coding is compared to the capacity optimal joint encoding.

Parallel MAC and BC are separable in the sense that separate encoding and decoding for each parallel channel achieves the capacity region. This does not hold for the general IC.

There is a large body of work on parallel multiuser ICs, stemming from applications in copper wire networks. It is usually referred to as dynamic spectrum management [55]. Basically, this refers to power control in parallel MAC, BC, and IC. Optimal spectrum management in multiuser ICs is reported in [56]. The power control problem in parallel IC is nonconvex and difficult [57]. There are frameworks to tackle the nonconvexity problem [58], based on monotonic optimization [59], sequential convex approximation [60], fractional programming [61], or monotonic programming [62].

Power control problems can be solved efficiently for parallel P2P, MAC, and BC if duality can be applied for the latter. Power control problems for general IC are more difficult.

For a recent survey of advances in optimization methods for wireless communication systems design, the interested reader is referred to [63].

B. Multiple Antenna Systems

In multiantenna systems, the additional degrees of freedom in the spatial dimension are very different from the spectral or temporal dimension. First, new capacity regions were required for P2P [64], MAC [65], BC [66], and IC [67]. Second, the reason for more complicated optimization problems is that the system model is described by channel matrices that do not always commute. As long as it is possible to decompose the matrix function into a diagonalized version, the problem becomes easier. More details on some results for multiuser MIMO can be found in [68].

The AWGN P2P MIMO capacity is given by

$$C_{\text{p2p}} = \max_{\boldsymbol{Q}\succeq \boldsymbol{0},\text{tr}(\boldsymbol{Q})\leq P} \log\det\left(\boldsymbol{I} + \rho \boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^H\right) \tag{27}$$

with $\rho = (1/\sigma^2)$, channel matrix $\boldsymbol{H}$ and transmit covariance matrix $\boldsymbol{Q}$. The optimal transmit strategy is to choose $\boldsymbol{Q} = \boldsymbol{V}\text{diag}(p_1,\ldots,p_n)\boldsymbol{V}^H$ with right eigenvectors $\boldsymbol{V}$ of the SVD of the channel matrix $\boldsymbol{H} = \boldsymbol{U}\text{diag}((\lambda_1)^{1/2},\ldots,(\lambda_n)^{1/2})\boldsymbol{V}^H$. Inserting the optimal $\boldsymbol{Q}$ into (27) leads to

$$C_{\text{p2p}} = \max_{p_i\geq 0,\sum p_i \leq P} \sum_{i=1}^{n} \log\left(1 + \rho p_i \lambda_i\right) \tag{28}$$

which corresponds to the optimal waterfilling solution again.

The AWGN MIMO MAC has a capacity region, which is achieved by SIC as in the standard MAC. The main difference it that the optimization involves the transmit covariance matrices $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$

$$\begin{aligned} R_1 &\leq \log\det\left(\boldsymbol{I} + \rho\boldsymbol{H}_1\boldsymbol{Q}_1\boldsymbol{H}_1^H\right)\\ R_2 &\leq \log\det\left(\boldsymbol{I} + \rho\boldsymbol{H}_2\boldsymbol{Q}_2\boldsymbol{H}_2^H\right)\\ R_1 + R_2 &\leq \log\det\left(\boldsymbol{I} + \rho\boldsymbol{H}_1\boldsymbol{Q}_1\boldsymbol{H}_1^H + \rho\boldsymbol{H}_2\boldsymbol{Q}_2\boldsymbol{H}_2^H\right). \end{aligned} \tag{29}$$

The channel matrices are $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$.

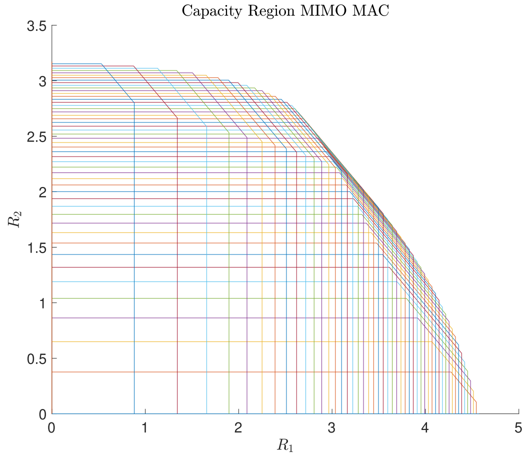

Fig. 11. ***MIMO MAC capacity region for 2 × 2 × 2 channel.***

The resulting MIMO MAC capacity region differs from the standard MAC region in that it no longer looks such as a pentagon but has a smooth boundary, as shown in Fig. 11 [69].

From Fig. 11, we can see that the decoding order has an effect on the achievable rate region, while for the sum rate, the decoding order does not matter, as can be seen in (29). For weighted sum rates to the left of the maximum sum rate, the optimal decoding order is 1 → 2, while for weighted sum rates to the right of the maximum sum rate, the optimal decoding order is 2 → 1. The corresponding programming problems are still convex programming problems, and the optimal weighted sum rate can be computed efficiently. Importantly, TS is still required to obtain all points of the capacity region.

For the general K-user MIMO MAC, the capacity region is characterized by the following inequalities for all subsets $\mathcal{S} \subseteq \{1, \ldots, K\}$:

$$\sum_{i \in \mathcal{S}} R_i \leq \log\det\left(\boldsymbol{I} + \rho \sum_{i \in \mathcal{S}} \boldsymbol{H}_i \boldsymbol{Q}_i \boldsymbol{H}_i^H\right). \tag{30}$$

Let us summarize the main findings for the MIMO MAC.

> For MIMO MAC, the maximum sum rate is achieved for any decoding order. For weighted sum-rate maximization, the decoding order is important. To reach any point within the capacity region, TS is required. The capacity region is smooth and consists of the union of pentagons.

The AWGN MIMO BC is an example of a nondegraded BC for which the capacity region is known [66]. From the MIMO MAC, the need for different encoding orders becomes important. Therefore, SIC is not sufficient to reach the capacity region. Instead, DPC is needed together with TS.

The capacity region of the AWGN MIMO BC is described by

$$\mathcal{R}_{\text{MIMO}}^{\text{BC}} = \text{cv}\left(\bigcup_{\pi, \boldsymbol{Q}_1, \boldsymbol{Q}_2} \left\{ R_i^{\text{DPC}}\left(\pi, \boldsymbol{Q}_1, \boldsymbol{Q}_2\right)\right)_{i=\{1,2\}} \right\}\right) \tag{31}$$

where cv is the convex closure of the union of the two decoding regions for the two different DPC coding orders π over all possible transmit covariance matrices $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ under a sum transmit power constraint $\text{tr}(\boldsymbol{Q}_1 + \boldsymbol{Q}_2) \leq P$. The rate of user i as a function of the coding order and the transmit covariance matrices is given by

$$\begin{aligned} R_i\left(\pi, \boldsymbol{Q}_{1,2}\right) = \log\det&\left(\boldsymbol{I} + \sum_{k=1}^{i} \boldsymbol{H}_{\pi(k)} \boldsymbol{Q}_{\pi(k)} \boldsymbol{H}_{\pi(k)}^H\right) \\ &- \log\det\left(\boldsymbol{I} + \sum_{k=1}^{i-1} \boldsymbol{H}_{\pi(k)} \boldsymbol{Q}_{\pi(k)} \boldsymbol{H}_{\pi(k)}^H\right). \end{aligned} \tag{32}$$

For illustration, Fig. 12 shows the capacity region of an AWGN MIMO BC from [70].

In Fig. 12, it can be seen that the maximum sum rate is again achieved by one of the two DPC orders. The TS between the maximum sum-rate point of the two DPC orders corresponds to the convex closure operation, and the other parts of the capacity region are achieved by varying the transmit covariance matrices under the sum power constraint. In stark contrast to the single-antenna case, where SC and SIC with one decoding order are sufficient to achieve the capacity region of BC, for the MIMO BC, which is a nondegraded BC, SC and SIC are not sufficient to reach the full capacity region. Furthermore, TS is also required to reach all points within the capacity region. In particular, the maximum sum-rate points require TS between the two DPC orders.

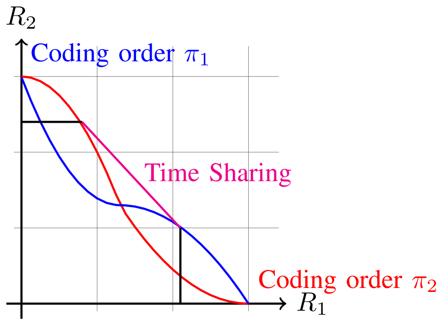

Fig. 12. ***Capacity region of MIMO BC channel. The coding order π_2 in red, coding order π_1 in blue, and the convex hull, i.e., the maximum sum-rate line in magenta. Furthermore, the region in which all rate points are achieved by maximizing the sum rate is indicated with black lines [70, Fig. 1].***

> For MIMO BC, the maximum sum rate is achieved for any decoding order. For weighted sum-rate maximization, the decoding order is important. To reach any point within the capacity region, DPC is required with both encoding orders and TS.

To conclude Section III-B, a few final remarks are in order. First, if we restrict ourselves to linear encoding and decoding schemes, e.g., only single-stream beamforming at the transmitter and receive beamforming with single-user decoding at the receivers, the MIMO BC is transformed into a single-antenna IC. This can be seen as follows. Denote the transmit beamforming vector for user i by $\boldsymbol{w}_i$ and the receive beamforming vector at receiver k by $\boldsymbol{v}_k$, and then, the resulting channels from the transmitter (data for user i) to the receiver k are given by

$$h_{ik} = \boldsymbol{v_k}^H \boldsymbol{H}_k \boldsymbol{w}_i \tag{33}$$

with channel coefficients h_{ik} for $i = 1, 2$ and $k = 1, 2$ for the corresponding IC.

Second, there exist a full elaborated duality theory under a sum transmit power constraint for the MIMO MAC and BC both for the capacity regions [71], the linear precoding and decoding [72], and MIMO MAC–BC duality with linear-feedback coding schemes [73].

Third, combining MIMO and parallel BC, the question on the separability of the coding and decoding across the parallel BC is studied, too. The separability of AWGN MIMO BC is not always given. There are cases in which inseparability occurs, such as when parallel AWGN MIMO BC is studied [74]. Also, under partial CSI, Joudeh and Clerckx [75] observe that parallel MISO BC is not separable.

C. Multiple-Cell Networks

The classical network architecture is network-centric and assigns users to BS, usually based on location, distance, or received signal strength. These *cellular* networks then consist of cells where a BS serves its assigned users in a number of sectors.

The main difference between cellular models and the single-cell multiuser models discussed above is the interference from other cells transmitting at the same time on the same frequency. Normally, the BSs belonging to different cells do not cooperate in the sense of sharing data and serving a user together. The cooperative scheme, where users are served simultaneously by more than one BS, was called COMP during the standardization of the fourth generation of mobile communications.

Early information-theoretic results for an idealized scenario are reported in [76]. While in practice, hexagonal BS arrays are often considered, the simplified 1-D linear model of [76] is used in many studies. This assumes that users only receive signals from their own BS and those in the immediate vicinity. This abstraction is intended to capture the locality of interference. The 1-D (or linear) version of this model, where all devices are located on the boundary of a large circle, is shown in Fig. 13. It is usually assumed that all users in the jth cell are jointly served by BS$_{j-1}$, BS$_j$, and BS$_{j+1}$. The accuracy of the Wyner model in modeling interference in cellular networks is studied in [77].

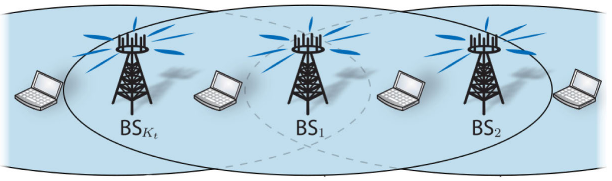

Fig. 13. ***Illustration of the multicell scenario called the onedimensional/linear Wyner model. Users are jointly served by the closest BS and its two neighbors (in a cyclic manner), and only experience interference from these three BSs (adapted from [78]).***

In traditional multicell systems, each user is served by one BS in one sector at a time, and resource allocation is performed unilaterally by its assigned BS. This is possible because the frequency reuse pattern is such that cell sectors using the same resources cause negligible interference to each other. In the next section, we describe coordinated multiple access variants that can be configured to reduce ICI through proper frequency planning.

In general, the additional ICI results in a channel model similar to the IC. The receiver can decide whether to treat the additional ICI as an additional AWGN or to try to decode it and perform SIC. For the latter, it must know the codebook used by the neighboring cells. Furthermore, if the SIC operation results in a reduction in the rate of the neighboring cells' data streams, it must inform the corresponding BS. This requires some form of coordination or co-operation.

It is important to distinguish between the different relationships that neighboring BS can have. These relationships are determined by the RAN architecture. They are summarized in Table 1. The difference between *coordination* and *cooperation* lies in the availability of the data. For coordination, the interface between BSs is used (e.g., X2 interface in 5G), while for cooperation, both the interface between BSs and the interface from BS to the core network (e.g., S1 interface in 5G) are required.

In the following, we briefly describe the four architectures mathematically and provide some achievable rates for the uplink (MAC) and the downlink (BC) transmission.

The distributed linear Wyner model in Fig. 13 leads to achievable SINR expressions of user k in cell ℓ in the downlink of the form

$$\text{SINR}_{k,\ell}^{dl} = \frac{|h_{\ell,k}|^2 p_{\ell,k}}{\sigma^2 + \sum_{m \neq k} |h_{\ell,k}|^2 p_{\ell,m} + \sum_{j \neq \ell} |h_{j,k}|^2 p_j} \tag{34}$$

where the nominator has three terms, the first being AWGN, the second being intracell interference, and the

Table 1 RAN Architectures for Multiple Cells, Their Description, and Signaling Overhead

Architecture	Description	Signalling
Distributed	Decisions are made locally at BSs without any signalling between neighbouring BSs. Depending on local CSI interference strategies are implemented locally only.	No additional overhead or signalling required.
Coordinated	Neighbouring BSs communicate to decide on their resource allocation and transmit strategies. This could be related to all dimensions: time, frequency, space but also interference temperature in specific directions. No user data is exchanged.	Signalling between BSs is required. Exchange of quantized local CSI or SINR.
Cooperative	Multiple BSs serve a single user, either coherently (joint transmission (JT)) or non-coherently. This leads to improved reliability by macro diversity. Additional coordination for interference management is possible, e.g., DPC.	User data has to be delivered to BS from core network, fronthaul links. For coordination additional load on BS to BS links.
Single cell	All BS are operated by a single central processing unit (CPU), Cloud-RAN (RAN) architecture, centralized processing	large demand on backhaul and fronthaul

third being ICI. For downlink transmission, it is important to note that this SINR expression is the basis for determining the decoding order for SC SIC. The ICI must be considered when designing the BC or MAC system.

