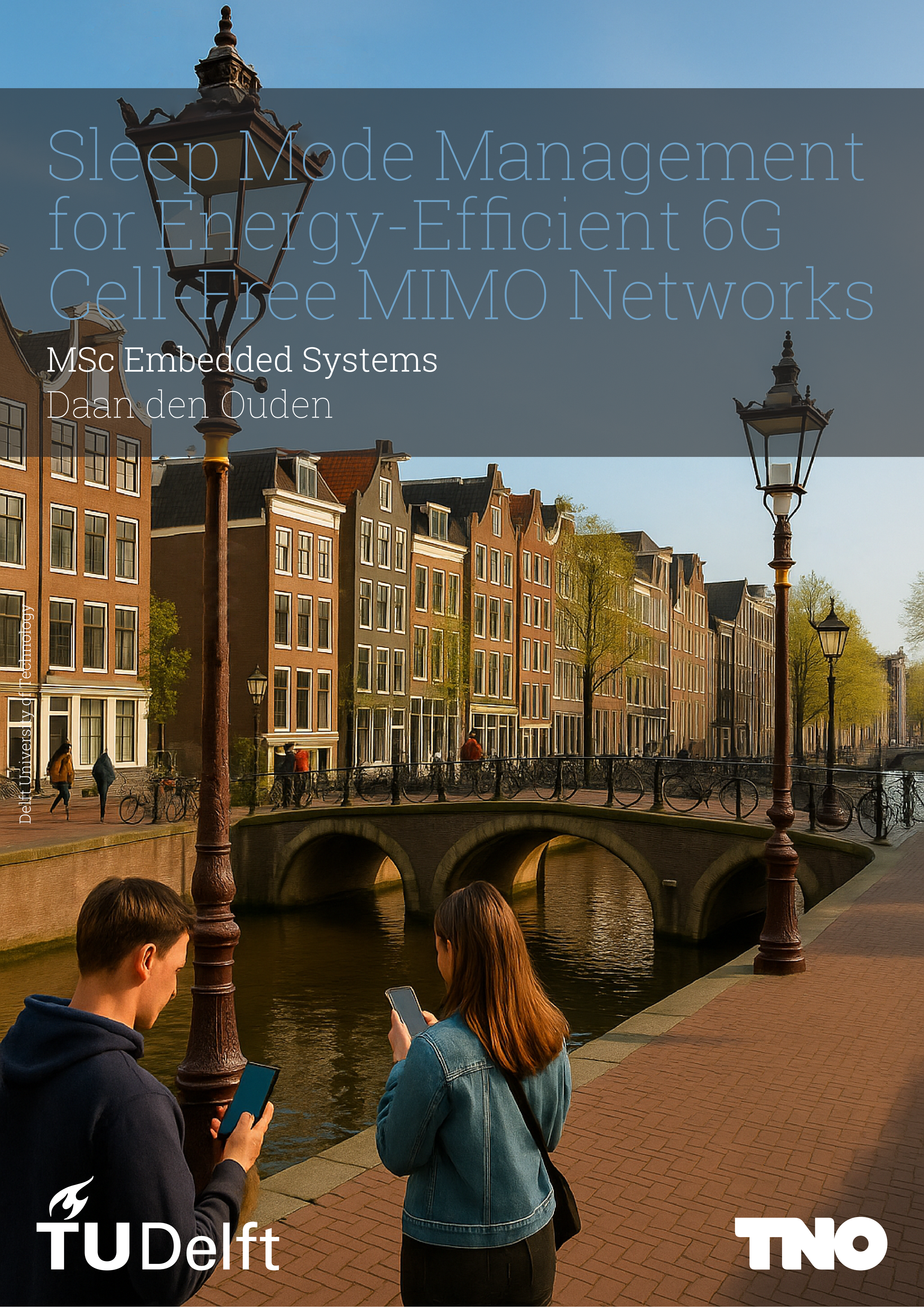
Sleep Mode Management for Energy-Efficient 6G Cell-Free MIMO Networks

MSc Embedded Systems

Daan den Ouden



Delft University of Technology

Sleep Mode Management for Energy-Efficient 6G Cell-Free MIMO Networks

by

Daan den Ouden

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Monday, July 14, 2025 at 15:00.

Student number:	4730070	
Project duration:	May 21, 2024 – July 14, 2025	
Thesis committee:	Dr. Remco Litjens MSc,	TU Delft/TNO, supervisor
	Prof. Dr. Haibin Zhang,	TNO, co-supervisor
	Dr. Qing Wang,	TU Delft

Preface

After more than a year of hard work, I am pleased to present my Master's thesis. Marking the end of more than twenty years of formal education, this work fulfills the last requirement necessary to obtain the degree of Master of Science in Embedded Systems at the Delft University of Technology. Over the course of the project, there have been many challenges. Fortunately, I did not have to face them alone.

First and foremost, I would like to extend my gratitude to my supervisors who have offered their valuable time to guide and support me. I want to thank my TU Delft supervisor Dr. Remco Litjens whose excellent Mobile Networks course started this thesis journey. Thank you, Remco, for always providing extensive and consistent feedback; always taking the time to discuss, brainstorm and explain. I also wish to thank my company supervisor Prof. Dr. Haibin Zhang for his invaluable comments, ideas and suggestions. Thank you, Haibin, for finding time in your busy schedule, even when halfway around the world. Furthermore, I would like to thank Dr. Maria Raftopoulou as the third, slightly less official supervisor. Thank you, Maria, for your participation in the numerous weekly thesis meetings and your feedback on my writing. Additionally, I would like to thank Dr. Qing Wang for assessing my work as a member of the thesis committee.

This research was conducted at The Netherlands Organisation for Applied Scientific Research (TNO) as part of the 6G Future Network Services (FNS) research program. I am very proud that part of my research has led to a patent application in the context of this program. I want to thank TNO, and my supervisors specifically, for providing me with this graduation project opportunity. I am grateful to all the wonderful and knowledgeable people at TNO's Networks department who have helped me tremendously over the course of the project. A special thanks to Sakshi Agarwal for sharing her knowledge on developing a system-level simulator. Thank you, to my fellow Networks interns, for providing your support and suggestions, even as you were all also incredibly busy with your own thesis projects.

A significant part of my research required simulations which could not have been performed if not for the considerable computing resources provided by both TNO and TU Delft. I specifically want to acknowledge the use of computational resources of the DelftBlue supercomputer provided by Delft High Performance Computing Centre [1], the compute cluster of the Q&CE department in the faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) at TU Delft and the Research Cloud of TNO's Networks department.

Finally, I would like to thank my friends and family who have all supported me immensely. Thank you for listening to my ramblings about simulator bugs, modeling choices and writing difficulties.

Daan den Ouden
The Hague, July 2025

Abstract

The sixth generation (6G) of mobile networks promises transformative capabilities in terms of, amount others, higher data rates, lower latency and ubiquitous coverage, but achieving these goals sustainably poses significant challenges. A promising solution lies in Cell-Free massive Multiple-Input Multiple-Output (CF-mMIMO) networks, where a dense deployment of geographically distributed Access Points (APs), coordinated by one or more Central Processing Units (CPUs), cooperatively serve User Equipments (UEs). While CF-mMIMO networks offer improved spectral efficiency and more spatially uniform service, their dense deployments can result in high energy consumption. As Mobile Network Operators (MNOs) typically design their networks such that Quality of Service (QoS) targets are met during peak hours, the network is left significantly over-dimensioned during off-peak hours. This thesis addresses this inefficiency by proposing a low-complexity, heuristic Sleep Mode Management (SMM) algorithm that reduces energy usage by switching off unneeded APs while maintaining acceptable QoS and without compromising coverage.

The proposed SMM algorithm supports multiple AP power states: active, light sleep and deep sleep. It incorporates realistic transition times between these states. Relying solely on practically available information such as long-term Channel State Information (CSI) and previously achieved data rates, the algorithm dynamically decides which APs can be temporarily put into a sleep mode. Importantly, it ensures population coverage is maintained and it preserves UE QoS based on a 10th UE throughput percentile target.

The proposed SMM algorithm is evaluated using a system-level simulator that models a realistic scenario based on the city center of Amsterdam, including lamppost-based AP deployments and a realistic basis for the spatial traffic distribution and daily traffic fluctuations. Simulation results demonstrate that the proposed SMM algorithm reduces the daily energy consumption of a CF-mMIMO network by up to 17.11% with the best overall configuration. If the parameters of the SMM algorithm are allowed to be adaptively tuned to the traffic load, the daily energy consumption can be reduced by up to 21.54%. This thesis not only contributes a novel SMM algorithm but also provides guidance for AP deployment strategies in 6G CF-mMIMO networks, determining that a higher number of lower-antenna-count APs can yield better QoS than fewer, more higher-antenna-count APs at the cost of energy efficiency.

Acronyms

3GPP	3rd Generation Partnership Project
ADC	Analog-to-Digital Converter
AE	Antenna Element
AI	Artificial Intelligence
AP	Access Point
ARZF	Adaptive Regularized Zero-Forcing
ASO	AP Switch On/Off
BPL	Building Penetration Loss
BS	Base Station
CBS	Centraal Bureau voor de Statistiek
CFN	Cell-Free Network
CF-mMIMO	Cell-Free massive Multiple-Input Multiple-Output
ChiS-ASO	Chi-Square-based ASO
CoMP	Coordinated Multi-Point
CPU	Central Processing Unit
CSI	Channel State Information
DAC	Digital-to-Analog Converter
DCC	Dynamic Cooperation Clustering
DDPG	Deep Deterministic Policy Gradient
DDQN	Double Deep Q-Network
DL	Downlink
DPP	Drift-Plus-Penalty
DPR	Dynamic Pilot Reuse
DRL	Deep Reinforcement Learning
DWCS	Distributed Wireless Communication System
eMBB	enhanced Mobile Broadband
FCC	Federal Communications Commission
FR1	Frequency Range 1
FR2	Frequency Range 2
FR3	Frequency Range 3
GB-KM	Geography-Based K-Means
GCA	Greedy Combining Algorithm
GoF	Goodness-of-Fit

GPA	Greedy Pilot Assignment
HF	Hard Fairness
FNS	Future Network Services
IB-KM	Interference-Based K-Means
ICR	Interference Contribution Ratio
IM-ASO	IMportance-based ASO
IMT	International Mobile Telecommunications
ITU-R	International Telecommunication Union Radiocommunication Sector
KS-ASO	Kolmogorov-Smirnov-based ASO
KPI	Key Performance Indicator
LBGPA	Location-Based Greedy Pilot Assignment
LSE-ASO	Logarithmic Statistical Energy-based ASO
LSFC	Large-Scale Fading Coefficient
LTE	Long Term Evolution
LoS	Line-of-Sight
MA-DDQN	Multi-Agent Double Deep Q-Network
MCL	Minimum Coupling Loss
MCS	Modulation and Coding Scheme
MD-ASO	Mixture Discrepancy-based greedy ASO
MG-ASO	Max effective channel Gain-based ASO
MIESM	Mutual Information Effective SINR Mapping
MIMO	Multiple-Input Multiple-Output
MI	Mutual Information
ML	Machine Learning
mMIMO	massive Multiple-Input Multiple-Output
MMSE	Minimum Mean Square Error
mmWave	millimeter wave
MNO	Mobile Network Operator
MR	Maximum Rate
MRC	Maximum-Ratio Combining
MRT	Maximum-Ratio Transmission
MU-MIMO	Multi-User Multiple-Input Multiple-Output
NLoS	Non-Line-of-Sight
NN-ASO	Nearest Neighbor-based ASO
NUM	Network Utility Maximization
O2I	Outdoor-to-Indoor
OG-ASO	Optimal energy efficiency-based Greedy ASO
PF	Proportional Fairness

PL	Path Loss
PL-ASO	Propagation Losses-aware ASO
PPO	Proximal Policy Optimization
RPA	Random Pilot Assignment
PRB	Physical Resource Block
PZF	Partial Zero-Forcing
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QuaDRiGa	Quasi-Deterministic Radio channel Generator
RAN	Radio Access Network
RDI	Rijksinspectie Digitale Infrastructuur
RFb-LPG-ASO	RF-beamformed Large-scale Propagation Gain ASO
RNG	Random Number Generator
RR	Round-Robin
RRM	Radio Resource Management
RS-ASO	Random Selection ASO
SCS	Subcarrier Spacing
SE	Spectral Efficiency
SF	Shadow Fading
SINR	Signal-to-Interference-plus-Noise Ratio
SMM	Sleep Mode Management
SMS	Short Message Service
SR-ASO	Spatial Regularity-based greedy ASO
SU-MIMO	Single-User Multiple-Input Multiple-Output
SOCP	Second-Order Cone Program
SUS	Semi-orthogonal User Selection
TDD	Time-Division Duplexing
TR	Technical Report
TS	Technical Specification
TTI	Transmission Time Interval
UDN	Ultra-Dense Network
UE	User Equipment
UL	Uplink
umMIMO	ultra-massive Multiple-Input Multiple-Output
UMTS	Universal Mobile Telecommunications System
ULA	Uniform Linear Array
UMi	Urban Micro
URLLC	Ultra-Reliable and Low-Latency Communications

WRC-23	World Radio Conference 2023
WRC-27	World Radio Conference 2027
ZF	Zero Forcing
ZFS	Zero-Forcing with Selection

Contents

1

Introduction

Since the invention of wireless communication technologies in the late nineteenth century, the world is now more connected than ever. In 2024, approximately 98 percent of the world population was covered by a mobile network [2]. Approximately every ten years a new mobile network generation is launched, as illustrated in Figure 1.1. This started with the first generation (1G) in the 1980s and has continued through to the fifth generation (5G) which launched in the late 2010s. Expected to continue the trend, 6G will most likely be launched around 2030. 6G is envisioned to enable transformative applications such as immersive extended reality, autonomous mobility, remote healthcare and smart cities. These applications will require even higher bit rates, lower latencies and better coverage than is currently possible with 5G. Before zooming into the requirements of 6G, it is useful to first introduce the terminology used in mobile networks.

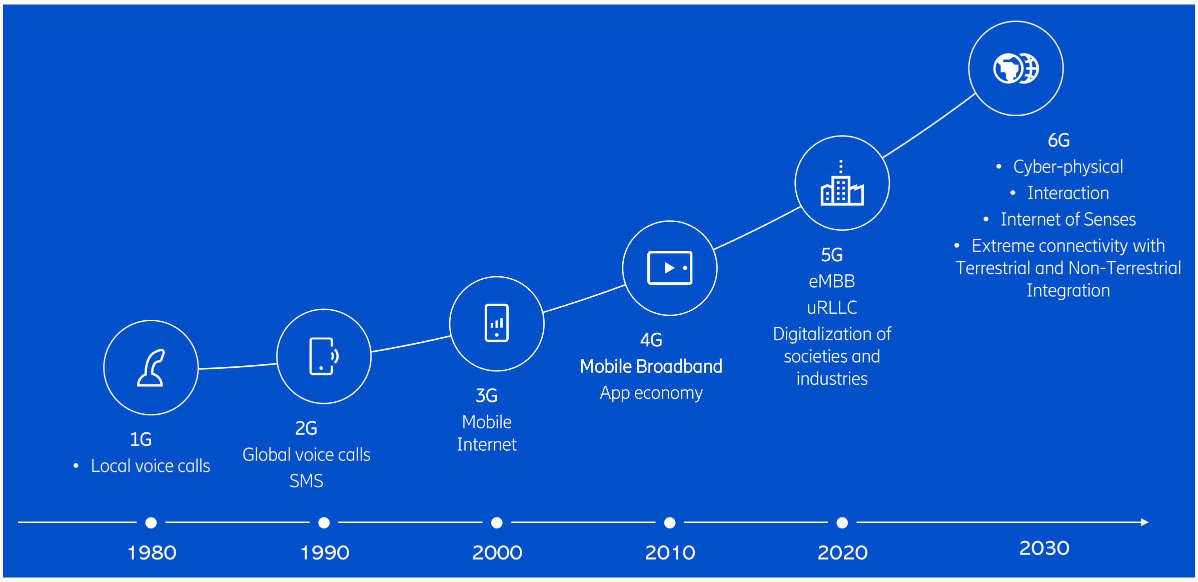


Figure 1.1: Evolution of mobile networks [3]. The abbreviations used in this figure are: Short Message Service (SMS), enhanced Mobile Broadband (eMBB) and Ultra-Reliable and Low-Latency Communications (URLLC).

In a traditional mobile network, there are Base Stations (BSs) at fixed locations that provide UEs with wireless access to telephony services and the Internet. A UE is any device that offers wireless connectivity to an end-user, such as the cellular modem of a smartphone, car or autonomous robot, for example. BSs are usually located on masts or rooftops and are equipped with high-powered transceivers to provide the UEs in its vicinity access to the aforementioned services. The area where a particular BS has the strongest signal for the UEs in that area is called a cell. Using many BSs, a given geographical area is divided into disjoint cells. This is why mobile networks are often called cellular networks. With this foundation established, attention can now turn to the ambitious performance targets envisioned for 6G.

In 2023, the International Telecommunication Union Radiocommunication Sector (ITU-R), an agency of the United Nations tasked with the international coordination of radio spectrum usage and telecom standards, published its vision on 6G titled ‘IMT-2030’ in Recommendation M.2160-0 [4]. Proposed requirements include:

- Peak data rate: maximum achievable data rates under ideal conditions of 50-200 Gb/s.
- Connection density: between 10^6 and 10^8 devices connected per km^2.
- UE experienced data rate: minimum UE experienced data rates of 300-500 Mb/s available ubiquitously across the coverage area.
- Sustainability: the energy efficiency should be “improved appropriately with the capacity increase in order to minimize overall power consumption” [4]. No specific bits/Joule target is stated.

As the process of 6G standardization is commencing, an important question is what technologies are needed to satisfy these requirements.

To meet the first requirement it is paramount to improve the achievable per-cell Spectral Efficiency (SE), which represents the number of bits of information that can be transmitted using the same time-frequency resources and is expressed in bit/s/Hz. The SE is an increasing function of the Signal-to-Interference-plus-Noise Ratio (SINR) of the link between a BS and a UE and is therefore limited by signal attenuation and interference from other transmissions utilizing the same time-frequency resources. One major approach taken in 4G and 5G is the use of Multiple-Input Multiple-Output (MIMO) technologies, where multiple antennas are used to send focused beams towards served UEs. This improves the SINR by increasing the received power and decreasing the interference from other transmissions. MIMO also allows for spatial multiplexing, where multiple UEs can be served simultaneously using the same time-frequency resources by creating multiple beams. For 6G, this concept can be exploited further going from massive MIMO (mMIMO) already common in 5G, with typical deployments of 32 to 64 antennas per antenna array, to ultra-massive MIMO (umMIMO) which could use hundreds or even thousands of antennas [5].

Solely increasing the number of antennas is likely not enough to also meet the second requirement. Traditionally, traffic growth has been handled by reducing the cell size by deploying more BSs and by allocating more spectrum. With regard to the former, deploying more BSs in the same area is also known as network densification. The area spectral efficiency, defined as the sum UE throughput normalized by the area and bandwidth, can be improved by deploying more BSs, as the signal strength is independent of the cell size. This is under the assumption that the S and I component of the SINR scale equally, as largely defined by their respective path loss exponents. Unfortunately, this does not appear to be the case as different path loss exponents apply at different distances. Even with a simple two-slope path loss model, one can already see that Ultra-Dense Networks (UDNs) can become interference-limited and therefore unable to handle more traffic [6]. As for allocating more spectrum, there are currently two main frequency ranges in use for mobile networks: Frequency Range 1 (FR1) [< 7 GHz] and Frequency Range 2 (FR2) [> 24 GHz]. FR1 has favorable propagation characteristics but is already so widely used there is little bandwidth left to allocate: a so-called ‘spectrum crunch’ [7]. FR2, more commonly branded as millimeter wave (mmWave) in 5G, still has a lot of spectrum available, but has the disadvantage of poor propagation characteristics. Signals using FR2 propagate almost exclusively via Line-of-Sight (LoS) paths and experience significant atmospheric attenuation. This leaves the section of the electromagnetic spectrum falling in between FR1 and FR2 at 7-24 GHz named Frequency Range 3 (FR3). It is being investigated for use in 6G as it has relatively favorable channel fading characteristics while also providing large frequency bands [8]. At the World Radio Conference 2023 (WRC-23), organized by the ITU-R, the 7.125–8.4 GHz and 14.8–15.35 GHz frequency ranges were identified for use in 6G and will be studied for WRC-27 [9]. In the US, the Federal Communications Commission (FCC) is also considering 12.7-13.25 GHz for 6G [10].

However, none of these options sufficiently address the third requirement. A considerable drawback of the current cellular design of mobile networks is the variability in service quality between UEs that are close to a BS and those at the cell edge. Not only does the signal strength decay rapidly with distance, but UEs at the edge of a cell also experience strong interference from neighboring BSs [11]. This leads to a significant difference in the experienced SINR, and therefore data rate, for UEs close to a BS versus those at the cell edge. In order to provide a certain minimum data rate at all locations in a cell, it would be desirable to suppress the interference at the cell edge to be able to provide more spatially uniform service.

A possible solution is combining the aforementioned technologies, mMIMO and UDNs, with Coordinated Multi-Point (CoMP) processing introduced in 4G. This concept, known as CF-mMIMO [12], abandons the idea of cells entirely. Instead, UEs are served by multiple APs coherently. These APs are 'stripped-down' versions of complete BSs, with fewer antennas and less transmission power. In the most extreme scenario, these APs only need to receive and transmit radio signals. Via a fronthaul network, all APs are connected to one or more CPUs, which are responsible for most of the actual signal processing. In this paradigm, each UE has its own cluster of APs that it is served by, which may or may not overlap with the clusters of APs of other UEs. Given that APs are cheaper and smaller than traditional BSs, and in line with the concept of UDNs, it is envisioned that many more will be deployed in a given area. A possible deployment scenario is to attach them to lampposts. This idea is borrowed from the traditional small cell context for which lamppost deployment is already marked as a scenario of interest by the 3rd Generation Partnership Project (3GPP), the most influential mobile networks standards organization, for 5G networks [13].

A potential drawback of deploying many APs, an associated fronthaul network and CPUs is that it conceivably conflicts with the fourth requirement: sustainability. Such a CF-mMIMO network with many densely-deployed APs could consume significant amounts of energy to meet the stringent 6G QoS demands. Fortunately, the upside is that the cell-free topology also provides a high degree of flexibility with regard to turning off APs when they are not needed. Their coverage area is relatively small and is designed to overlap with neighboring APs, meaning that during times of low traffic loads APs can be put to sleep with relatively little impact. How to decide which APs should serve which UEs and which APs can be put to sleep is the topic of this thesis.

1.1. Research questions

The objective of this thesis is to develop a heuristic SMM algorithm that saves energy by putting APs to sleep that are not needed to maintain coverage and a given QoS level. The algorithm should be applicable to a network where APs have a light and a deep sleep mode, where the latter has an associated non-zero transition time to go into and out of that state. It should rely only on practically available information, such as long-term statistical CSI, previous channel estimates and previously achieved data rates. Additionally, the algorithm must have a low computational complexity as it must be possible to create an implementation that runs in real-time. These objectives give rise to the following main research question:

- **How effective is a low-complexity heuristic SMM algorithm, making decisions based on practically available information, at reducing the total energy consumption in an urban user-centric CF-mMIMO network with APs characterized by a light and a deep sleep mode and a realistic non-zero transition time for the deep sleep mode, while preserving coverage and targeting a minimum QoS level?**

However, before this main research question can be answered, a determination needs to be made with regard to the best deployment scenario for APs in a CF-mMIMO network. Given that the total number of antennas is fixed, is it better to have many APs with a lower antenna count or fewer APs with a higher antenna count? This leads to the following research question:

- **Which AP deployment strategy yields better QoS and energy efficiency in a CF-mMIMO network: deploying a larger number of APs with fewer antennas each, or a smaller number of APs with more antennas each?**

1.2. Main contributions

The primary contribution of this thesis is the design of a heuristic SMM algorithm that advances the current state of the art by integrating the following key features:

1. Inclusion of multiple AP states: active, light sleep and deep sleep. With two sleep levels, each with their own transition time, there are more opportunities to conserve energy.
2. Dynamic determination of the number and set of APs that should enter a sleep state.
3. Use of a transition time between the different AP states.
4. Preservation of population coverage during operation.
5. Consideration of UE QoS through a defined 10th percentile throughput target.
6. Reliance solely on readily available information, such as CSI and previously achieved data rates.
7. Incorporation of user scheduling within the system framework.
8. Low computational complexity allowing for practical implementation.

Some subsets of these features have already been considered in the literature, but to the author's knowledge there has not been any work that considered all eight simultaneously. The current state of the art with regard to SMM algorithms for Cell-Free Networks (CFNs), and specifically CF-mMIMO networks, is detailed in Section 2.5. The shortcomings in the literature that these key features were inspired by are described in Section 2.6.

1.3. Outline

The remainder of this thesis is divided into five chapters. A thorough overview of the literature on CFNs with a focus on SMM algorithms is given in Chapter 2. Additionally, this chapter covers the relevant background knowledge and identifies the research gap. Chapter 3 specifies the modeling choices adopted in this thesis. The proposed heuristic SMM algorithm is detailed in Chapter 4. The simulation scenarios and results are discussed in Chapter 5. Chapter 6 provides an overview and presents the key conclusions, followed by recommendations for future work.

1.4. Notation

The following notation is adopted in this thesis. Any scalar value is represented using a regular typeface, such as l or L. Calligraphic letters, such as $\mathcal{A}$, denote a set. The notation $P_k(\mathcal{A})$ denotes the k-th percentile over the values in the set $\mathcal{A}$. Diacritics are used to indicate averages and estimates of the variables they appear on: averages are denoted with an overline ($^{-}$) and estimates with a caret ($\hat{}$). $\mathbb{C}$ and $\mathbb{R}$ are used to denote the set of complex numbers and the set of real numbers, respectively. $\mathbb{N}(\mu, \sigma)$ is used for a real-valued Gaussian distribution with a mean μ and variance σ.

Vectors are represented by lowercase letters in boldface such as $\mathbf{h}$. Bold uppercase letters are used for matrices such as the channel response matrix $\mathbf{H}$. $\mathbf{H}[i, j]$ denotes the element in the i^{th} row and j^{th} column of $\mathbf{H}$. Matrices with subscripts such as $\mathbf{H}_{a,u}$ indicate that the matrix is specific to an AP a and UE u. Matrices varying with time t and frequency f are indicated with $\mathbf{H}(t, f)$. The superscripts ^{T}, ^{H}, $^{-1}$, and †, denote the transpose, conjugate transpose, inverse, and Moore-Penrose pseudo-inverse, respectively.

2
Background

This chapter provides an introduction to, and an overview of the literature on, CFNs and more specifically a CFN-implementation called CF-mMIMO. Section 2.1 first introduces CFNs, providing historical context and detailing common modeling assumptions from the literature. Section 2.2 describes the process of UE-AP association and explains its necessity with regard to scalability and sustainability. Section 2.3 describes the different approaches for pilot assignment, which are often performed in tandem with UE-AP association. A concise overview of user scheduling algorithms is provided in Section 2.4. Shifting focus to the main topic of this thesis, namely energy savings, Section 2.5 provides a detailed literature review on the sleep mode management methodologies applicable to CF-mMIMO systems. Finally, the chapter concludes with a description of the research gap in Section 2.6.

