Ultra-Massive MIMO for Future Cell-Free Heterogeneous Networks - **MiFuture**

Radio Resource Management(RRM) & Scheduling

Deliverable 3.1

Grant Agreement Number: 101119643
Project Acronym: MiFuture
Project Title: ultra-massive MIMO for future cell-free heterogeneous networks
Call: HORIZON-MSCA-2022-DN-01
Type of action: HORIZON TMA MSCA Doctoral Networks - Industrial Doctorates
Granting authority: European Research Executive Agency
Project start date: 01/01/2024

VERSION DETAILS	
Original Delivery Date:	31/08/2025
Actual Delivery Date:	22/09/2025
Version:	V 0.3
Number of Pages:	44
Status:	Draft
Dissemination level:	Public

CONTRIBUTIONS	
Participant	**Name**
Instituto de Telecomunicações(IT)	Prof. Rui Dinis
Instituto de Telecomunicações(IT)	Prof.Marco Gomes
Instituto de Telecomunicações(IT)	Prabhat Gupta
Universidad Carlos III de Madrid (UC3M)	Rubén de Miguel Gil
Universidad Carlos III de Madrid (UC3M)	Masoumeh Jamshidi
Tampere University (TAU)	Mehmet Fatih Ayten
Ericsson	Josep Fernandez

DOCUMENT HISTORY			
Version	**Date**	**Responsible**	**Description**
3.0	22/09/2025	Prabhat Gupta	Third Draft
1.0	30/07/2025	Rubén de Miguel	First Draft
1.0	24/07/2025	Masoumeh Jamshidi	First Draft
1.0	10/08/2025	Mehmet Fatih	First Draft
1.0	25/07/2025	Josep Fernandez	First Draft

DELIVERY REVIEW			
Version	**Date**	**Reviewed By**	**Conclusion**
1.0	08/09/2025	Supervisory Board	Modify
2.0	10/09/2025	Supervisory Board	Modify
3.0	20/09/2025	Supervisory Board	Modify

*****e.g. Accepted, Develop, Modify, Re-edit, Update

Beneficiaries and Partners

Beneficiaries:

1. UNIVERSIDAD CARLOS III DE MADRID (UC3)
2. INSTITUTO DE TELECOMUNICACOES (IT)
3. TAMPEREEN KORKEAKOULUSAATIO SR (TAU)
4. LUNDS UNIVERSITET (ULUND)
5. UNIVERSITAT AUTONOMA DE BARCELONA (UAB)
6. VODAFONE INTELLIGENT SOLUTIONS ESPANA (VOIS)
7. OY L M ERICSSON AB (ERF)
8. ERICSSON AB (ERS)
9. NOKIA SOLUTIONS AND NETWORKS SP ZOO (NOP)
10. NOKIA SOLUTIONS AND NETWORKS OY (NOF)
11. ALTYS TECHNOLOGIES (ALT)
12. KEYSIGHT TECHNOLOGIES BELGIUM (KEY)

Partners:

1. UNIVERSIDADE NOVA DE LISBOA (UNL)
2. UNIVERSIDADE DE AVEIRO (UAv)
3. UNIVERSIDADE DE COIMBRA (UCo)
4. KATHOLIEKE UNIVERSITEIT LEUVEN (KUL)
5. KEYSIGHT TECHNOLOGIES UK LIMITED (KEY-UK)

Executive Summary

Efficiently utilising limited spectrum in an optimised and fair manner has always been a central challenge in wireless communications. Radio Resource Management (RRM) and scheduling are pivotal components of the network, responsible for allocating resources and meeting the diverse and stringent requirements of emerging services.

As we transition toward 6G and cell-free architectures, traditional RRM approaches are evolving to incorporate intelligent decision-making through artificial intelligence (AI). These advancements enable more accurate modelling of user behaviour and traffic patterns, leading to improved spectral efficiency and more adaptive resource allocation. However, they also introduce new challenges, particularly in power efficiency and Access Point (AP) selection, that must be addressed through holistic network design.

Power and spectral efficiency considerations now extend beyond the network core to include User Equipment (UE). Intelligent RRM and scheduling strategies at the UE level, including continuous radio link monitoring and adaptive AP connectivity, are essential to meet the timing and service constraints imposed by the network. In ultra-massive MIMO systems, non-coherent transmission schemes are being explored to further enhance throughput, reliability, and coordination between the physical and MAC layers.

Accurate user positioning, enabled by ultra-massive MIMO advanced sensing capabilities, plays a critical role in next-generation RRM. With real-time location awareness, networks can allocate resources more precisely, reduce latency, and ensure Quality of Service (QoS) for individual users. Native AI techniques support predictive mobility management and dynamic beamforming, making real-time AP selection and per-user power optimisation integral to the 6G RRM landscape.

State-of-the-art solutions already incorporate AI-driven functionality in the RAN Intelligent Controller (RIC), which hosts applications for network management, optimisation, and routing. Emerging research also focuses on leveraging positioning data for AP selection and integrating non-coherent waveforms in MIMO systems for enhanced flexibility and robustness.

These advances lay the foundation for 6G networks to shift from static, centralised management toward distributed, dynamic, and context-aware RRM and scheduling frameworks. This transition will support the vision of ubiquitous, high-performance, and energy-efficient communication in cell-free 6G networks.

Contents

List of Figures

List of Abbreviations

5G	Fifth Generation
5G NR	5G New Radio
AM	Amplitude Modulation
AP	Access Point
AR	Augmented Reality
BSR	Buffer Status Report
CP	Cyclic Prefix
CPU	Central Processing Unit
CSI	Channel State Information
DL	Downlink
DSA	Dynamic Spectrum Access
eMBB	enhanced Mobile BroadBand
EPC	Evolved Packet Core
FL	Federated Learning
gNB	gNodeB
GNN	Graph Neural Network
GSM	Global System for Mobile Communications
ID	identification
IoT	Internet of Things
IT	Instituto de Telecomunicações
JCAS	Joint Communication and Sensing
LCG	Logical Channel Group
LLM	Large Language Model
LTE	Long-Term Evolution
MAC	Medium Access Control

MIMO	Multiple-Input and Multiple-Output
mmW	millimeter Wave
MM	Mobility Management
ms	millisecond
MRT	Maximum Ratio Transmission
nea-RT RIC	Near-Real-Time RIC
non-RT RIC	Non-Real-Time RIC
NWDAF	Network Data Analytics Function
O-RAN	Open Radio Access Network
PBCH	Physical Broadcast Channel
PBR	Prioritized Bit Rate
PMI	Precoding Matrix Indicator
PSS	Primary Synchronization Signal
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RE	Resource Element
RI	Rank Indicator
RIC	RAN Intelligent Controller
RL	Reinforcement Learning
RLM	Radio Link Monitoring
RRM	Radio Resource Management
RS	Reference Signal
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RSSI	Received Signal Strength Indicator
SBA	Service-Based Architecture
SE	Spectral Efficiency
SINR	Signal-to-Interference & Noise Ratio
SLNR	Signal-to-Leakage & Noise Ratio
SNR	Signal-to-Noise Ratio

SR	Scheduling Request
SRS	Sounding Reference Signal
SS	Synchronization Signal
SSB	Synchronization Signal Block
SSS	Secondary Synchronization Signal
TA	Timing Advance
Te	Timing error
THz	Terahertz
UE	User Equipment
UL	Uplink
UmMIMO	ultra-massive MIMO
UMTS	Universal Mobile Telecommunications Service
URLLC	Ultra-Reliable Low-Latency Communication
VR	Virtual Reality
ZF	Zero-Forcing

Chapter 1

Introduction

Radio Resource Management (RRM) refers to the set of algorithms and protocols that manage the co-channel interference, how radio resources (like frequency, time, power and space) are allocated and used, User Equipment (UE) configurations and state control, and other radio transmission characteristics in wireless communication systems. Its main objective is to utilize the limited RF spectrum resources given and the radio network infrastructure as efficiently as possible. [3]

Scheduling refers to allocating these resources at every time interval in a way that each user receives the desired Quality of Service (QoS) and therefore the Quality of Experience (QoE) while utilizing the available resources, in the best possible manner and serving a maximum number of users at the same time. This introduces several optimisation problems linked to RRM. [4]

RRM is designed to ensure a smooth mobility experience for users, where both the UE and the network handle mobility management automatically, with no need for user involvement. To provide support for seamless mobility, the UE plays a central role in this process by performing tasks such as cell search and cell selection during idle mode, as well as signal measurements and reporting to the network for mobility decisions during connected mode. These operations are designed to be efficient, minimizing power consumption and complexity while maintaining high performance. [5]

To meet the growing demand for high-definition video, AR/VR, and cloud gaming, and to enable real-time applications such as autonomous driving and remote surgery, much higher bandwidths are required to achieve faster data rates and lower latencies. In 5G NR, a latency of 1 ms under ideal conditions and peak download data rates of 10 Gbps or more can be achieved. However, the stricter latency requirements, combined with the availability of Resource Blocks every 1 ms, have increased the challenges of scheduling and RRM. To address these challenges and leverage the diversity offered by MIMO, cell-free systems were introduced.

With the introduction of cell-free systems, a major challenge of RRM and scheduling has been addressed. In traditional cell-based systems, a large number of Access Points (APs) are required to support the increasing data rates and number of users, and the Signal-to-Noise Ratio (SNR) depends on the distance between the user and the serving AP. As a result, users located near the AP experience higher SNR, while those at the cell edge suffer from lower SNR. In contrast, cell-free systems eliminate these cell boundaries and adopt a user-centric approach. This approach identifies the most suitable APs for each user, allowing multiple APs to serve the user simultaneously in a coordinated manner, thereby ensuring the requested SNR and Quality of Service (QoS). The user-centric strategy can be implemented in two ways: a centralized approach, where each AP relays information to a Central Unit responsible for decision-making and resource allocation; or a decentralized approach, where each AP independently handles data encoding/decoding and performs its own resource management.