The coordinated architecture allows the BS to introduce interference coordination between neighboring cells. This scenario is illustrated in Fig. 14. Here, four BSs coordinate their beamforming strategies. Coordinated beamforming means that each BS has a disjoint set of users to serve with data but chooses transmission strategies jointly with all other BSs to reduce ICI. There can be any number of users in each cell. The special case with only one user per cell is the IC we studied earlier. In Fig. 14, the data clusters $\mathcal{D}_1, \ldots, \mathcal{D}_4$ correspond to the sets of users served by BSs $1, \ldots, 4$. The coordination cluster $\mathcal{C}_1 = \cdots = \mathcal{C}_4$ contains all users because there is a global beamforming coordination between the four BSs. In this case, one way to model the overall signal model in the system is described in [78, Sec. 1.3.2]. In the multicell scenario, the channel from all BSs to UE$_k$ is denoted $\boldsymbol{h}_k = [\boldsymbol{h}_{1k}^T \ldots \boldsymbol{h}_{K_t k}^T]^T \in \mathbb{C}^N$, where $\boldsymbol{h}_{jk} \in \mathbb{C}^{N_j}$ is the channel from BS$_j$.

DCC means that BS$_j$ has channel estimates for users in $\mathcal{C}_j \subseteq \{1, \ldots, K_r\}$, while the interference generated for users $k \notin \mathcal{C}_j$ is negligible and can be treated as part of the Gaussian background noise, and BS$_j$ provides data to users in $\mathcal{D}_j \subseteq \mathcal{C}_j$.

Based on the DCC, only certain channel elements of $\boldsymbol{h}_k$ will carry data and/or nonnegligible interference. These can be selected by the diagonal matrices $\boldsymbol{D}_k \in \mathbb{C}^{N \times N}$ and $\boldsymbol{C}_k \in \mathbb{C}^{N \times N}$, which are defined as

$$\boldsymbol{D}_k = \begin{bmatrix} \boldsymbol{D}_{1k} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{D}_{K_t k} \end{bmatrix} \tag{35}$$

where

$$\boldsymbol{D}_{jk} = \begin{cases} \boldsymbol{I}_{N_j}, & \text{if } k \in \mathcal{D}_j \\ \boldsymbol{0}_{N_j}, & \text{otherwise} \end{cases}$$

$$\boldsymbol{C}_k = \begin{bmatrix} \boldsymbol{C}_{1k} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{C}_{K_t k} \end{bmatrix} \tag{36}$$

where

$$\boldsymbol{C}_{jk} = \begin{cases} \boldsymbol{I}_{N_j}, & \text{if } k \in \mathcal{C}_j \\ \boldsymbol{0}_{N_j}, & \text{otherwise.} \end{cases}$$

Thus, $\boldsymbol{h}_k^H \boldsymbol{D}_k$ is the channel that carries data to UE$_k$ and $\boldsymbol{h}_k^H \boldsymbol{C}_k$ is the channel that carries nonnegligible interference. The symbol-sampled complex-baseband received signal at UE$_k$ is

$$y_k = \boldsymbol{h}_k^H \boldsymbol{C}_k \sum_{l=1}^{K_r} \boldsymbol{D}_l \boldsymbol{s}_l + n_k. \tag{37}$$

If all users perform TIN, the corresponding SINR for UE$_k$ in the downlink is

$$\text{SINR}_k = \frac{\boldsymbol{h}_k^H \boldsymbol{C}_k \boldsymbol{D}_k \boldsymbol{Q}_k \boldsymbol{D}_k^H \boldsymbol{C}_k^H \boldsymbol{h}_k}{\sigma_k^2 + \boldsymbol{h}_k^H \boldsymbol{C}_k \left(\sum_{l \neq k} \boldsymbol{D}_l \boldsymbol{Q}_l \boldsymbol{D}_l^H \right) \boldsymbol{C}_k^H \boldsymbol{h}_k}. \tag{38}$$

The SINR expression in (38) can then be used directly to formulate optimization problems for the transmit covariance matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_K$ or directly for the corresponding beamforming vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K$. They typically have the

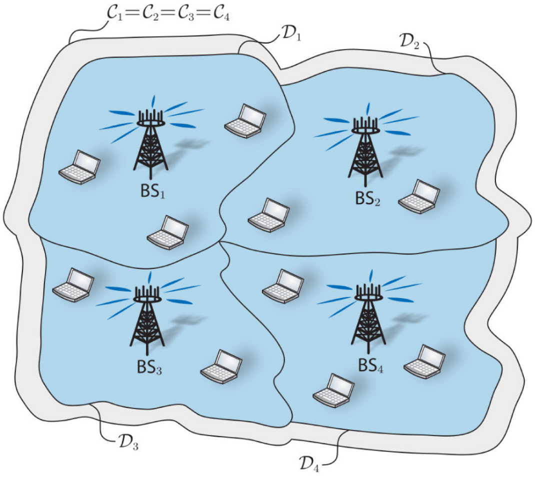

Fig. 14. ***Illustration of the multicell scenario of coordinated beamforming. Users are served by their own BS, while interference is coordinated by joint resource allocation between all BSs (adapted from [78]).***

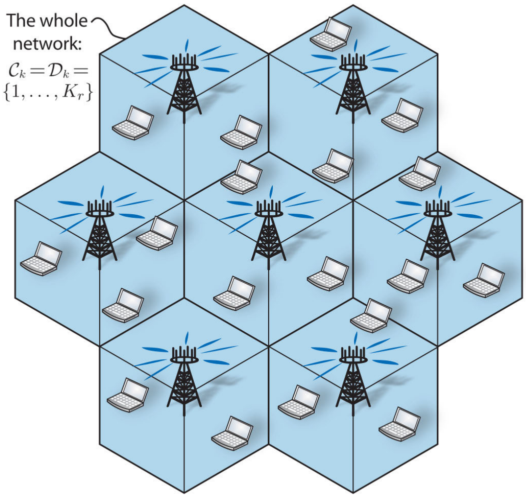

Fig. 15. ***Illustration of the JT scenario, where all cells and cell sectors are connected and perform JT to all users in the whole network (adapted from [78]).***

form

$$\begin{aligned}
&\max_{\boldsymbol{v}_1,\dots,\boldsymbol{v}_K} \quad f\left(g_1\left(\mathrm{SINR}_1\right), \dots, g_K\left(\mathrm{SINR}_K\right)\right) \\
&\text{s. t. } \mathrm{SINR}_k = \frac{|\boldsymbol{h}_k^H \boldsymbol{C}_k \boldsymbol{D}_k \boldsymbol{v}_k|^2}{\sigma_k^2 + \sum_{\mu \neq k} |\boldsymbol{h}_k^H \boldsymbol{C}_k \boldsymbol{D}_\mu \boldsymbol{v}_\mu|^2} \quad \forall k \\
&\qquad \sum_{k=1}^{K} \boldsymbol{v}_k^H \boldsymbol{Q}_{lk} \boldsymbol{v}_k \leq q_l \quad \forall l
\end{aligned} \tag{39}$$

where q_l and $\boldsymbol{Q}_{lk}$ can be used to model power constraints or other interference temperature constraints. The user performance functions $g_k(\cdot)$ are continuous and strictly monotonically increasing, while the system utility function $f(\cdot)$ is Lipschitz continuous and monotonically increasing.

It depends on the properties of the functions f and $g_1, \dots, g_K$ in (39), whether the programming problem can be solved efficiently or not. In general, the SINR expression in (38) is nonconvex with respect to $\boldsymbol{Q}_k$. For a recent survey on optimization methods for these interference networks, please see [63].

In the cooperative architecture, multiple BSs serve one user simultaneously. This can improve link availability, reliability, as well as coverage. However, it requires that the user data to be available at the serving BSs. The extreme case is illustrated in Fig. 15 where all cells and cell sectors perform JT. As can be seen from the datasets $\mathcal{D}_i$, it comprises all users in the whole network. The global JT is sometimes also called SFN and was first suggested for broadcast services such as DVB and its dynamic version in [79].

The mathematical model for the cooperative architecture is the same as for the coordinated architecture, i.e., the same SINR expression as in (38) applies. The architecture is reflected in the choice of matrices $\boldsymbol{C}_i$ and $\boldsymbol{D}_i$. Finally, the optimization problem in (39) can also be reused for cooperative and SFN.

A few final remarks about the multicell extension are in order. First, the description in this section has focused on the multicell downlink BC. However, a similar approach is possible for the multicell uplink MAC. In addition, nonlinear encoders or decoders, such as SIC or DPC, result in SINR expressions that look slightly different because some of the interference is removed. However, it also introduces another SINR constraint for the streams that are decoded first, i.e., additional min operations will appear in the corresponding programming problems. These could either be handled by a parametric approach by moving them to the constraints and solving them as two or more separate inequality constraints. Or the properties of the pointwise minimum of the objective functions can be exploited to derive an efficient algorithm.

Finally, as an alternative to the Wyner-based cellular model, stochastic geometry is also used to study cellular networks. For a standard reference, see [80], and for a recent introduction, the interested reader is referred to [81].

IV. COORDINATED MULTIPLE ACCESS VARIANTS

In this section, we explain the different variants of multiple access technologies currently used in wireless networks and relate their characteristics and performance to the fundamental limitations explained in Section II. In this tutorial, we will focus on so-called coordinated multiple access schemes, as opposed to random multiple access schemes such as ALOHA or CSMA with CD or CA. In coordinated multiple access, devices wishing to connect receive PRBs in response to their connection request. The task of the *scheduler* is to assign time–frequency–space PRBs to radio bearers for different services at different user terminals. The scheduling algorithms are not explicitly standardized, but their control channels and signaling are. Discussions are currently taking place on GFMA schemes for 6G, where the clear boundaries between coordinated and random multiple access will be resolved.

Traditionally, OMA schemes were used: in 2G [global system for mobile communications (GSM)] TDMA, 3G CDMA, and 4G and 5G OFDMA. With HSPA in 3G, MIMO started to lead to SDMA in LTE advanced. In addition, NOMA has received a lot of attention from the research community over the last decade. Having covered these multiple access schemes, we will end this section with GFMA, which is being considered for small data transfer, massive connectivity, and low latency.

A. Time-Division Multiple Access

Time-division multiplexing (TDM), or the more dynamic TDMA, was invented in [82]. In TDMA, the same frequency band can be shared by several devices. To avoid collisions on the channel, it divides the time duration into smaller time slots. Each time slot is assigned exclusively to one device. As the timing is well predictable, it is used in real-time machine communication protocols such as wireless highway addressable remote transducer (HART) [83]. In systems with information freshness requirements, AOI, TDMA shows good performance on average [84].

From an information-theoretic point of view, TDMA is suboptimal in both MACs and BCs. The degree of suboptimality depends on the power constraints. For peak power constraints (per codeword), the achievable rate region is the triangle spanned by the two single-user rate points. This is shown as the blue line in Fig. 6 for the MAC and in Fig. 8 for the BC. While the gain in the MAC by using SIC is significant, the gain of SC and SIC in the BC seems negligible. However, this changes dramatically when multiple-antenna systems are considered.

If we apply average power constraints over multiple codewords and adaptive power control [85], then shorter transmission times correspond to higher transmit power. In the MAC, this leads directly to a larger achievable rate range, as shown by the blue dashed line in Fig. 6. In the BC, there is only a power constraint at the single transmitter, so additional power control does not directly improve the achievable rate range.

In terms of control signaling and synchronization requirements, TDMA only requires symbol synchronization between nodes. For the downlink, this is easily achieved, while for the uplink, rough synchronization is required.

B. Orthogonal Frequency-Division Multiple Access

After the invention of TDMA, early interest turned to frequency-division multiplexing [86]. Intersymbol interference between adjacent symbols in TDMA led to the development of OFDMA, where orthogonality in the spectral domain allows multiple-carrier signals to be multiplexed simultaneously. Other advantages of OFDMA are that it is robust to multipath fading and narrowband interference only affects individual carriers.

The challenge with OFDMA is the PAPR. The superposition of different sinusoidal carrier signals in the time domain can produce large amplitude peaks. One option is to increase the power back-off required at the transmit power amplifiers, which leads to lower power efficiency. Alternatively, clipping of large amplitudes is possible, which leads to nonlinear signal distortion. Finally, subcarrier selection and precoding are alternative options to combat high PAPR. In addition, phase noise and carrier frequency offset are additional impairments that can occur in multicarrier transmission systems [87]. Nevertheless, OFDMA has been used in DAB since 1995, many DVB variants, IEEE 802.11a since 1999, and universal mobile telecommunication system (UMTS) and LTE since 2006. For more information on orthogonal frequency-division multiplexing (OFDM) in wireless communications, the interested reader is referred to [88] and more recently to MIMO-OFDM [89].

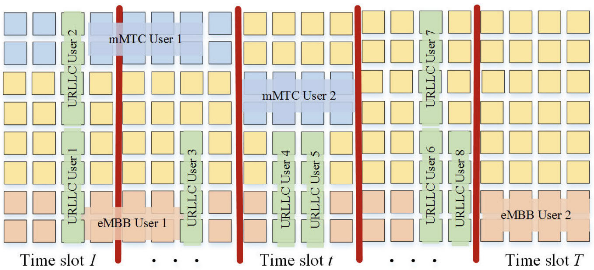

Fig. 16. ***Total PRB structure at cell 1: with $I = 8$ carriers in frequency domain and $J = 4$ time slots in time domain, there are 32 REs, for each mini-slot (adapted from [90, Fig.1.b)]).***

The two dimensions of time and frequency allow users and their services to be scheduled on the corresponding time–frequency PRBs. Recent work in [90] shows how different types of traffic can be efficiently multiplexed by ML on the time–frequency plane. Fig. 16 shows an example PRB structure for a 5G network supporting three different services, namely, URLLC, EMBB, and MMTC.