2.1. Cell-free networks

In a CFN, A geographically distributed APs are jointly serving the U UEs located in a given area. The APs are connected to one or multiple CPUs via a fronthaul network. The CPUs are responsible for coordinating the APs and are themselves connected to the core network via backhaul links. An example of a CFN topology is shown in Figure 2.1.

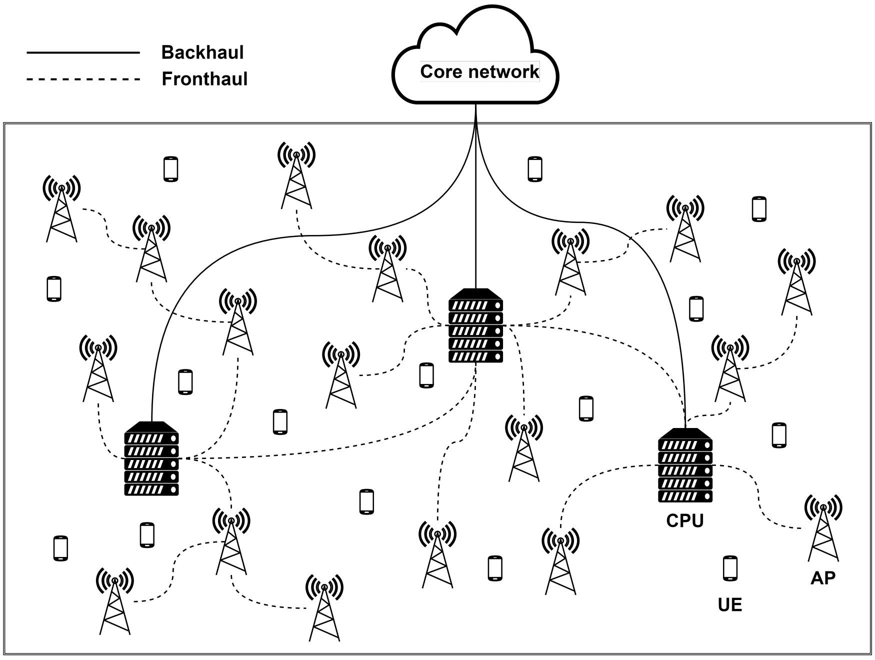


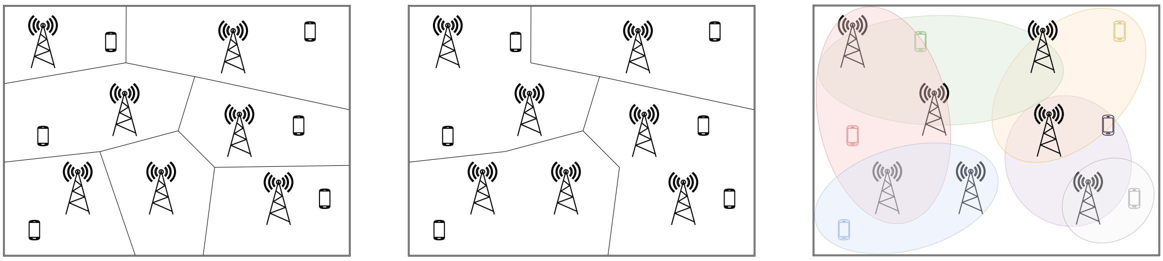
Figure 2.1: Illustration of a CFN.

It is important to note that the CPU(s) should be regarded as logical units which are not necessarily monolithic entities stationed at separate geographical locations, but can, for example, be made up of multiple local processors co-located at (some) APs. Additionally, the fronthaul might have a different topology than in the above depiction. The only requirement is that a fronthaul should exist and is able to perform the following four main tasks [11]:

1. Enabling phase synchronization between the APs.
2. Distributing the signal to be transmitted by the APs in the downlink.
3. Forwarding the received uplink signals to the CPU for processing.
4. Sharing CSI between the APs.

The term ‘cell-free’, first introduced by Yang and Marzetta [14] in 2013, indicates that there are no cell boundaries from the perspective of a UE. The origin of the CFN concept goes back much further than 2013, however. Wyner [15] published a paper on the benefits of joint signal processing in the uplink in 1994. An analysis of joint downlink transmission was first published by Shamai and Zaidel [16] in 2001. The first complete CFN architecture was the Distributed Wireless Communication System (DWCS) proposed by Zhou *et al.* [17] in 2003. In this system, UEs are served by a set of APs close to them. DWCS is an early example of a user-centric CFN, which is explained in the paragraph below. The first appearance of cooperative transmission in a practical system was the soft handover procedure used in UMTS (3G) [18] where UEs are temporarily jointly served by multiple BSs during the handover from one cell to another.

The concept of multi-cell cooperation was further developed for use in LTE-Advanced (4G) systems under the name Coordinated Multi-Point (CoMP) [19], an optional mode where UEs are served jointly by multiple BSs. 4G is not commonly considered a cell-free system, however, given that the use of CoMP is optional and the fact that CoMP is based on network-centric clustering. In network-centric clustering, APs are divided into disjoint sets that serve disjoint subsets of UEs, decided by the network. As a result, there are still cell boundaries, just with bigger cells. This approach can be contrasted with user-centric clustering, where the clusters are formed from a UE perspective and cluster overlap is allowed. The difference is visualized in Figure 2.2.



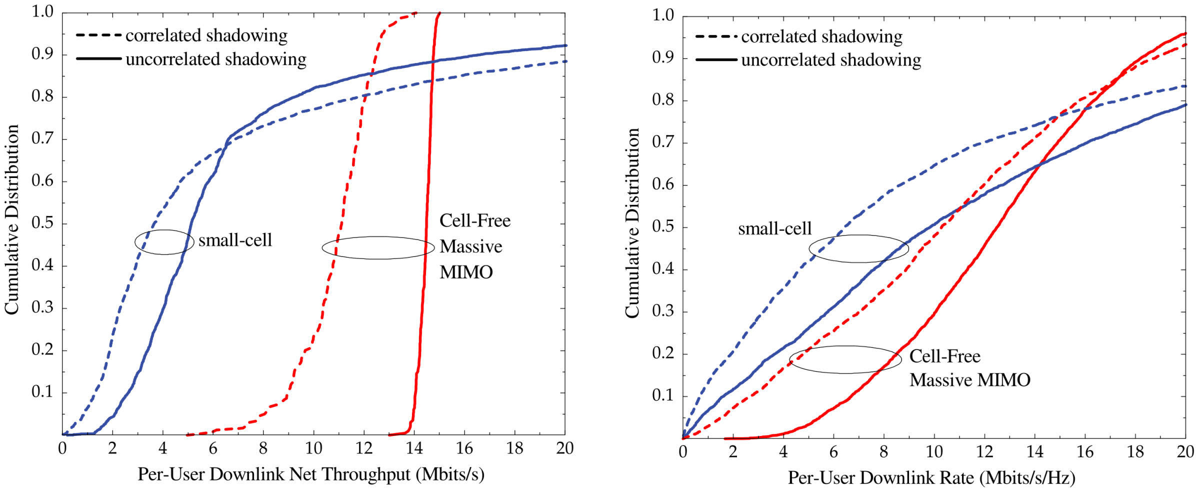
(a) Traditional cellular network. **(b)** Network-centric clustering. **(c)** User-centric clustering.

Figure 2.2: Illustration of different clustering types.

A true implementation of a CFN is the CF-mMIMO concept introduced in a conference paper by Ngo *et al.* in June 2015 [12]. The novelty with respect to earlier incarnations is the incorporation of massive MIMO where multiple antennas jointly serve either a single (SU-MIMO) or multiple (MU-MIMO) UEs using the same time-frequency resources. In CF-mMIMO, these multiple antennas are geographically distributed instead of being co-located. In Ngo *et al.*’s original formulation, all single-antenna APs cooperate phase-coherently to serve all UEs in the same time-frequency resource using Time-Division Duplexing (TDD) operation. In their vision, a CF-mMIMO system is also an UDN, i.e. a network with many more APs than simultaneously active UEs.

Ngo *et al.* show that the 95%-likely per-UE throughput of CF-mMIMO is significantly higher than a small-cell system. Importantly, the distribution of throughputs is also much more concentrated around the median, showing the potential of uniformly good service for most UEs. The same group of authors also published a more extensive journal article evaluating the performance of CF-mMIMO [20]. Figure

2.3, taken from this article, shows the cumulative distribution of the per-UE downlink throughput for a CF-mMIMO system versus a small-cell system.



(a) With max-min power control.

(b) With equal power sharing.

Figure 2.3: Cumulative distribution of the per-UE downlink throughput for correlated and uncorrelated shadow fading [20]. Setup: 100 APs and 40 UEs, 1x1 km^2 area, Greedy Pilot Assignment, Rayleigh fading.

In the CF-mMIMO literature [21], [22], TDD operation is often used in conjunction with the block fading model to characterize channel coherence in time and frequency. In this model, the available resources in time and frequency are divided up into blocks. Within each block, defined by a coherence time T^{c} in seconds and coherence bandwidth B^{c} in Hertz, the channel is assumed constant. In reality, wireless channels cannot be perfectly characterized using blocks. However, with an appropriate choice of T^{c} and B^{c} the block fading model can be a good approximation. “According to the Nyquist-Shannon sampling theorem, a signal that fits into this block is uniquely described by $\tau^{\mathrm{c}} = T^{\mathrm{c}}B^{\mathrm{c}}$ complex-valued samples” [21]. These samples, also called symbols, represent information in a wireless communication system. A coherence block’s τ^{c} symbols are divided up into three parts in the time domain: τ^{p} symbols are reserved for uplink pilot transmission, τ^{u} symbols for uplink data transmission and τ^{d} symbols for downlink data transmission. Naturally, $\tau^{\mathrm{c}} = \tau^{\mathrm{p}} + \tau^{\mathrm{u}} + \tau^{\mathrm{d}}$. The uplink pilot signals are transmitted to enable channel estimation. Downlink pilots are not needed as the use of TDD provides channel reciprocity. A visual representation of a coherence block is shown in Figure 2.4.

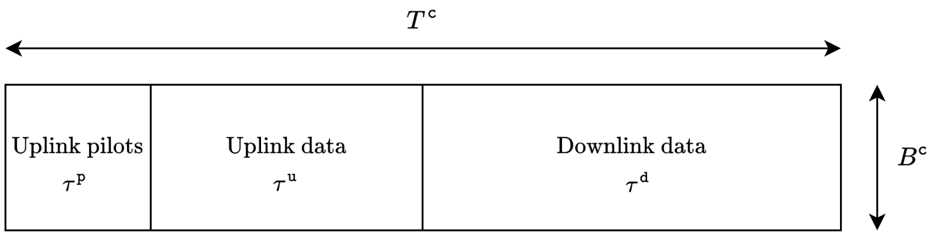


Figure 2.4: Structure of a coherence block.

2.2. UE-AP association

An important aspect of a CF-mMIMO system is the method used to determine which UEs will be served by which APs. Early papers did not address this question yet, requiring all APs to serve all UEs. However, this method is not scalable as the number of UEs grows large. According to Björnson and Sanguinetti [23], a CFN is scalable if and only if the following four tasks have a finite complexity for a given AP as the number of UEs tends to infinity:

1. Signal processing for channel estimation: computing channel estimates.

2. Signal processing for data reception and transmission: computing precoders and combiners.
3. Fronthaul signaling for data and CSI sharing: transmitting data to and from the CPU.
4. Power control optimization.

The model where all APs serve all UEs is not scalable with respect to any of these four tasks. The first proposal for a scalable CF-mMIMO system was made by Buzzi and D'Andrea [24] who introduced the concept of a user-centric CF-mMIMO network, where each UE is only served by a limited number of APs in its vicinity. This collection of APs is called a UE's cluster. The term clustering is commonly used to refer to the AP-UE association process. In their algorithm, UEs first transmit their pilot signals in order to let APs estimate the channels. Each AP then selects only the N UEs with the best channels as their clients for that coherence block, where N is a design parameter.

Scalability is not the only motivation, however, to consider using clustering. Ngo *et al.* [25] introduced the largest-large-scale-fading-based selection algorithm motivated by energy efficiency. For a given UE, most APs will be located far away and therefore do not meaningfully contribute to a UE's received signal, in essence wasting time-frequency resources. In the largest-large-scale-fading-based selection algorithm, each UE u is only associated with $|\mathcal{A}_u| < A$ APs that satisfy:

$$\sum_{a=1}^{|\mathcal{A}_u|} \frac{\bar{\beta}_{a,u}}{\sum_{a'=1}^{A} \beta_{a',u}} \geq \delta\% \tag{2.1}$$

where A is the total number of APs, $\mathcal{A}_u$ is the set of APs that will be serving u, $\beta_{a,u}$ is the Large-Scale Fading Coefficient (LSFC) between AP a and UE u, δ is a configurable parameter, and $\{\bar{\beta}_{a,u}\}$ is the sorted version of the set $\{\beta_{a,u}\}$ in descending order. In other words, each UE is only served by a subset of APs that together contribute at least a fraction of δ of the total average channel gain.

An alternative method that also uses the LSFCs is described by Braam *et al.* [26]. Their clustering strategy is to determine for each UE which AP it receives with the highest signal strength. Its cluster will then consist of that AP and all APs it receives with a signal strength that is at most some fraction less than the strongest AP. This method was also employed by Ito *et al.* [27].

2.3. Pilot assignment

In each coherence block, UEs transmit pilot signals to enable channel estimation. Ideally every UE should be assigned a pilot orthogonal to all other UE pilots to allow APs to disambiguate the pilots transmitted by different UEs. However, since there are only τ^{p} symbols designated for pilots, there are only τ^{p} orthogonal pilots available [21]. Reusing orthogonal pilots or using non-orthogonal pilots causes the channel estimates of different UEs to contaminate each other, a phenomenon aptly named pilot contamination. This interference not only increases the estimation error, but also causes channel estimates to become correlated whereas the actual channels are statistically independent [28]. Fortunately, the effects of pilot contamination can be mitigated by using a set of orthogonal pilots and only reusing pilots for UEs that are sufficiently distant from each other. This is possible because the pilots of distant UEs will not be received strongly due to path loss. The process of determining which pilot should be used by which UEs is called pilot assignment.

The simplest method, which often made an appearance in early works on CF-mMIMO, is Random Pilot Assignment (RPA) introduced by Ngo *et al.* [12], [20]. In RPA, each UE gets assigned a random pilot from a pool of orthogonal pilots. Its simplicity is unfortunately offset by mediocre performance as some neighboring UEs will use the same pilot resulting in a high degree of pilot contamination. An upgrade to RPA is the Greedy Pilot Assignment (GPA) algorithm also proposed by Ngo *et al.* [12], [20]. This algorithm starts with a random pilot assignment and then iteratively updates the pilot of the UE with the lowest rate. Zhang *et al.* [29] propose a variant on GPA named Location-Based GPA (LBGPA) where the random assignment initialization step is altered such that each UE has a certain radius within which another UE is not assigned the same pilot. To avoid getting stuck in a local optimum, a frequent problem for greedy approaches, several authors have explored the use of metaheuristic search methods commonly used for optimization problems such as Tabu-Search [30] and the Hungarian Algorithm [31]. A comparison of the performance of these pilot assignment methods is shown in Figure 2.5.

Motivated by the move towards user-centric CFNs, more advanced pilot assignment algorithms were designed with the clustering process in mind. In a user-centric architecture, the number of APs that need to clearly 'hear' a UE's pilot is limited to only those APs in a UE's cluster. This means pilot contamination is only an issue if an AP in the cluster of some UE also receives the same pilot signal from another UE with a high SINR. A useful insight in this regard is that UEs without cluster overlap are likely to be far away enough from each other to limit the effects of pilot contamination. The first pilot assignment algorithm that used this approach is Dynamic Pilot Reuse (DPR) introduced by Sabbagh *et al.* [32]. In DPR, pairs of UEs without any overlap in their cluster are allowed to use the same pilot. To maximize the so-called minimum reuse distance, defined as the minimum distance between two UEs sharing the same pilot, UEs are assigned a pilot by starting with the maximally distant UE pairs first. When there are less UEs left than pilots, each remaining UE can then be assigned a unique pilot.

The outcome of the clustering process is also explicitly used by Liu *et al.* who devised a pilot assignment method based on graph coloring [33]. In their approach, a graph is created where nodes represent UEs and edges are drawn between any two UE nodes that share an AP. A standard graph coloring algorithm such as DSatur [34] can then be used to assign colors, i.e. pilots, such that the UE nodes that share an edge, i.e. have cluster overlap, are not assigned the same color. Its performance in relation to RPA, GPA and some metaheuristic-based strategies is visualized in Figure 2.5.

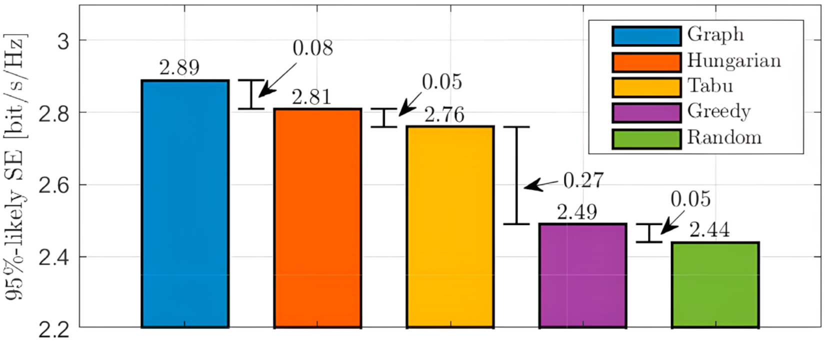


Figure 2.5: Comparison of the spectral efficiency for different pilot assignment strategies [22].

An important insight is that there exists a relationship between the processes of clustering and pilot assignment. Each AP should only serve at most τ^{p} UEs, otherwise some UEs would need to share the same pilot. This would cause strong pilot contamination. The clustering methods introduced by Buzzi and D'Andrea [24], Ngo *et al.* [20] and Ito *et al.* [27], do not limit the number of clients per AP, causing pilot contamination. Given this relation between clustering and pilot assignment, there have been a number of papers on algorithms that perform both tasks jointly. Björnson and Sanguinetti [23], for example, propose Dynamic Cooperation Clustering (DCC) as part of their scalable CFN framework. In DCC, each UE u first appoints a 'Master AP': its strongest received AP based on the LSFCs. Next, u's Master AP assigns u the pilot that currently has the least interference and asks neighboring APs if they can join in serving u. These neighboring APs will do so if they receive u's pilot signal with a higher gain than the UE that used that pilot thus far. A downside of DCC is that it can still happen that an AP must serve more than τ^{p} UEs as APs cannot refuse to be the Master AP of a UE. Björnson and Sanguinetti remark that if this, according to them, 'unlikely' situation occurs, multiple UEs can be assigned the same pilot if they are multiplexed in time or frequency.

Chen *et al.* [35] propose a clustering algorithm that *does* ensure that no AP serves more than τ^{p} UEs. The algorithm is still based on LSFCs, but includes an extra mechanism for the situation where a UE attempts to add a given AP a to their cluster that already serves τ^{p} UEs. Such an event triggers a competition in which a finds the UE with the weakest LSFC. This weakest UE puts a on its blacklist and continues trying to add other APs to its cluster. When a UE has $A-1$ APs on its blacklist, meaning it has lost all competitions so far, the last remaining AP must serve it. This mechanism prevents UEs from being abandoned, i.e. having no serving AP at all. Strictly speaking, this is not a joint algorithm, as no specific pilot assignment is included in this algorithm. It does ensure, however, that the strongest form of pilot contamination is not inevitable by only allowing τ^{p} UEs per AP. Also note that this method only ensures the clients of each individual AP can be assigned a unique pilot, but does not itself ensure that pilots are not reused between two physically proximal UEs that have non-overlapping clusters.

To complement their clustering strategy, Chen *et al.* describe three compatible pilot assignment methods: Geography-Based K-Means (GB-KM), Interference-Based K-Means (IB-KM) and User-Group pilot assignment [35]. All methods aim to maximize the minimum reuse distance to minimize pilot contamination. GB-KM, first introduced by Attarifar *et al.* [36], partitions the U users into τ^{p} disjoint subsets, with each subset containing U/τ^{p} UEs whose minimum distance is a large as possible. This is achieved by iteratively performing k-means clustering and designating the UEs closest to the centroids of these clusters as a pilot-sharing subset. IB-KM is a variation on GB-KM that uses an alternate distance metric for the k-means clustering process based on the level of interference that UEs would cause each other. The User-Group pilot assignment method is not based on k-means clustering, but instead aims to assign orthogonal pilots to UEs served by similar subsets of APs using the LSFC matrix. A comparison of the performance of DCC, GB-KM, IB-KM, User-Group and RPA is shown in Figure 2.6.

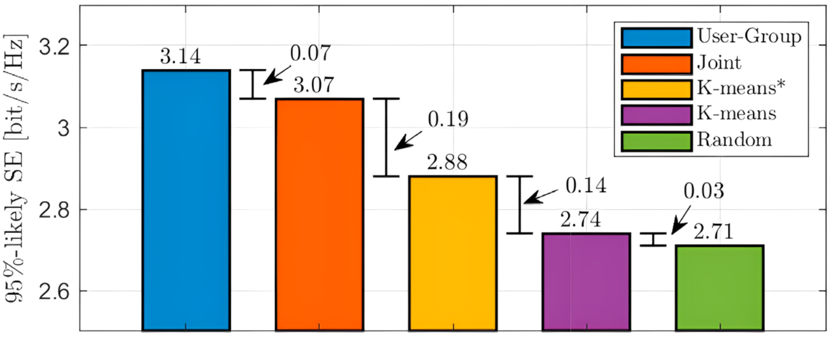


Figure 2.6: Comparison of the spectral efficiency for different pilot assignment strategies [22]. In this figure, 'Joint' refers to DCC, 'K-means*' to IB-KM and 'K-means' to GB-KM.

2.4. User scheduling

There are relatively few papers that consider user scheduling algorithms for CF-mMIMO systems. As noted in Section 2.1, early works on CF-mMIMO considered single-antenna APs ($M^{\mathtt{AP}} = 1$) and considered the network to be ultra-dense, i.e. $U < A$. As later works shifted the focus to multi-antenna APs ($M^{\mathtt{AP}} > 1$), this assumption was adapted as $A < U < A \cdot M^{\mathtt{AP}}$. In addition to assumptions on deployment density, nearly all works on CF-mMIMO consider all UEs to be active simultaneously and assess their performance in terms of ergodic rates, assuming no packet errors. Göttsch *et al.* [37] poses that these are not realistic assumptions. They argue the regime where $A < U$ is much more relevant when taking operator deployment cost into consideration, meaning that UEs need to be scheduled over the available time-frequency resources. With dynamic scheduling, they contend that the per-UE throughput rate is a much better performance metric given that ergodic rates are most likely not achievable given the discontinuous nature of packet scheduling.

In 2023, Mashdour *et al.* [38] investigated the use of user scheduling strategies to optimize the sum-rate performance in a network-centric clustering scenario with single-antenna APs. To this end, they extend the Zero-Forcing with Selection (ZFS) greedy scheduling algorithm first designed for traditional cellular systems. Zero Forcing (ZF) creates orthogonal, non-interfering channels between APs and UEs by inverting the channel response matrix H: the precoding matrix W can be found as $W = H^H(HH^H)^{-1}$. For a given AP-UE pair, ZF creates a beam that has nulls in the direction of other UEs in order to minimize interference. However, when $A < U$, HH^H becomes singular and cannot be inverted. Therefore, a selection algorithm is needed to decide which $U^* \leq A$ UEs to schedule. When selecting UEs to schedule, it is important to consider the co-orthogonality of their channels. The less orthogonal the channels are, the more beamforming gain needs to be sacrificed in order to create nulls in the direction of co-scheduled UEs.

As previously noted, Göttsch *et al.* [37] considered the regime where $A < U$. In their paper, they focus on optimizing for user fairness instead of maximizing the sum-rate performance. They formulate a Network Utility Maximization (NUM) optimization problem, which they solve using a Lyapunov Drift-Plus-Penalty (DPP) approach. Two fairness criteria are considered: Proportional Fairness (PF) and Hard Fairness (HF). Mashdour *et al.* [39] also studied fairness scheduling, but considered it jointly with clustering.

Shin *et al.* [40] combines a Lyapunov approach with the Semi-orthogonal User Selection (SUS) scheduling algorithm. The SUS algorithm considers UEs in order, based on their PF index, and selects those with favorable and semi-orthogonal channels. A more thorough explanation of the SUS algorithm is provided in Section 3.4.4.

2.5. Sleep mode management

There are several approaches to reduce the energy consumption of a mobile network. One possibility is to reduce the transmission power used by the APs in the network if the full power is not necessary to meet the performance requirements. Taking it one step further, one can also completely shut off an AP if it is not necessary to support the current number of active UEs. A variation on this is the concept of sleep modes, such as light or deep sleep, where one or more components of an AP are disabled to conserve energy, but can be remotely enabled again when necessary. An overview of the current literature and associated shortcomings can be found in Table 2.1.

Interest in the optimization of CF-mMIMO networks for sustainability appeared quickly after its introduction, with the first analysis on the energy efficiency of CF-mMIMO systems published by Ngo *et al.* in 2018 [25]. They derive a closed-form expression for the spectral efficiency of a CF-mMIMO network using conjugate beamforming under the assumption of Rayleigh fading. Using this expression, they formulate a non-convex optimization problem that optimally allocates the power coefficients such that the total energy efficiency is maximized under the constraint of a per-UE spectral efficiency requirement. To obtain a computationally feasible formulation, the problem is then approximated using a series of Second-Order Cone Programs (SOCPs) that can be solved by convex problem solvers such as MOSEK [57].

Van Chien *et al.* [41], [42] expanded on the above approach by not only considering the transmit power, but also allowing an AP to be turned off if it is not needed to satisfy performance demands. They start by formulating the total power minimization problem as a non-convex optimization problem and convert it to a mixed-integer SOCP that can by solved by convex problem solvers in similar fashion to Ngo *et al.*'s formulation. However, the authors note that this convex formulation is still too computationally complex for real-time applications. They develop an approximate, lower-complexity time solution by re-expressing the SOCP formulation as a sparse reconstruction problem and obtain a tractable problem using ℓ_p-norm relaxation combined with an iterative least squares approach. Additionally, a polynomial-time algorithm is developed determining the transmit powers and which APs should be on separately. This algorithm has a computational complexity of $\mathcal{O}(A^{3.5}U^{3.5})$, where A is the number of APs and U is the number of UEs, which is still prohibitive when real-time computation is required.

Femenias *et al.* [43] critique the practicality of Van Chien *et al.*'s method, asserting that the large-scale system parameters vary too quickly to be able to run the high-complexity selection algorithm in real-time. Instead, they focus on a bigger time scale, explaining that cellular networks are usually dimensioned to provide a minimum QoS during rush hour, leaving the network significantly underused during less busy periods. They propose several heuristic methods to dynamically turn on or off some APs based on the traffic load which they term AP Switch On/Off (ASO) strategies. However, they do not directly incorporate a mechanism to measure or use the current traffic load in their strategies. Rather, the strategies focus on *which* APs to turn off given that the number of APs to turn off has already been decided. The underlying assumption of these algorithms is that the spatial distribution of UEs is uniform over the coverage area. The following six strategies are introduced:

1. Random Selection ASO (RS-ASO): straightforward algorithm where each AP is equally likely to be turned off. Used to provide a lower bound on the energy efficiency that the other strategies can achieve.
2. Mixture Discrepancy-based greedy ASO (MD-ASO): iterative greedy algorithm based on the mixture discrepancy metric which quantifies the deviation between the actual AP distribution and a

perfectly uniform distribution. The idea is that under the assumption of a uniform UE distribution, the AP distribution should also be as uniform as possible. In each iteration, one AP is switched off that results in the highest decrease in the mixture discrepancy metric.