However, the cell-free architecture also introduces challenges, such as determining the most suitable APs to connect to in order to maximize spectral efficiency while simultaneously reducing signaling overhead and power consumption [6]. To address these issues, RRM solves various optimization problems, including AP selection and power allocation for users sharing the same time–frequency resources. This, however, leads to the challenge of channel estimation, which requires substantial overhead and can negatively affect the optimized performance [7] Another challenge is ensuring appropriate levels of cooperation among APs to support efficient resource allocation algorithms. These optimization algorithms may be designed with either a single objective or multiple objectives, typically focusing on maximizing the spectral efficiency of both individual users and the overall system [8]. Moreover, distributed large-scale antenna structures can be leveraged for sensing, enabling user positioning and tracking. This capability allows the user's location to be determined at each time instant, thereby enhancing resource management and AP selection [9].

Since cell-free networks are used to increase the throughput of the system, RRM with multi-user non-coherent systems becomes crucial. Non-coherent wireless communication systems can transmit information through the wireless medium without any channel estimation procedure. To achieve this, special modulation schemes are used [10]. MIMO systems, thanks to their channel hardening property brought by the high number of physical channels, bring a prolific environment for non-coherent modulation schemes.

The main problem that non-coherent schemes aim to address is the performance degradation in the uplink of coherent systems under high-mobility conditions, where rapid channel variations increase the frequency of channel estimation procedures and thereby waste time–frequency resources. Currently, three main types of non-coherent modulation schemes are used in MIMO systems: energy-based schemes, Grassmannian schemes, and differential schemes [2].

As previously mentioned, with a large number of antennas it is possible to achieve radar-like capabilities that allow tracking of the user and estimating the user's position within the network. Leveraging this information for scheduling and resource management can play a major role. The main advantages of knowing the user's position are:

- Knowing the best channels between the APs and the user.
- The APs can be preselected if the user's movement is predicted, reducing the latency.
- High interference zone can be known, and the best APs can be chosen accordingly to satisfy the user requirements.
- More precise and coordinated beamforming can be done towards the user
- Supporting faster handovers and beam reconfiguration.
- Improving fairness among users, as both edge users and central users are treated equally.
- Enhancing scalability by reducing signaling requirements.

Along with the user's location, the introduction to Artificial Intelligence (AI) in the network is highly influential in managing the resources. AI is now being embedded natively within network procedures and protocols. AI-native systems incorporate AI mechanisms intrinsically into their operations, enabling autonomous, adaptive, and proactive decision-making across multiple layers of the network. Open Radio Access Network (Open RAN) architecture, characterized by its openness, modularity, and standardization efforts led by the O-RAN Alliance, offers an ideal framework for deploying AI-native procedures. Open RAN disaggregates traditional RAN components and exposes open interfaces, thus enabling multi-vendor ecosystems and flexible deployment of AI functionalities. Moreover, the alignment with 3GPP standards, particularly Release 18, which introduces AI and data analytics frameworks such as Network Data Analytics Function (NWDAF), paves the way for scalable and interoperable AI-native protocol adoption. Embedding AI at the protocol and procedure level facilitates real-time resource optimization, predictive maintenance, enhanced security, and intelligent network slicing, all essential to fulfil the ambitious goals of current and future wireless systems.

To analyze RRM from different angles and perspectives, we will examine various strategies and state-of-the-art approaches for improved scheduling and management at the UE level, including the use of non-coherent systems, AI-driven methods, along with how position plays a major role in these scenarios.

Chapter 2

State of the Art

This chapter presents the state of the art in beyond-5G communication systems, with a focus on power and AP selection optimization in cell-free networks. It also examines how current systems address challenges through the use of non-coherent waveforms, the integration of AI, and user equipment (UE)-level strategies for RRM and scheduling.

2.1 Power Allocation in Cell-Free Networks

A key resource in cell-free networks is the transmit power allocated by each AP antenna to each User Equipment (UE) antenna in the downlink, or vice versa in the uplink. Power control algorithms regulate this allocation, optimizing system gains and enhancing overall network performance [11]. Power control schemes are generally governed by the Central Processing Unit (CPU). Conditions such as channel hardening and favorable propagation help to limit the transmission coefficients up to just large scale fading characteristics [9, 12]. In this way, distributed power control policies can handle the near–far effect between UEs and APs while also managing strong interference, both from other UEs and from estimation errors in the serving UE's channel.

Through these power control policies, specific objectives, such as maximizing spectral efficiency (SE) subject to power constraints, can be achieved, thereby optimizing network performance as the number of UEs and APs scales up. A common assumption is that the number of antennas exceeds the number of UEs [13]. The simplest strategy is to assume that all UEs transmit at maximum power, as is often considered in the design of several detection techniques [14, 15]. However, this approach is suboptimal as it leads to wasted power and excessive interference. Consequently, significant research has focused on developing efficient power control algorithms that balance the trade-offs between performance and computational complexity [16, 17].

Power allocation algorithms solve optimization problems subject to constraints such as the AP's link budget and the maximum transmit power supported by each antenna in both uplink and downlink. These constraints are often modeled using fairness criteria, which guarantee that each UE is served by at least one AP, with the decision variables being the power control coefficients. Optimization schemes include Max–Min Fairness (MMF), maximum sum rate, and total radiated power, among others. Such schemes may be combined to effectively solve the allocation problem [18]. As the network scales, complexity grows, and enforcing fairness becomes increasingly challenging [19]. To address this, multi-objective programming with meta-heuristic approaches can be employed, offering near-optimal solutions with reduced computational complexity [20, 21]. The key idea is to incorporate all relevant constraints into a multi-objective optimization framework to balance trade-offs while ensuring the system meets key performance indicators (KPIs).

User Power Optimization

Full-power transmission often results in excessive interference and inefficient energy use [9]. The objective is therefore to maximize metrics such as max–min fairness [8] or the network's maximum sum rate, while minimizing interference. Since higher power inevitably increases interference, careful selection of transmit power is critical for achieving the desired outcome, particularly under the assumption that all UEs share the same spectrum.

In the uplink, data reception occurs in two stages [15]. First, each AP estimates the channel within a coherent time–frequency block and applies a receiver-combining technique to generate local estimates of the UE's uplink data. In the second stage, these local estimates are forwarded to the CPU, which aggregates them through averaging into a final estimate. Importantly, the CPU does not require knowledge of the channel estimation, as the AP handles that locally. This approach significantly reduces signaling load on fronthaul and backhaul links between the APs and the CPU [22].

The local estimate for the k^{th} UE at AP l using a local combining vector $v_{l_k} \in \mathbb{C}N(0,1)$ is given by [23]:

$$\hat{s}_{l_k} = v_{l_k}^H Y_l = v_{l_k}^H H_{k_l} s_l + \sum_{i=1, i\neq k}^{i=K} v_{l_k}^H H_{l_i} s_{l_i} + v_{l_k}^H W_l \ . \tag{2.1}$$

Here, H_{kl} denotes the channel matrix, s_l is the signal from the desired UE, s_{l_i} are the the interfering signals from the remaining UEs and W_l is the the additive noise.

The most common design for the local combining vector is Maximum Ratio Combining (MRC), where $v_{l_k} = \hat{H}_{kl}$. MRC leverages favorable propagation and channel hardening to reduce inter-user interference while maximizing the desired signal [14]. After local combining, each AP forwards its estimate to the CPU, which performs a weighted averaging to maximize SE.

The achievable SE of a UE k can be expressed as [12]:

$$\text{SE}_k = \left(1 - \frac{\tau_p}{\tau_c}\right) log_2(1 + \text{SINR}_k) \tag{2.2}$$

where the Signal-to-Interference-plus-Noise Ratio (SINR) is:

$$\text{SINR}_k = \frac{\rho_k \mid \sum_{l=1}^{L} \mathbb{E}[v_{l_k}^H H_{k_l}] \mid^2}{\sum_{i=1,i\neq k}^{i=K} \rho_i \mathbb{E}[\mid \sum_{l=1}^{L} v_{l_k}^H H_{i_l} \mid^2] + \sigma^2 \sum_{l=1}^{L} \mathbb{E}[v_{l_k}^2]} \tag{2.3}$$

Considering a cell-free system, the uplink power optimization problem can be formulated as follows: the decision variables are the power coefficients allocated to each UE k, denoted as p_k. The constraint is that the power must not exceed the maximum uplink transmission power P_{max}. Power optimization is typically based on the max–min fairness scheme, which seeks to ensure balanced performance across all users [8]. This is achieved by optimizing p_k to maximize the spectral efficiency (SE) of the weakest UE, i.e., the one with the poorest channel conditions. In this way, the optimization avoids bias toward stronger UEs and guarantees fairness. The max–min fairness problem can be expressed as:

$$\begin{aligned} &\max_{\rho_k} \quad \min \ \text{SE}_k, \quad k = 1, \ldots, K \\ &\text{s.t.} \quad 0 \leq \rho_k \leq \rho_{\max}, \forall k. \end{aligned} \tag{2.4}$$

A similar formulation applies to downlink optimization. In this case, either the same SINR can be assumed considering it as a virtual uplink problem or the Signal-to-Leakage & Noise Ratio (SLNR) can be considered. In downlink transmission, power is allocated to beamforming vectors, which must be optimized and directed toward the intended UE to separate users and mitigate interference. This beamforming optimization problem introduces the concepts of eigenspaces and eigenvectors [24].

Different precoding strategies can be employed to balance the trade-off between high signal power and interference. The optimal solution generally lies between Zero-Forcing (ZF) and Maximum Ratio Transmission (MRT), with the latter being the downlink counterpart of Maximum Ratio Combining (MRC) in the uplink [24]. ZF represents one extreme, where beamforming vectors are designed to be orthogonal, thereby eliminating inter-user interference. MRT represents the opposite extreme, assuming no inter-user interference and only Gaussian noise in the system, and aims to maximize the desired signal power. In practice, optimization heuristics are often used to balance these two extremes, with ZF typically performing better in interference-limited scenarios [25].

The downlink beamforming problem can be visualized as projecting each user's signal into a large subspace, where each beam is aligned with the dominant eigenvector of the corresponding APs. Figure 2.1 illustrates this subspace for multiple users with respect to the beamforming direction of the intended user [26].