To support finer granularity for resource allocation, the PRBs are partitioned into multiple REs of equal duration and bandwidth. While EMBB services are allocated constant bandwidth over the whole time duration, MMTC services with intermittent traffic are allocated dynamically based on temporal demand, and the URLLC are allocated full bandwidth for short time instances to satisfy the low latency constraints. In the intelligent puncturing scheme, the PRBs are allocated to EMBB and MMTC users at each time slot and reallocated (punctured) to URLLC users at each mini-slot on demand. The aim of the puncturing scheme proposed in [90] is to minimize the negative impact of URLLC puncturing on the data rate of EMBB or MMTC users and to meet the reliability and latency requirements of URLLC users.

Since the fifth generation of mobile communications, a discussion about alternative waveforms instead to OFDM is ongoing [91]. For 6G, beside OFDM OTFS modulation [92], [93] is discussed as an alternative.

C. Code-Division Multiple Access

Another approach to scheduling more users over the available time–frequency resources is to assign codes or spreading sequences to different users and allow them to occupy the same spectrum at the same time. Wideband CDMA has emerged as the mainstream air interface solution for third-generation 3G networks [94], also called code-division CDMA.

Classical CDMA assigns spreading sequences of short chip duration (and hence large bandwidth) either deterministically based on codebooks of orthogonal or semi-orthogonal codes, or randomly. The main properties of the code sequences are their autocorrelation and cross correlation properties, which determine the self-user and interuser interference. The tradeoff between orthogonality and the number of sequences, and thus the number of users served, is studied in [95].

At the receiver side, the superposition of the user signals is separated by multiuser detection techniques [96]. More recently, SCMA has been proposed, where each user sends a sparse codeword (from a properly designed sparse codebook) corresponding to the instantaneous input message.

A number of variants of orthogonal, nonorthogonal, sparse, and dense CDMA-based schemes have been discussed in the standardization for 5G [97]. For a more recent review, see [98].

D. Spatial-Division Multiple Access

Another dimension for multiplexing, in addition to the time, frequency, and code domains, is the spatial domain, which can be controlled and managed by multiple-antenna systems. While the first ideas to use space to serve multiple users simultaneously date back at least to [99], with massive MIMO, the interest in SDMA schemes has increased significantly [100], [101]. For downlink SDMA, usually TDD is performed and channel estimates from uplink are used for downlink transmission.

Starting from the SINR expression in (38), the beamforming $\boldsymbol{v}_k$ for user k should be chosen to achieve a tradeoff between maximizing the received signal power at the intended user $|\boldsymbol{h}_k^H \boldsymbol{v}_k|^2$ and minimizing the interference power at other receivers $|\boldsymbol{h}_l^H \boldsymbol{v}_k|^2$. The two extreme cases are: maximizing the received power by MRT, i.e., $\boldsymbol{v}_k = (\boldsymbol{h}_k/||\boldsymbol{h}_k||)$, and minimizing the spurious power by ZF, i.e., $\boldsymbol{v}_k = (\Pi^{\perp}_{\boldsymbol{h}_l}\boldsymbol{h}_k/||\Pi^{\perp}_{\boldsymbol{h}_l}\boldsymbol{h}_k||)$ with projection into the orthogonal complement of the space spanned by $\boldsymbol{h}_l$, denoted by $\Pi^{\perp}_{\boldsymbol{h}_l}$. In [102], this tradeoff is studied to derive a characterization of the optimal beamforming parameterization. In [78], the framework is fully developed for multicell and multiuser MIMO systems.

When the number of antennas becomes large in massive MIMO, the advantages can be illustrated by considering the simple MRT. The received power is calculated as $\boldsymbol{v}_k = (1/M) \cdot (\boldsymbol{h}/||\boldsymbol{h}||)$, where the prefactor is derived from the transmit power constraint $P = 1$ since

$$S = \frac{1}{M} \frac{||\boldsymbol{h}_k^H \boldsymbol{h}_k||^2}{||\boldsymbol{h}_k||^2} = ||\boldsymbol{h}_k||^2 = \frac{1}{M} \sum_{n=1}^{M} |h_{kn}|^2 \tag{40}$$

with M antennas. Modeling the channel from each transmit antenna to the receiver k by IID zero-mean complex Gaussian random variables, taking the limit in the right-hand side of (40), leads to

$$\lim_{M\to\infty} \frac{1}{M} \sum_{n=1}^{M} |h_{kn}|^2 = \mathbb{E}\left[|h_{k1}|^2\right] = c \tag{41}$$

by the weak law of large numbers, with a constant c. The effect that the randomness of the fading channel disappears with increasing number of antennas is called *channel hardening* in [101]. The assumption to obtain the limit in (41) is that the channel statistics are standard IID Gaussian. In [103], the effect of channel hardening is compared for theory (IID), simulation (COST 2100 channel model), and measurements. The results show that the COST channel model well represents real scenarios, and the channel hardening effect is less pronounced in real measurements.

With respect to the interference caused to another user h_l, using the same beamformer $\boldsymbol{v}_k$, the expression is

$$I = \frac{1}{M} \frac{||\boldsymbol{h}_l^H \boldsymbol{h}_k||^2}{||\boldsymbol{h}_k||^2} \tag{42}$$

where the bound for $M \to \infty$ leads to interference avoidance if the users' channel coefficients are statistically independent. This is called *favorable propagation* in [101]. If the channels of users are statistically dependent, e.g., due to overlapping angular spreads, the interference term in (42) does not vanish with increasing M. For such scenarios, the NOMA approach discussed in the next section might be an option.

This example illustrates the challenges for massive MIMO and SDMA; CSI is required to perform the beamforming. To obtain CSI for all users K, pilot sequences are used. For a larger number of users, the use of orthogonal pilot sequences would lead to inefficient signaling. Therefore, nonorthogonal pilot sequences are used. This leads to *pilot contamination* [101, Sec. 3].

E. Nonorthogonal Multiple Access

During the development of 5G, a multiple access scheme called NOMA was introduced. Compared to classical OMA, where the information of multiple users can be retrieved via the low-complexity single-user detection algorithms, NOMA allows two users to be served simultaneously on the same frequency. For OMA, the number of users supported is limited by the number of orthogonal resource blocks available. Consequently, it is difficult for OMA techniques to support massive connectivity.

To tackle the above challenges, NOMA[1] has been proposed, which is based on the linear SC of multiple-user signals at the transmitter side combined with multiuser detection algorithms, such as SIC at the receivers side [104]. The main purpose of NOMA is to reap the benefits promised by information theory for the downlink and uplink transmissions, modeled by the BC and MAC, respectively [105]. As explained in Section II, the capacity region of the degraded BC is achieved by SC and SIC. Therefore, it is a wise decision to consider this transmission technique for the BC. In the uplink, for the MAC, it was explained in Section II that SIC at the receiver achieves the capacity region. Therefore, this technique is a good choice for the MAC.

[1] We consider only power-domain NOMA in this article, as it is the most popular NOMA variant.

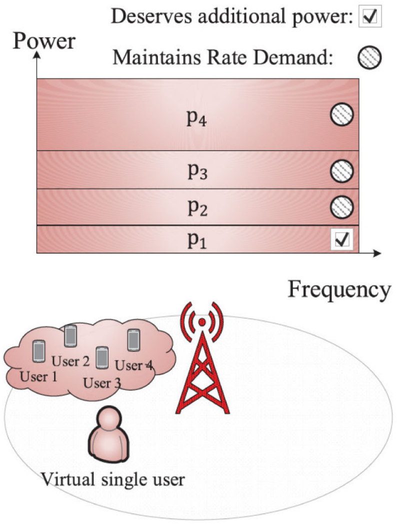

Fig. 17. *Single-carrier single-cell NOMA: capacity achieving for $U^{max} \geq K$. The term power-domain NOMA is visualized in the top of the figure.*

It is important to remember the constraints of SC and SIC for the degraded BC. If the channel is not degraded, an additional constraint is added for each SIC operation at the receivers. Furthermore, the SIC order is determined for the degraded case. In the nondegraded BC case, an optimization over the decoding orders and TS must be applied.

In this section, we describe how NOMA can be optimally incorporated into multicarrier and multicell networks from an information-theoretic point of view. Part of the exposition and further information and details can be found in [106] and [107]. In particular, the elements of [106, Fig. 1] will be discussed in detail. An important parameter for the following description is the NOMA cluster size $U^{\max}$, i.e., the maximum number of users in a NOMA cluster. Usually, small cluster sizes $U^{\max} = 2.3$ are assumed since the SIC complexity grows with the number of users in the cluster.

In Figs. 17–20, the markers for “maintain rate demand” and “deserves additional power” indicate which user will receive more transmit power and a higher data rate for the sum-rate maximization problem under rate constraints.

Fig. 17 shows the single-carrier and single-cell NOMA setup. As long as the cluster size is greater than or equal to the number of users, capacity will be achieved. If not, one option is to use OFDMA on top of NOMA.

Fig. 18 shows the combination of OFDMA and NOMA. Each NOMA cluster has cluster size two and the two clusters are separated in the frequency domain. It is also possible to combine NOMA with SDMA, see [108].

Fig. 19 shows the OMA case with pure OFDMA as a baseline scheme. For K users, at least the same number of subchannels is required. If more than K channels are available, additional multiplexing gain or spectral diversity can be realized.

Fig. 20 illustrates a hybrid scenario where four users are multiplexed per NOMA cluster and six NOMA clusters are multiplexed in the frequency domain. Interestingly, the optimal power allocation in the hybrid scenario can be solved efficiently by carefully exploiting the optimality conditions. It turns out that only one user per cluster—the cluster head—receives additional transmit power, while the other users only receive power to meet their rate requirements. Due to SIC, it is possible to iteratively compute the required transmit power for the noncluster head users in closed form.

The interested reader is referred to the recent surveys for NOMA provided in [109]. For COMP NOMA, an overview is provided in [110]. An example of resource allocation for NOMA-enabled multicarrier, multiantenna 6G networks is provided in [111]. For practical aspects of error rate analysis, see [112].

F. Grant-Free Multiple Access

Recently, GFMA techniques have received much attention from both industry and academia to effectively accommodate a large number of bursty devices transmitting short packets [113]. The basic principle of GFMA is to allow each device to communicate randomly with the BS, allowing multiple devices to share the same physical radio resources (time and frequency).

GFMA techniques completely avoid the immediate exchange of control signaling to establish a connection and may therefore be particularly attractive in scenarios with extremely tight latency constraints and/or a relatively large number of users to be served [11, Sec. 3.2].

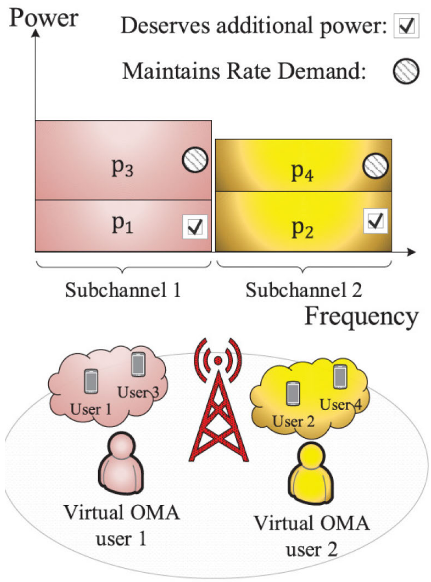

Fig. 18. *OFDMA NOMA: not capacity achieving, but feasible. SC and SIC are applied to each cluster.*

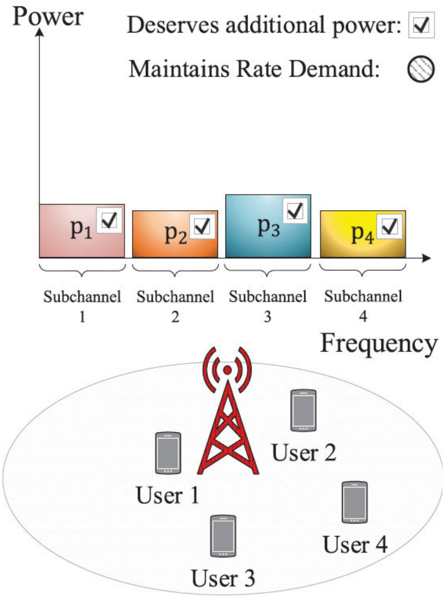

Fig. 19. ***Real OMA case with pure OFDMA.***

There are two main GFMA approaches [114]: contention-based (or uncoordinated) and contention-free (or coordinated). In the former, each user transmits its data in an *arrive-and-go* manner, using the nearest preconfigured grant-free PRBs, while in the latter, dedicated PRBs are dynamically pre-allocated to each user. Therefore, collisions can occur when using uncoordinated GFMA mechanisms, degrading one-shot reliability, while coordinated GFMA mechanisms promote more reliable performance at the cost of some inefficiency in resource utilization.

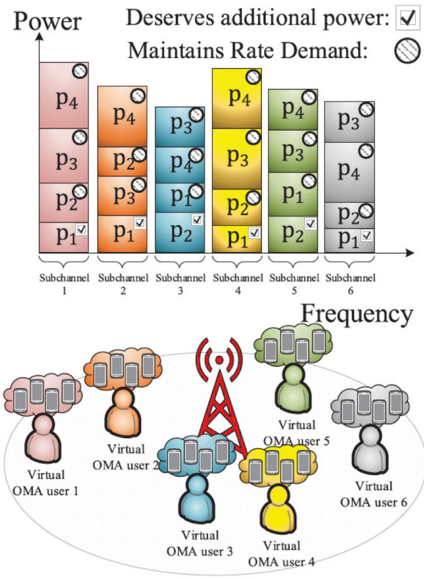

Fig. 20. ***Hybrid NOMA with K multiplexed users: OFDMA with six virtual OMA users and four users within each cluster.***

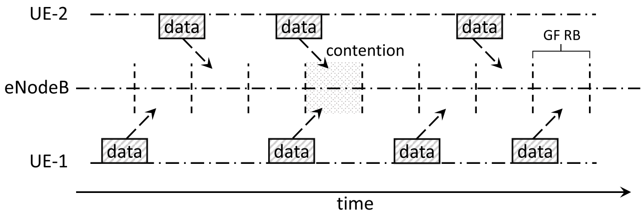

Fig. 21. ***Basic uncoordinated GFMA [119].***

For the contention-free GFMA variant, the information-theoretic limits outlined in Section II apply. For the contention-based GFMA, the information-theoretic limits on random multiple access, including variants of ALOHA, are formulated in terms of capacity and stability regions [115]. For the discrete memoryless random access channel (DM-RAC), Minero et al. [116] derive bounds on the capacity region and bounds the gap between achievable and converse rates. Another popular performance metric for uncoordinated random access is the AOI. In [117], modern random access protocols are studied and compared in terms of their AOI. For a recent overview of modern random access protocols, the reader is refereed to [118].