3. Spatial Regularity-based greedy ASO (SR-ASO): iterative greedy algorithm based on the metric of spatial regularity which measures the deviation of AP locations from a triangular lattice. Similarly to MD-ASO, in each iteration one AP is switched off: the one producing the highest increase in the spatial regularity metric.
4. Nearest Neighbor-based ASO (NN-ASO): iterative algorithm maximizing the minimum pairwise distance among the set of active APs. “At each step, out of the two APs that are nearer to each other, the one that is nearer to a third one is put in sleep mode” [43].
5. Propagation Losses-aware ASO (PL-ASO): iterative algorithm using variations in the large-scale propagation losses to turn off APs with poor channel qualities to UEs. Unlike the algorithms above, it is therefore not geometry based, but instead it dynamically utilizes shorter-term variations in the propagation environment. Starting from an empty set, APs that should be on are added iteratively as those with the minimum propagation losses to each of the UEs. In cases where the number of UEs is greater than the number of APs, k-means clustering is used to group the UEs first.
6. Optimal energy efficiency-based Greedy ASO (OG-ASO): higher-complexity iterative greedy algorithm used to derive an upper bound on the energy efficiency that the other strategies can achieve. Starting with all APs on, in each iteration the AP that increases the energy efficiency the most is turned off. This procedure is repeated until the energy efficiency cannot be improved further. This method assumes that one can know the UE throughput for a given set of active APs a priori, i.e. before making the actual decision which APs will be active.

The authors choose SR-ASO as the most suitable to be implemented in a practical CF-mMIMO network, citing its superior performance versus complexity trade-off. The PL-ASO algorithm has even better performance, but is not practical given the speed at which APs would be switched on or off. A graph of the average downlink energy and spectral efficiency for the different strategies is shown in Figure 2.7.

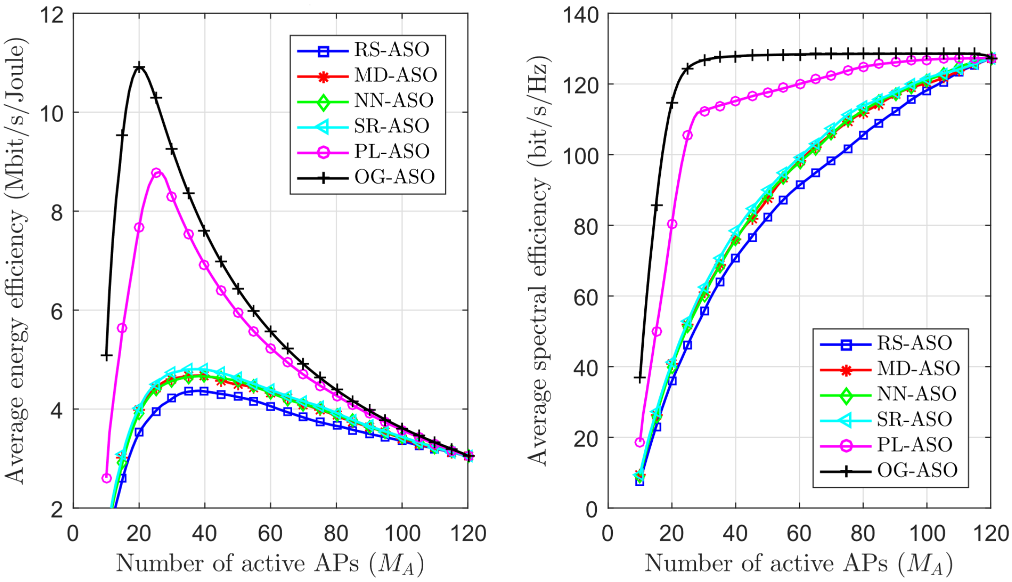


Figure 2.7: Comparison of the downlink energy and spectral efficiency of several heuristic SMM algorithms presented by Femenias *et al.* [43].

A similar group of authors, García-Morales *et al.* [44], expanded on these results by adding ASO strategies for non-uniform UE distributions in the context of a mmWave CF-mMIMO network. They present three ASO strategies based on a Goodness-of-Fit (GoF) approach where the spatial distribution of active APs is matched as closely as possible to the distribution of UEs. The three strategies are:

1. Chi-Square-based ASO (ChiS-ASO): an iterative greedy algorithm derived from the Chi-square

test.

2. Kolmogorov-Smirnov-based ASO (KS-ASO): an iterative greedy algorithm derived from the Kolmogorov-Smirnov test.
3. Logarithmic Statistical Energy-based ASO (LSE-ASO): an iterative greedy algorithm derived from the concept of statistical energy introduced by Aslan and Zech [58].

As in [43], the PL-ASO algorithm has the best showing of all implementable algorithms, but does need information on the large-scale fading conditions and is deemed impractical by the authors due to the sheer frequency of transitions. Of the novel GoF-approaches, LSE-ASO has the best performance. A graph of the average downlink energy and spectral efficiency for the different strategies is shown in Figure 2.8.

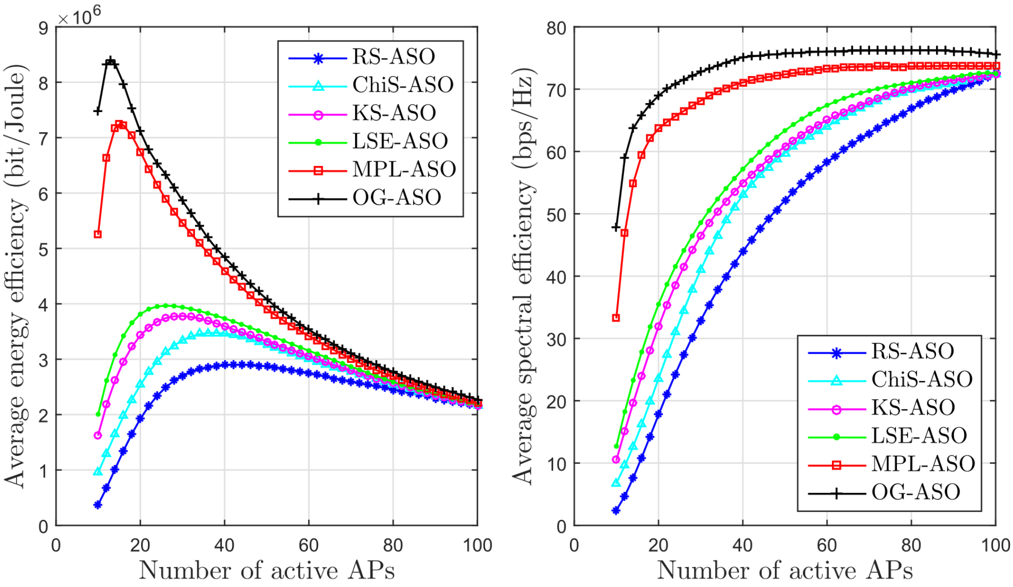


Figure 2.8: Comparison of the downlink energy and spectral efficiency of several heuristic SMM algorithms for networks with non-uniform UE distributions presented by Femenias *et al.* [43] and García-Morales *et al.* [44]. Note: in this graph, the PL-ASO strategy is named MPL-ASO.

In [45], Riera-Palou *et al.* extend their LSE-ASO algorithm for use in a user-centric CF-mMIMO network where UEs are served only by a subset of APs. In this new two time-scale approach termed MinCon ('Minimum Connectivity'), the LSE-ASO algorithm is used to adapt to the very long-term (VLT) effects, i.e. changes in the spatial distribution of UEs, while a new short-term (ST) adaptation component decides which APs will be serving which UEs based on the large-scale fading gains. Aiming to balance channel hardening with scalability, the ST component creates clusters such that each UE is connected to some minimum number of APs, but also that each AP only serves some maximum number of UEs. Starting from the UE-AP link with the maximum gain, AP a is included in UE u's cluster if a has not reached its maximum number of served UEs and the gain is significant with respect to the APs already serving u. This process continues until all APs are 'full' and all UEs have acquired their minimum number of serving APs or there are no more links left to consider.

In 2021, Jung and Hong presented a variant of PL-ASO called Max effective channel Gain-based ASO (MG-ASO) [46], where APs are selected by not merely considering the path loss, but all fading effects, i.e. the effective channel gain. They model a canonical CFN where all APs jointly serve all UEs. Similarly to the algorithms described by Femenias *et al.*, the number of APs that should be active is an input parameter to the algorithm. MG-ASO iteratively selects the APs with the largest sum effective channel gains to all UEs. Additionally, Jung and Hong also present IMportance-based ASO (IM-ASO) [46], where APs are iteratively selected based on their importance. The importance of an AP is defined as the effective channel gain of a UE-AP link divided by how many APs a UE receives stronger than some threshold. They show that both algorithms have similar performance, both slightly outperforming PL-ASO in terms of energy and spectral efficiency. Another variant of PL-ASO called RF-beamformed

Large-scale Propagation Gain ASO (RFb-LPG-ASO) was introduced by Hong and Na [47]. In this algorithm, the beamforming gain, based on centralized ZF beamforming, is subtracted from the path loss.

Ito *et al.* propose a Greedy Combining Algorithm (GCA) that simultaneously considers clustering and ASO [27]. Their clustering strategy is to determine for each UE which APs it receives strongest. Its cluster will then consist of that AP and all APs it receives with a signal strength that is at most some fraction less than the strongest AP. The ASO component of the algorithm starts by having all APs active. Next, one AP is chosen and excluded from the clustering process. The overall improvement is then quantified by taking a (tunable) weighted average of the improvement in spectral efficiency and energy efficiency. The AP with which the highest improvement is reached is then put to sleep. This iterative process continues until there is no AP with a positive overall improvement value. The authors show that their algorithm performs better than MinCon [45], while getting rid of the design parameter that determines how many APs need to be active. However, this method does assume that one can derive the spectral and energy efficiency for a particular configuration of APs a priori.

In 2023, Riera-Palou *et al.* [48] presented a greedy approach that, unlike their earlier heuristic algorithms [43]–[45], does not assume that the number of APs to turn off (or equivalently: keep on) is an input to the SMM algorithm. After obtaining an initial clustering, the algorithm iteratively puts APs to sleep whose LSFCs are the worst with respect to the overall set of UEs. The algorithm terminates when the energy efficiency no longer improves or the required UE throughput can no longer be satisfied. Again, this method assumes that one can derive UE bit rates and the total network energy efficiency a priori.

The approaches considered so far assume that switching APs on and off can be done at will and as often as desired. He *et al.* [49] argue that there are additional costs to these state transitions which should be considered. They classify two types of switching:

1. Node switching: turning on or off APs. Frequent on/off switching may cause thermal fatigue and shorten the lifespan of an AP's components.
2. Link switching: altering a UE's cluster. If a UE's cluster is changed frequently, there will be an increase in signaling overhead.

To limit the number of node and link switching, an SOCP optimization problem is formulated using an approach similar to Van Chien *et al.* [42]. Using two tunable weight parameters, α and β, one can control which switching type(s) should be limited and by how much.

Zheng *et al.* [50] further expand on the optimization problem-based approach by considering clustering, (hybrid) beamforming and ASO jointly. Due to the non-convex, intractable nature of the initial formulation, the authors use a combination of convex relaxing, the Dinkelbach method and exploiting the group sparsity structure to arrive at a tractable formulation for which an iterative algorithm is derived. Additionally, the authors also derive a lower-complexity algorithm based on an alternative optimization strategy which divides the problem into two subproblems that are alternately optimized until a convergence criterion is met. The first subproblem is concerned with clustering and beam selection, whereas the second subproblem derives the digital beamformers and performs ASO. They conclude that the lower-complexity algorithm closely matches the performance of the high-complexity algorithm, with both of them outperforming the state of the art.

A drawback of the methods that rely on the availability of large-scale propagation gains or CSI between all UEs and APs, such as the optimization problem-based approaches mentioned so far, is that no AP can ever go into a deep sleep state given that they must either measure or send out pilot signals periodically to facilitate channel characterization. Beerten *et al.* [51] presents a polynomial complexity algorithm based on [42] where sleeping APs are only activated if it is otherwise infeasible to satisfy a UE's SINR requirements, leaving some APs asleep. The proposed polynomial-complexity algorithm assumes that each UE's location is known and is divided into three phases: Initial Access, Feasible Set Search and Active APs Pruning. In the first phase, every UE connects to its closest APs, which is woken up if it was sleeping. In the second phase, an attempt is made to solve the optimization problem with the APs currently awake. If it is not feasible, additional APs are woken up one-by-one, chosen based on the value of their associated slack variables. In the third phase, considering APs in ascending order of their (assigned) transmit power, it is tested if the problem is still feasible without a given AP. If so, that AP is put to sleep. The proposed method approaches the performance of the SOCP-based solution from [42] while significantly less costly in terms of computational complexity.

Another heuristic approach, the (3 x E) algorithm, is presented by Kooshki *et al.* [52]. An important part of this algorithm is based on the concept of the Interference Contribution Ratio (ICR), which is an AP metric defined as the ratio of the strength of transmitted signals to all UEs not served by an AP (i.e. the interference caused) to the strength of transmitted signals to all UEs served by an AP (i.e. the desired signal). A high value means that an AP causes a lot of interference, degrading the performance and wasting energy. Putting such an AP to sleep would improve UE QoS and network energy efficiency. After an initial selection similar to [51], there is a two-phase sleep mode selection algorithm comprised of the 'certain' and 'conditional' phase. In the certain phase, APs with a high ICR are definitively turned off. In the conditional phase, only APs that improve the energy efficiency are allowed to remain active. With a time complexity of $\mathcal{O}(A \cdot U)$, where A is the number of APs and U the number of UEs, and an energy efficiency improvement of up to 60% over no sleep algorithm, it is an attractive heuristic approach. Unfortunately, the (3 x E) algorithm unrealistically assumes that the network bit rate for a particular AP configuration can be calculated a priori.

Over the past decade the popularity of Artificial Intelligence (AI) and Machine Learning (ML) has exploded as a result of the advancements in the field of deep learning. In line with this development, 3GPP and ITU-R are investigating the uses of AI for 6G [4], [59]. Of particular interest is the possibility of self-optimization. In the CFNs literature, there has also been some interest in the use of AI/ML for energy savings specifically. Mendoza *et al.* [53] propose a framework based on Deep Reinforcement Learning (DRL), specifically a centralized Double Deep Q-Network (DDQN). This framework uses the location of the UEs to determine which APs should be active. After this determination, the derivation of the power allocation is performed using a classical optimization problem-based approach. The DRL gets feedback using a reward function that incorporates the achieved SINRs and energy usage, with a tunable weight to focus either on better QoS or more energy savings. This approach is similar to the weights used in [27]. Several authors published follow-up papers using different DRL-based approaches such as: Deep Deterministic Policy Gradient (DDPG) [54], Multi-Agent DDQN (MA-DDQN) [55] and Proximal Policy Optimization (PPO) [56]. A comparison of the performance of these approaches is shown in Figure 2.9.

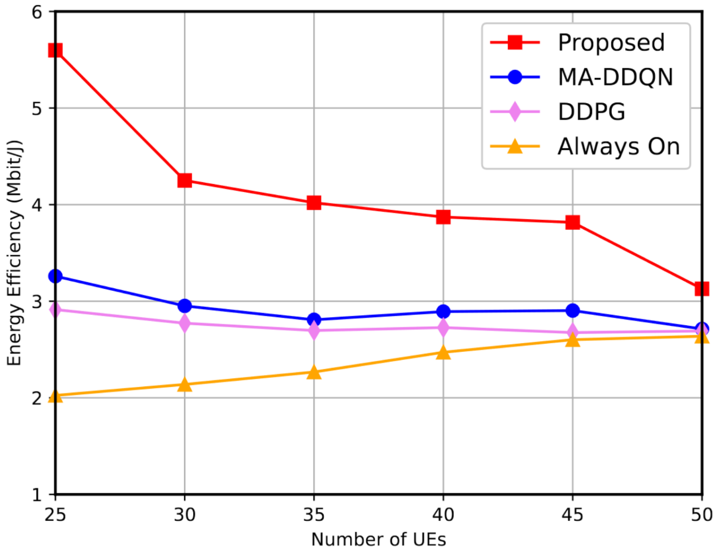


Figure 2.9: Comparison of the energy efficiency for various deep learning-based SMM strategies [56]. Note: The PPO approach is labeled 'Proposed' in this graph. 'Always On' refers to the scenario where no ASO algorithm is used.

2.6. Research gap

Eight shortcomings of current SMM algorithms for CF-mMIMO networks were identified during the literature review:

1. The use of only two states: APs are either on or off. With more states, each with their own transition times, there would be more flexibility to conserve energy on different time scales.
2. Assuming the number of APs to turn off is known. In reality, this information is not known and should be dynamically determined by the SMM strategy.
3. Using instantaneous transitions between states. When an AP is transitioning to a deep(er) sleep state, it will need some time to properly power down hardware components. Similarly, when an AP is woken up from a sleep state, it will take some amount of time to reactivate the hardware needed to serve UEs. This transition time will impact how fast the network can adapt to traffic variations and thus the performance of a given SMM algorithm.
4. Not accounting for coverage. If all APs in an area are shut down, coverage gaps will develop causing the network to be unavailable to serve UEs in that area.
5. No consideration of UE QoS. This impacts the tradeoff between performance and energy consumption. Without a target for a minimum throughput, for example, the best strategy would be to turn (almost) all APs off. This would be not acceptable from a user perspective, of course.
6. The use of information before it is available. Some papers assume they can calculate the SINR, throughput and/or energy efficiency of UEs for a particular AP configuration before actually turning on or off any APs to measure the impact. As pilots have not been transmitted yet, there is no information on the channels to make such calculations. This assumption is even more unlikely for sleeping APs, as it does not even have long-term statical CSI on UEs in its vicinity.
7. No user scheduling. When each UE must be served in every time slot, while in practice this is not necessary, this limits the number of UEs in the network that can be served.
8. High computational complexity. Any practical SMM strategy must be able to be executed in real-time for a decently-sized network. Algorithms reliant upon solving (multiple) SOCPs will definitely not be scalable, but even some of the heuristic approaches with polynomial runtime complexities still have big-O exponents higher than three, making them unsuitable for even moderately-sized networks.

An overview of SMM papers and the specific shortcomings that apply to each is shown in Table 2.1. Each shortcoming corresponds with a contribution of this work as described in Section 1.2.

Table 2.1: Overview of literature on SMM strategies. The numbers in the shortcomings column pertain to the identified research gaps and are detailed in Section 2.6.

Paper	Frequency (GHz)	Bandwidth (MHz)	Propagation model	Simulation area (km^2)	#APs	#UEs	Energy consumption model	Shortcomings
[25]	1.9	20	NLoS; three-slope path loss + shadow fading; Rayleigh fading	1 - 4	20-400	2 - 40	APs and fronthaul	1, 3, 4, 6, 7, 8
[41], [42]	2.0	?	NLoS; 3GPP UMi model; uncorrelated Rayleigh fading	1	20 - 50	20 - 40	APs and fronthaul	1, 3, 4, 6, 7, 8
[43]	2.0	20	LoS/NLoS; 3GPP UMi model; Rician fading	1	10-160	16-32	APs, UEs and fronthaul	1, 2, 3, 4, 5, 7
[44]	28.0	20	LoS/NLoS; 3GPP UMi model; Rician fading	0.25	10 - 100	8 - 24	APs, UEs and fronthaul	1, 2, 3, 4, 5, 7
[45]	?	20	LoS/NLoS; three-slope path loss + shadow fading; Rician fading	1	4 - 100	32	APs and fronthaul	1, 2, 3, 4, 5, 7
[46]	30.0	80	LoS/NLoS; path loss + shadow fading; Saleh-Valenzuela model	0.04	10 - 60	30	APs, fronthaul and CPU	1, 2, 3, 4, 5, 7
[47]	28.0	20	LoS/NLoS; 3GPP UMi model; Rician fading	0.04	20 - 100	30	APs, UEs and fronthaul	1, 2, 3, 4, 5, 7
[27]	3.5	20	LoS/NLoS; 3GPP TR 25.996; Rician fading	1	8 - 64	8 - 16	APs, UEs and fronthaul	1, 2, 3, 4, 6, 7
[48]	2.0	100	NLoS; path loss + shadow fading (3GPP TR 36.873); correlated Rayleigh fading	0.73	1 - 7	42 - 168	APs, fronthaul and CPU	1, 3, 4, 6, 7
[49]	?	20	NLoS; path loss + shadow fading (3GPP TS 36.814); correlated Rayleigh fading	0.16	? - 16	Based on traffic model	APs	1, 3, 4, 6, 7, 8
[50]	28.0	?	LoS/NLoS; path loss + shadow fading (“mmWave channel model”)	0.008	? - 16	4	APs and fronthaul	1, 3, 4, 6, 7, 8
[51]	?	20	NLoS; path loss + shadow fading; correlated Rayleigh fading	0.25	? - 15	7	APs	1, 3, 4, 6, 7, 8
[52]	4.0	20	LoS/NLoS; path loss + shadow fading (ITU-R Indoor Hotspot-eMBB)	2.92	? - 150	150 - 400	APs	1, 3, 4, 6, 7
[53]	2.0	?	LoS; path loss + shadow fading; uncorrelated Rayleigh fading	0.0036	? - 8	2 - 6	APs	1, 3, 4, 6, 7
[54]	?	?	Uncorrelated Rayleigh fading	1	? - 120	8 - 24	APs and fronthaul	1, 3, 4, 6, 7
[55]	?	?	Uncorrelated Rayleigh fading	0.01	20 - 50	15 - 25	APs	1, 3, 4, 6, 7
[56]	?	20	Uncorrelated Rayleigh fading	1	25 - 85	25 - 50	APs and fronthaul	1, 3, 4, 6, 7

3 Modeling

This chapter describes the modeling choices adopted in this work. It covers key aspects such as the network layout, channel model, UEs and APs characteristics, radio resource management and the energy consumption model.

This thesis considers a dense urban cell-free network of 500 x 500 square meters based on the city center of Amsterdam. A map indicating the location of the simulation area is shown in Figure 3.1.

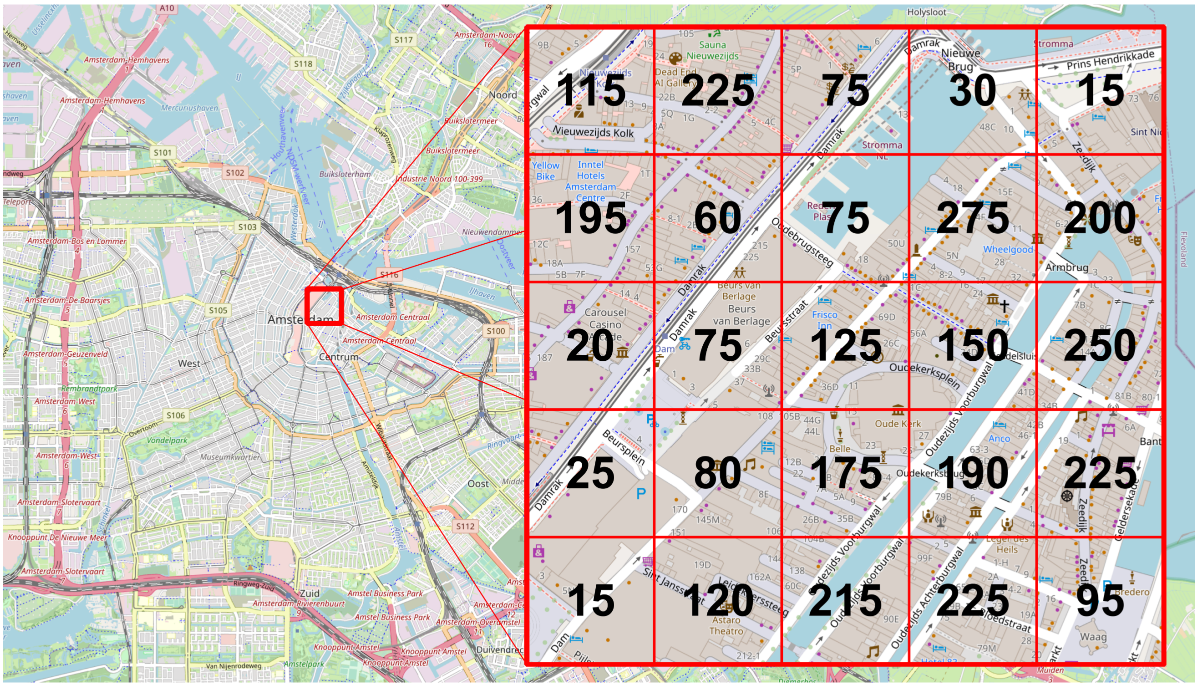


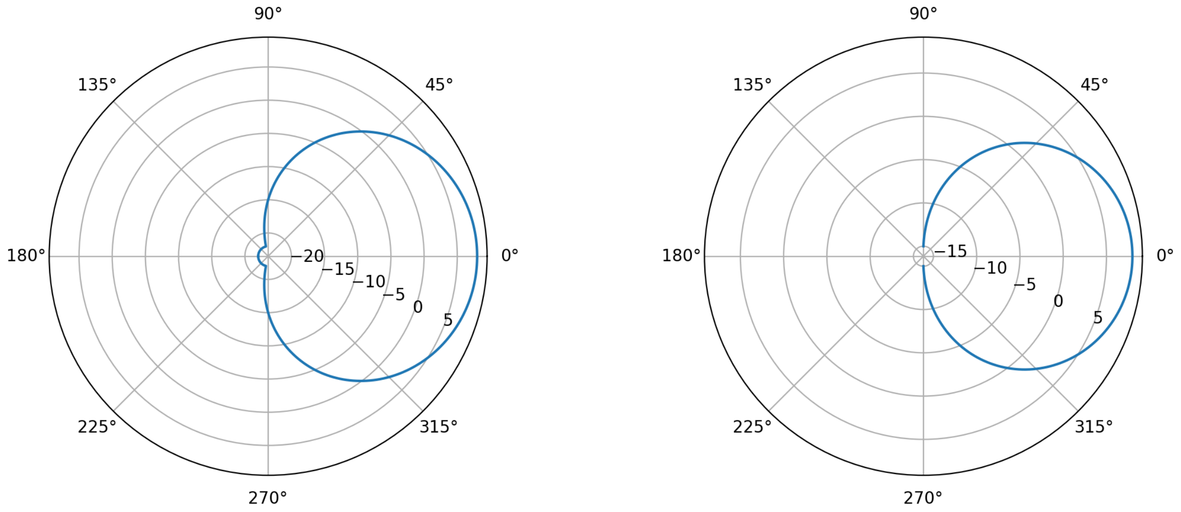
Figure 3.1: Map indicating the location of the simulation area. The magnified version shows the number of inhabitants of each 100 x 100 meter square according to the *Centraal Bureau voor de Statistiek* (CBS), the official government statistics agency of The Netherlands [60].

3.1. Network aspects

The network consists of A APs spread across the simulation area. Denote with $\mathcal{A} = \{0, ..., A-1\}$ the set of all APs. In line with an envisioned 6G deployment scenario of APs on lampposts, all actual locations of lampposts in the city center of Amsterdam are considered as potential locations for APs for the simulation. The height of the APs is set at the approximate height of lampposts in an urban area; $h^{\mathtt{AP}} = 6$ meters [61]. All APs are connected to a single CPU via an infinite-capacity fronthaul.