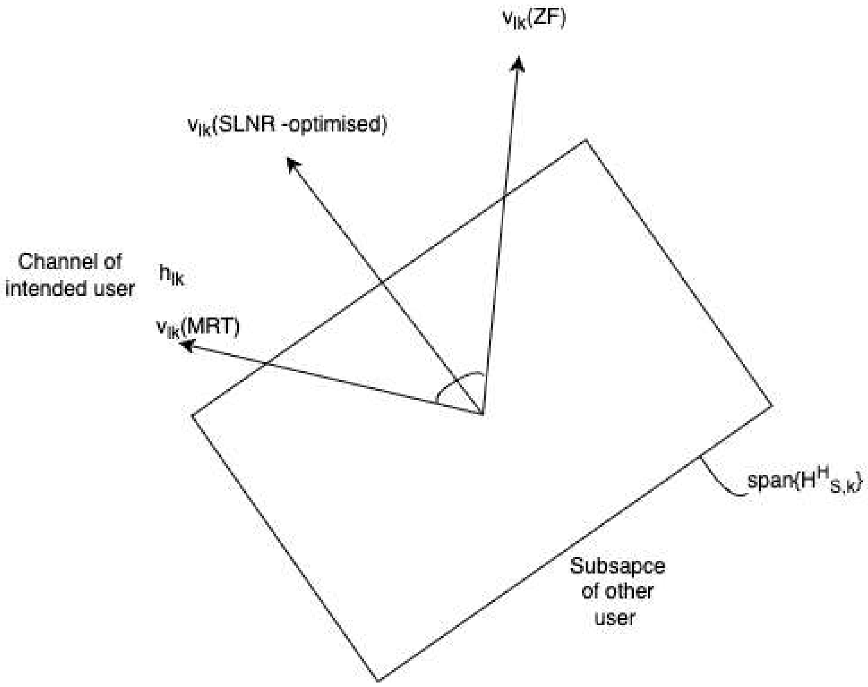

Figure 2.1: Eigen Subspace

The eigenvectors should be designed to maximize power for the intended user while maintaining orthogonality to other users. One way to quantify orthogonality between beamforming vectors is through the chordal distance, defined as [9]:

$$d_C(X, Y) = ||XX^H - YY^H||_F^2 \tag{2.5}$$

where X and Y are two different eigenvectors, and F denotes the Frobenius norm.

To optimize resources effectively and ensure efficient power allocation, it is essential to jointly consider AP–UE associations in a cell-free network, where the number of AP antennas typically exceeds the number of UEs. Selecting the most suitable APs for each UE is critical to achieving maximum system performance.

2.1.1 AP Selection

To evaluate performance, it is commonly assumed that the fronthaul has limited capacity, whereas the backhaul link has unlimited capacity and can handle all data transferred from the APs. However, to alleviate the burden on the CPU and reduce transmission and detection complexity, effective AP–UE associations are essential. Such approaches reduce fronthaul complexity and signaling requirements, ultimately lowering the processing complexity in the backhaul as well. In this context, optimized AP selection algorithms play a critical role [18].

Depending on network characteristics and UE distribution, the contribution of certain APs can result in high inter-user interference, thereby decreasing the SINR at the CPU [6]. In [19], a joint power allocation and AP selection scheme is proposed that minimizes energy consumption while preserving quality of service (QoS). Similarly, [6] presents an innovative AP selection and signal detection method that enhances UE throughput and reduces fronthaul load by considering both channel quality and effective channel gain to sequentially allocate APs to users. In [27], the authors demonstrate that AP selection schemes can significantly improve downlink spectral efficiency while balancing network load and reducing power consumption [22].

Thus, AP selection is crucial for minimizing fronthaul and backhaul overhead, reducing processing complexity, and ensuring load balancing, particularly in mobile scenarios. In [28], a cluster-centric approach for AP selection and power control is proposed. Meanwhile, [29] investigates the impact of different UE velocities on downlink SE, emphasizing the importance of pilot training up to a certain velocity and highlighting the effects of channel aging due to mobility. To mitigate channel aging, precise user tracking is required.

Most AP selection schemes aim to minimize the number of UEs connected to each AP in order to reduce signaling overhead and maximize spectral efficiency. This ensures buffer balance across APs, enabling each one to provide adequate QoS to its connected UEs [18]. To conserve power and improve energy efficiency, each UE is typically served by a subset of APs. This group can be selected based on criteria such as received power or large-scale fading characteristics [18].

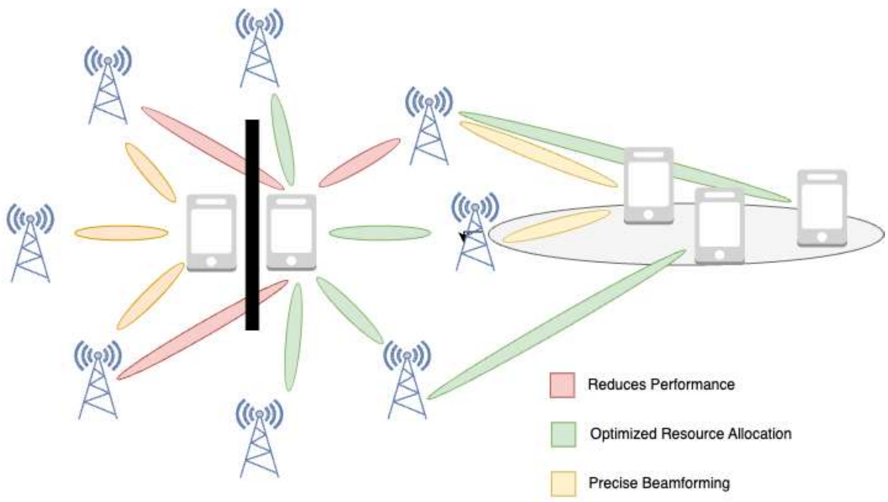

Figure 2.2: Positioning and Beamforming

As shown in [29], user positioning plays a critical role in AP selection. By leveraging location information, more accurate channel estimates between UEs and APs can be obtained, enabling the activation of only those APs with favorable channel conditions. This improves efficiency by reducing interference, facilitating faster handovers, and lowering signaling requirements. Furthermore, precise positioning enables more accurate beamforming, which increases SINR, enhances spectral efficiency, and excludes irrelevant APs. As a result, energy efficiency and throughput improve with optimized power allocation. Figure 2.2 illustrates how positioning, accurate beamforming, and optimized AP selection are essential for interference reduction and overall network performance. For instance, AP–UE connections represented by the red beams would increase system complexity, as these associations may correspond to APs with poor channel conditions, overloaded APs, or APs directing disproportionate power to users already experiencing favorable conditions, thereby reducing efficiency in a precise beamforming environment.

2.2 RRM activities at the UE side

In 5G NR, the UE plays an essential supporting role in both UL and DL scheduling while the base station, denoted as gNodeB (gNB), is ultimately responsible for the scheduling decisions. To enable efficient resource allocation and meet QoS requirements, the UE must periodically provide accurate feedback to the gNB.

In the UL, the gNB schedules the transmissions of the UE. To facilitate this, the UE sends Buffer Status Reports (BSRs), which inform the network of the amount of data awaiting transmission. When the UE has data to transmit but no resources have been allocated, it sends a Scheduling Request (SR) to the gNB so that it is aware that the UE has data to be transmitted.

The UE also assists the gNB in measuring uplink channel quality through the transmission of Sounding Reference Signals (SRSs). This information is used by the network to optimize uplink scheduling.

Additionally, the UE is responsible for logical channel prioritisation. When it has the resources to transit, it has to decide how to distribute them among its active logical channels. This decision is based on parameters such as Prioritized Bit Rates (PBRs) and Logical Channel Group (LCG) priorities. [30]

Measurements

A key responsibility of the UE in RRM is performing measurement procedures. These include assessing signal quality both within the same frequency band (intra-frequency) and across different frequency bands (inter-frequency). Examples of principal measurements performed by the UE include:

- Synchronization Signal - Signal-to-Interface-plus-Noise Ratio (SS-SINR)

- Synchronization Signal - Reference Signal Received Power (SS-RSRP)
- Synchronization Signal - Reference Signal Received Quality (SS-RSRQ)

These measurements are conducted either periodically or in response to specific events, such as when a neighboring cell offers better service than the current serving cell.

Apart from the measurements already mentioned, the UE oversees beam-specific measurements in scenarios involving beamforming. It must report Layer 1 RSRP values for individual beams, enabling the network to make informed decisions on beam management and mobility.

To facilitate accurate and timely measurements, the UE may be configured with measurement gaps. These gaps allow the UE to temporarily suspend data transmission or reception to scan other frequencies or Radio Access Technologies (RATs) without disrupting ongoing communication.

2.2.1 Mobility Management

Mobility Management (MM) consists of the UE's behaviour in both the idle and connected states. In idle mode, the UE performs cell reselection autonomously based on signal quality metrics and different priorities defined by the network. This process optimizes the battery usage of the UE and ensures that it remains on the most suitable cell all the time.

Meanwhile, in connected mode, the UE supports the network who controls the mobility through measurement reporting. It evaluates the quality of the serving ang neighbouring cells and report these findings to the network. Based on this information, the network may initiate the handover towards a neighbour cell with better quality services. The handover is performed by the UE which involve synchronizing with the new cell, performing random access procedures and resuming data transmission with minimal delay [31].

The UE has also the responsibility to detect and evaluate cells from other RATs like LTE or UMTS and support reselection if it is the case. This capability ensures seamless service continuity across heterogeneous network environments.

In high-speed mobility scenarios, the UE must support fast and reliable transitions between cells and RATs, including accurate timing, performing rapid measurements, and executing handovers with minimal interruption to ongoing services [32].

Radio Link Monitoring

Radio Link Monitoring (RLM) is a mechanism by which the UE assesses the quality of the downlink connection. The UE monitors reference signals, which may be based on Synchronization Signal Blocks (SSBs) or Channel State Information Reference Signals (CSI-RS), to determine whether the link quality meets predefined thresholds. These thresholds, known as Q_{out} and Q_{in}, represent the minimum acceptable and re-entry quality levels, respectively.

When the link quality is below Q_{out} for a specific duration, the UE must report an out-of-sync condition, indicating potential link failure. Otherwise, if the link quality is above Q_{in}, the UE must report an in-sync condition, signalling recovery.