The basic concept of GFMA is illustrated in Fig. 21 [120]. As multiple users transmit uncoordinated at the same time on the same frequency, collisions can occur, depending on the number of devices, the traffic density, and the number of PRB. Therefore, various approaches are proposed for improved CA and resolution, including message repetition.

GFMA can cause resource collision problems between active devices and the reliability of packet transmission can be degraded. NOMA techniques have been considered to solve such resource collision problems. For the combination of NOMA and GFMA, the interested reader is referred to [122].

Finally, since release 16, the 5G-NR standard has introduced the two-step-RACH (2SR) extension, which is a significant step toward GFMA [121], [123]. As shown in Fig. 22, the messages in the four-step RACH (4SR) procedure are named in time order as Msg1–Msg4, while in 2SR, the messages are named MsgA and MsgB. More specifically, the channel structure of MsgA includes the preamble (Msg1) and the data part (Msg3) in the physical uplink shared channel (PUSCH), and MsgB combines the random access response (Msg2) and the contention resolution (Msg4). As a result, only one round-trip cycle is required between the user equipment (UE) and the BS (gNB) to complete the 2SR procedure, instead of the two round-trip cycles required in 4SR.

G. Comparison of Coordinated Multiple Access Techniques

We summarize the coordinated MAC techniques in Table 2 and compare them in terms of occurrence in

Table 2 Comparison of Considered Coordinated MAC Techniques

MAC Technique	Standard	Scenario	Overhead and Requirements	Limitations
TDMA	GSM, wireless HART	2G mobile communications and data bus networks	requires time sync, high sync overhead, guard time	limited number of users, wait & latency, multi-path, intersymbol interference (ISI), not flexible
CDMA	IS-95, UMTS	2G and 3G mobile com.	time synchronization, bandwidth expansion, guard time and guard bands	close-far issue, multiple access interference (MAI)
OFDMA	LTE, new radio (NR), IEEE 802.11ax	4G and 5G mobile com.	cyclic prefix (CP), power back-off,	PAPR, time frequency offsets, Doppler, time-varying
SDMA	LTE-A, IEEE 802.11ah	4G, 5G, Multiuser MIMO in WLAN, TDD	CSI at transmitter, symbol-synchronization	time-varying channels, inter-cell interference
NOMA	under discussion for 6G, WiFi	dense nets, clustering opportunities, massive connectivity	user terminal processing SIC, downlink overhead for power allocation	SIC error propagation, robustness
GFMA	NR Rel. 16	5G mobile com., mMTC	activity detection, collision avoidance	collisions and service guarantees

communication standard, use-case and application scenario, overhead including signaling and complexity, and their limitations.

V. MODERN MULTIPLE ACCESS TECHNIQUES

The next generation of wireless networks will serve heterogeneous users with limited resources and stringent service requirements. It is therefore necessary to understand the fundamental limitations of different technologies and their optimal parameters. The optimal choice of architecture and configuration depends on the scenario and system parameters. Therefore, modern multiple access techniques need to be flexible, monitor the environment, and adapt their multiple access architecture accordingly.

In this section, we describe the current evolution of modern MAC techniques, their underlying network architectures, and recent results. Multiple access is implemented in the medium access control layer of the protocol stack. It is just above the physical layer and was implemented together in the BSs or user terminals.

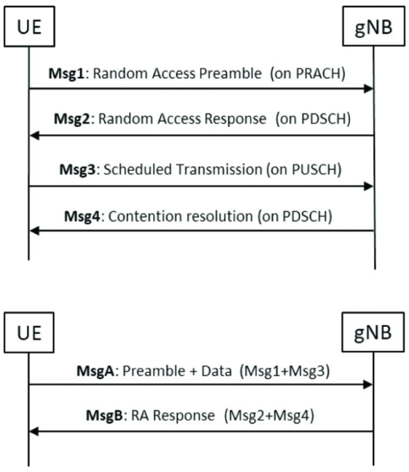
UE
gNB
Msg1: Random Access Preamble (on PRACH)
Msg2: Random Access Response (on PDSCH)
Msg3: Scheduled Transmission (on PUSCH)
Msg4: Contention resolution (on PDSCH)
UE
gNB
MsgA: Preamble + Data (Msg1+Msg3)
MsgB: RA Response (Msg2+Msg4)

Fig. 22. *Random access procedures operation in four-step RACH (top) and two-step RACH (bottom) (adapted from [121]).*

With the current developments toward a disaggregated RAN with open interfaces, called Open RAN, the medium access control layer is separated from the lower physical layer functions. Therefore, we need to consider the interfaces in the new disaggregated architecture when designing and optimizing multiple access techniques.

First, we describe the Open RAN architecture and its implications for next-generation multiple access schemes. We then consider three emerging architectures that have gained interest in the research community: cell-free or distributed antenna systems, unsourced massive random access, and combinations of these.

A. Open Radio Access Network Architectures

The Open RAN architecture simplifies and democratizes the development and operation of mobile networks. The open, standardized interfaces between devices and units, and the flexibility offered, allow different vendors and suppliers to participate and provide solutions. This leads to more competition, faster development cycles, and greater diversity and avoids monopolies or oligopolies.

Open RAN deployments are based on disaggregated, virtualized, and software-based components that are connected through open and standardized interfaces and are interoperable across different vendors [124]. SDN and NFV on the one hand and ML andAI on the MEC on the other hand, via C-RAN architectures, led to the idea of disaggregating BSs into RUs, DUs, and one or more CUs. Fig. 23 shows an overview of the basic building blocks of the Open RAN architecture.

Another innovation in the Open RAN standard is two so-called RICs. They manage network parameters and configurations in near-real-time (10 ms–1 s) and nonreal-time (more than 1 s) time scales. The Open RAN architecture defines interfaces between the different units and between the two RICs. They complement the 3GPP interfaces.

For a complete overview of Open RAN, the interested reader is referred to [125]. For our tutorial, the Open RAN architecture has important implications because the multiple access scheme affects all three disaggregated units,

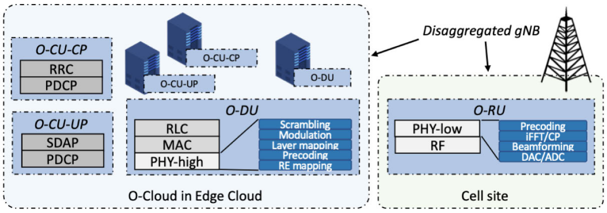

Fig. 23. ***Open RAN building blocks according to functional division 7.2. CU provides RRC, PDCP, and SDAP. DU supports RLC, medium access control, and the higher part of the physical layer. RU implements lower physical layer functions, such as frequency-domain functions, including scrambling, modulation, mapping, part of the precoding, and time-domain functions, including precoding, FFT with CP handling, spatial processing, and the RF chain (adapted from [124, Fig. 2]).***

RU, where the interference is generated and handled, e.g., via beamforming, DU, where access control, resource allocation, and scheduling are performed, and CU, where decisions on the configuration of the MAC protocol are made.

Furthermore, the Open RAN architecture impacts also the communication theoretic modeling of the multiple access network. This is shown in Fig. 24 where the wireless access MAC, BC, and IC are supplemented with the fronthaul links between DU and RUs. In Fig. 24, the midhaul links between DU and CU and the backhaul links from the CU to the core network are not shown.

We denote the wireless channel RU$_i$ to UE$_j$ by h_{ij}. We collect all wireless channel gains in the matrix $\boldsymbol{H}$. The midhaul rate constraints are denoted by c_1 and c_2. The fronthaul rates from DU$_k$ to RU$_i$ are denoted by d_{ki}.

Based on the example open RAN architecture is shown in Fig. 24, the entire network configuration, including midhaul, fronthaul, and wireless multiple access, is considered. While the links may have different characteristics in terms of data rate, reliability, latency, and energy efficiency depending on their link technologies, e.g., fiber, copper, and wireless, the overall network throughput between source (CU) and downlink destinations (users) can be computed by the cutset bound for graphical unicast networks [25, Sec. 15.2].

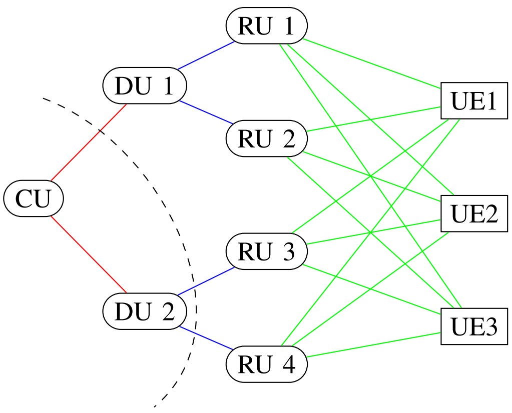

Fig. 24. ***Open RAN architecture with midhaul (red), fronthaul (blue), and wireless access (green) links.***

Table 3 Example Parameters for Link Capacities in the Open RAN Architecture From Fig. 24

Variable	Link	Rate
c_1	midhaul $CU - DU_1$	1.3 Gbit/s
c_2	midhaul $CU - DU_2$	1.5 Gbit/s
d_{11}	fronthaul $DU_1 - RU_1$	700 MBit/s
d_{12}	fronthaul $DU_1 - RU_2$	800 MBit/s
d_{23}	fronthaul $DU_2 - RU_3$	600 Mbit/s
d_{24}	fronthaul $DU_2 - RU_4$	500 Mbit/s

To illustrate the different schemes, let us assume the following configuration as shown in Table 3, where we focus on data rates and rate constraints.

1) Cutset Bound for Graphical Unicast Networks: In order to apply the cutset bound, the open RAN graph shown in Fig. 24 with nodes (CU, DUs, and RUs) and corresponding edges will be assigned a link capacity for each edge, e.g., the midhaul link from CU to DU1 has capacity $C_{c,d1}$ = 100 GBit/s. Then, the cutset bound [25, Th. 15.2] says that the capacity to the destination node UE1 is given by

$$C_1 = \min_{\substack{\mathcal{S} \subset \mathcal{N} \\ CU \in \mathcal{S}, \text{UE1} \in \mathcal{S}^c}} C(\mathcal{S}) \tag{43}$$

where $\mathcal{N}$ is the set of nodes, and the capacity of the cut $\mathcal{S}$ is given as

$$C(\mathcal{S}) = \sum_{\substack{(k,l) \in \mathcal{E} \\ k \in \mathcal{S}, l \in \mathcal{S}^c}} C_{kl} \tag{44}$$

where $\mathcal{E}$ is the set of all edges. The result is also called min-cut max-flow theorem. The capacity can be computed for different network architectures and it can also be combined with the capacity and achievable rate regions explained in Section II and the MAC schemes in Section IV.

For the example values in Table 3, the maximum sum rate for the midhaul and fronthaul is upper bounded by the cut illustrated in Fig. 23 as the dashed line. The maximum flow value for the cut is given by $1300 + 600 + 500 = 2400$ Mbit/s.

2) Classical OMA: Suppose that we use the classic orthogonal MAC schemes, where a user is served by only one RU on orthogonal resources only. Then, we need an assignment of users to RUs. In the simple example in Fig. 24, we consider the canonical assignment of RU$_i$ ↔ UE$_i$, $i = 1, 2, 3$.

Then, the resulting graph degenerates to a tree with the source CU and the three leaves UE$_1$–UE$_3$ and the unassigned RU. For the anecdotal example, we assume that the channel matrix $\boldsymbol{H}$, which describes the channel from

RUs to the users, is given as follows:

$$\boldsymbol{H} = \begin{bmatrix} 10 & 8 & 2 \\ 9 & 10 & 4 \\ 3 & 9 & 9 \\ 1 & 7 & 10 \end{bmatrix}. \tag{45}$$

Note that these channel gains can include the contributions of multiple antennas at the RUs. Then, the channel gain is computed as in the SINR expression in (38) as $h_{ij} = |\boldsymbol{h}_j^H \boldsymbol{v}_i|^2$.

For the wireless links, we further assume for simplicity that the transmit SNR is 10 dB and the total bandwidth is $B = 300$ MHz. For OFDMA, this given $W = 100$ MHz for each user. For the orthogonal wireless link from RU_1 to UE_1, this gives the P2P capacity in (1) calculated as

$$C_{11} = W \log_2(1 + 10 \cdot 10) = 692 \text{ Mbit/s}.$$

Similarly, we compute the other two rates $C_{22} = 692$ Mbit/s and $C_{33} = 651$ Mbit/s. The rate limits r_1–r_3 for the three users are computed as follows:

$$\begin{aligned} r_1 \le C_{11},\ r_2 \le C_{22},\ r_3 \le C_{33} \\ r_1 \le d_{11},\ r_2 \le d_{12},,r_3 \le d_{23} \\ d_{11} + d_{12} \le c_1,\ d_{23} \le c_2. \end{aligned} \tag{46}$$

The inequalities in (46) are linear in the rates and can be easily checked via linear programming. In the anecdotal example, the three users are limited by $r_1 + r_2 \le 1300$ Mbit/s and $r_3 \le d_{23} = 600$ Mbit/s. In total, a sum rate of 1900 Mbit/s is achievable. Compared to the upper bound from Section V-A1 there are 500 MBit/s left for improvements.

Note that in this example, we are only interested in the achievable rate and neglect both latencies and energy consumption. Furthermore, we assume homogeneous services for the three users. In a complete system design, the end-to-end performance for heterogeneous users can be optimized using network slicing [126]. The ML-based resource allocation under uncertainty is able to solve the corresponding nonconvex mixed-integer nonlinear programming problem.

In order to improve the data rates of users 1 and 2, the orthogonal resource allocation should be removed and NOMA schemes should be applied as we have discussed in Section IV.

One approach to remove the limitation of user 3 from the fronthaul constraint d_{23} would be to let the user connect to multiple RUs, in this case RU_3 and RU_4. The same approach can be useful for users 1 and 2 that share the spectrum in an orthogonal way. In the case of two connections, this scheme is called dual connectivity, while for more connections, it is called multiconnectivity.