Each AP is equipped with a single horizontal Uniform Linear Array (ULA) configured as either 2T2R, 4T4R or 8T8R, depending on the scenario. The antenna elements are linearly polarized and there are no sub-arrays. The resulting number of antennas per AP is $M^{\mathtt{AP}} = 2$, $M^{\mathtt{AP}} = 4$ or $M^{\mathtt{AP}} = 8$, respectively. Between antenna elements a spacing of $\lambda^c/2$ is used. The antenna array layout and gain per antenna element is modeled as specified in 3GPP Technical Report (TR) 38.901 [13]. The ULA is always angled towards the geometrical horizon, i.e. the downtilt or elevation angle is 0 degrees. The azimuth angle is chosen randomly for each AP.

The gain per antenna element is dependent on the horizontal and vertical angle between a UE and an AP's ULA. The resulting gain can be calculated using the formulas in Table 7.3-1 of 3GPP TR 38.901 [13], where the maximum gain is 8 dBi. The horizontal and vertical cut of the radiation pattern for a single antenna element is shown in Figure 3.2. The maximum transmit power of an AP is based on the number of antennas. The transmit power per antenna is defined as $P^{\mathtt{AP,ant,tx,max}} = 0.25$ W [25], [42], [62], [63]. This results in a transmit power of 0.5, 1 and 2 W for $M^{\mathtt{AP}} = 2$, $M^{\mathtt{AP}} = 4$ and $M^{\mathtt{AP}} = 8$, respectively. The transmit power is divided equally over all Physical Resource Blocks (PRBs).



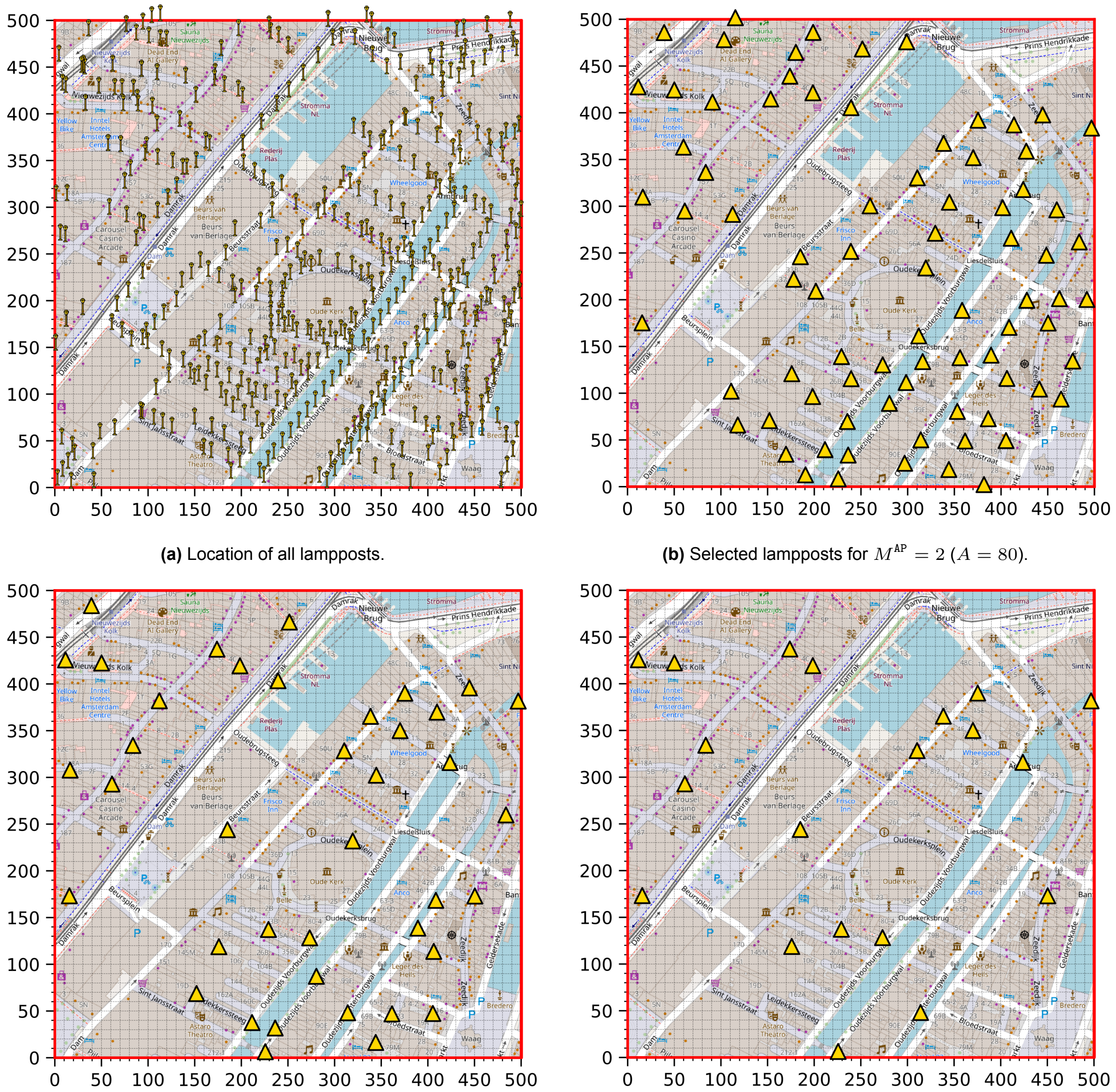
(a) Horizontal cut (elevation = 0°) of the antenna gain [dB].

(b) Vertical cut (azimuth = 0°) of the antenna gain [dB].

Figure 3.2: Horizontal and vertical cuts for an antenna element of an AP.

The location data for the lampposts is available on the city of Amsterdam's open geodata portal [64]. A map showing the locations of the 467 lampposts within the simulation area is depicted in Figure 3.3a. A subset of lampposts is chosen as AP locations such that 98% uplink population coverage is achieved, conforming to the standard of the Dutch mobile network regulator *Rijksinspectie Digitale Infrastructuur* (RDI) [65]. To assess population coverage, the simulation area is divided into 2,500 pixels of 10 x 10 meters. A pixel is considered to be covered if the average channel gain to a test UE placed at the pixel's midpoint is higher than a certain threshold based on a minimum required achievable uplink throughput of 1 Mbit/s. The derivation of this threshold is detailed in Section 3.4.1. The population coverage is then calculated by considering what fraction of pixels is covered, where each pixel is weighted by the CBS population data shown in Figure 3.1.

Using the population coverage as a guiding principle, an AP deployment is derived for each of the three deployment scenarios: $M^{\mathtt{AP}} = 2$, $M^{\mathtt{AP}} = 4$ and $M^{\mathtt{AP}} = 8$. It is important to note that the same total number of antennas must be deployed in each scenario to enable a fair comparison between the scenarios. To achieve this, the two-phase selection procedure is first performed for the $M^{\mathtt{AP}} = 8$ antenna case. Phase one consists of an iterative process where, in each iteration, the AP that increases the population coverage the most is added to the set of selected APs. Phase one ends when 98% population coverage is achieved. In phase two, 50% more APs are chosen to ensure that population coverage can be maintained when putting APs to sleep during low-traffic hours. As areas with a higher population density will need more APs to handle all of the traffic, APs that are closer to more UEs have a higher likelihood to be selected in this phase. For the $M^{\mathtt{AP}} = 2$ and $M^{\mathtt{AP}} = 4$ antenna cases, phase two does not select 50% extra APs on top of those already selected in phase one, but instead selects as many APs as needed such that the total number of antennas is the same as for the $M^{\mathtt{AP}} = 8$ antenna case. The result of the selection procedure is visualized in Figure 3.3.



(a) Location of all lampposts.

(b) Selected lampposts for $M^{\mathtt{AP}} = 2$ $(A = 80)$.

(c) Selected lampposts for $M^{\mathtt{AP}} = 4$ $(A = 40)$.

(d) Selected lampposts for $M^{\mathtt{AP}} = 8$ $(A = 20)$.

Figure 3.3: Overview of lampposts selected as AP deployment locations for the different AP configurations.

All APs share the same 100 MHz TDD carrier at $f^{\mathrm{c}} = 7.8$ GHz, with an associated wavelength $\lambda^{\mathrm{c}} \approx 3.84$ cm. This frequency falls into FR3 [7-24 GHz] which has been under exploration for use in 6G networks. This is right in between the existing FR1 [< 7 GHz] and FR2 [> 24 GHz]. As the use of FR3 has not been standardized yet, current 5G standards for FR1 and FR2 are used as a basis to select the numerology and associated parameters. The carrier is configured with numerology 2, meaning a Subcarrier Spacing (SCS) of 60 kHz. This numerology was chosen as it is the only one available both in FR1 and FR2 [66]. Numerology 2 has an associated Transmission Time Interval (TTI) duration of 0.25 milliseconds. The bandwidth $B = 100$ MHz is subdivided into $P = 134$ PRBs with an appropriately sized guardband at each end. For FR1 and FR2, there are 135 and 132 PRBs available at 100 MHz, respectively [66]–[68]. As 7.8 GHz is closer to FR1, 134 was chosen as an interpolation of 135 and 132. As is standard in 5G, each PRB consists of 12 subcarriers. With 134 PRB spaced at 60 kHz, there is some space left of the 100 MHz for guardband on either end.

To prevent boundary effects, but limit the computational complexity of the simulation, an infinite network is emulated by equipping the simulation area with toroidal wrap-around. This entails placing shifted versions of the simulation area around the actual simulation area. For the case of a square simulation area, this results in a total of nine versions of the network with the actual network in the middle surrounded by eight replicas. When determining propagation gains the best value is chosen from the nine options. For UEs at the edges of the simulation area, this extension of the network ensures that they can attain a serving cluster of comparable size to the UEs in the middle. This method is inspired by Recommendation M.2101-0 from the ITU-R [69]. Figure 3.4 visualizes the wrap-around effect for a simplified network with one UE and one AP. During the simulations, each UE can use one of the nine different versions of a given AP. This will be whichever projection provides the best average channel gain.

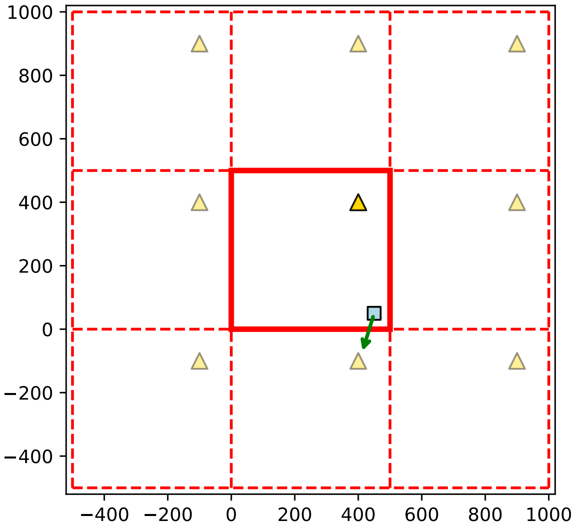


Figure 3.4: Visualization of the wrap-around effect for a simplified network containing one AP and one UE. The location of the UE is indicated by the square and the location of the AP and its virtual copies are indicates by the triangles. The green arrow indicates the AP projection that is closest to the UE in terms of euclidian distance.

3.2. User/traffic characteristics

Unfortunately, there is no public information available on the location of mobile phone users within the city center of Amsterdam. Therefore, population data is used as an approximation of the spatial distribution of UEs over the simulation area. To determine the location of a UE u, the simulation area is subdivided into 25 blocks of 100 x 100 meters. First, one of the 25 blocks is chosen in accordance with the population density. The probability that a specific block is chosen for u is equal to the fraction of people living within that block, given the total population of the simulation area. The second step is to choose the x and y coordinate offsets within the chosen block by drawing from a continuous uniform distribution. The simulation area and population data are shown overlaid on a map of Amsterdam in Figure 3.1.

In accordance with the CBS population data, there will be a total of $U = 3,250$ UEs in the simulation area. Let $\mathcal{U} = \{0, ..., U-1\}$ denote the set of all UEs. All UEs are assumed to be indoor and locally mobile at a pedestrian speed of 3 km/h, denoted $v \approx 0.83$ m/s. UE height is set at $h^{\text{UE}} = 1.5$ meters. The UE maximum transmit power is set at $P^{\texttt{UE,tx,max}} \approx 0.2$ W $(23$ dBm$)$, based on UE Power class 3 (Handheld UE) for 5G FR1 as defined in Table 6.2.1-1 from 3GPP Technical Specification (TS) 38.101-1 [67].

Each UE is configured with 1T4R antennas, meaning a total of four antennas where only one antenna is capable of transmission at any given time. As this thesis focuses on evaluating the downlink performance, the number of antennas is denoted $M^{\text{UE}} = 4$. However, one should keep in mind that the UEs are transmitting their uplink pilots using only one antenna. Each UE alternates on a fine timescale which antenna is used for pilot transmission, in order to allow the UEs to learn the full channel response matrix. The cross-polarized antennas are placed on the upper left and lower right corners of the UE.

The antenna gain is assumed to be 0 dBi.

UEs have two different states: they can be either active or inactive. Active UEs are assumed to be engaged in a full-buffer downlink data transfer. UEs transition between these two states based on two exponential distributions, defined by their scale parameters $\beta^{\texttt{active}}$ and $\beta^{\texttt{inactive}}$. At the start of a simulation, each UE will be initialized as either active or inactive. Throughout the simulation UEs will go through successive periods of inactivity and activity, where an active UE transitions to the inactive state after an active period with a duration sampled from an exponential distribution $Exp(\beta^{\texttt{active}})$. The same applies for the opposite transition, but with a duration sampled from an exponential distribution $Exp(\beta^{\texttt{inactive}})$. Note that in this notation the exponential distribution is defined by its scale parameter β instead of the rate parameter λ. β is the inverse of the rate λ ($\beta = 1/\lambda$) and is also the mean of the distribution. For fixed values of $\beta^{\texttt{active}}$ and $\beta^{\texttt{inactive}}$, these on/off transitions constitute a stationary process and as such the expected number of active UEs at any given time $\mathbb{E}[U^{\texttt{active}}]$ is wholly determined by the scale parameters:

$$\mathbb{E}[U^{\texttt{active}}] = U \cdot \frac{\beta^{\texttt{active}}}{\beta^{\texttt{active}} + \beta^{\texttt{inactive}}} \tag{3.1}$$

To simulate different traffic loads, the value of $\beta^{\texttt{inactive}}$ can be varied. Higher values of $\beta^{\texttt{inactive}}$ will mean UEs are inactive for longer and therefore fewer UEs will be active at the same time. Equation 3.1 is used to calculate the number of UEs that should start in the active state in order to prevent simulator warm up time: exactly $\mathbb{E}[U^{\texttt{active}}]$ UEs should start as active. If the simulation would be started with any other configuration of active and inactive, the first time blocks would not be representative of the chosen scale parameters as the state equilibrium would not have been reached yet.

The traffic load in a mobile network typically varies greatly over a day. Data from 2023 by Ericsson shows that there is an approximately seven-fold difference between peak and off-peak hours [70]. Figure 3.5 shows the Ericsson traffic curve for a dense urban area in Western Europe such as the city center of Amsterdam. The horizontal axis shows the hour of the day and the vertical axis shows the percentage of the total daily traffic. The graph shows that the busiest hour is between 1 and 2 pm, accounting for a total of 7.35% of the total daily traffic. To approximate this traffic curve, six distinct traffic levels are defined in terms of the expected number of active UEs: $\mathbb{E}[U^{\texttt{active}}] = \{20, 40, 60, 80, 100, 120\}$. First, the Ericsson data is scaled by setting the highest load of 7.35% equal to the highest defined traffic level $\mathbb{E}[U^{\texttt{active}}] = 120$. Each hour is then assigned the closest matching traffic level. The resulting quantized curve is overlaid on the Ericsson data in Figure 3.5. Using this quantized curve, the number of hours spent within each traffic level over the course of one day can be derived. This data is shown in Table 3.1 and will be used during the evaluation process of the proposed SMM algorithm.

Table 3.1: Overview of the traffic levels.

Traffic level	Expected number of active UEs	Hours
0	20	6
1	40	2
2	60	3
3	80	4
4	100	6
5	120	3

3.3. Channel characterization

The channel model adopted in this thesis is predominantly based on the 3GPP Urban Micro (UMi) model as described in 3GPP TR 38.901 [13] combined with the block fading model.

3.3.1. Coherence block

As is common in the CFN literature, the block fading model is utilized to characterize channel coherence in time and frequency. The coherence time T^{c} and coherence bandwidth B^{c} are determined using the

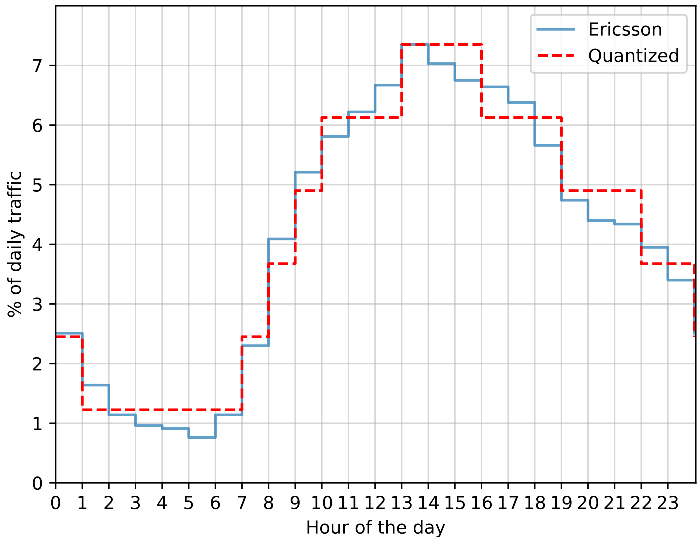


Figure 3.5: Traffic load variations over a typical day. The solid blue line shows the traffic curve for a dense urban area in Western Europe [70]. The red dotted line shows a quantized version of this data using eight traffic levels.

method described by Demir *et al.* [21]. Figure 3.6 illustrates the structure of the model. In practice, the size of a block where the frequency/time-domain fading is approximately flat is different for each UE.

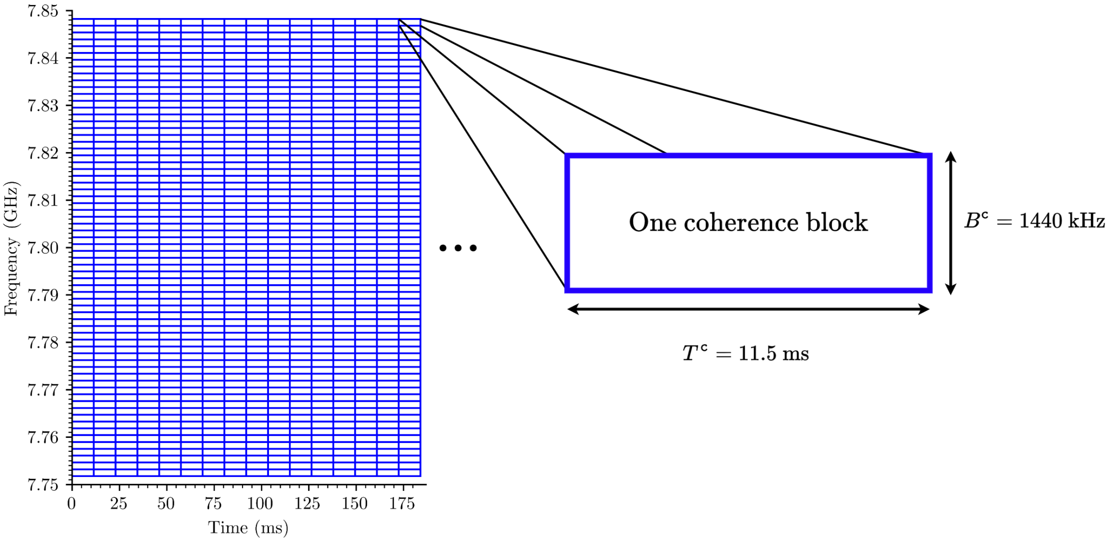


Figure 3.6: Layout of coherence blocks in time and frequency. In this thesis, the coherence bandwidth $B^c = 1440$ kHz and coherence time $T^c = 11.5$ ms. The frequency range runs from 7.75 GHz to 7.85 GHz for a total bandwidth $B = 100$ MHz. There are a total of 67 coherence blocks (134 PRBs) in the frequency domain, with the remaining space used as guardband.

To arrive at a common block size, a worst-case scenario is assumed based on 3GPP-provided simulation parameters. Given a dense urban scenario like the city center of Amsterdam, the 3GPP UMi channel model as described in 3GPP TR 38.901 [13] is deemed most appropriate. In line with the worst-case scenario assumption, the communication link is assumed to be Non-Line-of-Sight (NLoS). To calculate the coherence bandwidth, a value for the delay spread T^{ds} is needed. The delay spread is a measure of the 'multipathness' of a signal. Multipath fading describes the effect where a given transmitted signal can take multiple distinct paths to arrive at the receiver, with multiple copies of the signal arriving at different times. The delay spread reflects the difference between the arrival times of these copies. The coherence bandwidth is inversely proportional to the delay spread, meaning longer delay spreads result in shorter coherence bandwidths. Table 7.5-6 Part-1 of [13] defines a log-normal distribution for the delay spread. The 90th percentile of this distribution is used as the delay spread, which is dubbed the ‘long-delay profile’ in Section 7.7.3 of [13]. This results in a delay spread of approximately 313 nanoseconds. Using the rule of thumb from Demir *et al.* [21], $B^{\mathrm{c}} = 1/2T^{\mathrm{ds}}$, results in a coherence bandwidth $B^{\mathrm{c}} \approx 1596$ kHz.

The coherence time is determined by the UE velocity v and the carrier wavelength λ^c. Using a rule-of-thumb by Tse and Viswanath [71], T^{c} can be calculated as $T^{\mathrm{c}} = \lambda/4v$. Using this method, $T^{\mathrm{c}} \approx 11.53$ milliseconds. To simplify calculations, both figures are rounded to obtain an integer number of PRBs and TTIs in the frequency and time domain, respectively. To err on the side of caution, B^{c} and T^{c} are both rounded down. B^{c} is set at 1440 kHz, meaning 24 subcarriers at an SCS of 60 kHz, i.e. $P = 2$ PRBs. Given a total of 134 PRBs within the 100 MHz bandwidth, this results in 67 coherence blocks in the frequency domain. T^{c} is set at 11.5 ms, resulting in exactly 46 TTIs per coherence block. Using the Nyquist-Shannon theorem, this means that each coherence block consists of $\tau^{\mathrm{c}} = 1440$ kHz $\cdot 11.5$ ms $= 16560$ symbols. The total simulation time is denoted T, with the set of coherence blocks in time denoted as $\mathcal{T} = \{0, ..., T/T^{\mathrm{c}} - 1\}$. An individual block in the time domain is indexed t. The set of coherence blocks in frequency is termed $\mathcal{F} = \{0, ..., P/2 - 1\}$, with individual blocks in the frequency domain indexed as f.

A coherence block’s τ^{c} symbols are divided up into three parts in the time domain: τ^{p} symbols are reserved for uplink pilot transmission, τ^{u} symbols for uplink data transmission and τ^{d} symbols for downlink data transmission. Naturally in this TDD structure, $\tau^{\mathrm{c}} = \tau^{\mathrm{p}} + \tau^{\mathrm{u}} + \tau^{\mathrm{d}}$. The uplink pilot signals are transmitted to enable channel estimation. In the literature, commonly 5% of the coherence block is allocated for pilots. However, this is often in scenarios with higher than pedestrian UE speeds. In this work, 1 TTI is assigned for pilots, meaning $\tau^{\mathrm{c}} = 1440$ kHz $\cdot$ 0.25 ms $=$ 360 symbols. The other 45 TTIs are divided between the uplink and downlink based on the current 5G TDD frame split imposed by the RDI for the 5G 3.5 GHz band. Ignoring the guard time in the special timeslot, this Uplink (UL)-Downlink (DL) split is currently 1:3.25 [72]. Mimicking this ratio as closely as possible, 11 TTIs are assigned to the uplink and 34 TTIs will be used for the downlink. While time is allocated for uplink data transmission, the uplink performance is not evaluated, as this thesis is focused on the downlink. Figure 3.7 visualizes the subdivision within each coherence block.

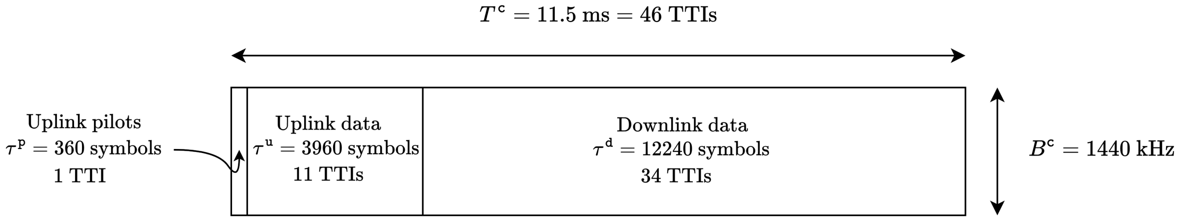

$T^{\mathrm{c}} = 11.5$ ms $= 46$ TTIs
Uplink pilots
$\tau^{\mathrm{p}} = 360$ symbols
1 TTI
Uplink data
$\tau^{\mathrm{u}} = 3960$ symbols
11 TTIs
Downlink data
$\tau^{\mathrm{d}} = 12240$ symbols
34 TTIs
$B^{\mathrm{c}} = 1440$ kHz

Figure 3.7: Subdivision of a coherence block; a TDD protocol.

3.3.2. Channel response matrix

For each coherence block at time t and frequency f, the channel between the antennas of a given AP a and a givenUE u is represented by the channel response matrix $\mathbf{H}_{a,u}(t, f) \in \mathbb{C}^{M^{\mathrm{UE}} \times M^{\mathrm{AP}}}$. This channel response matrix $\mathbf{H}_{a,u}(t, f)$ is determined by two elements: a large-scale and a small-scale fading component. The large-scale component, the average channel gain $G_{a,u} \in \mathbb{R}$, models large-scale effects such as path loss, shadowing, building penetration loss and antenna gain. The average channel gain is often termed the Large-Scale Fading Coefficient (LSFC) in the CFN literature. Due to the assumed local mobility of the UEs, it is constant in time and frequency for the duration of the simulation for each UE-AP link. The small-scale fading component, $\mathbf{H}_{a,u}^{\texttt{small-scale}}(t, f) \in \mathbb{C}^{M^{\mathrm{UE}} \times M^{\mathrm{AP}}}$, models the effects of multipath fading. These effects typically vary with time and frequency. However, adjacent coherence blocks, both in time and frequency, will have some level of correlation. The full channel response matrix $\mathbf{H}_{a,u}(t, f)$ combines the large-scale and small-scale components as follows:

$$\mathbf{H}_{a,u}(t, f) = \sqrt{G_{a,u}} \cdot \mathbf{H}_{a,u}^{\texttt{small-scale}}(t, f) \qquad \forall a \in \mathcal{A}, \forall u \in \mathcal{U}, \forall t \in \mathcal{T}, \forall f \in \mathcal{F} \tag{3.2}$$

The wideband equivalent channel $\bar{\mathbf{H}}_{a,u}(t)$, used by the scheduler, is obtained by averaging over all coherence blocks in the frequency domain:

$$\bar{\mathbf{H}}_{a,u}(t) = \frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}}|\mathbf{H}_{a,u}(t,f)| \cdot \frac{\sum_{f\in\mathcal{F}}\mathbf{H}_{a,u}(t,f)}{|\sum_{f\in\mathcal{F}}\mathbf{H}_{a,u}(t,f)|} \qquad \forall a\in\mathcal{A},\forall u\in\mathcal{U},\forall t\in\mathcal{T} \tag{3.3}$$

In this averaging method, the average of the absolute channel gains over all coherence blocks in frequency is multiplied by the normalized sum over all coherence blocks in frequency to preserve the magnitude.