The layers involved in Radio Link Monitoring and how the flow of the messages can be seen in figure 2.3.

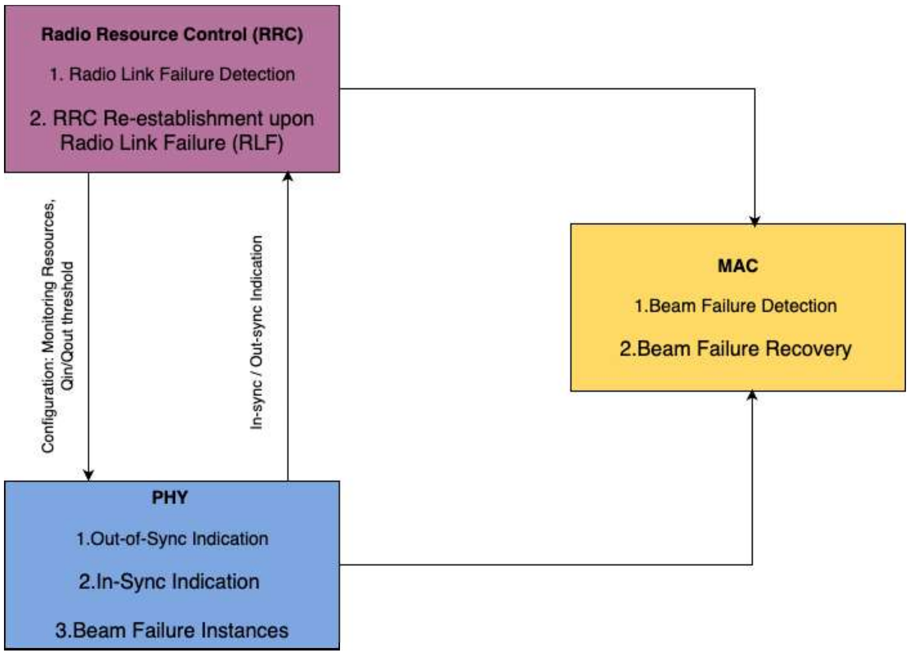

Figure 2.3: Radio Link Monitoring

Conformance Requirements

The UE must meet a comprehensive set of conformance requirements to ensure performance across devices and network deployments. These requirements are validated through standardized test scenarios that simulate various operational conditions, including different deployment architectures, frequency ranges and mobility scenarios.

The tests assess the UE's ability to perform measurements, report accurately, maintain timing alignment, and respond appropriately to variations in link quality. Some requirements are evaluated using statistical methods, with confidence levels ensuring the reliability of test outcomes.

Key performance metrics include measurement accuracy, reporting delay, handover execution time, and link recovery duration [33].

Timing Accuracy

In UL transmissions, keeping precise timing is essential for the UE. The UE must align its transmission timing with the DL transmissions timing of the serving cell using Timing Advance (TA) commands received from the network. The initial transmission timing error must remain within strict limits, typically defined in terms of a timing error threshold (Te) [34].

The UE has to adjust its timing in response to changes in the DL timing. These adjustments must adhere to defined step sizes and rates, ensuring smooth transitions without causing interference. The UE's ability to perform these adjustments is verified through the transmission of SRSs, which serve as a reference for measuring timing accuracy [35, 36].

2.3 Non-Coherent Schemes for MIMO systems

2.3.1 Grassmannian schemes

Grassmannian schemes are closely related to Unitary Space–Time Modulation (USTM) [37], which is a coherent scheme that proposes composing constellations of symbols which span several transmissions and spatial channels (antennas). USTM is optimal if the coherence time T is bigger than the number of transmitting antennas M [38].

As long as the narrowband assumption is holds [4], the effect of the channel $\mathbf{H}$ on an UTSM symbol can be modeled as a linear transformation. If the coherence time is longer than the number of transmissions spanned by the USTM symbol, the linear transformation exerted by the channel will be identical for all the rows (uses of the channel) of the USTM symbol. Consequently, the subspace spanned by the columns of the received USTM symbol is the same as that spanned by the columns of the transmitted USTM symbol. Grassmannian modulation schemes exploit this property by embedding information in the subspaces spanned by the USTM symbol columns, which remain invariant to the channel $\mathbf{H}$ [39].

These schemes are typically designed for MIMO systems with relatively few antennas. However, as the number of antennas increases, constellation design becomes computationally prohibitive. This scalability issue makes Grassmannian schemes a suboptimal choice for ultra-massive MIMO systems.

2.3.2 Energy-based schemes

Energy-based schemes for ultra-massive MIMO rely on the same principles as Amplitude Modulation (AM) systems, i.e., they use only the amplitude of the transmitted signal to transmit information. The way to use AM in multiple antenna systems consists of averaging the received signal energy in each of the antennas of the system according to [40]

$$\frac{1}{N}\|\mathbf{y}\|^2 = \frac{1}{N}\sum_{n=1}^{N} |y_n|^2, \tag{2.6}$$

where $\mathbf{y}$ is the received signal and N is the number of receive antennas. This averaging process exploits the channel hardening property of multi-antenna systems, whereby the effective channel becomes increasingly deterministic as N grows. This helps mitigate the problems AM systems have when dealing with fast fading channels.

While energy-based schemes exhibit attractive capacity scaling with increasing numbers of antennas, they face challenges in multiplexing large numbers of users. This limitation makes them less suitable for ultra-massive MIMO deployments, where the ability to simultaneously serve very high numbers of users is a fundamental requirement.

2.3.3 Differential schemes

Differential schemes can be seen as an evolution of energy-based schemes, as they also exploit the phase of the received signal, thereby achieving higher spectral efficiency. These schemes rely on the assumption that the channel coherence time is greater than twice the symbol duration, ensuring that the channel can be considered constant over two consecutive symbols. This condition is typically satisfied in most wireless communication systems. Differential schemes exploit it by encoding information in the difference between consecutive symbols. In this approach, the transmitted symbols are precoded with the previously transmitted symbol:

$$x[n] = x[n-1]\, s[n], \quad n > 1, \tag{2.7}$$

where $x[n]$ is the transmitted symbol at time n, and $s[n]$ is the information symbol being differentially encoded. The first transmitted symbol is predefined and serves as the starting point of the differentially encoded chain. Decoding is performed by multiplying the received symbol by the complex conjugate of the previously received symbol:

$$z[n] = y[n-1]^*\, y[n], \tag{2.8}$$

where $z[n]$ is the decoded symbol and $y[n]$ is the received symbol at time n. It is important to note that there is no error propagation thanks to (2.8) being only dependent on two consecutive received symbols (rather than previously decoded symbols). Differential modulation and demodulation can also be done in the frequency domain by applying the encoding across contiguous OFDM subcarriers [2].

Several adaptations of differential schemes to MIMO systems have been proposed, with the first being Differential Unitary Space–Time Modulation (DUSTM) [37, 41]. This modulation scheme, like Grassmannian signaling, uses USTM symbols. The difference is that here, instead of coding information into subspaces, the information symbols are transmitted through differential USTM symbols according to

$$\mathbf{X}_\tau = \mathbf{S}_\tau \mathbf{X}_{\tau-1}, \tag{2.9}$$

where $\mathbf{X}_\tau \in \mathbb{C}^{M\times M}$ is the τ^{th} transmitted DUSTM symbol, with M denoting the number of transmitting antennas, and $\mathbf{S}_\tau \in \mathbb{C}^{M\times M}$ is the τ^{th} USTM information symbol. Equation (2.9) is no more than the USTM expansion of (2.7), and thus requires the coherence time to be at least twice the symbol duration. In the case of USTM, this spans $2M$ transmissions. As a result, the higher the number of antennas (and therefore transmissions) used by the USTM symbol, the higher the coherence time of the channel must be in order to maximize the throughput. Because of this, in DUSTM, the condition $T \geq 2M$ must be met [42]. This limitation, together with the user multiplexing problems of USTM, makes this scheme a suboptimal choice for ultra-massive MIMO.

Problems with multiplexing multiple users and high computational complexity also affect Differential Space–Time Block Codes (DSTBC), the differential counterpart of the Alamouti scheme [43], making them unsuitable for ultra-massive MIMO systems.

Another adaptation of (2.8) to MIMO systems consists on averaging across the receiving antennas in order to exploit channel hardening [44]

$$z[n] = \frac{1}{N}\sum_{i=1}^{N} y_i[n-1]^* y_i[n]. \tag{2.10}$$

The effect of channel hardening provides significant noise and interference reduction, which scales with the number of receiving antennas N. Moreover, the simplicity of (2.10) enables efficient operation with large antenna arrays without requiring excessive computational resources.

However, a major drawback of all differential schemes is that they require constellations whose points form a closed group under matrix multiplication; in other words, the product of two constellation points must also belong to the constellation. Otherwise, (2.7) would generate symbols outside the intended set. This restriction forces differential schemes to rely on M-PSK constellations, which considerably limits their performance at high SNR compared to coherent schemes that can employ high-spectral-efficiency QAM constellations.

Among all non-coherent schemes, differential non-coherent modulation based on averaging is the best non-coherent modulation scheme for ultra-massive MIMO. This is because of its low computational complexity, together with its attractive performance scaling with increasing numbers of antennas.