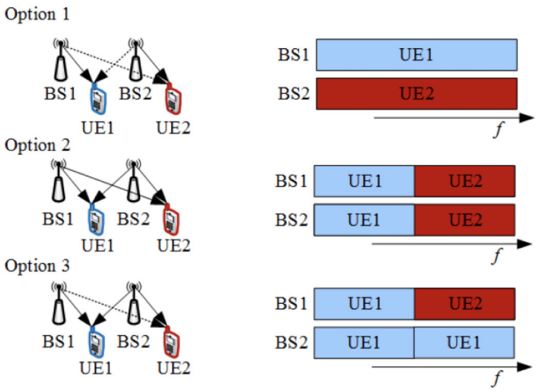

Fig. 25. ***Connectivity options in a two BSs and two users scenario. Solid lines are desired links and dashed lines represent the nondesired, interfering links (adapted from [127]).***

3) Multiconnectivity: Multiple BSs operating on the same carrier frequency transmit simultaneously to the same user so that interfering BSs become wanted BSs, resulting in improved SINR [127]. This is one option for cooperation between BSs, as also explained in Section IV. Another possibility is to serve a user with different data sources in an incoherent way.

Fig. 25 shows different options for the dual connectivity. Option 1 is the situation for our example above where RU_1 serves user 1 and RU_2 serves user 2. In the orthogonal resource allocation, they are using different frequency bands.

In Fig. 25, they use the same frequency band and cause interference. This leads to the IC situation described in Section II-E. As it can be seen from the channel matrix $\boldsymbol{H}$ in (45), the interference between the two links is moderate. Let us calculate the achievable rates with TIN as in (18) with $W = 200$ MHz: $R_{11} = 200 \log_2(1 + 10 \cdot 10/(1 + 10 \cdot 9)) = 214$ Mbit/s and $R_{22} = \log_2(1 + 10 \cdot 10/(1 + 10 \cdot 8)) = 232$ Mbit/s. Due to the large interference power, using TIN does not outperform TDMA.

The second option in Fig. 25 allows both users to be supported from both RU_1 and RU_2 and the interference to be resolved by OFDMA. In this case, the achievable rates for user 1 using $W = 100$ MHz for noncoherent JT are calculated as $R_1 = 100 \log_2(1 + 10 \cdot 10 + 10 \cdot 9) = 758$ Mb/s, while the second user gets $R_2 = 100 \log_2(1 + 10 \cdot 10 + 10 \cdot 80) = 750$ Mb/s. Again, the midhaul and fronthaul links will limit the total throughput.

The third option in Fig. 25 is heterogeneous multiconnectivity, where user 1 is served by both RUs on both frequency bands, while user 2 is served by only one RU on frequency band 2. This could be a useful option if the service requirements of the users are different.

An important challenge is how to match users to RUs. This could be solved using matching theory [128], which leads to stable user assignments [129]. Alternatively, assignments can be found by combinatorial auctions [130].

4) Multiconnectivity With Rate Splitting: For UE_3 in our anecdotal example from Fig. 23, multiconnectivity could improve the data rate if it is combined with RS. Let us perform RS at DU_2 and then send the two different messages via RU_3 and RU_4 to user 3. Similar to the derivation in Section II-B, we can realize a data rate (ignoring the interference caused from RU_1 and RU_2 for simplicity) with $W = 100$ MHz

$$\begin{aligned} R_{33} &\leq WC\left(\frac{10\cdot 9}{1+10\cdot 7}\right) = 118\,\text{Mbit/s} \\ R_{43} &\leq WC\,(10\cdot 10) = 666\,\text{Mbit/s} \\ R_3 &= R_{33} + \min\left(R_{43}, d_{24}\right) = 618\,\text{Mbit/s}. \end{aligned} \tag{47}$$

The sum rate achieved is 1918 Mbit/s.

5) Cloud-RAN: Another special case of the Open RAN architecture in Fig. 24 is the C-RAN, where all RUs are directly connected to a CU that centrally operates and controls all parameters [131], resource allocation and PHY optimization. The main idea behind C-RAN is to combine the BBUs of several BSs into a centralized BBU pool for statistical multiplexing gain while shifting the burden to high-speed wired transmission of in-phase and quadrature (IQ) data [132]. The idea of moving the necessary transmission and processing resources for a wireless access network to the cloud has already been formulated in [133].

In this case, the midhaul link capacities in Fig. 24 would be set to infinity. This makes the entire network act as a massive distributed antenna system covering the entire network coverage area. This case was illustrated in Fig. 15, where all BSs transmit together to serve the users. From an information-theoretic point of view, the downlink would correspond to a two-hop or BC, while the uplink is the two-hop MAC. The two-hop comes from the transmission from the CU to the RUs.

The achievable rates can be calculated based on the fundamental limits from Section II and constrained by the fronthaul rate limits. Since all RUs serve all users, they require the data for all users sent over the fronthaul links. The total sum data rate is limited by $\min(d_{ij})$ over all i, j. This shows that this network architecture requires significant computing and processing power from the CU, including very high data links for the fronthaul, especially with the massive increase in the number of UEs per unit area in beyond 5G/6G networks.

In the heterogeneous C-RAN, a combination of distributed RUs controlled by the CU and macro BC is used. The offloading of macrocell users to the C-RAN and efficient resource allocation is discussed in [134]. The coexistence between macro BS and C-RAN can be modeled by cognitive radio approaches where LSA is negotiated between primary and secondary users [135].

The combination of RS and C-RAN is studied in [136] and [137]. Statistical CSI at the transmitter is considered and the problem of stochastic coordinated beamforming for ergodic sum-rate maximization is proposed and solved. A gain of up to 27% of RS over TIN and NOMA is reported.

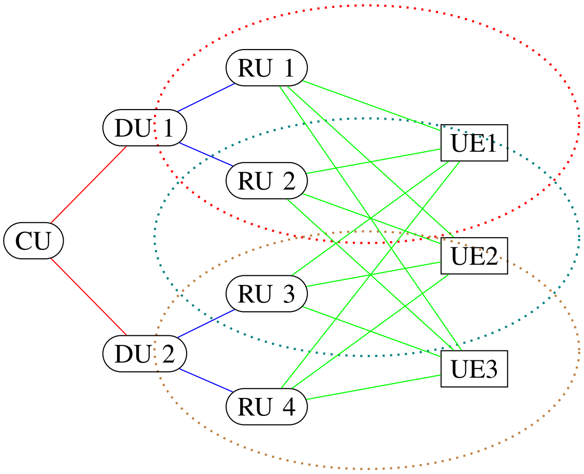

Fig. 26. ***Open RAN architecture with cell-free configuration and user clusters.***

B. Cell-Free Multiple Access

The CFMA network architecture is a user-centric design. Each user connects to a number of RUs required to obtain the service. This is a departure from the network-centric design of all the schemes discussed above. The RUs serving a user form a cooperation cluster for that particular user. One of the first references to CFMA with massive numbers of RUs is [138].

The three clusters in Fig. 26 can be interpreted as three distributed antenna systems or three multiple-antenna cells serving the three users. Depending on the functional split chosen, part of the signal processing is performed at the RUs, and part is performed at the DUs or the CU [139].

The main challenge is to achieve the benefits of cell-free operation practically, with computational complexity and fronthaul requirements that are scalable to enable massively large networks with many mobile devices [140]. The monograph [140] describes the state-of-the-art signal processing algorithms for channel estimation, uplink data reception, and downlink data transmission with either centralized or distributed implementation.

Let us first concentrate on one user, e.g., user 1, and its corresponding cluster, i.e., RU_1 and RU_2 (red dashed ellipsoid in Fig. 26. Let us denote the channel coefficients as h_{11} and h_{21}, which are complex numbers with attenuation (amplitudes) and delay (phase). This looks like an MISO channel with two transmit antennas and one receive antenna. With perfect CSI at the transmitter and with a sum power constraint $P = 2$, the optimal beamforming strategy is MRT and the resulting effective channel is $h_{12}^{mrt} = |h_{11}|^2 + |h_{21}|^2$. Usually, each RU has its own power constraint $P = 1$. Then, the optimal beamforming strategy is to adjust the phase of the two signals arriving at the user (similar to equal gain transmission) and obtain the effective channel $h_{12}^{egt} = 1/2 \cdot (|h_{11}| + |h_{21}|)^2$. In the literature, this is also called coherent beamforming or joint beamforming. Without accurate phase information at the two RUs, noncoherent beamforming can be applied, which achieves an effective channel $h_{12}^{nch} = 1/2 \cdot (|h_{11}|^2 + |h_{21}|^2)$. Clearly

$$h_{12}^{mrt} \geq h_{12}^{egt} \geq h_{12}^{nch} \tag{48}$$

where equality is achieved for the first inequality with $|h_1| = |h_2|$ and for the second inequality if at least one channel is zero. For our example channel matrix $\boldsymbol{H}$, we have the following inequalities for the three transmission schemes: $19 \geq 18.99 \geq 8.5$. This shows clearly that coherent beamforming achieves a significant gain in terms of received signal power because the channels to user 1 have similar gains.

The downlink SINR in (38) can be specialized to the example. The SINR expression for user 1, including power control and interference (teal and brown dashed clusters in Fig. 26) from the other users' codewords, is given by

$$\text{SINR}_1 = \frac{\left(|h_1|\sqrt{p_{11}} + |h_2|\sqrt{p_{21}}\right)^2}{\sigma^2 + |h_{21}|^2 p_{22} + |h_{31}|^2 (p_{32} + p_{33}) + |h_{41}|^2 p_{43}}. \tag{49}$$

Similar expressions for the SINRs of the other users can be derived from the power allocation. Next, various power control problems can be formulated, including min–max power control [138], [141], sum-rate maximization [142], energy efficiency maximization [143], and power minimization under rate constraints [144]. There are also results on power control under fronthaul and midhaul constraints. There are also differentiated results for uplink and downlink operation and for the pilot signal phase.

The combination of RS at the RUs to improve the achievable data rates of the wireless is performed in [145] for specific beamformers and in terms of sum rate. In [146], max–min power control for RS in cell-free MIMO is performed. Robustness against pilot contamination is reported. Finally, Zheng et al. [147] report on asynchronous cell-free massive MIMO with RS and its robustness to hardware impairments.

In the case of multiple-antenna RUs, there are several proposals for beamforming optimization. Let us denote the local channels at the multiple-antenna RU by $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_L$ for L users. The conjugate beamformer is then given by $\boldsymbol{w}_k = \boldsymbol{h}_k^*$ [141]. Variants include extended normalized conjugate beamforming $\boldsymbol{w}_k = \boldsymbol{h}_k^* \| |\boldsymbol{h}_k^*||$ [148], modified conjugate beamforming [149], local partial ZF precoding [150], and team MMSE precoding [151].

Note that we have not discussed the challenges associated with obtaining the CSI at the transmitter. This is usually achieved by TDD operation and exploitation of channel reciprocity. The interested reader is referred to the monograph [140].

As we have seen in our anecdotal example in Section V-A2, an optimal design must take into account the constraints and limitations of both midhaul, fronthaul, and wireless access. In [152], joint fronthaul load balancing and compute resource allocation is performed. We follow a slightly different approach for simplicity. For a holistic design, fronthaul constraints could be modeled by simple rate constraints, as explained in Section V-A2. Let us conclude the anecdotal example with a combination of OFDMA, CFMA, and RS. We allocate 100 MHz to each user (OFDMA), assign users according to the clusters shown in Fig. 24, and perform RS for all three users.[2]

1) *Cluster user 1 (decoding order):* First, codewords received from RU$_2$, and then, codewords received from RU$_1$, i.e.,

$$\begin{aligned} R_{21} &= W \cdot C\left(\frac{10 \cdot 9}{1 + 10 \cdot 10}\right) = 92\,\text{Mbit/s} \\ R_{11} &= W \cdot C\,(100) = 666\,\text{Mbit/s} \\ R_1 &= 666 + 92 = 758\,\text{Mbit/s}. \end{aligned} \tag{50}$$

2) *Cluster user 2:* The sum rate is directly computed as

$$R_2 = R_{22} + R_{32} = W \cdot C\,(100 + 90) = 758\,\text{Mbit/s} \tag{51}$$

and by using TS, the rate can be split between the two decoding orders, as shown in Fig. 6. The total rate for both RUs is divided equally into 379 Mb/s each.

3) *Cluster user 3 (decoding order):* First, codewords received from RU$_3$ followed by codewords from RU$_4$. Note also that the fronthaul between DU$_2$ and RU$_4$ is limited to 500 Mbit/s. Therefore, the power required for RU$_4$ can be reduced to achieve exactly the maximum of 500 Mb/s, which is SNR$_4 = 3.1$

$$\begin{aligned} R_{33} &= W \cdot C\left(\frac{10 \cdot 9}{1 + 3.1 \cdot 10}\right) = 193\,\text{Mbit/s} \\ R_{43} &= W \cdot C\,(10 \cdot 3.1) = 500\,\text{MBit/s} \\ R_3 &= R_{33} + R_{43} = 693\,\text{MBit/s}. \end{aligned} \tag{52}$$

4) *Fronthaul constraints:* $R_{11} = 666 \leq 700 = d_{11}$, $R_{21} + R_{22} = 92 + 379 = 471 < 800 = d_{12}$, $R_{32} + R_{33} = 379 + 193 = 572 < d_{23}$, $R_{43} = 500 = d_{34}$, and the midhaul constraints: $666 + 471 = 1137 < c_1$ and $572 + 500 = 1072 < 1500$.

The total sum rate achieved is 2209 Mbit/s, which is much closer to the upper limit of 2400 Mbit/s than the baseline scheme with single-user allocations and OFDMA, which

[2] Note that we could also optimize powers and rates, but this is beyond the scope of this tutorial. Initial results are reported in [153].

achieved 1900 Mbit/s. This is an improvement of over 300 Mbit/s.

C. Anecdotal Two-User Example

To illustrate the performance of the different options of fronthaul architecture, coding scheme, and resource allocation, we consider an anecdotal example with a minimum number of devices, RUs, and antennas (adapted from [153]). The CU is connected with two RUs via the fronthaul links. The rate constraints on the two fronthaul links are $r_1 > 0$ and $r_2 > 0$. The two RUs serve two receivers over a standard IC [154]. The channel gain of the direct channels is normalized, while the two ICs have a channel gain of a. The two transmitters transmit the codewords with transmit power p_1 and p_2. The two encoders receive their data D_1 and D_2 from the CU. The inverse noise power is denoted by $\rho = (1/\sigma^2)$.

The received signal at user i is given by

$$Y_i = \sqrt{p_i} X_i + \sqrt{p_j} a X_j + Z_i \tag{53}$$

where $j \neq i$ and $1 \leq i, j \leq 2$. We assume that the fronthaul links from the CU to the RUs are error-free and the control signaling is neglected, i.e., only required user data are considered at the RU when load on the fronthaul links is computed.