The components and associated calculations of the average channel gain are detailed in Section 3.3.3. The modeling of the small-scale effects is explained in Section 3.3.4.

3.3.3. Average channel gain

The average channel gain or Large-Scale Fading Coefficient (LSFC) models the large-scale fading effects that are not frequency selective within the considered bandwidth and do not change over the duration of the simulation. It includes the path loss $L^{\mathtt{p}}_{a,u}$, shadow fading $L^{\mathtt{sf}}_{a,u}$, per antenna element gain $G^{\mathtt{ae}}_{a,u}$, body loss $L^{\mathtt{b}}_{a,u}$, and building penetration loss $L^{\mathtt{bp}}_{a,u}$, all expressed in dB. To avoid unrealistically small losses, a minimum coupling loss $L^{\mathtt{mc}}_{a,u}$ of 53 dB is used as prescribed in Section 4.4 of 3GPP TS 38.104 [66]. The average channel gain $G_{a,u}$ in dB is calculated as:

$$G_{a,u}[\text{dB}] = -\max(L^{\mathtt{mc}}_{a,u},\ L^{\mathtt{p}}_{a,u} + L^{\mathtt{sf}}_{a,u} + L^{\mathtt{bp}}_{a,u} + L^{\mathtt{b}}_{a,u} - G^{\mathtt{ae}}_{a,u}) \qquad \forall a\in\mathcal{A},\forall u\in\mathcal{U} \tag{3.4}$$

The average channel gain used to test for coverage $G^{\mathtt{ul}}_{a,u}$ is specific to the uplink as it includes an estimate of the AP combining gain. It is defined as:

$$G^{\mathtt{ul}}_{a,u}[\text{dB}] = -\max(L^{\mathtt{mc}}_{a,u},\ L^{\mathtt{p}}_{a,u} + L^{\mathtt{sf}}_{a,u} + L^{\mathtt{bp}}_{a,u} + L^{\mathtt{b}}_{a,u} - G^{\mathtt{ae}}_{a,u} - 10\log_{10}(M^{\mathtt{AP}})) \qquad \forall a\in\mathcal{A},\forall u\in\mathcal{U} \tag{3.5}$$

Path loss

The calculation of the distance-based path loss $L^{\mathtt{p}}_{a,u}$ is based on the UMi NLoS scenario as defined in Table 7.4.1-1 of 3GPP TR 38.901 [13]. The adopted path loss model differs from the 3GPP standard in three ways. First, the minimum UE-AP distance of 10 meters is ignored given that the usage of a minimum coupling loss already ensures that unrealistically small losses are prevented. Second, the prescribed AP height of 10 meters is not adhered to. However, since the difference with the assumed $h^{\mathtt{AP}} = 6$ meters is small, the effects are assumed to be negligible. Third, the final term for $L^{\mathtt{p,NLoS}}_{a,u}$, $0.3(h^{\mathtt{UE}} - 1.5)$, has been removed as it is redundant given that $h^{\mathtt{UE}} = 1.5$ meters. The modeled path loss is then given by:

$$L^{\mathtt{p,base}}_{a,u} = \max(L^{\mathtt{p,LOS}}_{a,u}, L^{\mathtt{p,NLoS}}_{a,u}) \qquad \forall a\in\mathcal{A},\forall u\in\mathcal{U} \tag{3.6}$$

$$L^{\mathtt{p,LOS}}_{a,u} = \begin{cases} 32.4 + 21\log_{10}(d^{\mathtt{3d}}_{a,u}) + 20\log_{10}(f^{\mathtt{c}}) & \text{if } d^{\mathtt{2d}}_{a,u} \le d^{\mathtt{bp}} \\ 32.4 + 40\log_{10}(d^{\mathtt{3d}}_{a,u}) + 20\log_{10}(f^{\mathtt{c}}) & \text{else} \\ \quad -9.5\log_{10}((d^{\mathtt{bp}})^2 + (h^{\mathtt{AP}} - h^{\mathtt{UE}})^2) & \end{cases} \qquad \forall a\in\mathcal{A},\forall u\in\mathcal{U} \tag{3.7}$$

$$L^{\mathtt{p,NLoS}}_{a,u} = 22.4 + 35.3\log_{10}(d^{\mathtt{3d}}_{a,u}) + 21.3\log_{10}(f^{\mathtt{c}}) \qquad \forall a\in\mathcal{A},\forall u\in\mathcal{U} \tag{3.8}$$

$$d^{\mathtt{bp}} = 4\cdot(h^{\mathtt{AP}} - 1)\cdot(h^{\mathtt{UE}} - 1)\cdot f^{\mathtt{c}}/c \tag{3.9}$$

where $d^{\mathtt{2d}}_{a,u}$ is the 2D UE-AP distance, $d^{\mathtt{3d}}_{a,u}$ is the 3D UE-AP distance, $d^{\mathtt{bp}}$ is the breakpoint distance and c is the speed of light. The breakpoint distance is a model parameter that determines the distance at which the path loss exponents change.

In addition to the base path loss, Table 7.4.3-2 of [13] defines extra path losses for signals going outside to inside, also known as Outdoor-to-Indoor (O2I) losses. Per the 3GPP method, the distance that a signal travels indoor is determined by taking the minimum of two uniformly distributed random variables between 0 and 25 meters. Given that the total UE-AP distance might be smaller than 25 meters, the total distance is used as a limit on the maximum indoor distance.

$$\begin{aligned} d_{a,u}^{\texttt{2d,indoor,1}} &\sim U(0,25) \qquad \forall a \in \mathcal{A}, \forall u \in \mathcal{U} \\ d_{a,u}^{\texttt{2d,indoor,2}} &\sim U(0,25) \qquad \forall a \in \mathcal{A}, \forall u \in \mathcal{U} \end{aligned} \tag{3.10}$$

$$L_{a,u}^{\texttt{p,indoor}} = 0.5 \cdot \min(d_{a,u}^{\texttt{2d}},\ \min(d_{a,u}^{\texttt{2d,indoor,1}},\ d_{a,u}^{\texttt{2d,indoor,2}})) \qquad \forall a \in \mathcal{A}, \forall u \in \mathcal{U} \tag{3.11}$$

The total path loss $L_{a,u}^{\texttt{p}}$ is given by:

$$L_{a,u}^{\texttt{p}} = L_{a,u}^{\texttt{p,base}} + L_{a,u}^{\texttt{p,indoor}} \qquad \forall a \in \mathcal{A}, \forall u \in \mathcal{U} \tag{3.12}$$

Body loss

Body loss is the attenuation of a signal due to the presence of a human in the path between an AP and a UE. While the body loss depends on which part and how much of the body is in a signal's path, an average value of $L_{a,u}^{\texttt{b}} = 6$ dB [73] is chosen.

Shadowing

For each AP a, the Shadow Fading (SF), a type of slow fading, is modeled at each integer $(x,\ y)$-location by a Gaussian random variable $S_a^{\texttt{x,y}}$ with a mean of 0 and a standard deviation of $\sigma^{\texttt{sh}} = 7.82$ dB. This value is in accordance with the UMi NLoS model as defined in Table 7.4.1-1 from 3GPP TR 38.901 [13]. The SF values are assumed to be correlated, both in the spatial domain and between APs. The SF value for a given UE-AP link is dependent on the specific AP and the 2D UE location. Each integer $(x,\ y)$-location in the simulation area has an associated SF value as defined in the shadowing map. The shadowing maps are created by first subdividing the simulation area into pixels. Each pixel is then assigned a value which defines the shadow fading (in dB) for UEs located within that pixel. While each AP has its own unique shadowing map, all A shadowing maps are mutually correlated to model the fact that the signals transmitted to or from APs are influenced by the same physical obstacles.

Shadowing maps that model spatial and inter-AP correlation are generated using the method created by Fraile *et al.* [74]. This method starts with a single randomly generated Gaussian map which is used as a basis for the AP-specific shadowing maps. The inter-AP correlation parameter determines how much of the base map is mixed into the AP-specific maps. Spatial correlation is created by convolving each map with an appropriately sized filter, where the filter size is defined by the decorrelation distance. The convolution process that produces the spatial correlation is slightly altered to ensure that the toroidal topology is taken into account.

The decorrelation distance is set at 13 meters following the recommendation for the UMi NLoS scenario in Table 7.5-6 Part-1 from 3GPP TR 38.901 [13]. The value for the inter-AP correlation, 0.5, is taken from [74], where it is termed site-to-site cross correlation. In summary, the SF loss value for a particular UE-AP link $L_{a,u}^{\texttt{sf}}$ is defined by the value of the pixel that contains UE u's location in AP a's shadowing map. An example of a shadowing map generated using the method and parameters described in this section is shown in Figure 3.8.

Building penetration loss

As all UEs are considered to be indoor and all APs outdoor, all UE-AP links experience building penetration loss. When a signal traverses from outside to inside, the glass and concrete in a building's walls will cause signal degradation. To model this effect, the low-loss model defined in Table 7.4.3-2 from 3GPP TR 38.901 [13] is used. In this model, each wall is a mix of 30% glass and 70% concrete. The building penetration loss for a given UE-AP link is modeled as a Gaussian random variable with mean $\mu = L^{\texttt{bpl,wall}}$ and standard deviation $\sigma = 4.4$:

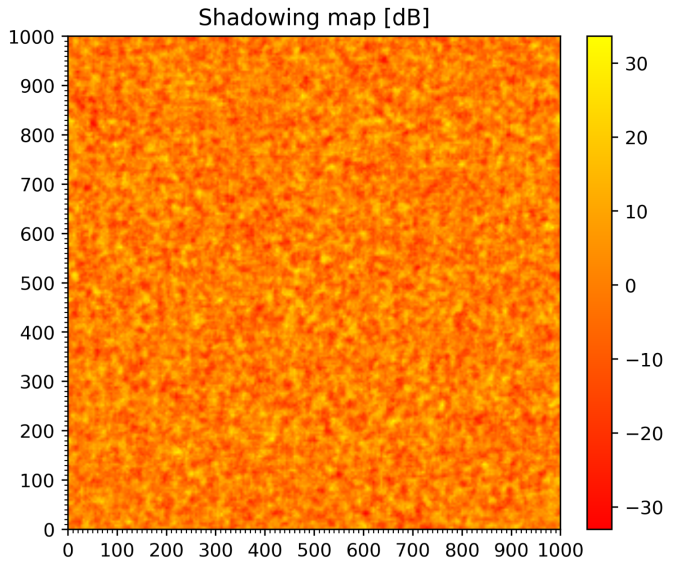


Figure 3.8: Example of an AP's shadowing map.

$$
\begin{aligned}
L^{\texttt{bpl,glass}} &= 2 + 0.2 \cdot f^{\texttt{c,GHz}} \\
L^{\texttt{bpl,concrete}} &= 5 + 4 \cdot f^{\texttt{c,GHz}} \\
L^{\texttt{bpl,wall}} &= 5 - 10 \log_{10}(0.3 \cdot 10^{-\frac{L^{\texttt{bpl,glass}}}{10}} + 0.7 \cdot 10^{-\frac{L^{\texttt{bpl,concrete}}}{10}})
\end{aligned}
\tag{3.13}
$$

$$
L^{\texttt{bp}}_{a,u} \sim \mathbb{N}(L^{\texttt{bpl,wall}}, 4.4) \qquad \forall a \in \mathcal{A}, \forall u \in \mathcal{U}
\tag{3.14}
$$

where $f^{\texttt{c,GHz}} = 7.8$, i.e. the carrier frequency $f^{\texttt{c}}$ in GHz.

3.3.4. Small-scale fading

The small-scale fading variable $\mathbf{H}^{\texttt{small-scale}}_{a,u}(t, f)$ models the effects of multipath fading and should be generated such that adjacent coherence blocks in time and frequency exhibit a degree of coherence. The UMi NLoS model from 3GPP TR 38.901 [13], which underpins the modeling choices in this thesis, features both of these aspects. A set of channel impulse responses over the entire frequency band for a particular duration and a particular UE-AP link is called a channel trace. QuaDRiGa [75], a tool developed by Fraunhofer HHI, implements the UMi NLoS model and is used to generate these traces.

Using QuaDRiGa, a set of channel traces is generated for use in the simulator. Recall that UEs are locally mobile at a speed of $v \approx 0.83$ m/s. This local mobility is implemented as UEs moving along a circular track with a radius of 1 meter. Ideally, each UE-AP link would be modeled by its own channel trace. However, storing these traces would require excessive amounts of memory. For example, the channel traces between 3,250 UEs and 100 APs, consisting of 67 coherence blocks in frequency and 870 coherence blocks in time ($\sim$ 10 seconds), would require $870 \cdot 67 \cdot 3{,}250 \cdot 100 \cdot 8 \cdot 4 \cdot 128 \approx 8.8$ TiB of storage when using 128-bit complex floats.

As a pragmatic solution, a smaller, but representative set of generic traces is generated in the following way. QuaDRiGa is used to generate statistics for 1,000,000 traces such as the delay spread and azimuth/elevation angular spread of arrival/departure. A sample of twenty traces is then selected from the 1,000,000 traces such that the distribution of the sample statistics closely matches the distribution of the population statistics. For the simulation, one of these twenty generic traces is randomly selected for each UE-AP link. To optimally exploit the limited entropy available in the small number of generic traces, the trace selected for a given UE-AP link is time-shifted by a randomly chosen offset. The traces are generated such that they wrap around at the end. This is possible because all UEs move in a circular pattern and therefore end up where they started.

In order to support arbitrary UE locations, the generated set of QuaDRiGa traces is made phaseless, i.e. the phase shift due to the distance between each transmitter and receiver antenna is removed from the trace. The following formula is applied to each trace:

$$\mathbf{H}^{\texttt{QuaDRiGa,phaseless}}[k,l] = e^{j2\pi \cdot d^{\text{3d}}_{k,l}/\lambda^c} \cdot \mathbf{H}^{\texttt{QuaDRiGa,phased}}[k,l] \qquad \forall k \in \{0,...,M^{\texttt{AP}}-1\}, \forall l \in \{0,...,M^{\texttt{UE}}-1\} \tag{3.15}$$

where $d^{\text{3d}}_{k,l}$ is the 3D distance between the kth AP antenna and the lth UE antenna in the QuaDRiGa simulation. During the actual simulation, phase shifts due to the distance are reintroduced using the actual distances between the kth AP antenna and the lth UE antenna for each UE-AP link:

$$\mathbf{H}^{\texttt{small-scale}}[k,l] = e^{-j2\pi \cdot d^{\text{3d}}_{k,l}/\lambda^c} \cdot \mathbf{H}^{\texttt{QuaDRiGa,phaseless}}[k,l] \qquad \forall k \in \{0,...,M^{\texttt{AP}}-1\}, \forall l \in \{0,...,M^{\texttt{UE}}-1\} \tag{3.16}$$

3.3.5. Noise

The noise is modeled as a combination of thermal noise and noise caused by hardware imperfections. The latter is captured in the noise figure which models all signal degradation within the components of the RF chain.

The thermal noise is defined as $N^{\text{t}} = k \cdot T \cdot B$ (in Watt), where k is the Boltzmann constant, T the temperature in Kelvin and B the bandwidth in Hz. The temperature is set at a typical 290° K (≈ 17° C) [76]. This results in $N^{\text{t}} = -93.98$ dBm $(4.004 \cdot 10^{-13}$ W) for the 100 MHz carrier.

The noise figure of the UE signal chain $N^{\texttt{f,UE}}$ is set at 8 dB. The noise figure at the AP side $N^{\texttt{f,AP}}$ is set at 3 dB. In the context of downlink transmissions, the total noise power N^{dl} therefore consists of thermal noise N^{t} and UE noise figure $N^{\texttt{f,UE}}$, resulting in $N^{\text{dl}} = -85.98$ dBm $(2.526 \cdot 10^{-12}$ W) over the carrier. For the uplink transmissions, the total noise power N^{ul} consists of thermal noise N^{t} and AP noise figure $N^{\texttt{f,AP}}$, resulting in $N^{\text{ul}} = -90.98$ dBm $(7.989 \cdot 10^{-13}$ W) over the carrier.

3.4. Radio resource management

Efficient Radio Resource Management (RRM) is critical to harness the full potential of a CF-mMIMO network. This section outlines the modeling approach of key RRM aspects, including the strategies for clustering, pilot assignment, scheduling, precoding, and combining. Furthermore, the section details how to compute the resulting SINR and throughput for each UE. The remaining key RRM aspect, Sleep Mode Management (SMM), is covered separately in Chapter 4 as it is the central focus of this thesis and entails a newly proposed solution.

For notational simplicity, the sets $\mathcal{U}$ and $\mathcal{A}$ are redefined in this section to only include the active UEs and active APs in a given coherence block.

3.4.1. Coverage criterion

A coverage criterion is needed to model whether or not a UE experiences coverage at a given location. Additionally, it is used to create an appropriately dimensioned network and during the clustering process. As the coverage of a network is primarily determined by the limited transmit power available at the UE-side, a minimum gain threshold $G^{\texttt{ul,min}}$ is derived from an uplink perspective for use as a coverage criterion. The following definition of coverage is used: a UE is said to have coverage if a minimum uplink throughput of 1 Mb/s can be attained. The rest of this section details the steps to obtain a minimum gain threshold $G^{\texttt{ul,min}}$ that corresponds with the minimum uplink throughput requirement $R^{\texttt{ul,min}} = 1$ Mb/s.

First, the required SINR $\texttt{SINR}^{\texttt{ul,min}}$ to achieve 1 Mb/s is calculated using the Shannon–Hartley theorem [77], [78]:

$$\texttt{SINR}^{\texttt{ul,min}} = 2^{R^{\texttt{ul,min}}/B} - 1 = 6.956 \cdot 10^{-3}\ (-21.58\ \text{dB}) \tag{3.17}$$

Next, to convert the SINR requirement to a specific gain, the noise and interference need to be known. The uplink noise power was previously defined in Section 3.3.5 as $N^{\mathtt{ul}} = -90.98$ dBm. The interference used in this calculation is defined through the noise rise. The noise rise $N^{\mathtt{r}}$ refers to the increase in noise power at a receiver due to interference from other UEs as compared to the baseline noise. The noise rise is calculated as:

$$N^{\mathtt{r}} = \frac{I^{\mathtt{ul}} + N^{\mathtt{ul}}}{N^{\mathtt{ul}}} \tag{3.18}$$

where $I^{\mathtt{ul}}$ is the interference in the uplink. In this work, a noise rise of $N^{\mathtt{r}} = 3$ dB is assumed. The minimum gain threshold $G^{\mathtt{ul,min}}$ can now be calculated as:

$$G^{\mathtt{ul,min}} = \mathtt{SINR}^{\mathtt{ul,min}} \cdot \frac{I^{\mathtt{ul}} + N^{\mathtt{ul}}}{P^{\mathtt{UE,tx,max}}} = 5.557 \cdot 10^{-14} \; (-132.55 \text{ dB}) \tag{3.19}$$

3.4.2. Clustering

The clustering method is inspired by Braam *et al.* [26]. The set of APs serving a given UE u, denoted $\mathcal{A}_u$ is determined in two steps. First, each UE u determines its best serving AP a_u^* as the AP with the highest LSFC:

$$a_u^* = \underset{a \in \mathcal{A}}{\arg\max}\, G_{a,u} \qquad \forall u \in \mathcal{U} \tag{3.20}$$

In [26], the cluster is then defined to include all APs whose LSFCs are at most ψ dB lower than that experienced towards AP a_m^*. However, in this work, an extra check is applied to ensure coverage (as defined in Section 3.4.1). Only the APs for which $G_{a,u}^{\mathtt{ul}} > G^{\mathtt{ul,min}}$ are included in the cluster:

$$\mathcal{A}_u = \left\{ a \in \mathcal{A} \;\middle|\; 10 \log_{10} \frac{G_{a_u^*,u}}{G_{a,u}} \leq \psi \;\wedge G_{a,u}^{\mathtt{ul}} > G^{\mathtt{ul,min}} \right\} \qquad \forall u \in \mathcal{U} \tag{3.21}$$

where the clustering threshold ψ is a configurable parameter. The UL subscript in the average channel gain $G_{a,u}^{\mathtt{ul}}$ indicates that an estimate for the AP combining gain is included (see Equation 3.5) as coverage is evaluated from an uplink perspective. Note that when $\psi = 0$ dB, the network effectively becomes a traditional cellular network in the sense that each UE is served by a single AP. From an AP a's perspective, the set of UEs they will be serving, denoted $\mathcal{U}_a$, is defined as:

$$\mathcal{U}_a = \{ u \in \mathcal{U} \mid a \in \mathcal{A}_u \} \qquad \forall a \in \mathcal{A} \tag{3.22}$$

A UE u served by a is also referred to as a *client* of a. Similarly, the set of UEs that a is serving $\mathcal{U}_a$ are collectively referred to as the clients of a.

3.4.3. Pilot assignment

At the start of each coherence block, UEs transmit pilot signals to enable channel estimation. Recall from Section 3.3.1 that at the start of each coherence block one TTI, or 360 symbols, are allocated for these uplink pilots. As explained in Section 2.3, this implies that there are 360 orthogonal pilots available to assign to UEs. To prevent pilot contamination, it is important to not reuse the same pilots for UEs that are physically proximal.

The pilot assignment algorithm used in this work is based on the graph coloring method described by Liu *et al.* [33]. There are two basic steps to the algorithm: creating the so-called 'conflict graph' Γ and applying a vertex graph coloring algorithm to Γ in order to derive a pilot assignment.

In the conflict graph $\Gamma = (\mathcal{U}, \mathcal{E})$, the set of vertices represent the set of UEs $\mathcal{U}$. $\mathcal{E}$ denotes the set of edges, where an edge $e \in \mathcal{E}$ is drawn between the nodes representing UEs u^* and u' if the clusters of u^* and u' have at least one AP in common. The intuition behind this condition is that if two UEs share an AP and use the same pilot, severe pilot contamination is caused at that AP, making it unable to estimate both channels. In formal mathematical notation, the set of edges $\mathcal{E}$ is defined as:

$$\mathcal{E} = \big\{\{u^*, u'\} \subseteq \mathcal{U} \mid \mathcal{A}_{u^*} \cap \mathcal{A}_{u'} \neq \emptyset\big\} \tag{3.23}$$

The DSatur vertex coloring algorithm devised by Brélaz [34] is used to assign colors, i.e. pilots, such that the UE nodes that share an edge in the conflict graph are not assigned the same color. When the traffic load is sufficiency high, it may occur that the DSatur algorithm produces a coloring that exceeds the maximum number of available pilots, i.e. 360 in this thesis. Given that the maximum expected number of active UEs considered in this thesis is lower than 360, this is not a cause of concern for this work. The set of pilots is actually unnecessarily large given the traffic load considered, however, as the TTI is used as the smallest time unit, one TTI is the lowest possible allocation. In practice, if more pilots are required than are available, a possible solution is to increase the time allocated for pilot transmission, thereby creating more orthogonal pilots. This does mean that there is less time available for data transmission, leading to a decrease in the UE experienced throughput. Another option is to tolerate a certain level of non-orthogonality among the set of pilots, thereby creating more pilots. This will increase the level of pilot contamination in the network, thereby also leading to a decrease in the UE experienced throughput.

In this work, the time allocated for pilots transmission is assumed to be fixed, but if in practice a situation arises where to number of available pilots is insufficient, the time allocated for pilot transmission could be dynamically extended, thereby creating more orthogonal pilots. Given that the coherence block length remains constant, this approach does mean that the time left for uplink and downlink data transmission is reduced. This indicates a trade-off: by lengthening the time allocated for pilots the channel estimation accuracy is improved, possibly resulting in an improvement in the UE throughput performance, but the UE throughput performance is simultaneously negatively effected by the reduction in time allocated for data transmission.

3.4.4. Scheduling

Instead of serving all UEs in every time slot, which is a strategy employed in a significant portion of papers on CF-mMIMO, network resources can be exploited more efficiently by scheduling a subset of UEs over the available downlink TTIs. By utilizing spatial multiplexing, multiple UEs can be scheduled to use the same time-frequency resources. A detailed rationale for the use of scheduling in CF-mMIMO networks is given in Section 2.4. To reduce complexity, each scheduled UE is served by only one layer and wideband scheduling is used. In wideband scheduling, scheduled UEs are assigned all PRBs in a given downlink TTI t. The wideband equivalent channel response matrix $\bar{\mathbf{H}}_{a,u}(t)$ is defined in Equation 3.3.

There are several popular scheduling types used in mobile networks, each with their own objective. The three most common types are: Round-Robin (RR), Maximum Rate (MR) and Proportional Fairness (PF) scheduling. The RR scheduler divides the time-frequency resources over the UEs in a fixed, cyclic order. This ensures resource fairness as each UE is assigned an equal amount of resources. However, substantial performance potential is left untapped given that the individual channel conditions are not taken into account. MR scheduling prioritizes UEs with the best channel conditions, maximizing the network throughput. This technique ensures efficient use of network resources and is especially effective if UE channel conditions vary significantly. However, the disadvantage is that UEs with poor channel conditions will rarely, if ever, be scheduled. This phenomenon is called starvation. PF scheduling strikes a balance between fairness and efficiency by considering both a UE's instantaneously achievable bit rate based on the channel conditions and its achieved throughput so far, ensuring that UEs with weaker channels are still served while maintaining good overall network performance.

Proportional Fairness scheduling

As PF scheduling strikes a good balance between fairness and efficiency, and is commonly used in operational 3/4/5G networks, a multi-user scheduling algorithm based on PF scheduling is adopted in this work. The specific implementation is described below.

Scheduling decisions are governed by the PF ratio $\texttt{PF}_u(t)$, which is defined as:

$$\texttt{PF}_u(t) = \frac{R_u^{\texttt{MRT}}(t)}{\bar{R}_u(t-1)} \qquad \forall u \in \mathcal{U} \tag{3.24}$$

where $R_u^{MRT}(t)$ is the bit rate that u could be served at in TTI t by its cluster $\mathcal{A}_u$ if it was the only UE in the network and Maximum-Ratio Transmission (MRT) precoding was used. $\bar{R}_u(t-1)$ denotes the average throughput experienced by u so far. After each TTI t, $\bar{R}_u(t)$ is updated as:

$$\bar{R}_u(t) = \alpha \cdot R_u(t) + (1-\alpha)\bar{R}_u(t-1) \qquad \forall u \in \mathcal{U} \tag{3.25}$$

where $R_u(t)$ is the bit rate experienced by u in TTI t and $\alpha = 0.01$ is the exponential smoothing factor [79]. If u is not scheduled in TTI t, $R_u(t) = 0$. When u has just become active, including at the start of the simulation, $\bar{R}_u(t)$ is initialized to 0.

SUS algorithm

In order to schedule multiple UEs using spatial multiplexing, an adaptation of the Semi-orthogonal User Selection (SUS) algorithm [80] is employed. For each downlink TTI t, the scheduler first ranks the UEs by their PF ratio $\texttt{PF}_u(t)$ in descending order. The scheduler then goes through this ordered list and schedules a UE u if and only if its channel is sufficiently orthogonal to those of the UEs that have already been scheduled and there are enough AP antennas to support u in addition to all already scheduled UEs. The pseudocode is provided in Algorithm 1 and a description of the algorithm is given below. In the pseudocode, $\mathcal{U}^C$ is used to denote the set of candidate UEs that can be scheduled. The set of scheduled UEs is denoted $\mathcal{U}^S$.