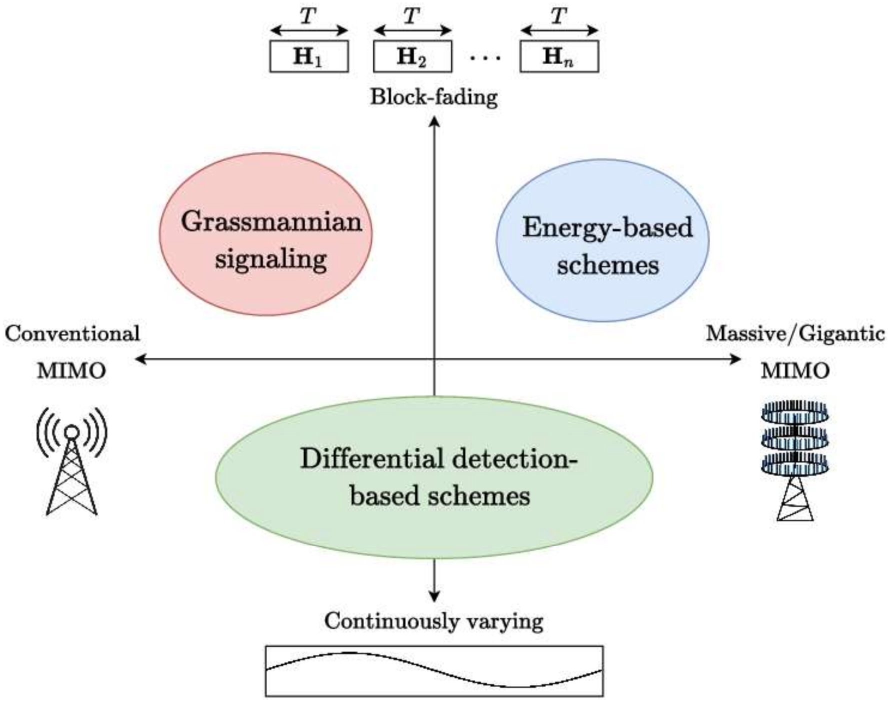

Figure 2.4: Summary of Non-Coherent Schemes [1]

Figure 2.4 summarizes the main non-coherent schemes. Differential approaches stand out for their simplicity and scalability in scenarios with large antenna arrays and continuous waveforms, provided that the channel coherence time is at least twice the symbol duration.

2.4 Use of AI in RRM & Scheduling

Recent advancements in 5G networks have underscored the critical need for flexible and open architectures that can seamlessly integrate AI-native procedures and protocols. Traditional network architectures often impose constraints on scalability, interoperability, and the rapid deployment of intelligent solutions. In this context, Open Radio Access Network (O-RAN) emerges as a transformative paradigm that fosters modularity, openness, and vendor diversity, creating an ideal platform for the adoption and acceleration of AI-driven innovations. This section explores the role of O-RAN in enabling the integration of AI-native protocols to enhance network performance, adaptability, and automation [45].

The Medium Access Control (MAC) layer in 5G networks plays a critical role in resource scheduling and allocation. AI-native MAC protocols utilize advanced Reinforcement Learning (RL) algorithms to adaptively schedule radio resources, taking into account dynamic traffic patterns, user mobility, and interference conditions. Such adaptive scheduling improves throughput, latency, and fairness in multi-user scenarios.

Furthermore, AI-empowered Dynamic Spectrum Access (DSA) protocols enable real-time detection and exploitation of available spectrum holes, boosting spectral efficiency, especially in dense urban deployments.

AI-native Routing and Control Protocols

Efficient routing in heterogeneous 5G networks is challenged by diverse service requirements and fluctuating network conditions. Graph Neural Networks (GNNs) have been proposed to model network topologies and predict link quality, enabling intelligent, adaptive routing that dynamically balances loads and minimizes congestion [46].

RL-based distributed control agents continuously update routing policies to optimize end-to-end network performance.

AI-native Network Management Protocols

AI-native network management incorporates AI for tasks such as fault detection, QoS and QoE prediction, and configuration management. Federated Learning (FL) frameworks enable distributed AI model training across multiple RAN nodes without centralizing sensitive user data, thus preserving privacy while improving model accuracy.

Such AI-native management techniques lead to proactive maintenance, dynamic resource allocation, and automated network optimization [47].

AI-native Security Protocols

AI-native security protocols embed machine learning algorithms to detect anomalies, intrusions, and attacks in real-time. The systems employ AI to dynamically adapt security policies and countermeasures. However, the vulnerability of AI models to adversarial attacks requires robust defense mechanisms to ensure network reliability and trustworthiness [48].

Disaggregated Architecture and Open Interfaces

O-RAN architecture decomposes the traditional RAN into Central Unit (CU), Distributed Unit (DU), and Radio Unit (RU), interconnected through standardized open interfaces such as F1, E2, and A1, as defined by the O-RAN Alliance [49].

This disaggregation allows multi-vendor deployments and facilitates integration of AI-native components at various functional points in the network.

RAN Intelligent Controller (RIC)

The RIC is a key enabler of AI-native functionality within O-RAN, divided into two parts:

- Near-Real-Time RIC (near-RT RIC): Operates at latency scales between 10 ms and 1 s, hosting xApps that execute AI-native control functions such as adaptive scheduling, beamforming, interference management, and mobility optimization. These xApps use RL and other AI methods to optimize radio resources dynamically.
- Non-Real-Time RIC (non-RT RIC): Works at timescales beyond 1s and hosts rApps responsible for AI/ML model training, network analytics, policy management, and lifecycle management of AI applications [50]. FL techniques enable distributed, privacy-preserving model training across network nodes.

AI-native xApps and rApps

xApps and rApps represent the programmable AI-native applications that realize the intelligence of the RIC. While xApps focus on near real-time control, reacting to immediate network events, rApps provide strategic, long-term optimization, monitoring AI model performance, and updating policies accordingly.

A generic framework for RAN intelligence can be seen in figure 2.5. The data is collected in the cloud, where it is trained and the model is sent back for real-time implementations.

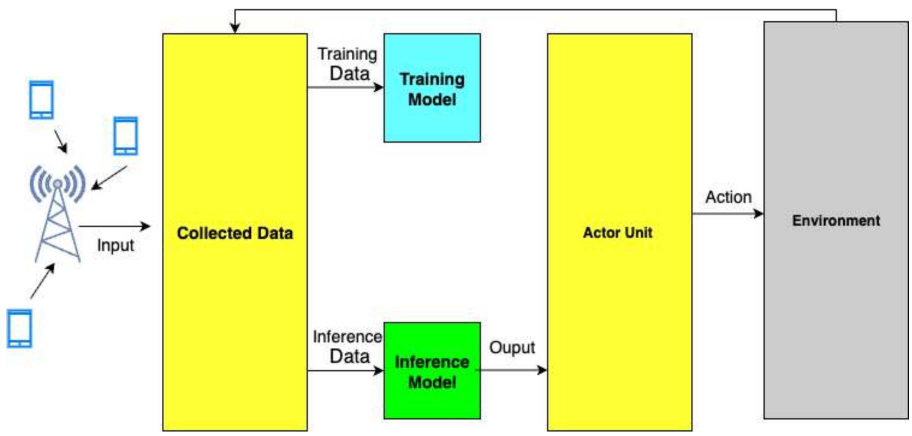

Figure 2.5: Generic Framework for RAN Intelligence

Compliance with 3GPP Standards

O-RAN AI-native procedures conform to 3GPP Release 18 specifications, including the Network Data Analytics Function (NWDAF) framework for standardized data collection, analytics, and AI service orchestration.

This standardization ensures interoperability, vendor-neutral deployments, and scalability of AI-native solutions in commercial networks.

2.4.1 Key AI-native Methods and Algorithms

To enable intelligent and autonomous behavior in next-generation wireless systems, various AI-native methods and algorithms have been integrated into different layers of the network architecture. These techniques offer powerful capabilities in learning, adaptation, prediction, and decision-making, particularly in highly dynamic and resource-constrained environments like O-RAN. This section highlights key methods, ranging from reinforcement learning and federated learning to graph-based models and transformer architectures—that are shaping the design and operation of AI-native protocols.[45]

Reinforcement Learning (RL) and Deep RL

RL allows agents to learn optimal policies through interactions with the environment, receiving rewards or penalties. Deep RL uses deep neural networks to handle complex state spaces, making it suitable for real-time network control challenges like adaptive scheduling and handover decisions in O-RAN.

Federated Learning (FL)

FL distributes AI model training across multiple RAN components, enabling collaboration without exposing raw data, crucial for user privacy. FL is applied in non-RT RIC rApps for predictive maintenance, fault detection, and network optimization.

Graph Neural Networks (GNNs)

GNNs effectively model the network topology as graphs, capturing spatial relationships and dependencies. They are utilized for adaptive routing, link quality prediction, and anomaly detection, which are critical for maintaining network performance in dynamic conditions.

Transformers and Large Language Models (LLMs)

Emerging use of transformer architectures in network management tasks enables handling complex temporal dependencies and automated policy formulation. LLMs assist in generating natural language explanations for AI decisions and enabling intuitive network control interfaces.[51]

2.4.2 Case Studies and Projects

While AI-native procedures and protocols have been widely discussed in theoretical and architectural contexts, several real-world initiatives are actively implementing these concepts. This section presents selected case studies and research projects that demonstrate how AI is being integrated into current and future wireless systems, from O-RAN deployments and 6G research to standardization efforts within 3GPP. These examples provide practical insights into the maturity, challenges, and directions of AI-native networking.

O-RAN and the Role of RIC

The O-RAN Alliance's RIC concept implements AI-native control by hosting xApps and rApps that apply AI/ML techniques in near real-time and non-real-time scenarios, respectively. These applications enable dynamic, autonomous control of network resources, improving performance and reducing operational costs [52].

6G Research Projects

Leading projects such as Hexa-X [53] explore AI-native techniques for ultra massive MIMO systems with positioning information integration, focusing on improving spatial multiplexing, beamforming, and latency-sensitive scheduling.

3GPP Standard Contributions

3GPP Release 18 introduces enhancements to support AI-native procedures, including standardized data models and analytics interfaces within NWDAF, essential for AI orchestration and cross-vendor compatibility [54].

2.4.3 Open Challenges and Future Directions

Despite progress, several critical challenges remain for AI-native procedures in current systems:

- Explainability and Interpretability: AI-native decisions must be transparent and explainable to network operators to build trust and facilitate debugging.
- Data Quality and Availability: Scarcity of labeled training data and heterogeneous data sources complicate effective AI model training, especially in federated settings.

- Real-time Constraints: Near-RT RIC operations demand AI methods that satisfy stringent latency requirements while maintaining high accuracy.
- Robustness to Adversarial Attacks: Ensuring AI models are resilient to malicious manipulations is vital for network security.
- Interoperability and Standardization: Integrating AI-native protocols seamlessly across multi-vendor and legacy systems remains a complex task.

Future research is poised to focus on the fusion of AI-native protocols with edge computing, optimization of AI lifecycle management, and extension of AI-native principles to 6G technologies such as terahertz communications and integrated sensing.