The considered network architecture and transmission schemes are based on so-called random P2P coding schemes [28]. Gaussian codebooks are applied for all data streams for all users. Hence, we will operate with Shannon capacities and achievable rates of the following type:

$$R_i = \log\left(1 + \mathrm{SINR}_i\right) \tag{54}$$

where SINR is typically given by

$$\mathrm{SINR}_i = \frac{\sum_{\ell} h_{\ell i} p_{\ell i}}{\sigma^2 + \sum_{\ell} \sum_{j \neq i} h_{\ell i} p_{\ell j}} \tag{55}$$

where the contributions of the received signal powers are added for the signal of interest as well as the interference. Note that this SINR expression differs from the one obtained for coherent beamforming in cell-free systems, e.g., in [140]. The reason for being more conservative with the received signal power computation is that the requirements in terms of synchronization at the RUs are very demanding. Furthermore, RS can be easily applied to achieve the SINR expression in (55). We have seen in Section II-B that even for a P2P channel, RS can be applied to achieve the same sum rate as using a single Gaussian codebook.

In the baseline schemes, we are not using any cell-free or distributed MIMO architecture, but the standard IC. RU 1 serves user 1 and RU 2 serves user 2 on the same time and frequency. The next set of schemes serves both users using both distributed antennas or RUs. The transmitter does not apply RS. Noncoherent transmission and either TIN or SIC is performed at the receiver.

The idea of the cell-free architecture with RS is to split the messages for both users at the CU into two parts each. The two messages for each user are then sent over the two RUs. The users can then decide how to decode the superposition of the four codewords. In order to obtain the two intended messages, at least the two codewords belonging to the own messages should be decoded. In addition, the interference from the other user's codewords could be decoded, too.

The simulation is carried out for different IC gains a and the results are presented in Fig. 27. The noise power is normalized to one and the transmit power is varied to operate at certain SNR regions. Moreover, we distinguish between the following three regions of a: $a < 0.5$, $0.5 \leq a \leq 1$, and $1 < a$ are scenarios with low, medium, and high interference, respectively. The performance metric is the achievable sum rate of the two receivers.

As for the baseline schemes (see Section II-E), TIN performs well at low interference and a drop in performance can be seen in medium interference scenarios, whereas the performance of the SIC scheme increases with interference and achieves best performance at high interfering scenarios. This can be understood as TIN is optimal for low interference in terms of sum capacity [38]. For strong interference, i.e., $a > 1$, SIC at both receivers achieves the complete capacity region. For stronger interference and using TIN, TDMA is optimal, i.e., in order to maximize the sum rate, only one user will be supported.

Both CF (without RS) schemes are limited by the fronthaul constraint because the sum user data rate is transmitted over the fronthaul to both RUs. Similar to their baseline counterparts, TIN achieves better sum rates at low interference. However, for medium and high interferences, the difference in performance is negligible.

As expected, the best-performing scheme in this scenario is *CF with RS*. It outperforms all introduced schemes in any interference region. Due to RS, it is not limited by the fronthaul and also performs better at higher interference. This performance gain is achieved by applying RS already at the CU. The additional flexibility of allocating the rates of the two data streams per user is used to fully exploit the fronthaul rates.

Even though this shows only an anecdotal example, the results motivate to consider the design of midhaul, fronthaul, and wireless access jointly.

D. Unsourced Massive Random Access

In Section II, the model contained a predefined set of transmitters that want to communicate with a predefined set of receivers. In contrast, in massive random access, there exists a large number of wireless nodes, which are not all always active. Not all transmitters are active and they also do not always have data to transmit, and the packets could be rather short. Neither the receiver nor the

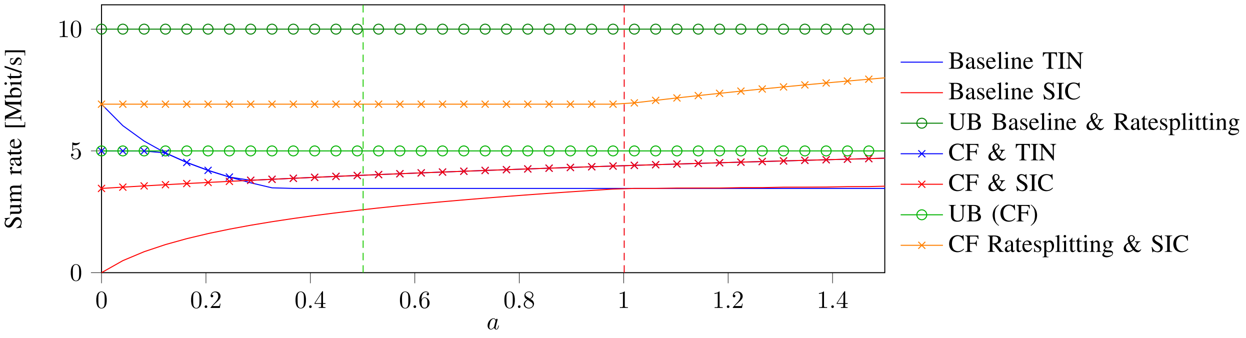

Fig. 27. ***Comparison of sum-rate performance of the introduced schemes for various ICs a with fronthaul load constraint ($r_1 = r_2 = 5$ Mbit/s) and inverse noise power $\rho = 1$ for power constraints $P_1 = P_2 = 10$ W (adapted from [153]).***

transmitters know the activity. This situation can occur in IoT or WSN scenarios.

If the classical approach MAC outlined in Section II is applied in these scenarios, the achievable rate for the users is very small: As an example, if there are 100 nodes deployed and the transmit power is normalized to one, each transmitter could transmit 0.03 $= (1/100)C(100P)$ BPCU, which is 1.

Therefore, new approaches to achieve higher rates in massive random access have been considered recently. In [155], the UMAC model is proposed, where a number of active transmitters from the set of all transmitters use the same codebook to send their messages. In UMAC systems, all transmitters share an identical codebook and the amount of data transmitted at each transmitter is the same. The receiver then tries to decode the set of messages sent but cannot distinguish between users, i.e., the source of the messages, hence, the name *unsourced* multiple access.

The setup is quite different from the traditional MAC. Let K be the total number of senders, where k is the number of active senders. M is the number of messages. In agnostic RACH, neither the sender nor the receiver knows the active set of senders. All transmitters have the same common codebook. The decoder's task is then to estimate the number of active users k and the set of corresponding messages. The first step could be done by energy detection, while the second step could be done by information density threshold detector [156], [157].

In [158], the first-order capacity is studied when the number of users is some function of the block length, and users use individual codebooks for identification and an identical codebook for information transmission. However, the achievable rates vanish as the number of transmitters grows asymptotically. Therefore, the energy efficiency of synchronous UMAC with PUPE is investigated.

Since the distributed transmitters in UMAC are not necessarily synchronized and due to the different channel delays in the uplink, the asynchronous UMAC is recently studied. In [159], T-fold ALOHA, and [160], OFDM, the time-asynchronous problem is transformed into a frequency-shift problem. The maximum delay in [159] must be less than the length of the CP. Chen et al. [161] use a sparse OFDMA scheme and compressed sensing-based algorithms to reliably identify arbitrary asynchronous devices and decode messages. Finally, Wu et al. [162] derive worst case bounds on the PUPE.

It is important to note that recent code designs for UMAC include sparse regression codes [163], sparse graph codes [164], and polar codes [165]. Furthermore, Schiavone et al. [166] study the design of the enhanced spread spectrum ALOHA in the context of asynchronous UMAC. Another recent development is the combination of UMAC within the cell-free architecture [167]. Finally, Agostini et al. [168] summarize possible evolutions of grant-free random access in the next generation of the 3GPP wireless cellular standard.

VI. CONCLUSION AND OPEN CHALLENGES

We conclude the tutorial with a summary of the main topics and results that were explained and highlighted followed by a list of open challenges.

A. Conclusion

Modern multiple access techniques need to be flexible, adaptive, efficient, and scalable. They also need to be designed together with the network architecture and from a cross-layer perspective, i.e., physical layer techniques such as channel coding and decoding, spatial and spectral signal processing together with power control, resource allocation, and user assignment, taking into account network architecture constraints such as fronthaul and midhaul constraints. The basis for systematic system design is network information theory. Sound mathematical modeling and problem formulation is essential.

Therefore, the tutorial has started with a comprehensive review of results from network information theory, highlighting a number of key insights. The extension from single antennas to multiple antennas, multiple carriers, and multiple cells is described in terms of a mathematical framework. Within this framework, the classical coordinated multiple access techniques are introduced and compared. We then describe the modern multiple access techniques, including multiple connectivity, C-RAN, and CFMA. For completeness, recent developments in grant-free and random massive access are briefly introduced.

B. Open Challenges

There are a number of interesting and rich research problems in modern multiple access. These include fundamental information-theoretic research, signal processing for communications, optimization, software-defined radio and networking, and ML for communications.

It is important to stress that some fundamental capacity regions, e.g., for the general BC, for the IC with moderate interference, are not known and remain unsolved. SIC is a fundamental building block for the receivers in MAC, BC, and IC. While the first-order capacity analysis of SIC is well established, the second- or third-order analysis of joint decoding is only partially solved for the MAC [169] and for the BC [170]. In addition, at the boundary between coordinated and uncoordinated multiple access, the idealized system model must take into account the time dynamics of multiple packets (transmissions and retransmission attempts). This introduces memory and possibly feedback into the channel model. A unified analysis of the fundamental limitations of these models is left for future work.

In the tutorial, we have not discussed information-theoretic secure transmission. In particular, the wiretap channel model and its variants, including secure multiple access, are not addressed. In this area, there is ongoing work, which exploits physical layer security to address confidential, authenticated, and anonymous multiple access [171]. There are many open questions in the area of security and safety for modern multiple access techniques, too.

For fast fading channels with larger Doppler, OTFS modulation is currently being studied. Only a few papers have considered OTFS MAC. In [172], delay–Doppler resource blocks are carefully allocated to users to avoid MAI. In [173], resource allocation in the delay–Doppler domain is also studied. Pilot signal patterns are designed and the performance of OTFS MAC with OFDMA and single-carrier FDMA.

Many transceiver schemes for modern MAC require some degree of synchronization (see Table 2). In particular, coherent cell-free massive MIMO beamforming requires phase synchronization between distributed RUs [174]. Synchronization in distributed networks between mobile nodes is an ongoing research challenge.

The joint optimization of beamforming, power control, and allocation under midhaul, fronthaul, and wireless access constraints is relevant and challenging. Liu et al. [175] propose a multidimensional intelligent MAC optimization problem. The corresponding programming problems may be nonconvex mixed-integer programming problems that cannot be solved globally [63]. Therefore, the optimization and configuration of network parameters could be found by ML techniques [176].

Considering the computing power and storage at DUs and RUs, the corresponding programming problem becomes more complex. Storage at intermediate nodes could include the ability to cache content to reduce latency and improve efficiency. Finally, the reconfiguration of the network architecture by SDN controlled by VNF could lead to a native cloud network [177].

Finally, the integration of ML with distributed and intelligent decision-making for resource allocation in MAC is a growing area of research. General design guidelines for wireless communication in edge learning, collectively referred to as learning-driven communication, are outlined in [178]. Specifically for MAC, Liu et al. [179] propose the use of ML-based algorithms to solve the multidimensional intelligent MAC optimization problem. Finally, Wu et al. [180] apply ML-based algorithms to optimize the RSMA strategy in reconfigurable intelligent surface (RIS)-assisted wireless systems. The integration of the various MAC coding, decoding, beamforming, and resource allocation options into a distributed ML-based framework is lacking. There are several open research questions related to the efficient ML-based optimization of modern MAC schemes.

Acknowledgment

The author would like to thank the anonymous reviewers for their suggestions to improve the presentation, clarity, and comprehensiveness of the tutorial. The discussion with all colleagues from the 6G-research and innovation cluster (6G-RIC) is acknowledged. ■

REFERENCES

[1] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proc. IEEE*, vol. 109, no. 7, pp. 1166–1199, Jul. 2021.

[2] C.-X. Wang et al., "On the road to 6G: Visions, requirements, key technologies, and testbeds," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 905–974, 2nd Quart. 2023.

[3] *Future Technology Trends of Terrestrial International Mobile Telecommunications Systems Towards 2030 and Beyond*, document ITU M.2516, ITU, 2022.

[4] O. Aydin, E. A. Jorswieck, D. Aziz, and A. Zappone, "Energy-spectral efficiency tradeoffs in 5G multi-operator networks with heterogeneous constraints," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5869–5881, Sep. 2017.

[5] C. D. Alwis et al., "Survey on 6G frontiers: Trends, applications, requirements, technologies and future research," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 836–886, 2021.

[6] C. Tian, J. Chen, S. N. Diggavi, and S. S. Shamai, "Optimality and approximate optimality of source-channel separation in networks," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 904–918, Feb. 2014.

[7] J. Bartelt, P. Rost, D. Wubben, J. Lessmann, B. Melis, and G. Fettweis, "Fronthaul and backhaul requirements of flexibly centralized radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 105–111, Oct. 2015.

[8] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018.

[9] G. P. Fettweis and H. Boche, "6G: The personal tactile internet—And open questions for information theory," *IEEE BITS Inf. Theory Mag.*, vol. 1, no. 1, pp. 71–82, Sep. 2021.

[10] M. Soleymani, I. Santamaria, E. Jorswieck, and S. Rezvani, "NOMA-based improper signaling for multicell MISO RIS-assisted broadcast channels," *IEEE Trans. Signal Process.*, vol. 71, pp. 963–978, 2023.

[11] N. H. Mahmood, I. Atzeni, E. A. Jorswieck, and O. L. A. Lόpez, "Ultra-reliable low-latency communications: Foundations, enablers, system design, and evolution towards 6G," *Found. Trends Commun. Inf. Theory*, vol. 20, nos. 5–6, pp. 512–747, 2023, doi: 10.1561/0100000129.

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

[13] V. Y. F. Tan and O. Kosut, "On the dispersions of three network information theory problems," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 881–903, Feb. 2014.

[14] N. Abramson, "Multiple access in wireless digital networks," *Proc. IEEE*, vol. 82, no. 9, pp. 1360–1370, Sep. 1994.

[15] S. S. Shamai and A. D. Wyner, "Information-theoretic considerations for symmetric, cellular, multiple-access fading

channels. I," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1877–1894, Nov. 1997.