Before any UE is scheduled, the scheduler first finds the best-targeted antenna $\tilde{m}_u$ for each UE as:

$$\tilde{m}_u = \underset{m=1,\ldots,M^{\texttt{UE}}}{\arg\max}\ \bar{\mathbf{h}}_{\mathcal{A}_u,u,m}(t)\bar{\mathbf{h}}^H_{\mathcal{A}_u,u,m}(t) \qquad \forall u \in \mathcal{U} \tag{3.26}$$

where $\bar{\mathbf{h}}_{\mathcal{A}_u,u,m}(t)$ is the row-concatenated combined channel response vector between the m^{th} antenna of UE u to all antennas of the APs in u's cluster:

$$\bar{\mathbf{h}}_{\mathcal{A}_u,u,m}(t) = \left[\bar{\mathbf{H}}_{a,u}(t)[m]\right]^{\texttt{row}}_{a\in\mathcal{A}_u} \in \mathbb{C}^{|\mathcal{A}_u|\cdot M^{\texttt{AP}}} \tag{3.27}$$

For each UE-AP link, the channel associated with the best-targeted UE antenna is denoted $\bar{\mathbf{h}}_{a,u,\tilde{m}_u}(t) \in \mathbb{C}^{M^{\texttt{AP}}}$. The need to derive this best antenna stems from the fact that the applied beamforming algorithm, be it ZF or MRT, will target the beam at a specific UE antenna. A logical choice for this antenna is the one that provides the highest channel gain.

The channel orthogonality between a considered UE u and an already scheduled UE u^* for a specific AP a in u's cluster $\mathcal{A}_u$, denoted $\gamma_{u,u^*,a}$, can now be calculated as:

$$\gamma_{u,u^*,a} = \frac{|\bar{\mathbf{h}}_{a,u,\tilde{m}_u}(t)\bar{\mathbf{h}}^H_{a,u^*,\tilde{m}_{u^*}}(t)|}{\|\bar{\mathbf{h}}_{a,u,\tilde{m}_u}(t)\| \cdot \|\bar{\mathbf{h}}_{a,u^*,\tilde{m}_{u^*}}(t)\|} \tag{3.28}$$

Note that $\bar{\mathbf{h}}_{a,u,\tilde{m}_u}(t)$ in Equation 3.28 refers to the individual channel between all antennas of AP a and antenna $\tilde{m}_u$ of UE u, whereas $\bar{\mathbf{h}}_{\mathcal{A}_u,u,m}(t)$ in Equation 3.26 refers to the combined channel from all antennas of the APs in u's cluster $\mathcal{A}_u$ to antenna m of UE u.

A UE-under-consideration u is only scheduled if, for each AP a in its cluster $\mathcal{A}_u$, the orthogonality of the channel between each already scheduled UE u^* of AP a and the channel between u and a is below the channel orthogonality threshold $\gamma^{\texttt{th}}$. A typical setting of $\gamma^{\texttt{th}} = 0.5$ is assumed [80]. Additionally, the total number of already scheduled UEs by the APs in u's cluster must be fewer than the combined number of antennas at those APs. This requirement is imposed to prevent a situation where more UEs are scheduled than there are AP antennas to serve them. As detailed in Section 2.4, this limit is necessary to enable the use of ZF precoding.

Algorithm 1 Scheduling algorithm

```
Input: $t$; $\gamma^{\mathtt{th}}$; $\mathcal{U}^C = \mathcal{U}$; $\mathcal{PF} = \left\{\mathtt{PF}_u(t) \mid u \in \mathcal{U}\right\}$; $\mathcal{H} = \left\{\bar{\mathbf{h}}_{a,u,\tilde{m}_u}(t) \mid u \in \mathcal{U}, a \in \mathcal{A}_u\right\}$
Output: $\mathcal{U}^S$

▷ Go through each candidate UE ◁
while $\mathcal{U}^C \neq \emptyset$ do
    $u^* \leftarrow \arg\max_{u \in \mathcal{U}^C} \mathtt{PF}_u(t)$    ▷ Select the candidate UE with the highest PF index

    ▷ For each AP serving $u^*$, check the channel orthogonality to already scheduled UEs ◁
    for all $a \in \mathcal{A}_{u^*}$, $u' \in \mathcal{U}^S \cap \mathcal{U}_a$ where $u' \neq u^*$ do
        if $\left|\bar{\mathbf{h}}_{a,u^*,\tilde{m}_{u^*}}(t)\bar{\mathbf{h}}^H_{a,u',\tilde{m}_{u'}}(t)\right| > \gamma^{\mathtt{th}} \cdot \left\|\bar{\mathbf{h}}_{a,u^*,\tilde{m}_{u^*}}(t)\right\| \cdot \left\|\bar{\mathbf{h}}_{a,u',\tilde{m}_{u'}}(t)\right\|$ then
            continue while loop    ▷ Continue to the next UE if any check fails

    ▷ If $u^*$'s cluster also has enough available antennas, then $u^*$ is scheduled ◁
    if $\sum_{a \in \mathcal{A}_{u^*}} |\mathcal{U}^S \cap \mathcal{U}_a| \geq |\mathcal{A}_{u^*}| \cdot M^{\mathtt{AP}}$ then
        $\mathcal{U}^S \leftarrow \mathcal{U}^S \cup \{u^*\}$
        $\mathcal{U}^C \leftarrow \mathcal{U}^C \setminus \{u^*\}$
```

3.4.5. Precoding

To coherently serve a UE u in a downlink TTI t, the APs in u's cluster must coordinate their transmissions such that constructive interference is created at u's location. Given that each AP is equipped with multiple antennas, it is possible to send directed beams using MIMO beamforming techniques. In order to steer the beam, the signal is transmitted with a different phase and amplitude at each antenna. These phases and amplitudes can be represented by a vector of complex numbers called a precoding vector or precoder. The precoder used by AP a to serve u is denoted $\mathbf{w}_{a,u} \in \mathbb{C}^{M^{\mathtt{AP}}}$. In the CFN literature, several different beamforming methods have been utilized. The most popular beamforming methods include: Maximum-Ratio Transmission (MRT), Zero Forcing (ZF) and Minimum Mean Square Error (MMSE). In this study, both MRT and ZF are utilized.

Recall that wideband scheduling is adopted in this work, meaning that the set of scheduled UEs are assigned all 67 coherence blocks (or equivalently: all 134 PRBs) within the bandwidth B in a particular TTI t. Note, however, that precoders are derived separately for each coherence block. For notational convenience, the time and frequency indications are left out in this section. For example, the channel response matrix $\mathbf{H}_{a,u}(t, f)$ between u and a at a particular downlink TTI t and with frequency f is simply denoted $\mathbf{H}_{a,u}$.

Maximum ratio transmission

Before the precoders can be derived, the scheduler, detailed in Section 3.4.4, must first decide which set of UEs will be scheduled. An important input to the scheduling algorithm is the estimated rate that each UE can be served at if considered in isolation. The optimal precoder for a UE served in isolation can be derived using MRT.

As the beam for a UE will only be targeting one particular UE antenna, the first step is to derive, for every AP, which UE antenna provides the best channel. As this information is also required by the scheduler, the derivation can be found in Equation 3.26. The estimated channel between each AP a and the best-targeted antenna $\tilde{m}_u$ of UE u is represented as a row vector $\hat{\mathbf{h}}_{a,u} \in \mathbb{C}^{M^{\mathtt{AP}}}$. For each UE u, the combined channel from all APs in its cluster to the best-targeted antenna of u is represented as a row-concatenated vector:

$$\hat{\mathbf{h}}_u = \left[\hat{\mathbf{h}}_{a,u}\right]^{\mathtt{row}}_{a \in \mathcal{A}_u} \in \mathbb{C}^{|\mathcal{A}_u| \cdot M^{\mathtt{AP}}} \qquad \forall u \in \mathcal{U} \tag{3.29}$$

The combined MRT precoder used by u's cluster $\mathbf{w}_u^{\mathtt{MRT}}$ can be derived from $\hat{\mathbf{h}}_u$ as:

$$\mathbf{w}_u^{\mathtt{MRT}} = \frac{\hat{\mathbf{h}}_u^H}{\|\hat{\mathbf{h}}_u\|} \qquad \forall u \in \mathcal{U} \tag{3.30}$$

Zero-forcing

In this study, network-wide ZF will be used to derive the actual precoding vectors after the scheduler has decided which UEs will be served in a particular downlink TTI t. As ZF is applied over the whole network, the first step is to define the combined estimated channel response matrix of all scheduled UEs:

$$\hat{\mathbf{H}}_{\mathcal{U}^S} = \Big[\big[\hat{\mathbf{h}}_{a,u} \cdot 1\{a \in \mathcal{A}_u\} \big]^{\texttt{row}}_{\mathcal{A}} \Big]^{\texttt{col}}_{\mathcal{U}^S} \in \mathbb{C}^{|\mathcal{U}^S| \times (A \cdot M^{\texttt{AP}})} \tag{3.31}$$

In this matrix, each row represents one scheduled UE and each column contains either the estimated channel response vector $\hat{\mathbf{h}}_{a,u}$ if a serves u or a zero vector of length $M^{\texttt{AP}}$ if a does not serve u. ZF creates orthogonal, non-interfering channels between APs and UEs by taking the pseudo-inverse of the combined channel response matrix:

$$\mathbf{W}_{\mathcal{U}^S} = \hat{\mathbf{H}}^{\dagger}_{\mathcal{U}^S} = \hat{\mathbf{H}}^{H}_{\mathcal{U}^S} (\hat{\mathbf{H}}_{\mathcal{U}^S} \hat{\mathbf{H}}^{H}_{\mathcal{U}^S})^{-1} = \begin{bmatrix} \mathbf{w}_{0,u} \\ \vdots \\ \mathbf{w}_{A-1,u} \end{bmatrix}^{\texttt{row}}_{u \in \mathcal{U}^S} \in \mathbb{C}^{(A \cdot M^{\texttt{AP}}) \times |\mathcal{U}^S|} \tag{3.32}$$

In the matrix $\mathbf{W}_{\mathcal{U}^S}$, each column consists of the concatenated, non-normalized precoding (column) vectors $\mathbf{w}_{a,u}$ used by each of the APs for a scheduled UE $u \in \mathcal{U}^S$. Naturally, each AP a only uses the precoders $\mathbf{w}_{a,u}$ for those UEs that they actually serve, i.e. $u \in \mathcal{U}_a$. The non-normalized precoders are subsequently normalized over the antennas at each AP:

$$\mathbf{w}'_{a,u} = \frac{\mathbf{w}_{a,u}}{||\mathbf{w}_{a,u}||} \qquad \forall u \in \mathcal{U}^S, \forall a \in \mathcal{A}_u \tag{3.33}$$

Each AP is assumed to use equal power sharing to divide its maximum transmit power over the UEs that it serves. The transmission power is incorporated into the precoders as such:

$$\tilde{\mathbf{w}}_{a,u} = \sqrt{\frac{1}{|\mathcal{U}^S_a|} \cdot P^{\texttt{AP,ant,tx,max}} \cdot M^{\texttt{AP}}} \cdot \mathbf{w}'_{a,u} \qquad \forall u \in \mathcal{U}^S, \forall a \in \mathcal{A}_u \tag{3.34}$$

where $\mathcal{U}^S_a$ denotes the scheduled set of clients of AP a.

It must be noted that practical systems do not typically derive their precoders network-wide for scalability reasons. More distributed precoding schemes such as local Partial Zero-Forcing (PZF) [81] are better suited to larger networks.

3.4.6. Combining

At the UE side, Maximum-Ratio Combining (MRC) is used to combine the signals received at each of the $M^{\texttt{UE}} = 4$ antennas. The combiner (row) vector used by UE u is denoted $\mathbf{v}_u$ and is defined as:

$$\mathbf{v}_u = (\mathbf{H}_u \tilde{\mathbf{w}}_u)^{\dagger} \in \mathbb{C}^{M^{\texttt{UE}}} \qquad \forall u \in \mathcal{U}^S \tag{3.35}$$

where $\mathbf{H}_u$ is the combined channel from all AP antennas in u's cluster to all antennas of u and $\tilde{\mathbf{w}}_u$ is the combined precoding vector used by the APs in u's cluster:

$$\mathbf{H}_u = [\mathbf{H}_{a,u}]^{\texttt{row}}_{a \in \mathcal{A}_u} \in \mathbb{C}^{M^{\texttt{UE}} \times (|\mathcal{A}_u| \cdot M^{\texttt{AP}})} \qquad \forall u \in \mathcal{U}^S \tag{3.36}$$

$$\tilde{\mathbf{w}}_u = [\tilde{\mathbf{w}}_{a,u}]^{\texttt{col}}_{a \in \mathcal{A}_u} \in \mathbb{C}^{|\mathcal{A}_u| \cdot M^{\texttt{AP}}} \qquad \forall u \in \mathcal{U}^S \tag{3.37}$$

3.4.7. SINR and throughput calculation

With the precoders and combiners known, the experienced SINR for each UE u can now be determined as:

$$\mathtt{SINR}_u = \left(\mathbf{v}_u \mathbf{v}_u^H \cdot N^{\text{dl}} + \sum_{u^* \in \mathcal{U}_u^I} \mathbf{c}_{u,u^*} \mathbf{c}_{u,u^*}^H \right)^{-1} \qquad \forall u \in \mathcal{U}^S \tag{3.38}$$

where u^* represents a co-scheduled UE that causes interference to UE u and recall that N^{dl} is the effective noise power in the downlink. The set of UEs causing interference is equal to the set of scheduled UEs excluding u itself: $\mathcal{U}_u^I = \mathcal{U}^S \setminus \{u\}$. The amount of interference that a given co-scheduled UE u^* causes to u per receive antenna can be expressed as a vector $\mathbf{c}_{u,u^*}$:

$$\mathbf{c}_{u,u^*} = \mathbf{v}_u (\mathbf{H}_{u^*} \tilde{\mathbf{w}}_{u^*}) \in \mathbb{C}^{M^{\text{UE}}} \qquad \forall u \in \mathcal{U}^S, \forall u^* \in \mathcal{U}_u^I \tag{3.39}$$

with $\mathbf{H}$ and $\tilde{\mathbf{w}}$ as introduced in Equation 3.36 and Equation 3.37, respectively.

Recall that each of these matrices and vectors apply to a particular moment in time t and a particular frequency f. These labels were temporarily removed for notational convenience, but are now reintroduced: the SINR for user u at time t and frequency f, as derived in Equation 3.38, is denoted $\mathtt{SINR}_u(t, f)$. As the scheduled UEs are assigned the entire carrier, the individual coherence block-specific narrowband SINRs $\mathtt{SINR}_u(t, f)$ need to be mapped to a wideband SINR. This can be achieved using the Mutual Information Effective SINR Mapping (MIESM) method introduced by Brueninghaus *et al.* [82]. With MIESM, the effective wideband SINR $\mathtt{SINR}_u(t)$ over all 67 coherence blocks (or equivalently: 134 PRBs) can be determined in two steps. First, the narrowband SINRs are mapped into Mutual Information (MI) values $\mathtt{MI}_u(t, f)$ and averaged over the carrier to obtain $\mathtt{MI}_u(t)$. Second, the MI $\mathtt{MI}_u(t)$ value can be mapped back into a wideband SINR $\mathtt{SINR}_u(t)$. The mapping between MI and SINR is based on a curve dependent on the Modulation and Coding Scheme (MCS). These conversion curves are shown in Figure 3.9. When applying the MIESM method, 256-QAM curve is assumed in this work. Given the pre-convergence overlap of the different curves, this assumption is mostly non-consequential.

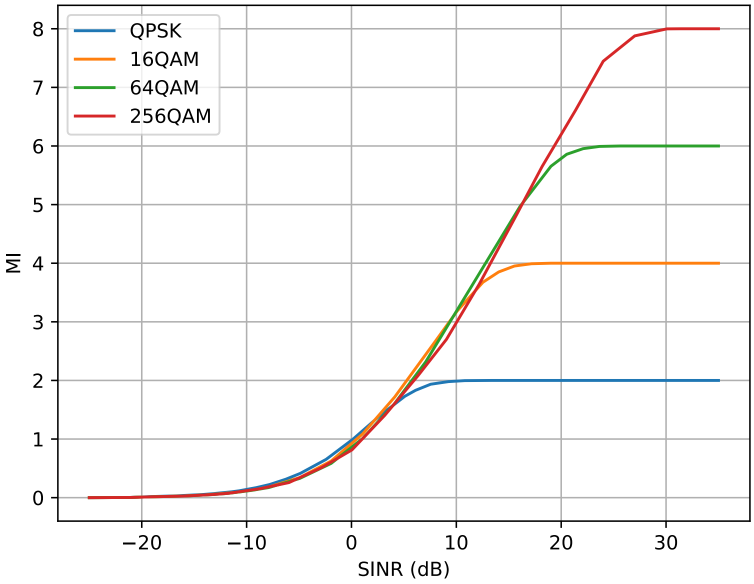


Figure 3.9: MI versus SINR curves used by MIESM for different modulation orders [82].

Finally, the experienced bit rate of UE u for a downlink TTI t can be estimated using a truncated form of the Shannon bound as described in Section 4.2.7 of 3GPP TR 38.921 [83]. The Shannon bound, derived from the Shannon–Hartley theorem [77], [78], represents the maximum theoretical throughput that can be achieved for a given SINR. The truncated Shannon rate is given by:

$$R_u(t) = \min(R^{\text{max}},\ C^{\text{f}} \cdot B \cdot \log_2(1 + \texttt{SINR}_u(t))) \tag{3.40}$$

where R^{max} is the maximum attainable bit rate and C^{f} is a correction factor to account for signaling overhead and inaccuracies in channel estimation. $R^{\text{max}} = 460.56$ Mb/s, as calculated using the method described by Section 5.1.3.2 of 3GPP TR 38.214 [84]. R^{max} is derived based on the number of PRBs, numerology, fraction of time allocated for downlink data transmission, maximum coding rate, maximum modulation order and number of transmitted layers. A maximum coding rate of $948/1024$ is assumed for the purpose of deriving R^{max}. The correction factor is set to $C^{\text{f}} = 0.75$ based on [85].

3.5. Energy consumption

Three energy consuming entities are modeled in this thesis: the APs, the fronthaul and the CPU. The total energy consumption of the network in Watt is defined as:

$$P^{\texttt{total}} = \sum_{a \in \mathcal{A}} P_a^{\texttt{AP}} + \sum_{a \in \mathcal{A}} P_a^{\texttt{FH}} + P^{\texttt{CPU}} \tag{3.41}$$

where $P_a^{\texttt{AP}}$, $P_a^{\texttt{FH}}$ and $P^{\texttt{CPU}}$ are defined in their respective sections below. UEs are not included in the energy consumption model given that the focus of this thesis is on Radio Access Network (RAN) energy savings, where APs are put to sleep to conserve energy. This approach is also common in the CF-mMIMO SMM literature, see for example the energy consumption models used by Jung and Hong [46] and Riera-Palou *et al.* [48]. The chosen parameter values and sources underlying these choices can be found in Table 3.2.

3.5.1. APs

As will be described in more detail in Chapter 4, in any given coherence block, an AP is in one of seven different states. There are three primary states and four transitional states. The primary states are: active, light sleep and deep sleep. Active APs can transition to light sleep and vice versa without any transition time. However, transitioning from active to deep sleeping or the other way around has an associated transition time $T^{\text{t}} = 0.5$ seconds, which gives rise to the first two transitional states: active-to-deep-sleeping (A2D) and deep-sleeping-to-active (D2A). The chosen transition time is taken from Debaillie *et al.* [88] who describe four sleep modes where the deepest sleep mode is termed ‘long-term sleep’ with an associated symmetric transition time of 0.5 seconds. This transition time also applies to the transition from light sleep to deep sleep and vice versa, giving rise to two more transitional states: light-sleeping-to-deep-sleeping (L2D) and deep-sleeping-to-light-sleeping (D2L).

The deep sleep mode consumes the lowest amount of energy and is used when an AP is not necessary to maintain network performance in terms of the experienced UE throughput and coverage. At the beginning of each coherence block, the CPU decides which APs should transition into or out of deep sleep mode. Afterward, all currently active and light sleeping APs participate in the cluster formation

Table 3.2: Energy consumption model parameters.

Parameter	Symbol	Value	Unit	Sources
AP fixed power	$P^{\texttt{AP,fixed}}$	8	W	[27], [43]
AP fixed power per RF chain	$P^{\texttt{AP,chain}}$	0.2	W	[27], [42], [43], [62]
AP maximum transmit power per antenna	$P^{\texttt{AP,ant,tx,max}}$	0.25	W	[25], [42], [62], [63], [86]
AP power amplifier efficiency	η	0.39	-	[27], [43], [87]
Fronthaul fixed power	$P^{\texttt{FH,fixed}}$	0.825	W	[42], [48], [63]
Fronthaul traffic-dependent power	$\zeta^{\texttt{FH}}$	0.25	W/Gbps	[27], [42], [87]
CPU fixed power	$P^{\texttt{CPU,fixed}}$	5	W	[48], [62], [86]
CPU traffic-dependent power	$\zeta^{\texttt{CPU}}$	0.1	W/Gbps	[48]

process. If any light sleeping APs are chosen to be in a UE's cluster, they are woken up. Any non-deep sleeping AP that is not included in any UE cluster is put into light sleep mode for that coherence block.

To facilitate participation in the clustering process for light sleeping APs and for the purposes of the SMM algorithm for deep sleeping APs, basic knowledge of the current channel conditions are required. As both light and deep sleeping APs do not transmit pilots, an alternative solution is required. To ensure that the CPU still has some knowledge with regard to the channel conditions, both light and deep sleeping APs are assumed to have the capability to periodically broadcast synchronization signals [23] that enable estimation of the average channel gain.

For active APs, the energy consumption of an AP is determined by a fixed part for the circuit power, a fixed part per RF chain and the transmit power. The circuit power includes the power used for the control system, cooling system and power supply. The RF chain includes the power required for the components that generate, transmit/receive and filter the RF signals. It includes oscillators, mixers and the Digital-to-Analog Converters (DACs)/Analog-to-Digital Converters (ADCs). The energy consumption for sleeping APs only includes the fixed parts. It is based on the energy usage for active APs but scaled down by the sleep power savings coefficient ξ_a. The complete energy consumption model for an AP is defined as:

$$P_a^{\mathtt{AP}} = \begin{cases} P^{\mathtt{AP,fixed}} + C \cdot P^{\mathtt{AP,chain}} + \frac{1}{\eta} P_a^{\mathtt{AP,tx}} & \text{if } a \text{ is active} \\ (1-\xi_a)(P_a^{\mathtt{AP,\ fixed}} + C \cdot P_a^{\mathtt{AP,\ chain}}) & \text{if } a \text{ is sleeping} \end{cases} \tag{3.42}$$

where $C = M^{\mathtt{AP}}$ is the number of RF chains at each AP, $P^{\mathtt{AP,fixed}}$ is the fixed power of an AP, $P^{\mathtt{AP,chain}}$ is the power used per RF chain, η is the power amplifier efficiency and $P_a^{\mathtt{AP,tx}}$ is the transmit power used by AP a. When an AP is actively transmitting in a given TTI, $P_a^{\mathtt{AP,tx}} = P^{\mathtt{AP,ant,tx,max}} \cdot M^{\mathtt{AP}}$ as transmitting APs always use the maximum transmit power. For any active AP that does not have any UEs to serve in a given TTI, $P_a^{\mathtt{AP,tx}} = 0$. The sleep power savings coefficient ξ has two possible values: 0.3 for light sleep and 0.9 for deep sleep. When APs are in one of the transitional states, active-to-deep-sleeping or deep-sleeping-to-active, the 0.3 light sleep value is applied when calculating their energy consumption.

The idea of scaling the fixed parts in this manner to model the energy usage of multiple sleep stages is taken from Riera-Palou *et al.* [48]. The 0.3 value for the light sleep mode is derived from a study considering an actual Huawei-based 5G deployment [89] and is also used in [48]. The value of 0.9 for the deep sleep mode is inspired by studies that consider only a deep sleep mode [27], [43], [44].

3.5.2. Fronthaul

The energy consumption model of the fronthaul considers each AP-CPU link separately. When an AP is active, the energy consumption of its fronthaul connection to the CPU consists of a fixed part $P_a^{\mathtt{FH,fix}}$ and a traffic-dependent part based on the sum throughput of all the UEs it serves in a given TTI. For an AP a, its fronthaul throughput $R_a^{\mathtt{FH}}$ is calculated as:

$$R_a^{\mathtt{FH}} = \sum_{u \in \mathcal{U}_a} R_u \tag{3.43}$$

where R_u is the throughput of UE u and recalling that $\mathcal{U}_a$ is the set of UEs served by AP a. To calculate the traffic-dependent energy consumption, the throughput in Gb/s is multiplied by the power needed to transfer one Gb of data over the fronthaul $\zeta_a^{\mathtt{FH}}$ expressed in Watt/Gb/s. When an AP is sleeping, its fronthaul power consumption is reduced to 10% of the fixed power needed for active APs. This value is inspired by the 0.9 sleep power savings coefficient for the deep sleep mode as used by Femenias *et al.* [43] and García-Morales *et al.* [44]. Mathematically, the energy consumption of AP a's fronthaul connection in Watt is then expressed as:

$$P_a^{\mathtt{FH}} = \begin{cases} P_a^{\mathtt{FH,fixed}} + R_a^{\mathtt{FH}} \cdot \zeta_a^{\mathtt{FH}} & \text{if } a \text{ is active} \\ 0.1 \cdot P_a^{\mathtt{FH,fixed}} & \text{if } a \text{ is sleeping} \end{cases} \tag{3.44}$$

3.5.3. CPU

The energy consumption model of the CPU consists of a fixed part and a traffic-dependent part. A similar model is used by Riera-Palou *et al.* [48] and Chen *et al.* [62]. The fixed part models tasks such as running the AP sleep algorithm, scheduling and deriving the precoders. In this simplified model, these tasks are assumed to take a constant amount of energy, however, in practice they will take longer, and therefore use more energy, when there are more active UEs in the network. The traffic-dependent part models the fact that data needs to be processed by the CPU, e.g. data must be encoded and decoded, and is defined by the sum throughput of all UEs:

$$R^{\text{CPU}} = \sum_{u \in \mathcal{U}} R_u \tag{3.45}$$

Since the CPU must always manage the network, it cannot be put to sleep and therefore its energy model does not include any sleep or power reduction aspects. The CPU energy model is defined as:

$$P^{\text{CPU}} = P^{\text{CPU},\texttt{fixed}} + R^{\text{CPU}} \cdot \zeta^{\text{CPU}} \tag{3.46}$$

4
SMM algorithm

This chapter presents the proposed SMM algorithm. Section 4.1 provides a description of the algorithm and contains the pseudocode. Section 4.2 presents variants of the SMM algorithm that were tested but ultimately rejected.

4.1. Description

The proposed SMM algorithm is divided into two parts which are executed at two distinct moments within a given coherence block. At the beginning of each coherence block, the first part of the SMM algorithm decides which APs should go into deep sleep and which APs should transition from the deep sleep to the active mode. Recall that both going into and out of deep sleep takes some time: the transition time $T^{\mathrm{t}} = 0.5$ seconds. Note that transitioning APs are not considered to be in either of the three primary states: active, light sleep and deep sleep. While transitioning, the SMM algorithm can take no actions on such APs as state transitions cannot be interrupted. After the first part of the SMM algorithm has concluded, all currently light sleeping APs are woken up to participate in the clustering process along with all already active APs. After the clustering process, the second part of the SMM algorithm puts all APs that have not been chosen to be part of any UE's cluster into the light sleep mode.