Chapter 3

Planned Activities on RRM and Scheduling

3.1 Research Directions at the UE Level

The evolution from 5G to 6G is expected to bring transformative changes to wireless communications, particularly through the use of terahertz frequencies and the adoption of ultra-massive MIMO (UmMIMO) technologies. While much research has focused on base station capabilities, the UE side presents equally critical challenges and opportunities for meeting the demands of 6G networks, including higher data rates, ultra-low latency, and enhanced reliability.

With sub-terahertz frequency bands, the wavelength becomes small enough to allow miniaturization of antenna elements and their integration into compact UE designs. Smaller antennas enable larger arrays, which in turn increase spectral efficiency and support higher data rates. Furthermore, beamforming can enhance transmit power efficiency, improve robustness against fading and noise, and strengthen interference management. The implementation of hybrid beamforming techniques is essential to address the high path loss and directional propagation characteristics of sub-THz signals [55].

The design of antenna array configurations and the selection of appropriate modulation schemes are crucial for optimizing performance while maintaining energy efficiency. To achieve the high throughput targets of 6G, the UE must support multiple independent data streams, which requires sophisticated spatial multiplexing techniques. This, in turn, demands precise control over antenna geometry, inter-element spacing, and beamforming capabilities to minimize interference and maximize channel capacity. Moreover, the modulation schemes must be adaptive, capable of dynamically adjusting to channel conditions to balance spectral efficiency and power consumption. Higher-order modulations such as 256-QAM or beyond may be employed under favorable conditions, while lower-order schemes ensure robustness in challenging environments. Together, these design choices enable UEs to sustain high data rates without compromising battery life or thermal stability, which are critical for mobile and wearable devices operating in dense, high-frequency 6G networks [56].

UmMIMO operations demand high-speed signal processing and multiple RF chains which will increase the power consumption significantly. Battery life in mobile devices and thermal dissipation, especially in compact enclosures, need to be considered. Moreover, the integration of these elements in the UE will make the signal processing and hardware design more complex. Currently, there are studies related to to optimize the spectral efficiency while minimizing the energy consumption of the communication system and Towards Power Efficient 6G Sub-THz Transmission [57, 58].

In 6G networks, integrating advanced sensing and localization capabilities into UEs will be pivotal, especially for immersive technologies such as augmented reality (AR) and virtual reality (VR). These applications demand centimeter-level localization accuracy to ensure seamless alignment between digital content and the physical environment. Sub-terahertz frequencies, with their short wavelengths and high spatial resolution, offer the potential for ultra-precise positioning. By leveraging joint communication and sensing (JCAS) techniques, 6G UEs can simultaneously transmit data and extract environmental information, enabling real-time tracking of user movement and surroundings. This precision is essential for minimizing motion-to-photon latency and avoiding spatial disorientation in AR/VR experiences. Furthermore, accurate localization enhances context-awareness, allowing devices to adapt content dynamically based on user position and orientation, which is critical for industrial, medical, and entertainment applications in the 6G era.

As 6G networks push into higher frequencies and more directional communication paradigms, ensuring robust security and privacy at the user equipment level becomes increasingly critical. The use of sub-terahertz bands and ultra-massive MIMO introduces new vulnerabilities but also opens opportunities for physical layer security techniques that exploit the unique characteristics of the wireless channel. For instance, beam-based communication enables spatial confinement of signals, reducing the risk of eavesdropping and unauthorized access. Additionally, location-aware encryption and channel-based key generation can provide lightweight, context-sensitive security mechanisms that are well-suited for mobile and resource-constrained devices. These innovations are particularly important in dense urban environments and for applications involving sensitive data, such as healthcare, finance, and industrial automation. As 6G UEs become more intelligent and interconnected, embedding security into the physical and MAC layers will be essential to protect user data and ensure trust in next-generation wireless systems [59].

3.2 Research directions for Non-Coherent Systems

Multi-user non-coherent schemes

Initial MIMO schemes were designed with the goal of maximizing capacity in a single-user MIMO setup. Non-coherent modulation schemes are no exception and such is the case of Grassmannian schemes, DUSTM, and STBC. However, these schemes do not scale well with increasing numbers of users, and become unfeasible with the expected number of multiplexed users in UmMIMO. As a result, conceptually simpler schemes, like energy-based and differential schemes based on averaging are more desirable. Among these, differential averaging schemes are particularly attractive because they exploit the phase of the incoming signal, making them more robust and enabling the possibility of spatial processing.

The extension of differential schemes based on averaging to a multi-user scenario has been done by multiplexing users in the constellation domain, in a similar manner to multi-user energy based schemes but with the advantage of having the complex plane (a two-dimensional space), instead of just the received energy (a one-dimensional space). The symbols transmitted from several users will sum over the air, forming joint symbols $z[n]$ (as shown in Figure 3.1), which stem from the extension of (2.10) to the multi-user case as

$$\begin{aligned} z[n] =& \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N} y_{k,i}[n-1]^* y_{k,i}[n] \\ =& \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N} |h_{ki}|^2 \beta_k s_k[n] + \zeta[n], \end{aligned} \tag{3.1}$$

where K is the total number of users, h_{ki} is the complex channel coefficient from user k to antenna i, β_k is the transmitted power by user k, and ζ represent the interference and noise terms. It can be shown that, if the different users have uncorrelated channels, ζ decreases as N increases [44].

The joint constellation (as the one from Figure 3.1) must be known by the receiving base station to assign the correct bits to each user. On top of that, stable power control is required by the user terminals being multiplexed, so that the joint constellation does not degrade excessively.

In summary, the main differences between the coherent and the non-coherent RRM optimization problem are:

- More resource elements (REs) in the OFDM grid are available for data transmission in non-coherent schemes, since no pilots are required.
- Spatial layers and beam resources are lost due to the absence of a channel sensing protocol.

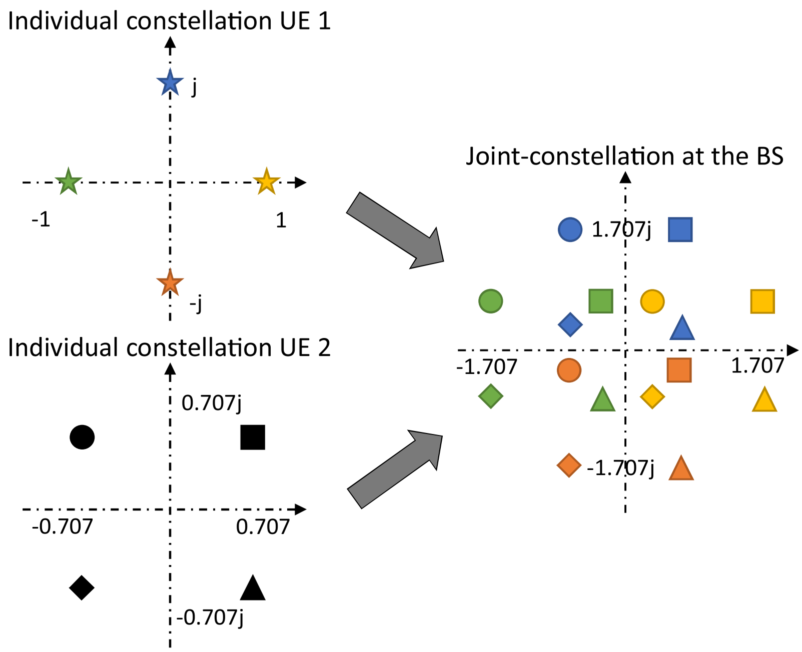

Figure 3.1: Example of a joint constellation for two users from [2].

- Users can be multiplexed in the constellation domain [60].
- Data from the same user cannot be scattered across the OFDM grid, as (2.7) requires the channel to remain constant, which is only a reasonable assumption for contiguous time-frequency resources.

RRM for non-coherent systems

Radio Resource Management in non-coherent schemes shares its objectives with coherent RRM, i.e., maximize the amount of data transmitted by the system while maintaining a minimum QoS for all users.

Provided the users have spatially uncorrelated channels, non-coherent differential modulation offers the possibility of constellation multiplexing [60]. This form of multiplexing does not scale well with the number of users, due to the points of the joint constellation being very close to each other. To deal with this, constellation multiplexing can be combined with scheduling users in orthogonal time-frequency resources. For instance, [61] demonstrates the benefits of combining constellation multiplexing with TDMA, but it does not address the fundamental limitation of non-coherent schemes being restricted to M-PSK constellations, which significantly limits throughput at high SNR compared to coherent schemes.

To mitigate the lack of spatial resources that non-coherent systems suffer from, [62] proposes a passive spatial filtering approach for Rician fading that enables the exploitation of spatial resources without any previous knowledge of the channel, which removes the need for constellation multiplexing and allows for more separation between constellation points. The angles of the different users are part of the constraints of the RRM procedure and, since the sensing in [62] is completely passive, the angles can be known during the initial connection of each user with the BS through the RACH procedure, which happens before the scheduling of data time-frequency resources. This can provide useful information for the RRM process. Despite its promising results, passive spatial filtering is limited to Rician channels, i.e., scenarios with a dominant line-of-sight (LOS) component. This makes it less practical for real-world deployments, where users are expected to switch seamlessly between Rayleigh (NLOS) and Rician (LOS) conditions. Multiplexing users in the angular domain under Rayleigh fading is significantly more difficult and remains an open research challenge.

Finally, as in coherent communications, spatially correlated channels are undesirable, since neither constellation multiplexing (3.1) nor spatial filtering an be effectively applied. Therefore, one viable RRM strategy is be to group users that have a substantial spatial separation between them in the same time-frequency resources, while reserving orthogonal subcarriers or time slots for users that suffer from high spatial correlation.

3.3 Research Direction at AI-native techniques

AI-native Scheduling for Ultra Massive MIMO

Efficient resource scheduling in UM-MIMO systems requires managing large antenna arrays and diverse user demands in real time. Traditional rule-based scheduling methods become infeasible due to the complexity and scale. AI-native scheduling uses advanced RL and Deep RL to dynamically allocate resources by learning from environmental feedback, improving throughput, and reducing latency. Positioning information further enhances scheduling decisions by predicting user mobility and channel quality variations, enabling proactive resource allocation and improved reliability. Recent studies have shown that AI-driven scheduling methods outperform classical approaches, especially in dense urban and highly mobile scenarios [63].