[16] O. Somekh and S. S. Shamai, "Shannon-theoretic approach to a Gaussian cellular multiple-access channel with fading," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1401–1425, Jul. 2000.

[17] A. Jamalipour, T. Wada, and T. Yamazato, "A tutorial on multiple access technologies for beyond 3G mobile networks," *IEEE Commun. Mag.*, vol. 43, no. 2, pp. 110–117, Feb. 2005.

[18] M. Aldababsa, M. Toka, S. Gökçeli, G. K. Kurt, and O. Kucur, "A tutorial on nonorthogonal multiple access for 5G and beyond," *Wireless Commun. Mobile Comput.*, vol. 2018, Jun. 2018, Art. no. 9713450, doi: 10.1155/2018/9713450.

[19] Q.-V. Pham et al., "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116974–117017, 2020.

[20] A. F. M. S. Shah, A. N. Qasim, M. A. Karabulut, H. Ilhan, and M. B. Islam, "Survey and performance evaluation of multiple access schemes for next-generation wireless communication systems," *IEEE Access*, vol. 9, pp. 113428–113442, 2021.

[21] S. Chaturvedi, Z. Liu, V. Bohara, A. Srivastawa, and P. Xiao, "A tutorial on decoding techniques of sparse code multiple access," *IEEE Access*, vol. 10, pp. 58503–58524, 2022.

[22] B. Clerckx et al., "A primer on rate-splitting multiple access: Tutorial, myths, and frequently asked questions," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1265–1308, May 2023.

[23] B. Clerckx et al., "Multiple access techniques for intelligent and multi-functional 6G: Tutorial, survey, and outlook," 2024, *arXiv:2401.01433*.

[24] Y. Wu, X. Gao, S. Zhou, W. Yang, Y. Polyanskiy, and G. Caire, "Massive access for future wireless communication systems," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 148–156, Aug. 2020.

[25] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[26] E. Biglieri, J. Proakis, and S. S. Shamai, "Fading channels: Information-theoretic and communications aspects," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.

[27] F. Baccelli, A. El Gamal, and D. Tse, "Interference networks with point-to-point codes," in *IEEE Int. Symp. Inf. Theory*, Jul. 2011, pp. 435–439.

[28] B. Bandemer, A. El Gamal, and Y. Kim, "Optimal achievable rates for interference networks with random codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6536–6549, Dec. 2015.

[29] S. N. Diggavi and T. M. Cover, "The worst additive noise under a covariance constraint," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 3072–3081, Nov. 2001.

[30] D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[31] A. Tajer, A. Steiner, and S. S. Shamai, "The broadcast approach in communication networks," *Entropy*, vol. 23, no. 1, p. 120, Jan. 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/120

[32] H. H.-J. Liao, "Multiple access channels," Ph.D. dissertation, Dept. Elect. Eng., Univ. Hawaii Honolulu, Honolulu, HI, USA, 1972.

[33] T. Cover, "Broadcast channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 2–14, Jan. 1972.

[34] T. Cover, "Comments on broadcast channels," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2524–2530, Oct. 1998.

[35] N. Jindal and A. Goldsmith, "Dirty-paper coding versus TDMA for MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 5, pp. 1783–1794, May 2005.

[36] A. S. Motahari and A. K. Khandani, "Capacity bounds for the Gaussian interference channel," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 620–643, Feb. 2009.

[37] X. Shang, G. Kramer, and B. Chen, "A new outer bound and the noisy-interference sum-rate capacity for Gaussian interference channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 689–699, Feb. 2009.

[38] V. S. Annapureddy and V. V. Veeravalli, "Gaussian interference networks: Sum capacity in the low-interference regime and new outer bounds on the capacity region," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3032–3050, Jul. 2009.

[39] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 1, pp. 49–60, Jan. 1981.

[40] R. H. Etkin, D. N. C. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5534–5562, Dec. 2008.

[41] I. Sason, "On the corner points of the capacity region of a two-user Gaussian interference channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 7, pp. 3682–3697, Jul. 2015.

[42] A. Vahid, M. A. Maddah-Ali, A. S. Avestimehr, and Y. Zhu, "Binary fading interference channel with no CSIT," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3565–3578, Jun. 2017.

[43] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1991.

[44] E. A. Jorswieck and H. Boche, *Majorization and Matrix Monotone Functions in Wireless Communications* (Foundations and Trends in Communications and Information Theory), vol. 3, S. Verdú, Ed. Boston, MA, USA: Now, 2007.

[45] C. Xing, Y. Jing, S. Wang, S. Ma, and H. V. Poor, "New viewpoint and algorithms for water-filling solutions in wireless communications," *IEEE Trans. Signal Process.*, vol. 68, pp. 1618–1634, 2020.

[46] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[47] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.

[48] R. S. Cheng and S. Verdu, "Gaussian multiaccess channels with ISI: Capacity region and multiuser water-filling," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 773–785, May 1993.

[49] S. Vishwanath, W. Rhee, N. Jindal, S. Jafar, and A. Goldsmith, "Sum power iterative waterfilling for Gaussian vector broadcast channels," in *Proc. IEEE ISIT*, Jun. 2003, p. 467.

[50] A. J. Goldsmith and M. Effros, "The capacity region of broadcast channels with intersymbol interference and colored Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 47, no. 1, pp. 219–240, Jan. 2001.

[51] D. N. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 1997, p. 27.

[52] P.-J. Wan and P. Chen, "Maximizing weighted sum-rate over Gaussian broadcast channels," *IEEE Trans. Inf. Theory*, vol. 70, no. 4, pp. 2922–2935, Apr. 2024.

[53] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.

[54] V. R. Cadambe and S. A. Jafar, "Parallel Gaussian interference channels are not always separable," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 3983–3990, Sep. 2009.

[55] J. M. Cioffi, S. Jagannathan, M. Mohseni, and G. Ginis, "CuPON: The copper alternative to PON 100 Gb/s DSL networks," *IEEE Commun. Mag.*, vol. 45, no. 6, pp. 132–139, Jun. 2007.

[56] Y. Zhao and G. J. Pottie, "Optimal spectrum management in multiuser interference channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 8, pp. 4961–4976, Aug. 2013.

[57] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.

[58] J.-S. Pang, G. Scutari, F. Facchinei, and C. Wang, "Distributed power allocation with rate constraints in Gaussian parallel interference channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3471–3489, Aug. 2008.

[59] L. P. Qian and Y. J. Zhang, "S-MAPEL: Monotonic optimization for non-convex joint power control and scheduling problems," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1708–1719, May 2010.

[60] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[61] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[62] B. Matthiesen, C. Hellings, E. A. Jorswieck, and W. Utschick, "Mixed monotonic programming for fast global optimization," *IEEE Trans. Signal Process.*, vol. 68, pp. 2529–2544, 2020.

[63] Y.-F. Liu et al., "A survey of recent advances in optimization methods for wireless communications," 2024, *arXiv:2401.12025*.

[64] I. E. Telatar and R. G. Gallager, "Combining queueing theory with information theory for multiaccess," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 963–969, Aug. 1995.

[65] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 684–702, Jun. 2003.

[66] H. Weingarten, Y. Steinberg, and S. S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.

[67] X. Shang, B. Chen, G. Kramer, and H. V. Poor, "Capacity regions and sum-rate capacities of vector Gaussian interference channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 5030–5044, Oct. 2010.

[68] B. Clerckx and C. Oestges, *MIMO Wireless Networks: Channels, Techniques and Standards for Multi-Antenna, Multi-user and Multi-cell Systems*. New York, NY, USA: Academic, 2013.

[69] E. A. Jorswieck, *Transmission Strategies for the MIMO MAC*. London, U.K.: Hindawi, 2005, ch. 21, pp. 423–442.

[70] E. Jorswieck and S. Rezvani, "On the optimality of NOMA in two-user downlink multiple antenna channels," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 831–835.

[71] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.

[72] R. Hunger, *Dualities for the MIMO BC and the MIMO MAC With Linear Transceivers*. Berlin, Germany: Springer, 2013, pp. 17–53, doi: 10.1007/978-3-642-31692-0_3.

[73] S. Belhadj Amor, Y. Steinberg, and M. Wigger, "MIMO MAC-BC duality with linear-feedback coding schemes," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 5976–5998, Nov. 2015.

[74] C. Hellings and W. Utschick, "On the inseparability of parallel MIMO broadcast channels with linear transceivers," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6273–6278, Dec. 2011.

[75] H. Joudeh and B. Clerckx, "On the separability of parallel MISO broadcast channels under partial CSIT: A degrees of freedom region perspective," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4513–4529, Jul. 2020.

[76] A. D. Wyner, "Shannon-theoretic approach to a Gaussian cellular multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1713–1727, Nov. 1994.

[77] J. Xu, J. Zhang, and J. G. Andrews, "On the accuracy of the Wyner model in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, pp. 3098–3109, Sep. 2011.

[78] E. Björnson, "Optimal resource allocation in coordinated multi-cell systems," *Found. Trends Commun. Inf. Theory*, vol. 9, nos. 2–3, pp. 113–381, 2013, doi: 10.1561/0100000069.

[79] M. Eriksson, "Dynamic single frequency networks," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 10, pp. 1905–1914, Oct. 2001.

[80] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[81] J. G. Andrews, A. K. Gupta, A. Alammouri, and H. S. Dhillon, *An Introduction to Cellular Network Analysis Using Stochastic Geometry*. Cham, Switzerland: Springer, 2023.

[82] K. Beauchamp, *History of Telegraphy*. Edison, NJ, USA: IET, 2001.

[83] J. Song et al., "WirelessHART: Applying wireless technology in real-time industrial process control," *Proc. IEEE Real-Time Embedded Technol. Appl. Symp.*, Apr. 2008, pp. 377–386.

[84] H. Pan and S. C. Liew, "Information update: TDMA or FDMA?" *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 856–860, Jun. 2020.

[85] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1986–1992, Nov. 1997.

[86] S. B. Weinstein, "The history of orthogonal frequency-division multiplexing [history of communications]," *IEEE Commun. Mag.*, vol. 47, no. 11, pp. 26–35, Nov. 2009.

[87] C. Shahriar et al., "PHY-layer resiliency in OFDM communications: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 292–314, 1st Quart., 2015.

[88] T. Hwang, C. Yang, G. Wu, S. Li, and G. Ye Li, "OFDM and its wireless applications: A survey," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1673–1694, May 2009.

[89] H. Harkat, P. Monteiro, A. Gameiro, F. Guiomar, and H. Farhana Thariq Ahmed, "A survey on MIMO-OFDM systems: Review of recent trends," *Signals*, vol. 3, no. 2, pp. 359–395, Jun. 2022. [Online]. Available: https://www.mdpi.com/2624-6120/3/2/23

[90] M. R. Abedi, M. R. Javan, M. Pourghasemian, N. Mokari, and E. A. Jorswieck, "AI-assisted dynamic frame structure with intelligent puncturing schemes for 5G networks," *IEEE Access*, vol. 11, pp. 113995–114012, 2023.

[91] B. Farhang-Boroujeny and H. Moradi, "OFDM inspired waveforms for 5G," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2474–2492, 4th Quart., 2016.

[92] L. Gaudio, G. Colavolpe, and G. Caire, "OTFS vs. OFDM in the presence of sparsity: A fair comparison," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4410–4423, Jun. 2022.

[93] S. K. Mohammed, R. Hadani, A. Chockalingam, and R. Calderbank, "OTFS—Predictability in the delay-Doppler domain and its value to communication and radar sensing," *IEEE BITS Inf. Theory Mag.*, early access, Sep. 27, 2023, doi: 10.1109/MBITS.2023.3319595.

[94] R. Prasad and T. Ojanpera, "An overview of CDMA evolution toward wideband CDMA," *IEEE Commun. Surveys*, vol. 1, no. 1, pp. 2–29, 1st Quart., 1998.

[95] V. V. Veeravalli and A. Mantravadi, "The coding-spreading tradeoff in CDMA systems," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 2, pp. 396–408, Feb. 2002.

[96] S. Verdú, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[97] Y. Shan, C. Peng, L. Lin, Z. Jianchi, and S. Xiaoming, "Uplink multiple access schemes for 5G: A survey," *ZTE Commun.*, vol. 15, no. S1, pp. 31–40, Jun. 2017.

[98] L. Yu et al., "Sparse code multiple access for 6G wireless communication networks: Recent advances and future directions," *IEEE Commun. Standards Mag.*, vol. 5, no. 2, pp. 92–99, Jun. 2021.

[99] C. Farsakh and J. A. Nossek, "Application of space division multiple access to mobile radio," in *Proc. 5th IEEE Int. Symp. Pers., Indoor Mobile Radio Commun., Wireless Netw., Catching Mobile Future.*, vol. 2, 1994, pp. 736–739.

[100] T. L. Marzetta, E. G. Larsson, and H. Yang, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[101] E. Björnson, J. Hoydis, and L. Sanguinetti, *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency*. Boston, MA, USA: Now, 2017.

[102] E. A. Jorswieck, E. G. Larsson, and D. Danev, "Complete characterization of the Pareto boundary for the MISO interference channel," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5292–5296, Oct. 2008.

[103] S. Willhammar, J. Flordelis, L. Van Der Perre, and F. Tufvesson, "Channel hardening in massive MIMO: Model parameters and experimental assessment," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 501–512, 2020.

[104] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf. (VTC Spring)*, Jun. 2013, pp. 1–5.

[105] M. Vaezi and H. V. Poor, *NOMA: An Information Theoretic Perspective*. Cham, Switzerland: Springer, 2019, pp. 167–193, doi: 10.1007/978-3-319-92090-0_5.

[106] S. Rezvani, E. A. Jorswieck, R. Joda, and H. Yanikomeroglu, "Optimal power allocation in downlink multicarrier NOMA systems: Theory and fast algorithms," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1162–1189, Apr. 2022.

[107] S. Rezvani and E. Jorswieck, "Optimal resource allocation for NGMA," in *Next Generation Multiple Access*. Piscataway, NJ, USA: Wiley, 2024, pp. 71–100.

[108] Z. Ding, "NOMA beamforming in SDMA networks: Riding on existing beams or forming new ones?" *IEEE Commun. Lett.*, vol. 26, no. 4, pp. 868–871, Apr. 2022.

[109] M. Vaezi, G. A. A. Baduge, Y. Liu, A. Arafa, F. Fang, and Z. Ding, "Interplay between NOMA and other emerging technologies: A survey," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 900–919, Dec. 2019.