The core of the algorithm is concerned with deciding which APs should go into or out of deep sleep as these transitions have an associated non-zero transition time, as opposed to the transition from the active to the light sleep mode, or vice versa, which is instantaneous. This core is divided into two distinct parts: determining which APs should go into deep sleep and determining which APs should wake up from deep sleep. The main approach that governs both is the same, however. First, APs are ordered by some metric which indicates the importance of their availability to serve UEs. Second, these APs are addressed iteratively by the algorithm; a decision is made per AP whether it should go into or awaken from deep sleep by using the metric and an a priori configured threshold. Additionally, before an AP can transition into deep sleep, a check is performed to ensure the 98% population coverage is maintained if that AP were to be put to sleep. The metric and threshold used for each of the two parts are discussed in the sections below. The pseudocode for the entire SMM algorithm is provided in Algorithm 2.

In this section, the set of all currently active APs is represented as $\mathcal{A}^{\mathrm{on}}$, the set of all currently light sleeping APs as $\mathcal{A}^{\mathrm{ls}}$ and the set of all currently deep sleeping APs as $\mathcal{A}^{\mathrm{ds}}$. The set of APs currently transitioning to deep sleep is denoted $\mathcal{A}^{\mathrm{2ds}}$. This set encompasses both the APs transitioning to deep sleep from the active mode (A2D) and the APs transitioning to deep sleep from the light sleep mode (L2D). The set of APs currently transitioning to the active mode is denoted $\mathcal{A}^{\mathrm{2on}}$. This set only encompasses the APs transitioning from the deep sleep mode to the active mode (D2A). While a transition from the deep sleep to the light sleep mode (D2L) also exists, it is not used in the proposed SMM algorithm. The motivation for this choice is rooted in the fact that the transitions between the active and light sleep modes have no associated transition time. The omission of the D2L transmission is irrelevant, as any non-deep sleeping APs that is not included in any UE's cluster is automatically transitioned to the light sleep mode at no cost.

A time index i is used when referring to the state of a particular vector, matrix or set at a particular coherence block i. For example, $\mathcal{A}^{\text{on}}(i)$ refers to the set of active UEs in coherence block i. If no index is used, the current state is meant.

4.1.1. Transition to deep sleep mode

To decide whether a particular *active* AP a should go to deep sleep, the algorithm should estimate the impact of a's potential absence on the UEs currently served by a. This is not straightforward given that the UEs served by a will likely also have other APs that serve them. As the resultant SINR of a served UE u is not just determined by a but by all APs serving u, a method is needed to quantify the contribution of AP a in relation to the rest of the APs serving u. As the SMM algorithm is executed at the beginning of the coherence block, no pilots have been transmitted yet. This means there is unfortunately no information yet on the UE channels or the bit rates they can attain in that coherence block. This information is therefore also not available to use as a way to estimate the effect of a's removal from UE u's cluster when transitioning to deep sleep mode.

Instead, the algorithm will use the channel estimates and attained bit rates from previous coherence blocks. More precisely, for each active AP a, the algorithm looks at the average throughput experienced by each UE that a has served over a window consisting of the previous k coherence blocks. For each UE u, two throughput metrics are determined: the overall average experienced throughput over the window $R_u^{\texttt{window}}$ and an estimate for the average experienced throughput over the window if u was served by a exclusively, denoted $R_{a,u}^{\texttt{window}}$. As formalized in Equation 4.2, for each coherence block i, the fraction of the attained bit rate that is attributable to AP a can be estimated by multiplying the overall bit rate by the channel gain that AP a experiences towards UE u relative to the total channel gain experienced by u. Recall that UEs go through cycles of active and inactive periods. As a consequence, u might not have been active during the entire window. In such coherence blocks, $R_u(i) = 0$. To solve this, only the $l_u \leq k$ coherence blocks in which u was active count towards the average throughput metrics:

$$R_u^{\texttt{window}}(t) = \frac{1}{l_u} \cdot \sum_{i=t-k}^{t-1} R_u(i) \tag{4.1}$$

$$R_{a,u}^{\texttt{window}}(t) = \frac{1}{l_u} \cdot \sum_{i=t-k}^{t-1} R_u(i) \cdot \frac{\hat{\mathbf{h}}_{a,u}(i)\hat{\mathbf{h}}_{a,u}^H(i)}{\sum_{a' \in \mathcal{A}_u(i)} \hat{\mathbf{h}}_{a',u}(i)\hat{\mathbf{h}}_{a',u}^H(i)} \tag{4.2}$$

where the row vector $\hat{\mathbf{h}}_{a,u} \in \mathbb{C}^{M^{\text{AP}}}$ represents the estimated response of the channel between AP a and the best-targeted antenna of UE u. This specific vector is also used by the scheduler and to derive the precoders. The derivation of the best-targeted antenna for each UE can be found in Section 3.4.4. $R_u(i)$ is the bit rate attained by UE u in coherence block i.

By subtracting $R_{a,u}^{\texttt{window}}$ from $R_u^{\texttt{window}}$, an estimate can be made of the remaining throughput experienced by UE u in case AP a were to be put into deep sleep. The basic idea is to verify whether the worst remaining throughput still exceeds the configured minimum throughput threshold $R^{\texttt{min,sleep}}$. Rather than focusing on the absolute worst case, the algorithm looks at the 10th percentile of the estimated remaining throughputs among all UEs served by a in the past k coherence blocks. This metric is termed the to-deep-sleep-metric $R_a^{\text{on}}(t)$ and is defined as:

$$R_a^{\text{on}}(t) = P_{10}\big(\{R_u^{\texttt{window}}(t) - R_{a,u}^{\texttt{window}}(t) \mid u \in \mathcal{U}_a^{\texttt{window}}(t)\}\big) \qquad \forall a \in \mathcal{A}^{\text{on}} \tag{4.3}$$

where $\mathcal{U}_a^{\texttt{window}}$ is the set of clients that AP a has served for at least one of the past k coherence blocks.

For each AP a, the algorithm checks whether $R_a^{\text{on}}(t)$ exceeds the to-deep-sleep-threshold $R^{\texttt{min,sleep}}$. If so, turning off a is deemed to not unacceptably degrade the experienced throughput of its clients and so a could potentially go to sleep. AP a will only actually be put into deep sleep mode if the population coverage requirement of 98% is still satisfied without a.

It may occur that an AP a has not been continuously active in the last k coherence blocks. Of course, it is possible to reduce the window to only consider the coherence blocks in which a was active, but this would make the to-deep-sleep-metric less numerically stable. To prevent this situation, only the APs that have been active for at least k coherence blocks are eligible to go into deep sleep and are denoted by the set $\mathcal{A}^{\text{on},\texttt{window}}$.

It is important to note that $R_a^{\text{on}}(t)$ should be recalculated during the execution of the SMM algorithm for AP a if for some other AP a^* it has already been decided that it should go to sleep and these APs shared clients in one or more of the last k coherence blocks. This is because the sum in the denominator of the fraction in Equation 4.2 is now different as a^*'s contribution to the sum should be removed given that it has been decided that a^* should transition to deep sleep. In this manner, the impact of already decided switch-offs is taken into account when making new decisions in the same coherence block.

Given the fact that previous decisions will impact future ones, this also implies that the order in which the SMM algorithm addresses the APs is important. For this reason, a systematic ordering rule is required. In the proposed SMM algorithm, the APs in the set $\mathcal{A}^{\text{on},\texttt{window}}$, i.e. the APs that are eligible to go to deep sleep, are ordered by their $R_a^{\text{on}}(t)$ value, highest first. The intuition for this particular ordering is that that APs with the highest to-deep-sleep-metric values are those whose absence has the least chance of deterioting their client UE's experienced throughput to below some given throughput performance target.

A separate mechanism has been devised to allow *light sleeping* APs to transition to the deep sleep mode. Recall that the light sleep mode is used for APs that are not included in any UE's cluster. Given that these AP's did not have any clients, the above metric cannot be used to determine whether they should go into deep sleep. Instead, the following rule is applied: if an AP a has been light sleeping during all of the last k coherence blocks and the coverage requirement is still satisfied without a, then a is transitioned into the deep sleep mode. The set of APs that have been light sleeping for at least the last k coherence blocks is denoted $\mathcal{A}^{\texttt{ls},\texttt{window}}$.

All active and light sleeping APs that were excluded from the sets $\mathcal{A}^{\text{on}}$ and $\mathcal{A}^{\texttt{ls}}$ because they were not in these respective modes for at least the last k coherence blocks will remain in their current mode.

4.1.2. Transition to active mode

Deciding whether a particular deep sleeping AP should be woken up and brought back to active mode is more challenging than deciding whether an active AP should go into deep sleep. For a given deep sleeping AP a, it is difficult to estimate the impact that waking up a will have on the performance experienced by UEs in a's vicinity. It is not possible to use the attained bit rates of a's clients, as it is sleeping and therefore does not have any clients. Additionally, because a is sleeping, the channels between a and the UEs in its vicinity are not known and therefore also cannot be used.

In order to obtain some information with regard to the channel conditions, it is assumed that a deep sleeping AP a has the capability to periodically broadcast synchronization signals [23]. Active UEs can use these signals to estimate the average channel gain to a, denoted $G_{a,u}$, and forward this information to the CPU via the active APs that are serving them. Using these average channel gains, the algorithm can determine which UEs are in a's neighborhood. This neighborhood is formally defined as the set of UEs with an average channel gain larger than $G^{\texttt{nh},\texttt{min}}$ dB to a deep sleeping AP a:

$$\mathcal{U}_a^{\texttt{nh}} = \{u \in \mathcal{U} \mid G_{a,u} > G^{\texttt{nh},\texttt{min}}\} \qquad \forall a \in \mathcal{A}^{\texttt{ds}} \tag{4.4}$$

For a deep sleeping AP a, this neighborhood can be used as a substitute for an active AP's clients. Analogous to the mechanism used to decide which APs should go into deep sleep, for each UE in a's neighborhood, the experienced average throughput over the window consisting of the past k coherence blocks $R_u^{\texttt{window}}(t)$, as defined in Equation 4.1, is used as a basis. Again, the algorithm will look at the 10th percentile over this set of throughputs. This metric is termed the to-active-metric $R_a^{\texttt{ds}}(t)$ and is defined as:

$$R_a^{\texttt{ds}}(t) = P_{10}(\{R_u^{\texttt{window}}(t) \mid u \in \mathcal{U}_a^{\texttt{nh}}(t)\}) \qquad \forall a \in \mathcal{A}^{\texttt{ds}} \tag{4.5}$$

For each AP a, the algorithm checks whether $R_a^{\texttt{ds}}(t)$ is below the to-active-threshold $R^{\texttt{max},\texttt{wake}}$. If so, the throughput performance of the UEs in a's vicinity is deemed to be too low and a is woken up in an attempt to improve it. Similar to the procedure used to decide which active APs should go into deep sleep and to prevent a ping-pong effect, where APs are repeatedly switched on and off, only APs that have been in deep sleep for at least k coherence blocks are eligible to wake up. The set of eligible APs is denoted $\mathcal{A}^{\texttt{ds,window}}$.

4.1.3. Time complexity

While not a formal proof of the complexity of the proposed algorithm, a simple analysis of the pseudocode reveals that it has a time complexity of at most $\mathcal{O}(A^2 \cdot U \cdot k)$. To compute the metrics, for each AP, all client UEs are addressed for each coherence block in the window of size k. This suggests a worst-case time complexity of $\mathcal{O}(A \cdot U \cdot k)$ for that particular step. In each iteration of the loops starting at lines 8, 13, 24, 31 and 42 AP are addressed at most once, with the exception of the loop at line 13 which also has an inner loop that potentially addresses all other APs, i.e. $\mathcal{O}(A^2)$. This worst case can occur when each UE is served by all APs. Combining this double loop from line 13 with the time complexity to calculate the metrics, this results in an overall worst-case time complexity of $\mathcal{O}(A^2 \cdot U \cdot k)$. This is a substantionally lower time complexity than most of the SMM algorithms described in the literature, which are often exponential or have higher exponents in their polynomials. Of the SMM algorithms analyzed for the literature review, only the (3 x E) algorithm presented by Kooshki *et al.* [52] has a lower time complexity with $\mathcal{O}(A \cdot U)$.

Algorithm 2 Sleep Mode Management (SMM) algorithm

```
Input: $t$; $\mathcal{T} = \{i \in \mathbb{Z} \,|\, t-k \le i \le t-1\}$; $R^{\texttt{min,sleep}}$; $R^{\texttt{max,wake}}$; $k$; $G^{\texttt{nh,min}}$;
    $\forall i \in \mathcal{T}$: $\mathcal{A}^{\text{on}}(i)$; $\mathcal{A}^{\texttt{ls}}(i)$; $\mathcal{A}^{\text{ds}}(i)$; $\mathcal{A}^{\text{2ds}}(i)$; $\mathcal{A}^{\text{2on}}(i)$; $\hat{\mathbf{h}}_{a,u}(i), a \in \mathcal{A}(i), u \in \mathcal{U}_a(i)$;
        $R_u(i), a \in \mathcal{A}(i), u \in \mathcal{U}_a(i)$
Output: Updated $\mathcal{A}^{\text{on}}(t)$; $\mathcal{A}^{\texttt{ls}}(t)$; $\mathcal{A}^{\text{ds}}(t)$; $\mathcal{A}^{\text{2ds}}(t)$; $\mathcal{A}^{\text{2on}}(t)$

▷ Determine which APs have been in the same state for at least the past $k$ coherence blocks ◁
Determine $\mathcal{A}^{\texttt{on,window}}$
Determine $\mathcal{A}^{\texttt{ls,window}}$
Determine $\mathcal{A}^{\texttt{ds,window}}$

▷ Initialize max-priority queue $\mathcal{Q}$ with $a \in \mathcal{A}^{on,window}$ as elements and $R_a^{on}(t)$ as the priority ◁
$\mathcal{Q} \leftarrow \texttt{PriorityQueue}()$
for all $a \in \mathcal{A}^{\texttt{on,window}}$ do
    Compute $R_a^{\text{on}}(t)$    ▷ Using Equations 4.1, 4.2 and 4.3
    $\mathcal{Q}.\texttt{insert}(a, R_a^{\text{on}}(t))$

▷ Decide which active APs should go into deep sleep ◁
while $\mathcal{Q} \neq \emptyset$ do
    $a \leftarrow \mathcal{Q}.\texttt{extract\_max}()$
    if $R_a^{\text{on}}(t) > R^{\texttt{min,sleep}}$ and $\texttt{PopulationCoverage}(\mathcal{A}^{\text{on}} \cup \mathcal{A}^{\texttt{ls}} \setminus \{a\}) \ge 0.98$ then
        ▷ Start transition to deep sleep for AP $a$ ◁
        $\mathcal{A}^{\text{on}} \leftarrow \mathcal{A}^{\text{on}} \setminus \{a\}$
        $\mathcal{A}^{\text{2ds}} \leftarrow \mathcal{A}^{\text{2ds}} \cup \{a\}$
        for all $a^* \in \mathcal{A}^{\text{on}}$ that had shared clients with $a$ in any of the past $k$ coherence blocks do
            Recompute $R_{a^*}^{\text{on}}(t)$    ▷ Using Equations 4.1, 4.2 and 4.3
            $\mathcal{Q}.\texttt{update\_priority}(a^*, R_{a^*}^{\text{on}}(t))$

▷ Decide which light sleeping APs should go into deep sleep ◁
for all $a \in \mathcal{A}^{\texttt{ls,window}}$ do
    if $\texttt{PopulationCoverage}(\mathcal{A}^{\text{on}} \cup \mathcal{A}^{\texttt{ls}} \setminus \{a\}) \ge 0.98$ then
        ▷ Start transition to deep sleep for AP $a$ ◁
        $\mathcal{A}^{\texttt{ls}} \leftarrow \mathcal{A}^{\texttt{ls}} \setminus \{a\}$
        $\mathcal{A}^{\text{2ds}} \leftarrow \mathcal{A}^{\text{2ds}} \cup \{a\}$

▷ Decide which deep sleeping APs should wake up ◁
for all $a \in \mathcal{A}^{\texttt{ds,window}}$ do
    Determine $\mathcal{U}_a^{\text{nh}}$    ▷ Using Equation 4.4
    Compute $R_a^{\text{ds}}(t)$    ▷ Using Equation 4.5
    if $R_a^{\text{ds}}(t) < R^{\texttt{max,wake}}$ then
        ▷ Start transition to active mode for AP $a$ ◁
        $\mathcal{A}^{\text{ds}} \leftarrow \mathcal{A}^{\text{ds}} \setminus \{a\}$
        $\mathcal{A}^{\text{2ds}} \leftarrow \mathcal{A}^{\text{2ds}} \cup \{a\}$

▷ Perform clustering and put APs without any clients into light sleep mode ◁
$\mathcal{A}^{\text{on}} \leftarrow \mathcal{A}^{\text{on}} \cup \mathcal{A}^{\texttt{ls}}$
Form clusters $\mathcal{A}_u$ for each active UE $u$ using each AP $a \in \mathcal{A}^{\text{on}}$    ▷ As described in Section 3.4.2
for all $\{a \in \mathcal{A}^{\text{on}} \,|\, \mathcal{U}_a = \emptyset\}$ do
    $\mathcal{A}^{\text{on}} \leftarrow \mathcal{A}^{\text{on}} \setminus \{a\}$
    $\mathcal{A}^{\texttt{ls}} \leftarrow \mathcal{A}^{\texttt{ls}} \cup \{a\}$
```

4.2. Rejected variations

This section describes variations on the proposed SMM algorithm that were tested but ultimately rejected.

4.2.1. Importance threshold

In the proposed SMM algorithm there are currently two checks that decide which active APs should go into deep sleep: a throughput and a coverage check. The first check is aimed at ensuring that an AP a is only turned off if the other APs that are also serving some UE u are able to provide a satisfactory throughput for u without a. However, this check is based on a 10th percentile over all UEs, meaning that there might be some UEs for which a was their only AP, potentially causing severe performance degradation if a was shut off.

This potential flaw was the motivation to introduce a third check based on an importance metric. This importance metric $\mathtt{IM}_{a,u}(t)$ quantifies for each UE u that AP a serves what fraction of u's total received signal power is contributed by AP a. The mathematical definition is similar to Equation 4.2, but does not involve the bit rate of u:

$$\mathtt{IM}_{a,u}(t) = \frac{\hat{\mathbf{h}}_{a,u}(t)\hat{\mathbf{h}}^H_{a,u}(t)}{\sum_{a' \in \mathcal{A}_u(t)} \hat{\mathbf{h}}_{a',u}(t)\hat{\mathbf{h}}^H_{a',u}(t)} \qquad \forall a \in \mathcal{A}^{\text{on}}(t), \forall u \in \mathcal{U}_a(t) \tag{4.6}$$

A particular AP a is then only allowed to go to sleep if its importance to all of its clients is below the importance threshold $\mathtt{IM}^{\mathtt{max,sleep}}$ in all past k coherence blocks. The importance metric is used to judge if a is too important to a particular UE u's throughput for it to be put into the deep sleep mode. For example, the extreme setting of $\mathtt{IM}^{\mathtt{max,sleep}} = 1$ would mean a must not be the only AP for any of its clients. A setting of $\mathtt{IM}^{\mathtt{max,sleep}} = 0.8$ would mean that any UE served by a must get less than 80% of its signal power from a in order for a to still be allowed to go into deep sleep. Mathematically, the condition a has to satisfy to be put into deep sleep mode is formulated as:

$$\mathtt{IM}_{a,u}(i) < \mathtt{IM}^{\mathtt{max,sleep}} \text{ for } \forall u \in \mathcal{U}_a(t), i{=}t{-}k, ..., t{-}1 \qquad \forall a \in \mathcal{A}^{\text{on}} \tag{4.7}$$

When an implementation of the SMM algorithm including this condition was tested on the simulator, the algorithm appeared to be overly cautious with regard to switching off APs. This led to a relatively high energy consumption. The root cause of this cautiousness is that the condition must be satisfied in all of the past k coherence blocks. If there is only one block in which this condition is not met, then an AP cannot go into deep sleep. It is also doubtful how necessary this condition is. Assume that the importance threshold $\mathtt{IM}^{\mathtt{max,sleep}} = 1$ and say UE u is only served by AP a. Recall that this situation is a result of the clustering process and means that there is no other AP that could serve u that falls within the clustering threshold ψ. However, that does not mean that there could not be some other AP that could adequately serve u falling just outside of the clustering threshold. Due to these two reasons, it was decided to remove the importance condition from the SMM algorithm.

4.2.2. No-wake-up radius

In the proposed SMM algorithm, APs are woken up if the throughput of UEs in their vicinity is too low. Given the fact that the transition from deep sleep to active mode takes $T^{\mathtt{t}} = 0.5$ seconds, the impact of deciding to wake up an AP can only be measured after the transition has been completed. This measurement delay might cause, for example, a situation where multiple APs are woken up when the UE throughput in a given area is low before discovering that only waking up one would have sufficed.

Therefore, it might be better if the network would wait until one AP in a given area is fully active in order to measure its impact on the throughput of UEs in its vicinity before making the decision to wake up more APs. This hypothesis was the motivation behind the introduction of a 'no-wake-up radius'. For each AP a^* that is transitioning to the active mode, any other AP a must be further than $d^{\mathtt{min,wake}} = 100$ meters away from a^* to be allowed to wake up. Mathematically, the condition a has to satisfy to be woken up is formulated as:

$$d_{a,a^*} > d^{\texttt{min,wake}} \text{ for } \forall a^* \in \mathcal{A}^{\text{d2a}} \qquad \forall a \in \mathcal{A}^{\text{ds}} \tag{4.8}$$

where $\mathcal{A}^{\text{d2a}}$ denotes the set of APs transitioning to the active mode and d_{a,a^*} denotes the distance between AP a and AP a^* in meters.

During the development process of the SMM algorithm it was discovered, however, that this extra condition did not significantly impact the resulting energy consumption. A probable reason for this is that another AP can still be woken up directly after the AP blocking it has finished transitioning. Additionally, this method might be less effective during a simulation where the traffic load is constant. For a constant traffic load, an equilibrium for the number of active APs is expected to be reached fairly quickly. After this state equilibrium, APs will likely only transition in and out of deep sleep mode sporadically, obviating the need for a condition that blocks APs from waking up simultaneously within a window of 0.5 seconds and only if they are close together. For these reasons, it was decided to remove the no-wake-up radius from the SMM algorithm.

5 Scenarios and results

The performance of the proposed SMM algorithm is evaluated using simulations. This chapter presents the scenarios that were tested in these simulations and analyzes the corresponding results. Section 5.1 presents the derivation of the best AP configuration in terms of the number of AP antennas $M^{\mathtt{AP}}$ and the clustering threshold ψ. This configuration is subsequently used as the testing scenario for the SMM algorithm. Different parameter values are evaluated and a best configuration for the SMM algorithm is derived in Section 5.2.

The values for the fixed parameters used in all simulations can be found in Chapter 3. A comprehensive overview of these values is presented in Table A.1 in Appendix A.

5.1. AP configuration

Before the proposed SMM algorithm can be evaluated, first the best AP configuration for the CF-mMIMO network must be selected in terms of the number of antennas per AP $M^{\mathtt{AP}}$ and the clustering threshold ψ. This addresses the research question on whether deploying a larger number of APs with fewer antennas each or a smaller number of APs with more antennas each is better. To answer this question, three main scenarios are tested: $M^{\mathtt{AP}} = 2$, $M^{\mathtt{AP}} = 4$ and $M^{\mathtt{AP}} = 8$. Using the AP deployment method described in Section 3.1, this results in a deployment of $A = 80$, $A = 40$ and $A = 20$ APs, respectively. For each of these scenarios, a clustering threshold of $\psi = \{0, 3, 6, 9, 12\}$ dB is evaluated for the range of traffic loads. Recall that a clustering threshold of $\psi = 0$ dB essentially turns the network into a regular cellular network as each UE will only connect to the strongest-received AP. This threshold value is included to assess the benefits of a CFN architecture over a traditional cellular architecture. Table 5.1 shows an overview of the parameters and the range of values that were tested. For these tests, the SMM algorithm is essentially turned off. Only the part that allows for light sleep for APs without clients is still enabled. Note that the use of light sleep only affects the energy consumption, not the throughput performance or the network capacity given that APs without clients by definition cannot contribute to any UE's throughput. Additionally, as opposed to the transitions to the deep sleep mode, there is no transition time associated with the transition from the active to light sleep mode or vice versa. Therefore, as there are no downsides, the mechanism that controls the use of light sleep is left enabled for this scenario.

Table 5.1: Configurations tested to select the best AP configuration.

Parameter	Symbol	Value	Unit
Number of AP antennas	$M^{\mathtt{AP}}$	2, 4, 8	-
Number of APs	A	80, 40, 20	-
Clustering threshold	ψ	0, 3, 6, 9, 12	dB
Expected number of active UEs	$\mathbb{E}[U^{\mathtt{active}}]$	15, 30, 45, 60, 75, 90, 105, 120, 135, 150	-

To compare the performance of the different AP antenna configurations and the impact of the clustering threshold ψ, six Key Performance Indicators (KPIs) will be utilized:

- 10th UE throughput percentile (Mb/s): a key metric commonly used by MNOs to assess the network's performance. It measures the throughput performance for the UEs with the worst-case experience and reflects how well the network serves UEs with the least unfavorable channel conditions or those located in the busiest areas where APs are more likely to be highly-loaded. Ensuring that this bottom-tier experience is still acceptable is critical for customer satisfaction. It is also particularly relevant for CFNs as a candidate technology for 6G, as one of the proposed requirements prescribes a certain minimum UE experienced data rate that should be available "ubiquitously across the coverage area" [4]. This is related to the concept of spatial throughput fairness, where, in an ideal CFN, the experienced throughput of a UE is not dependant on its position in relation to its serving APs. The throughput is defined as the time-average throughput experienced by each individual UE during their active periods. The 10th UE throughput percentile is then determined by calculating the 10th percentile value of all of these UE-specific time-averaged throughputs.
- Average UE throughput (Mb/s): another metric to characterize a UE's experienced QoS. The throughput is defined as the time-average throughput experienced by each individual UE during their active periods. The average UE throughput is then determined by calculating the average value over all these UE-specific time-averaged throughputs.
- Average network throughput (Gb/s): defined as the average sum throughput of all UEs in the network. It is used to assess the performance of the network as a whole.
- Average AP PRB utilization: a value between 0 and 1 expressing, on average, what fraction of PRBs are utilized by APs. Given the use of wideband scheduling, the PRB utilization in a given downlink TTI is 1 if an AP is serving at least one UE and 0 if it is not. The average PRB utilization is therefore equal to the fraction of time that an AP is serving at least one UE. If an AP is in the light sleep mode in a given TTI, it has a PRB utilization of 0.
- Average network required power (W): the mean power required by the network during operation, measured in Watts. Unlike energy in Joules, which depends on the simulation duration, power provides a time-independent measure of the energy consumption of the network. It is expected that the $M^{\texttt{AP}} = 2$ setting will consume the most energy since the circuit power often dominates the overall energy consumption [90] and the highest number of APs are deployed in the $M^{\texttt{AP}} = 2$ scenario.
- Average network energy efficiency (Mb/J): defined as the ratio of the total amount of data transmitted over the network to the total energy needed to facilitate said transmissions. A higher value indicates that more data is transmitted per unit of energy consumed, which is desirable from both an economic and a sustainability perspective. The average network throughput (Mb/s) and average network required power (W = J/s) KPIs can be used as substitutes for the total amount of data transmitted and the network required energy, respectively, when reasoning about the average network energy efficiency given that both utilize the same time unit in this work, viz. seconds.