AI-native Beamforming in Ultra Massive MIMO

Beamforming is vital for focusing transmissions and minimizing interference in Um-MIMO. However, the large antenna count complicates real-time beam management. AI-native beamforming utilizes deep learning techniques such as Transformer models and GNNs to infer optimal beam patterns from channel state and positioning data. These models enable faster and more accurate beam selection compared to traditional optimization algorithms, and they adapt to dynamic propagation conditions effectively. Leveraging positioning allows beamforming to anticipate user movements and adjust beams proactively, enhancing link stability and throughput [64].

AI-native Channel Estimation for Ultra Massive MIMO

Channel estimation in UmMIMO suffers from pilot contamination and high training overhead due to massive antenna arrays. AI-native methods employ machine learning models that capture spatial-temporal correlations in channel data, reducing pilot signal requirements while maintaining high accuracy. FL frameworks facilitate decentralized training across multiple nodes, preserving user privacy and scaling efficiently.

Incorporating positioning information further refines channel models by relating physical user locations with propagation characteristics, aiding in handling blockage and multipath effects [65]. Experimental evaluations confirm that AI-native channel estimators outperform conventional methods under diverse and dynamic channel conditions.

Integration of Positioning in AI-native Ultra Massive MIMO Protocols

Positioning plays a pivotal role in enhancing AI-native UmMIMO procedures. Joint communication and sensing paradigms allow fusion of location, velocity, and environmental context data with network information, improving scheduling, beamforming, and channel estimation decisions. This synergy enables ultra-reliable and low-latency communications by facilitating faster handovers, better interference management, and dynamic resource optimization. Leading 6G research projects emphasize the integration of positioning to realize fully AI-native UmMIMO systems capable of adapting to complex environments [66]

Conclusion

AI-native techniques that incorporate positioning information are essential to meet the stringent performance demands of UmMIMO in future wireless networks. These techniques enable adaptive scheduling, intelligent beamforming, and efficient channel estimation, addressing challenges posed by large antenna arrays and user dynamics. Future work should focus on developing scalable AI models with enhanced privacy features and deep integration of sensing and positioning capabilities to achieve next-generation communication goals.

3.4 Research Directions in Spatial Resource Allocation and AP selection

Spatial resource allocation can be achieved through the design of beamforming vectors, the selection of suitable APs for these beamforming vectors, and optimized power allocation and scheduling, all supported by accurate user positioning.

User position can be estimated using either distance-based or angle-based approaches. The simplest method involves Time Difference of Arrival (TDoA) combined with the Least Mean Squares (LMS) technique, which provides accurate and fast results with low complexity, but is not well-suited for continuous user tracking. For mobility scenarios, filtering techniques such as Kalman and Extended Kalman filters can be employed. Since beamforming design often benefits from angular information, Angle of Arrival (AoA)-based methods are particularly useful.

Once the user's position is known, beamforming vectors can be designed by incorporating channel conditions. This can be accomplished by analyzing eigenvalues in the joint eigenspace of users and APs. Based on these eigenvalues, the most suitable APs can be selected, thereby reducing the overall network complexity. Subsequently, power allocation from each AP to the user can be optimized to maximize efficiency and performance.

Figure 3.2 illustrates a scenario where each user is connected to multiple APs, which contribute different amounts of power and direct their beams accordingly. Some APs provide lower power contributions, while others contribute more, depending on the channel conditions between the AP and the user.

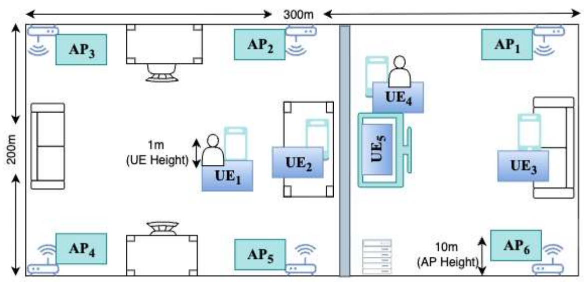

Figure 3.2: Illustration Scenario

Figure 3.3 shows the power required from APs when all the APs are connected to serve the user. Connecting all the APs increases complexity and computation, thus it is benefecial to effectively select only the optimal subset of APs which would jointly serve the user. For instance, AP-6 contributes significantly only to User-3, while its contributions to other users are negligible. Since each AP has a total power budget, forcing AP-6 to serve users with poor channel gains would require disproportionately high transmit power. A more efficient strategy is to deactivate AP-6 for those users and concentrate its available power on User-3, where the channel gains are stronger, as transmit power is proportional to the square of the signal amplitude.

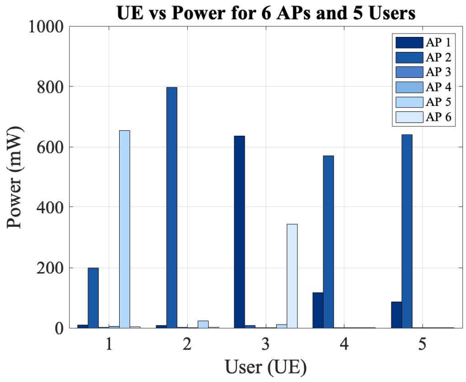

Figure 3.3: Power Allocated from each AP to different users

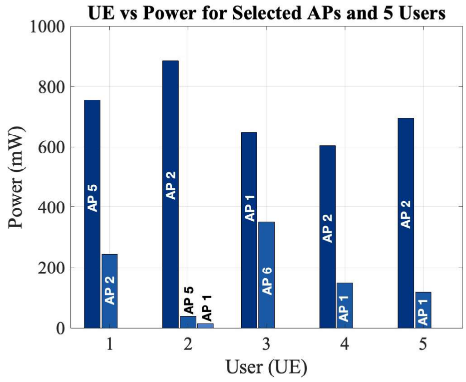

Figure 3.4: Power Allocated from selected APs to different users

Optimised Power and AP selection

As shown in Figure 3.4, low-contributing APs are switched off, and only APs that provide effective beamforming for each user remain active, while still meeting user QoS requirements. This optimization is possible by leveraging user location, channel gains, and beamforming vector design.

In summary, incorporating user location into spatial resource allocation allows more effective planning, reduces the overall network burden, improves latency and throughput, and enables dynamic and energy-efficient resource allocation.

Conclusion

With the support of state-of-the-art methods, we are exploring new directions such as integrating sensing and localization capabilities into UE to enable advanced applications like AR and digital twins through JCAS. For 6G to reach its full potential, extensive research is required on non-coherent systems, including the design of energy-based and differential schemes. Equally important is the development of advanced beamforming and resource allocation strategies, such as eigenvector-based methods, which can make effective use of user location information. Achieving these goals also depends on AI-native techniques, which can be leveraged to estimate channel models and anticipate scheduling needs. Together, these research directions are complementary and essential, forming a roadmap toward the successful realization of 6G and cell-free systems in the coming years.

References

[1] K.-H. Ngo, D. Cuevas, R. de Miguel Gil, V. M. Baeza, A. G. Armada, and I. Santamaria, "Noncoherent mimo communications: Theoretical foundation, design approaches, and future challenges," 2025. [Online]. Available: https://arxiv.org/abs/2505.23172

[2] K. Chen Hu, Y. Liu, and A. García Armada, "Non-Coherent massive MIMO-OFDM for communications in high mobility scenarios," *ITU Journal on Future and Evolving Technologies*, Dec. 2020.

[3] everything RF, "What is radio resource management?" https://www.everythingrf.com/community/what-is-radio-resource-management, 2024, accessed: 2024-05-28.

[4] E. Björnson and Ö. T. Demir, *Introduction to Multiple Antenna Communications and Reconfigurable Surfaces*. Now Publishers, Inc., Jan. 2024.

[5] F. Boixadera, "Radio resource management," in *LTE – The UMTS Long Term Evolution*, S. Sesia, I. Toufik, and M. Baker, Eds. John Wiley & Sons Ltd., 2009, pp. 301–320.

[6] H. T. Dao and S. Kim, "Effective channel gain-based access point selection in cell-free massive mimo systems," *IEEE Access*, vol. 8, pp. 108 127–108 132, 2020.

[7] F. Conceição, M. Gomes, V. Silva, and R. Dinis, "Streamlined swift allocation strategies for radio stripe networks," 2024. [Online]. Available: https://arxiv.org/abs/2412.07732

[8] F. Conceição, C. H. Antunes, M. Gomes, V. Silva, and R. Dinis, "Max-min fairness optimization in uplink cell-free massive mimo using meta-heuristics," *IEEE Transactions on Communications*, vol. 70, no. 3, pp. 1792–1807, 2022.

[9] E. Björnson, J. Hoydis, and L. Sanguinetti, *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency*, 2017.

[10] A. Manolakos, M. Chowdhury, and A. J. Goldsmith, "Constellation design in noncoherent massive SIMO systems," in *2014 IEEE Global Communications Conference*, Dec. 2014, pp. 3690–3695.

[11] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive mimo systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4445–4459, 2017.

[12] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge University Press, 2016.

[13] B. E. Q. N. H. Interdonato, G., "Ubiquitous cell-free massive mimo communications," *J Wireless Com Network*, 2019.

[14] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive mimo versus small cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, 2017.

[15] E. Björnson and L. Sanguinetti, "Making cell-free massive mimo competitive with mmse processing and centralized implementation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 77–90, 2020.

[16] Y. Zhao, I. G. Niemegeers, and S. H. De Groot, "Power allocation in cell-free massive mimo: A deep learning method," *IEEE Access*, vol. 8, pp. 87 185–87 200, 2020.

[17] T. H. Nguyen, T. K. Nguyen, H. D. Han, and V. D. Nguyen, "Optimal power control and load balancing for uplink cell-free multi-user massive mimo," *IEEE Access*, vol. 6, pp. 14 462–14 473, 2018.

[18] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive mimo," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 1, pp. 25–39, 2018.