[110] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 176–183, Oct. 2017.

[111] A. Zakeri, A. Khalili, M. R. Javan, N. Mokari, and E. Jorswieck, "Robust energy-efficient resource management, SIC ordering, and beamforming design for MC MISO-NOMA enabled 6G," *IEEE Trans. Signal Process.*, vol. 69, pp. 2481–2498, 2021.

[112] H. Yahya, A. Ahmed, E. Alsusa, A. Al-Dweik, and Z. Ding, "Error rate analysis of NOMA: Principles, survey and future directions," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1682–1727, 2023.

[113] N. H. Mahmood et al., "White paper on critical and massive machine type communication towards 6G," 2020, *arXiv:2004.14146*.

[114] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G new radio," in *Proc. 16th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2019, pp. 607–612.

[115] J. Luo and A. Ephremides, "On the throughput, capacity, and stability regions of random multiple access," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2593–2607, Jun. 2006.

[116] P. Minero, M. Franceschetti, and D. N. C. Tse, "Random access: An information-theoretic perspective," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 909–930, Feb. 2012.

[117] A. Munari, "Modern random access: An age of information perspective on irregular repetition slotted Aloha," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3572–3585, Jun. 2021.

[118] M. Berioli, G. Cocco, G. Liva, and A. Munari, "Modern random access protocols," *Found. Trends Netw.*, vol. 10, no. 4, pp. 317–446, 2016, doi: 10.1561/1300000047.

[119] S. Xing, X. Xu, Y. Chen, Y. Wang, and L. Zhang, "Advanced grant-free transmission for small packets URLLC services," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2019, pp. 1–5.

[120] L. Zhang and J. Ma, *Grant-Free Multiple Access Scheme*. Cham, Switzerland: Springer, 2019, pp. 515–533, doi: 10.1007/978-3-319-92090-0_16.

[121] E. Peralta, T. Levanen, F. Frederiksen, and M. Valkama, "Two-step random access in 5G new radio: Channel structure design and performance," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Apr. 2021, pp. 1–7.

[122] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1805–1838, 3rd Quart., 2020.

[123] C.-C. Tseng, H.-C. Wang, J.-R. Chang, L.-H. Wang, and F.-C. Kuo, "Design of two-step random access procedure for URLLC applications," *Wireless Pers. Commun.*, vol. 121, no. 2, pp. 1187–1219, Nov. 2021, doi: 10.1007/s11277-021-09060-4.

[124] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 1376–1411, 2nd Quart., 2023.

[125] M. Polese et al., "Empowering the 6G cellular architecture with open RAN," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 2, pp. 245–262, Feb. 2024.

[126] A. Gharehgoli, A. Nouruzi, N. Mokari, P. Azmi, M. R. Javan, and E. A. Jorswieck, "AI-based resource allocation in end-to-end network slicing under demand and CSI uncertainties," *IEEE Trans. Netw. Service Manag.*, vol. 20, no. 3, pp. 3630–3651, Sep. 2023.

[127] M. Simsek, T. Hößler, E. Jorswieck, H. Klessig, and G. Fettweis, "Multiconnectivity in multicellular, multiuser systems: A Matching- based approach," *Proc. IEEE*, vol. 107, no. 2, pp. 394–413, Feb. 2019.

[128] E. A. Jorswieck, "Stable matchings for resource allocation in wireless networks," in *Proc. IEEE DSP*, Jul. 2011, pp. 1–8.

[129] T. Hößler, P. Schulz, E. A. Jorswieck, M. Simsek, and G. P. Fettweis, "Stable matching for wireless URLLC in multi-cellular, multi-user systems," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5228–5241, Aug. 2020.

[130] D. Csercsik and E. Jorswieck, "Preallocation-based combinatorial auction for efficient fair channel assignments in multi-connectivity networks," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 8407–8422, Nov. 2023.

[131] S. Park, S. Jeong, J. Na, O. Simeone, and S. S. Shamai, "Collaborative cloud and edge mobile computing in C-RAN systems with minimal end-to-end latency," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 7, pp. 259–274, 2021.

[132] A. Checko et al., "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.

[133] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," *IBM J. Res. Develop.*, vol. 54, no. 1, pp. 4:1–4:12, Jan. 2010.

[134] N. Moosavi, M. Sinaie, P. Azmi, P.-H. Lin, and E. Jorswieck, "Cross layer resource allocation in H-CRAN with spectrum and energy cooperation," *IEEE Trans. Mobile Comput.*, vol. 22, no. 1, pp. 145–158, Jan. 2023.

[135] A. H. Zarif, P. Azmi, N. M. Yamchi, M. R. Javana, and E. A. Jorswieck, "AoI minimization in energy harvesting and spectrum sharing enabled 6G networks," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 4, pp. 2043–2054, Dec. 2022.

[136] A. A. Ahmad, Y. Mao, A. Sezgin, and B. Clerckx, "Rate splitting multiple access in C-RAN," in *Proc. IEEE 31st Annu. Int. Symp. Pers., Indoor Mobile Radio Commun.*, Aug. 2020, pp. 1–6.

[137] A. A. Ahmad, Y. Mao, A. Sezgin, and B. Clerckx, "Rate splitting multiple access in C-RAN: A scalable and robust design," *IEEE Trans. Commun.*,

vol. 69, no. 9, pp. 5727–5743, Sep. 2021.

[138] E. Nayebi, A. Ashikhmin, T. L. Marzetta, and H. Yang, "Cell-free massive MIMO systems," in *Proc. 49th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2015, pp. 695–699.

[139] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–13, Dec. 2019.

[140] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Foundations of user-centric cell-free massive MIMO," *Found. Trends Signal Process.*, vol. 14, nos. 3–4, pp. 162–472, 2021, doi: 10.1561/2000000109.

[141] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, Jul. 2017.

[142] M. Zaher, Ö. T. Demir, E. Björnson, and M. Petrova, "Learning-based downlink power allocation in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 174–188, Jan. 2023.

[143] T. C. Mai, H. Q. Ngo, and L.-N. Tran, "Energy efficiency maximization in large-scale cell-free massive MIMO: A projected gradient approach," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6357–6371, Aug. 2022.

[144] L. Miretti, R. L. G. Cavalcante, S. Stanczak, M. Schubert, R. Böhnke, and W. Xu, "Closed-form max-min power control for some cellular and cell-free massive MIMO networks," in *Proc. IEEE 95th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2022, pp. 1–7.

[145] A. R. Flores, R. C. De Lamare, and K. V. Mishra, "Rate-splitting meets cell-free MIMO communications," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2022, pp. 657–662.

[146] A. Mishra, Y. Mao, L. Sanguinetti, and B. Clerckx, "Rate-splitting assisted massive machine-type communications in cell-free massive MIMO," *IEEE Commun. Lett.*, vol. 26, no. 6, pp. 1358–1362, Jun. 2022.

[147] J. Zheng, J. Zhang, J. Cheng, V. C. M. Leung, D. W. K. Ng, and B. Ai, "Asynchronous cell-free massive MIMO with rate-splitting," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1366–1382, May 2023.

[148] G. Interdonato, H. Q. Ngo, and E. G. Larsson, "Enhanced normalized conjugate beamforming for cell-free massive MIMO," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 2863–2877, May 2021.

[149] M. Attarifar, A. Abbasfar, and A. Lozano, "Modified conjugate beamforming for cell-free massive MIMO," *IEEE Commun. Lett.*, vol. 8, no. 2, pp. 616–619, Apr. 2019.

[150] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, "Local partial zero-forcing precoding for cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4758–4774, Jul. 2020.

[151] L. Miretti, E. Björnson, and D. Gesbert, "Team MMSE precoding with applications to cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6242–6255, Aug. 2022.

[152] Z. Li, F. Göttsch, S. Li, M. Chen, and G. Caire, "Joint fronthaul load balancing and computation resource allocation in cell-free user-centric massive MIMO networks," 2023, *arXiv:2310.14911*.

[153] E. Jorswieck, L. Kunz, R. Raghunath, and B. Peng, "Rate-splitting downlink transmission in cell-free networks under fronthaul constraints," in *Proc. IEEE ISWCS*, 2024.

[154] A. B. Carleial, "Interference channels," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 1, pp. 60–70, Jan. 1978.

[155] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2523–2527.

[156] R. C. Yavas, V. Kostina, and M. Effros, "Gaussian multiple and random access channels: Finite-blocklength analysis," *IEEE Trans. Inf. Theory*, vol. 67, no. 11, pp. 6983–7009, Nov. 2021.

[157] R. C. Yavas, V. Kostina, and M. Effros, "Random access channel coding in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 67, no. 4, pp. 2115–2140, Apr. 2021.

[158] X. Chen, T.-Y. Chen, and D. Guo, "Capacity of Gaussian many-access channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3516–3539, Jun. 2017.

[159] S. S. Kowshik, K. Andreev, A. Frolov, and Y. Polyanskiy, "Short-packet low-power coded access for massive MAC," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Nov. 2019, pp. 827–832.

[160] O. Ordentlich and Y. Polyanskiy, "Low complexity schemes for the random access Gaussian channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2528–2532.

[161] X. Chen, L. Liu, D. Guo, and G. W. Wornell, "Asynchronous massive access and neighbor discovery using OFDMA," *IEEE Trans. Inf. Theory*, vol. 69, no. 4, pp. 2364–2384, Apr. 2023.

[162] J.-S. Wu, P.-H. Lin, M. A. Mross, and E. A. Jorswieck, "Worst-case per-user error bound for asynchronous unsourced multiple access," 2024, *arXiv:2401.14265*.

[163] A. Fengler, P. Jung, and G. Caire, "SPARCs for unsourced random access," *IEEE Trans. Inf. Theory*, vol. 67, no. 10, pp. 6894–6915, Oct. 2021.

[164] A. Fengler, G. Liva, and Y. Polyanskiy, "Sparse graph codes for the 2-user unsourced MAC," in *Proc. 56th Asilomar Conf. Signals, Syst., Comput.*, Oct. 2022, pp. 682–686.

[165] E. Marshakov, G. Balitskiy, K. Andreev, and A. Frolov, "A polar code based unsourced random access for the Gaussian MAC," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2019, pp. 1–5.

[166] R. Schiavone, G. Liva, and R. Garello, "Design and performance of enhanced spread spectrum Aloha for unsourced multiple access," 2024, *arXiv:2402.12101*.

[167] B. Çakmak, E. Gkiouzepi, M. Opper, and G. Caire, "Joint message detection and channel estimation for unsourced random access in cell-free user-centric wireless networks," 2023, *arXiv:2304.12290*.

[168] P. Agostini et al., "Evolution of the 5G new radio two-step random access towards 6G unsourced MAC," 2024, *arXiv:2405.03348*.

[169] E. MolavianJazi and J. N. Laneman, "A second-order achievable rate region for Gaussian multi-access channels via a central limit theorem for functions," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6719–6733, Dec. 2015.

[170] D. Tuninetti, P. Sheldon, B. Smida, and N. Devroye, "On second order rate regions for the static scalar Gaussian broadcast channel," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 1982–1999, Jul. 2023.

[171] M. Mitev, A. Chorti, H. V. Poor, and G. P. Fettweis, "What physical layer security can do for 6G security," *IEEE Open J. Veh. Technol.*, vol. 4, pp. 375–388, 2023.

[172] V. Khammammetti and S. K. Mohammed, "OTFS-based multiple-access in high Doppler and delay spread wireless channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 528–531, Apr. 2019.

[173] G. D. Surabhi, R. M. Augustine, and A. Chockalingam, "Multiple access in the delay-Doppler domain using OTFS modulation," 2019, *arXiv:1902.03415*.

[174] U. K. Ganesan, R. Sarvendranath, and E. G. Larsson, "BeamSync: Over-the-air synchronization for distributed massive MIMO systems," 2023, *arXiv:2311.11070*.

[175] Y. Liu, X. Wang, G. Boudreau, A. B. Sediq, and H. Abou-Zeid, "A multi-dimensional intelligent multiple access technique for 5G beyond and 6G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1308–1320, Feb. 2021.

[176] Q. Wang et al., "Resource allocation based on radio intelligence controller for open RAN toward 6G," *IEEE Access*, vol. 11, pp. 97909–97919, 2023.

[177] Q. Li, "6G cloud-native system: Vision, challenges, architecture framework and enabling technologies," *IEEE Access*, vol. 10, pp. 96602–96625, 2022.

[178] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, Jan. 2020.

[179] Y. Liu, X. Wang, J. Mei, G. Boudreau, H. Abou-Zeid, and A. B. Sediq, "Situation-aware resource allocation for multi-dimensional intelligent multiple access: A proactive deep learning framework," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 116–130, Jan. 2021.

[180] M. Wu, Z. Gao, Y. Huang, Z. Xiao, D. W. K. Ng, and Z. Zhang, "Deep learning-based rate-splitting multiple access for reconfigurable intelligent surface-aided tera-hertz massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1431–1451, May 2023.

ABOUT THE AUTHORS

Eduard Axel Jorswieck (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science from Technische Universität Berlin (TU Berlin), Berlin, Germany, in 2004.

From 2006 to 2008, he was a Postdoctoral Fellow and an Assistant Professor with Signal Processing Group, KTH Stockholm. From 2008 to 2019, he was heading the Chair for Communication Theory with TU Dresden. He is currently the Managing Director of the Institute of Communications Technology, the Head of the Chair for Communications Systems, and a Full Professor with Technische Universität Braunschweig, Brunswick, Germany. He has published more than 180 journal articles, 17 book chapters, one book, four monographs, and some 300 conference papers. His main research interests include communications, applied information theory, and signal processing for networks.

Dr. Jorswieck was a co-recipient of the IEEE Signal Processing Society Best Paper Award. He and his colleagues were also a recipients of the Best Paper Awards and the Best Student Paper Awards from the IEEE CAMSAP 2011, IEEE WCSP 2012, IEEE SPAWC 2012, IEEE ICUFN 2018, PETS 2019, and ISWCS 2019, and IEEE ICC 2024. Since 2017, he has been the Editor-in-Chief of the *EURASIP Journal on Wireless Communications and Networking*. Since 2022, he has been on the editorial board of IEEE TRANSACTIONS ON COMMUNICATIONS. Since 2024, he has been an Editor for IEEE TRANSACTIONS ON INFORMATION THEORY. He was on the editorial boards of the IEEE SIGNAL PROCESSING LETTERS, IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.