[19] T. X. Vu, S. Chatzinotas, S. ShahbazPanahi, and B. Ottersten, "Joint power allocation and access point selection for cell-free massive mimo," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[20] A. Pereira, F. Rusek, M. Gomes, and R. Dinis, "Deployment strategies for large intelligent surfaces," *IEEE Access*, vol. 10, pp. 61 753–61 768, 2022.

[21] F. Conceição, M. Gomes, V. Silva, R. Dinis, and C. H. Antunes, "Joint spectral and power efficiency optimization in uplink radio stripes," *IEEE Transactions on Communications*, vol. 72, no. 8, pp. 5209–5225, 2024.

[22] F. Conceição, L. Martins, M. Gomes, V. Silva, and R. Dinis, "Access point selection for spectral efficiency and load balancing optimization in radio stripes," *IEEE Communications Letters*, vol. 27, no. 9, pp. 2383–2387, 2023.

[23] C. Pesantez Diaz, F.; Estevez, "Performance evaluation of cf-mmimo wireless systems using dynamic mode decomposition," *Telecom*, 2024.

[24] E. Björnson and E. Jorswieck, *Optimal Resource Allocation in Coordinated Multi-Cell Systems*, 2013.

[25] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive mimo systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4445–4459, 2017.

[26] E. Björnson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes]," *IEEE Signal Processing Magazine*, vol. 31, no. 4, pp. 142–148, 2014.

[27] T. V. Chien, E. Björnson, and E. G. Larsson, "Optimal design of energy-efficient cell-free massive mimo: Joint power allocation and load balancing," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 5145–5149.

[28] F. F. J. W. Y. Lin, X.; Xu, "Resource allocation for tdd cell-free massive mimo systems," *Electronics*, 2022.

[29] R. C. Abhinav Anand, Chandra R. Murthy, "Impact of mobility on the downlink performance of cell-free massive mimo systems," *Physical Communication*, 2023.

[30] T. Korhonen, "Scheduling," https://www.3gpp.org/technologies/scheduling, May 2023, accessed: 2025-06-03.

[31] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5g-enabled tactile internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, 2016.

[32] E. Gures, I. Shayea, A. Alhammadi, M. Ergen, and H. Mohamad, "A comprehensive survey on mobility management in 5g heterogeneous networks: Architectures, challenges and solutions," *IEEE Access*, vol. 8, pp. 195 883–195 913, 2020.

[33] ETSI, "5G; NR; User Equipment (UE) conformance specification; Radio Resource Management (RRM) (3GPP TS 38.533 version 16.4.0 Release 16)," 3GPP, Valbonne, France, Technical Specification 3GPP TS 38.533 V16.4.0, 2020, available online: https://www.3gpp.org/ftp/Specs/archive/38_series/38.533/38533-g40.zip.

[34] NR Explained, "Timing advance (ta) in 5g nr," https://www.nrexplained.com/ta, 2024, accessed: 2025-06-05.

[35] MathWorks, "Nr uplink channel state information estimation using srs," https://se.mathworks.com/help/5g/ug/nr-uplink-channel-state-information-estimation-using-srs.html, 2024, accessed: 2025-06-05.

[36] TechSpec, "16.1 synchronization procedures — 3gpp ts 36.213," https://itecspec.com/spec/3gpp-36-213-16-1-synchronization-procedures/, 2025, accessed: 2025-06-05.

[37] B. M. Hochwald and T. L. Marzetta, "Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading," *IEEE transactions on Information Theory*, vol. 46, no. 2, pp. 543–564, 2000.

[38] T. Marzetta and B. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Transactions on Information Theory*, vol. 45, no. 1, pp. 139–157, Jan. 1999.

[39] K.-H. Ngo, D. Cuevas, R. de Miguel Gil, V. Monzon Baeza, A. Garcia Armada, and I. Santamaria, "Noncoherent MIMO Communications: Theoretical Foundation, Design Approaches, and Future Challenges," *Submitted to IEEE Communications Surveys and Tutorials*, 2025.

[40] A. Manolakos, M. Chowdhury, and A. Goldsmith, "Energy-Based Modulation for Noncoherent Massive SIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7831–7846, Nov. 2016.

[41] B. Hughes, "Differential space-time modulation," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2567–2578, Nov. 2000.

[42] B. Hochwald and W. Sweldens, "Differential unitary space-time modulation," *IEEE Transactions on Communications*, vol. 48, no. 12, pp. 2041–2052, Dec. 2000.

[43] V. Tarokh and H. Jafarkhani, "A differential detection scheme for transmit diversity," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 7, pp. 1169–1174, Jul. 2000.

[44] A. G. Armada and L. Hanzo, "A non-coherent multi-user large scale SIMO system relaying on M-ary DPSK," in *2015 IEEE International Conference on Communications (ICC)*, Jun. 2015, pp. 2517–2522.

[45] W. Azariah, F. A. Bimo, C.-W. Lin, R.-G. Cheng, N. Nikaein, and R. Jana, "A survey on open radio access networks: Challenges, research directions, and open source approaches," *Sensors*, vol. 24, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/3/1038

[46] P. Almasan, J. Suárez-Varela, K. Rusek, P. Barlet-Ros, and A. Cabellos-Aparicio, "Deep reinforcement learning meets graph neural networks: Exploring a routing optimization use case," *Computer Communications*, vol. 196, p. 184–194, Dec. 2022.

[47] H. Zhang, H. Zhou, and M. Erol-Kantarci, "Federated deep reinforcement learning for resource allocation in o-ran slicing," 2022. [Online]. Available: https://arxiv.org/abs/2208.01736

[48] M. H. Meng, G. Bai, S. G. Teo, Z. Hou, Y. Xiao, Y. Lin, and J. S. Dong, "Adversarial robustness of deep neural networks: A survey from a formal verification perspective," *IEEE Transactions on Dependable and Secure Computing*, p. 1–1, 2024. [Online]. Available: http://dx.doi.org/10.1109/TDSC.2022.3179131

[49] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding o-ran: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Communications Surveys Tutorials*, vol. 25, no. 2, pp. 1376–1411, 2023.

[50] P. K. Thiruvasagam, C. T, V. Venkataram, V. R. Ilangovan, M. Perapalla, R. Payyanur, S. M. D, V. Kumar, and K. J, "Open ran: Evolution of architecture, deployment aspects, and future directions," 2023. [Online]. Available: https://arxiv.org/abs/2301.06713

[51] Z. Liu, B. Liu, A. Valcarce, and X. Chu, "Llm-based emulation of the radio resource control layer: Towards ai-native ran protocols," 2025. [Online]. Available: https://arxiv.org/abs/2505.16821

[52] "O-ran research reports," accessed: July 24, 2025. [Online]. Available: https://www.o-ran.org/research-reports

[53] H. Wymeersch, A. Stavridis, K. Schindhelm, H. Chen, H. Guo, M. F. Keskin, S. Lindberg, J. M. Mateos Ramos, M. H. Moghaddam, M. A. Nazari, I. Perera, A. Ramirez, R. Schroeder, T. Svensson, A. Wolfgang, and V. Yajnanarayana, "Hexa-x reports for localisation and sensing," accessed: July 24, 2025. [Online]. Available: https://hexa-x.eu/new-deliverable-about-final-models-and-measurements-for-localisation-and-sensing-has-been-publis

[54] "3gpp release 18," accessed: July 24, 2025. [Online]. Available: https://www.3gpp.org/specifications-technologies/releases/release-18

[55] R. Dilli, "Design and feasibility verification of 6g wireless communication systems with state of the art technologies," *International Journal of Wireless Information Networks*, vol. 29, pp. 93–117, 2022, published online 28 December 2021.

[56] ——, "Design and analysis of sixth-generation (6g) microstrip patch antenna," *SN Applied Sciences*, vol. 4, no. 3, pp. 1–18, 2022.

[57] Z. Gao, Z. Wan, Y. Mei, K. Ying, and K. Wang, *Millimeter-Wave/Sub-Terahertz Ultra-Massive MIMO Transmission Technology*, 1st ed. Singapore: Springer Singapore, 2023.

[58] M. Younes, P. Desombre, and Y. Louet, "Trade-off between spectral and energy efficiency in thz frequency bands for 6g," in *2022 1st International Conference on 6G Networking (6GNet)*, 2022, pp. 1–5.

[59] I. F. Akyildiz, A. Kak, and S. Nie, "6g and beyond: The future of wireless communications systems," *IEEE Access*, vol. 8, pp. 133 995–134 030, 2020.

[60] M. J. Lopez-Morales, K. Chen-Hu, A. Garcia-Armada, and O. A. Dobre, "Constellation Design for Multiuser Non-Coherent Massive SIMO Based on DMPSK Modulation," *IEEE Transactions on Communications*, vol. 70, no. 12, pp. 8181–8195, Dec. 2022.

[61] V. M. Baeza and A. G. Armada, "Orthogonal versus Non-Orthogonal multiplexing in Non-Coherent Massive MIMO Systems based on DPSK," in *2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, Jun. 2021, pp. 101–105.

[62] R. de Miguel Gil, K. Chen-Hu, and A. Garcia Armada, "Angle-Domain Multi-User Multiplexing for Non-Coherent Massive MIMO in Rician Channels," in *Submitted to 2025 IEEE Global Communications Conference (GLOBECOM)*, 2025.

[63] Q. An, S. Segarra, C. Dick, A. Sabharwal, and R. Doost-Mohammady, "A deep reinforcement learning-based resource scheduler for massive mimo networks," *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 1, pp. 242–257, 2023.

[64] N. Shlezinger, M. Ma, O. Lavi, N. T. Nguyen, Y. C. Eldar, and M. Juntti, "Ai-empowered hybrid mimo beamforming," 2023. [Online]. Available: https://arxiv.org/abs/2303.01723

[65] X. Zhou, L. Liang, J. Zhang, P. Jiang, Y. Li, and S. Jin, "Generative diffusion models for high dimensional channel estimation," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2025.

[66] G. Pan, Y. Gao, Y. Gao, Z. Zhong, X. Yang, X. Guo, and S. Xu, "Ai-driven wireless positioning: Fundamentals, standards, state-of-the-art, and challenges," 2025. [Online]. Available: https://arxiv.org/abs/2501.14970