5.1.1. Impact of the traffic load

Recall from Section 3.2 that the traffic load is expressed in terms of the expected number of active UEs $\mathbb{E}[U^{\texttt{active}}]$. The 10th UE throughput percentile and the average UE throughput is shown versus the expected number of active UEs $\mathbb{E}[U^{\texttt{active}}]$ for the distinct cases of $M^{\texttt{AP}}$ in Figure 5.1 and Figure 5.2, respectively. In these figures and the subsequent figures in Section 5.1, the graphs use a solid line to indicate the average value over three simulation seeds. A shaded region around each line indicates the range between the minimum and maximum encountered value over all simulation seeds. There are separate graphs for the different cases of $M^{\texttt{AP}}$, with the lowest value of $M^{\texttt{AP}} = 2$ antennas on the left and the highest value of $M^{\texttt{AP}} = 8$ antennas on the right. Each graph contains five colored lines: one for each of the five clustering thresholds ψ. This section analyzes the impact of the traffic load by looking at the average trend across all cases of $M^{\texttt{AP}}$ and all clustering thresholds ψ. The specific impacts of $M^{\texttt{AP}}$ and ψ will be discussed in later sections.

As the traffic load increases, both the 10th UE throughput percentile and the average UE throughput

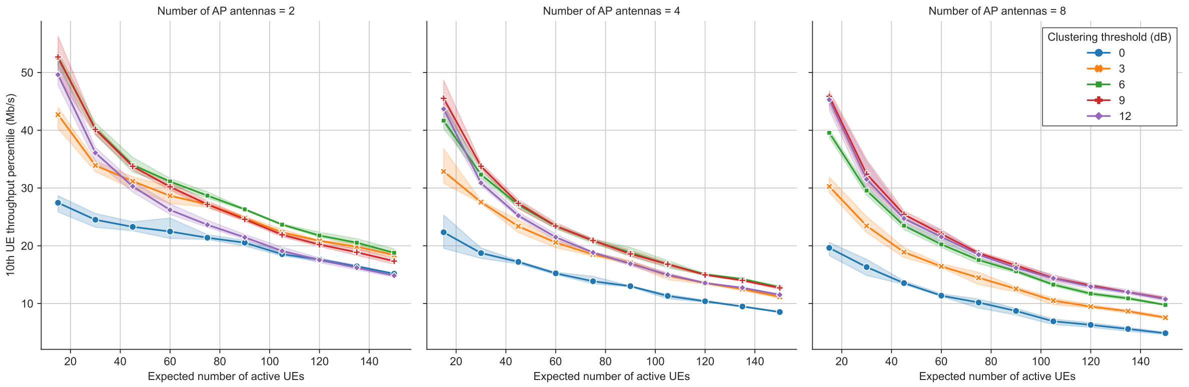


Figure 5.1: 10th UE throughput percentile versus the traffic loads for each clustering threshold ψ and the $M^{\mathtt{AP}} = 2$, $M^{\mathtt{AP}} = 4$ and $M^{\mathtt{AP}} = 8$ antenna configurations.

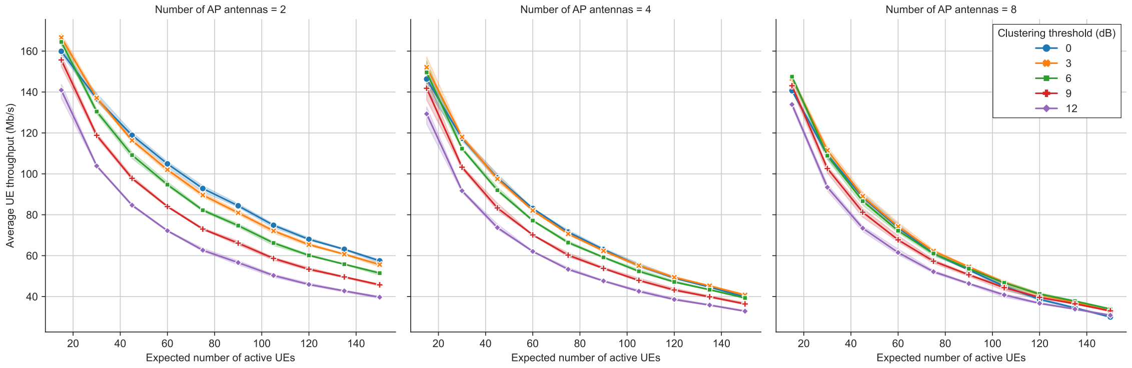


Figure 5.2: Average UE throughput versus the traffic loads for each clustering threshold ψ and the $M^{\mathtt{AP}} = 2$, $M^{\mathtt{AP}} = 4$ and $M^{\mathtt{AP}} = 8$ antenna configurations.

go down. As the same time-frequency resources need to be shared with more UEs, the share of resources that is assigned to any one UE goes down and therefore the UE throughput also goes down. More precisely, as evidenced in the graphs, there is an inversely proportional relationship where $R_u \propto 1/\mathbb{E}[U^{\mathtt{active}}]$.

The average AP PRB utilization is shown versus the expected number of active UEs $\mathbb{E}[U^{\mathtt{active}}]$ for the distinct cases of $M^{\mathtt{AP}}$ in Figure 5.3. As the traffic load grows, the average AP PRB utilization increases. Recall that given the assumption of wideband scheduling the PRB utilization essentially measures what fraction of downlink TTIs an AP was serving at least one UE. At low traffic loads, fewer UEs are active and therefore there are also fewer total APs involved in serving them. As the number of active UEs increases, the number of APs that have at least one UE to serve grows, hence the PRB utilization increases. At lower traffic loads, this increase is faster than at the higher traffic loads. This is because, as the traffic load increases, it becomes ever less likely that the newly added active UEs would include an AP in their cluster that is not already serving at least one UE, given that the number of APs that do not serve any UE is decreasing. Of course, the PRB utilization can only keep increasing with the traffic load as long as there are still APs left that do not already serve at least one UE in at least some of the TTIs. After that point, the PRB utilization equals 1 and can no longer increase. The graph shows that this occurs for each of the clustering thresholds as the traffic load exceeds a certain level for the $M^{\mathtt{AP}} = 8$ antenna case, while for the other $M^{\mathtt{AP}}$ settings this would occur for loads beyond the considered traffic load range.

The average network throughput is shown versus the expected number of active UEs $\mathbb{E}[U^{\mathtt{active}}]$ for the distinct cases of $M^{\mathtt{AP}}$ in Figure 5.4. As the traffic load grows, the average network throughput generally increases. This is because of two reasons. First, there is multi-user diversity gain where the scheduler is able to better exploit the individual channel conditions of UEs as the number of UEs in the system increases. Secondly, a more important reason is that the higher the load, the higher the

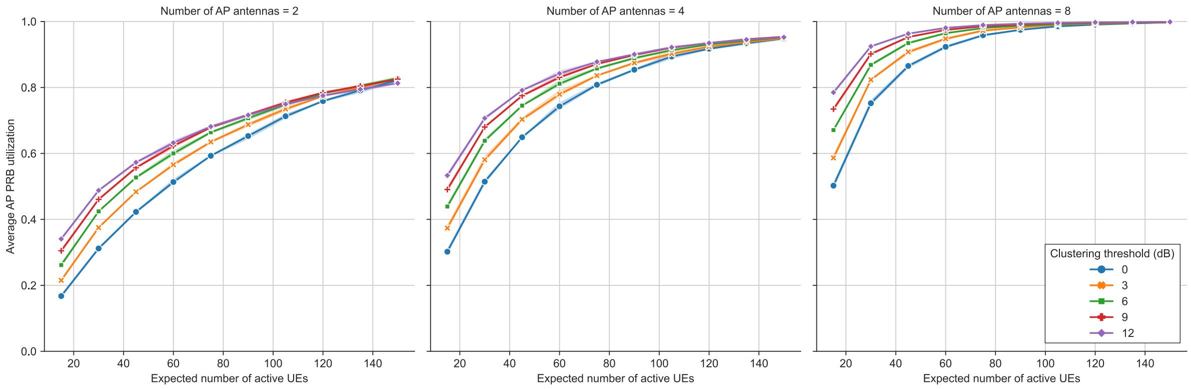


Figure 5.3: Average AP PRB utilization versus the traffic loads for each clustering threshold ψ and the $M^{\mathtt{AP}} = 2$, $M^{\mathtt{AP}} = 4$ and $M^{\mathtt{AP}} = 8$ antenna configurations.

chance that an AP is serving at least one UE, hence more resources can be utilized to enhance the network throughput. This is evidenced by Figure 5.3 which shows that the PRB utilization increases as the traffic load grows. The increase in the average network throughput tapers off at the highest traffic loads for two reasons. First, the additional multi-user diversity gains decrease if there are ever more UEs to choose between. Given that PF scheduling compels the scheduler to also serve these UEs with poorer channel conditions at some point, the advantage of having more UEs to choose from yields ever lower additional gains. Second, at some point all AP resources are in use and therefore no further gain is possible in that regard.

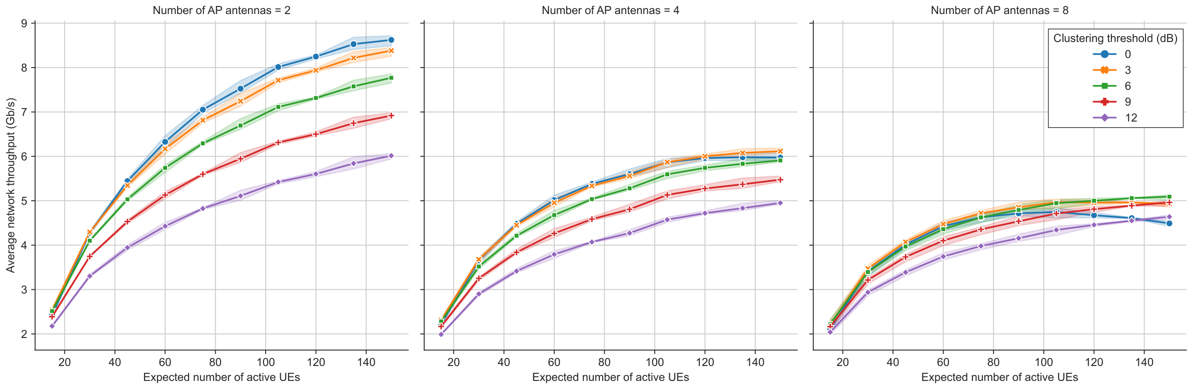


Figure 5.4: Average network throughput versus the traffic loads for each clustering threshold ψ and the $M^{\mathtt{AP}} = 2$, $M^{\mathtt{AP}} = 4$ and $M^{\mathtt{AP}} = 8$ antenna configurations.

For some combinations of the clustering threshold and number of antennas, the graph shows that the average network throughput not only tapers off at higher traffic loads, but actually starts to decrease. The case where the number of antennas $M^{\mathtt{AP}} = 8$ and the clustering threshold $\psi = 0$ dB is the clearest example of this trend. It is unclear from the graphs whether this trend will occur across all clustering thresholds and antenna cases. However, it appears that it at least also occurs for the $M^{\mathtt{AP}} = 8$ antenna case with a clustering threshold of $\psi = 3$ dB and ostensibly also for the case where $M^{\mathtt{AP}} = 4$ and $\psi = 0$ dB. This means that the average UE throughput declines at a faster rate than can be compensated for by the fact that there are more UEs in the network. A likely cause for this effect is the use of SUS combined with the specific choice for an orthogonality threshold of $\gamma^{\mathtt{th}} = 0.5$. Assuming that the fraction of UEs that have sufficiently orthogonal enough channels to be co-scheduled stays approximately constant, as the expected number of active UEs increases, the number of co-schedulable UEs also increases. This means that ever more UEs will be scheduled as the expected number of active UEs increases. However, due to the use of ZF, this also means that ever more beamforming gain needs to be sacrificed in order to create nulls in the direction of ever more other co-scheduled UEs. This effect is likely non-linear as the traffic load increases, meaning that after a certain point, there is a disproportionate decline in the

SINR and therefore UE throughput, causing the observed decline in the average network throughput.

To verify this hypothesis, extra simulations were performed with a SUS orthogonality threshold of 0 and 1 for the $M^{\text{AP}} = 8$ antenna case. Figure 5.5 shows the average network throughput versus the expected number of active UEs $\mathbb{E}[U^{\text{active}}]$ for SUS orthogonality thresholds of $\gamma^{\text{th}} = \{0, 0.5, 1\}$. A value of 0 effectively disables multi-user scheduling, whereas a value of 1 schedules all UEs simultaneously. If the above explanation is valid, then for a value of $\gamma^{\text{th}} = 0$, the average network throughput should not decrease, but only plateau, given that no beamforming gain needs to be sacrificed due the use of single-user scheduling. This is indeed what the graph shows. Likewise, the graph shows that when even more UEs are co-scheduled than in the main scenario, i.e. when $\gamma^{\text{th}} = 1$ instead of $\gamma^{\text{th}} = 0.5$, there is a more dramatic decrease in the average network throughput after a certain traffic load. This effect can likely be overcome by choosing a different precoding scheme such as Adaptive Regularized Zero-Forcing (ARZF) [91] or MMSE [92], which effectively pursue a better balance between attained beamforming gain and the level of inter-UE interference. Exploring such precoding alternatives falls outside the scope of this study.

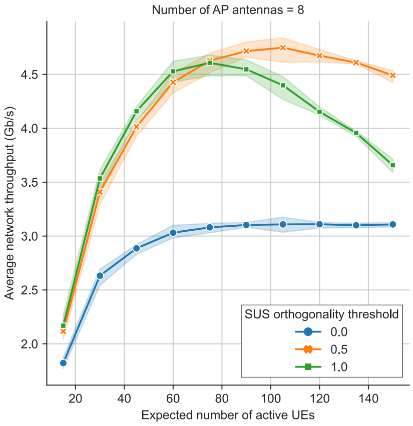


Figure 5.5: Average network throughput versus the traffic loads for SUS clustering thresholds of $\gamma^{\text{th}} = \{0, 0.5, 1\}$.

Moving on, Figure 5.6 shows the average network required power versus the expected number of active UEs $\mathbb{E}[U^{\text{active}}]$ for the distinct cases of M^{AP}. As the expected number of active UEs increases, the required power also grows. There are two main causes that explain this intuitively expected trend. Firstly, for the lower traffic loads, it is much more likely that some APs have no clients in a particular coherence block and are therefore put into light sleep, saving energy. As the traffic load increases, ever fewer APs have no clients, meaning that light sleep cannot be used as often, causing an increase in the average network required power. This reasoning also applies to APs that *do* have clients in a coherence block, but whose clients are not scheduled during a TTI. Those APs do not save energy by using light sleep, but simply because they are not transmitting. As the traffic load increases, ever fewer APs have no scheduled UEs during a TTI, as shown by the average PRB utilization curves in Figure 5.3. The slope of the average network required power curve flattens at the higher traffic loads as the average PRB utilization curve also flattens as it approaches its maximum value. Secondly, as the expected number of active UEs grows, so does the average network throughput, as shown in Figure 5.4. Given that both the fronthaul and CPU energy consumption models incorporate UE throughput, this also leads to an increase in the network required power. As the traffic load grows, the rate of growth slows as the network approaches the maximum attainable network throughput. The curve for the average network required power does not decrease after a certain traffic load for any of the clustering thresholds or any of the distinct cases of M^{AP} such as the average network throughput does for some cases. The first effect is apparently more influential than the second effect.

The average network energy efficiency versus the expected number of active UEs $\mathbb{E}[U^{\text{active}}]$ for the distinct cases of M^{AP} is depicted in Figure 5.7. The top and bottom sets of graphs visualize the same data. The difference is that the bottom graphs each use their own vertical axis scaling in order to make it easier to see the general trend in the average network required power as the number of UEs increases. Broadly speaking, the network energy efficiency increases as the traffic load grows. In this

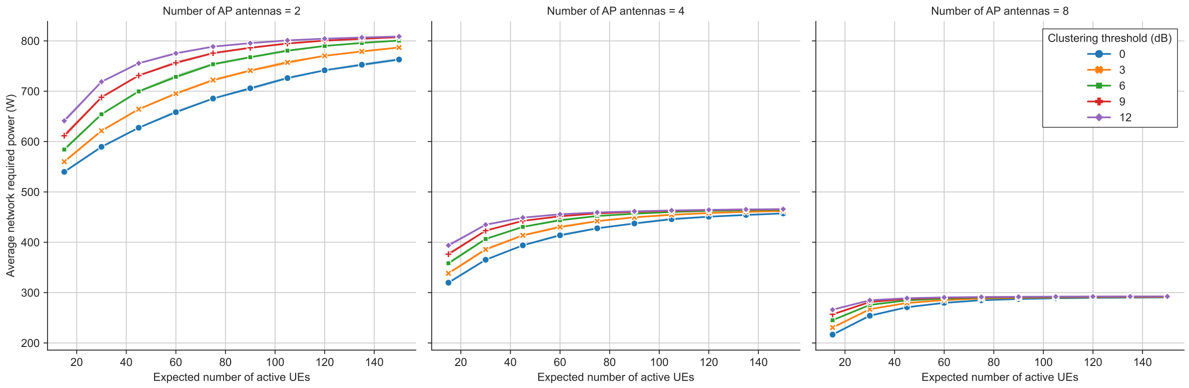


Figure 5.6: Average network required power versus the traffic loads for each clustering threshold ψ and the $M^{\mathtt{AP}} = 2$, $M^{\mathtt{AP}} = 4$ and $M^{\mathtt{AP}} = 8$ antenna configurations.

scenario where all APs are on, the fixed part of an AP's energy consumption dominates the overall energy consumption, as evidenced by Figure 5.6. While the average network required power rises slowly, the average network throughput rises more sharply as the scheduler can now more optimally exploit the channel conditions offered by a larger group of active UEs. The result is that the energy efficiency increases as the traffic load grows. For some cases, the average network energy efficiency start to decrease at the higher traffic loads. This is for the same reason that the average network throughput starts to decline for those same cases as explained earlier.

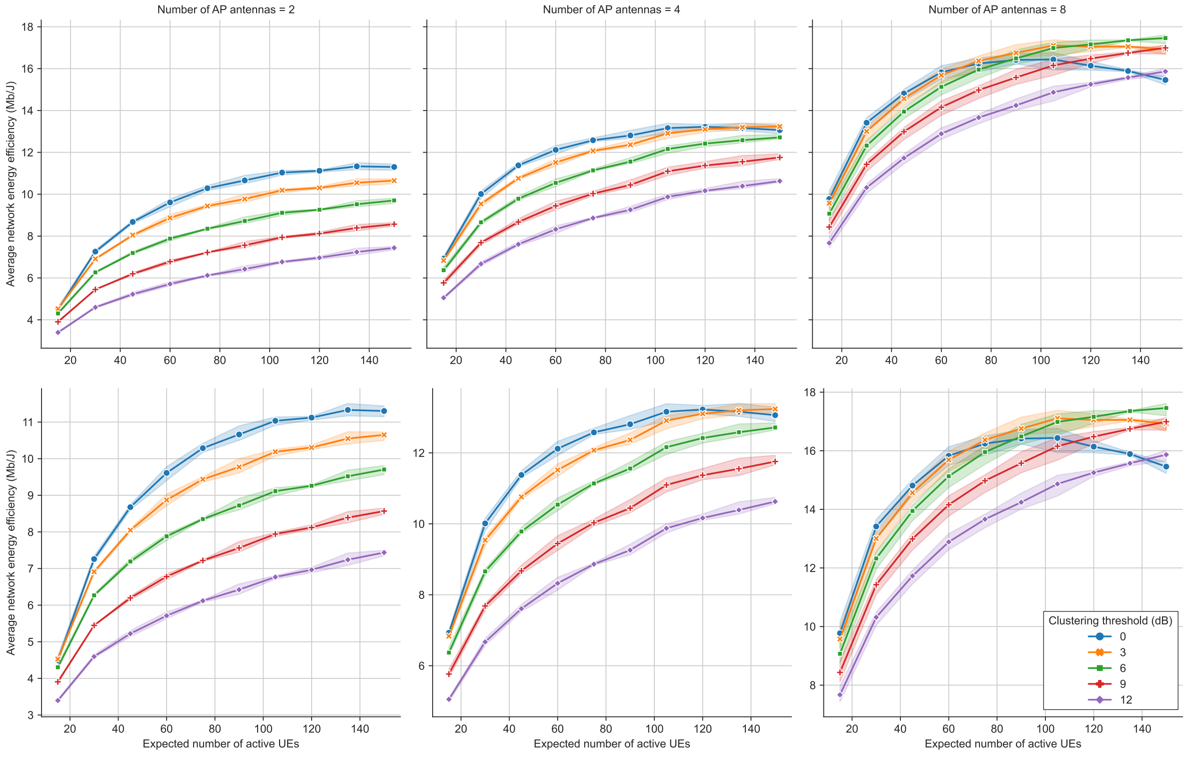


Figure 5.7: Average network energy efficiency versus the traffic loads for each clustering threshold ψ and the $M^{\mathtt{AP}} = 2$, $M^{\mathtt{AP}} = 4$ and $M^{\mathtt{AP}} = 8$ antenna configurations. The top and bottom graphs visualize the same data. The top graphs have a shared vertical axis, whereas the bottom graphs each have an individually scaled vertical axis.

5.1.2. Impact of the clustering threshold

The graph of the 10th UE throughput percentile, presented in Figure 5.1, shows the advantage of a CFN over a traditional cellular network. The case where the clustering threshold $\psi = 0$ dB, emulating a traditional cellular network, clearly results in worse throughputs for the UE with poor channel conditions versus the cases where $\psi > 0$ dB, i.e. when the network operates as a CFN. For every antenna configuration M^{AP}, going from $\psi = 0$ dB to the optimal ψ results in an approximately 50-100% improvement of the 10th UE throughput percentile for low(er) loads. It must be noted, however, that the improvements in the 10th UE throughput percentile come at the cost of the average UE throughput as evidenced by Figure 5.2. For most antenna configurations M^{AP} and traffic loads, a clustering threshold of $\psi = 0$ dB actually has the highest average UE throughput. Generally, each step up in the clustering threshold ψ corresponds with a decline in the average UE throughput for a given traffic load. As the average network throughput is simply the sum throughput over all UEs, the same trend can be seen in Figure 5.4. Nevertheless, the decrease in the average UE throughput is less dramatic than the increase in the 10th UE throughput percentile. For example, when looking at the corresponding decrease in the average UE throughput for that 50-100% improvement in the 10th UE throughput percentile between $\psi = 0$ dB and the optimal ψ dB, it is only approximately 5-10%. Essentially, these results confirm that a CFN sacrifices some performance of the UEs with the best channel conditions to ensure that the UEs with poorer channel conditions are guaranteed a better throughput, which matches with the claim that a CF-mMIMO network can provide a more spatially uniform quality of service.

The improvement in the 10th UE throughput percentile is largest when going from $\psi = 0$ dB to $\psi = 3$ dB. Recall from Section 3.4.2 that a clustering threshold of $\psi = 3$ dB means that a UE will be served by the strongest-received AP and all other APs that it receives with a signal strength that is at most 3 dB less than the strongest-received AP. As the difference in signal strength within the UE's cluster is at most 3 dB, all APs in the cluster have a relatively similar channel gain to the UE, resulting in a roughly equal contribution to the experienced SINR, and therefore throughput, of the UE. Going up to $\psi = 6$ dB further increases the UE's throughput, while the step up to $\psi = 9$ dB only results in a slight improvement of the UE's throughput or even a slight decline. At $\psi = 12$ dB, one can see a noticeable decrease in the 10th UE throughput percentile. This is due to the fact that the APs in the UE's cluster that have a signal strength that is 9 to 12 dB worse than the strongest-received AP only contribute a small fraction of the resulting SINR. In fact, by forcing these 'weak' APs to serve the UE, time-frequency resources are taken away from serving UEs for which those APs have a more favorable channel. The effect is a lower overall throughput for the average UE.

As depicted in Figure 5.3, the PRB utilization of APs increases with higher clustering thresholds. This can be explained by the fact that a higher clustering threshold ψ means that each UE will be served by a larger cluster of APs. Looking from the opposite perspective, this also means that an AP will have more UEs to serve. This increases the likelihood that an AP will be serving at least one UE in a given downlink TTI, therefore increasing the average AP PRB utilization. The same reasoning also applies to the average network required power visualized in Figure 5.6. For a higher clustering threshold ψ, the required power grows as more APs will be involved in data transmissions in a given TTI. At the higher traffic loads, the difference in the average network power required between the clustering thresholds decreases. The same goes for the difference in the average PRB utilization. This is because initially, at low traffic loads, there are still APs left that are not serving UEs in a particular TTI and going to a higher clustering threshold means that more of those 'inactive' APs will be participating, causing a significant jump in the average PRB utilization. However, as the traffic load increases, there are ever fewer APs left that do not serve at least one UE during a TTI, even for a clustering threshold of $\psi = 0$. Therefore, using a higher clustering threshold will cause an ever smaller increase in average PRB utilization as the traffic load grows.

Recall that the average network energy efficiency, depicted in Figure 5.7, is calculated using two other KPIs: the average network throughput and the average network required power. The graph shows that for the lower traffic loads, the lower clustering thresholds ψ have a higher energy efficiency. Starting with the lowest clustering threshold, at the higher traffic loads, the energy efficiency starts to decline, implying that the higher clustering thresholds are more efficient at the higher traffic loads. Before this decline sets in, it appears that the gap between the different clustering thresholds actually increases. This is the result of two effects:

- The average network throughput (see Figure 5.4) is generally higher for smaller clustering threshold values, however, as previously explained, due to the loss in beamforming gain there is a certain point in the traffic load after which the average network throughput starts to decrease.

The traffic load at which this decrease occurs seems to be lowest for the lowest clustering threshold. Therefore, the graph implies that at some traffic load, the order will have completely reversed, i.e. the average network throughput will be highest for the highest clustering throughputs.

- The average network required power (see Figure 5.6), is highest for the highest clustering threshold, as explained above. However, the relative gap between the clustering threshold does become smaller as the traffic load grows.

The combination of these effects mean that at lower traffic loads, the network is more energy efficient for lower clustering thresholds, with the relative gap widening as the traffic load increases due to the relative gap narrowing in the average network required power. However, the decline in the average network throughput has a more significant effect size for the cases where it occurs (most notably for $M^{\mathtt{AP}} = 8$ antennas), causing a decline in the energy efficiency for the lower clustering thresholds first. As previously mentioned, it is not certain that this effect will also occur for the case where $M^{\mathtt{AP}} = 2$ or to what extent. Assessing this would require simulations at traffic loads beyond $\mathbb{E}[U^{\mathtt{active}}] = 150$. For this study, this was deemed infeasible given the significant run time of such simulations.

5.1.3. Impact of the AP antenna configuration

In terms of the UE throughput performance, it is clear that the $M^{\mathtt{AP}} = 2$ antenna configuration has the best performance. Both the 10th UE throughput percentile, depicted in Figure 5.1, as well as the average UE throughput, depicted in Figure 5.2, show this. This is the net result of the interaction between two opposite effects:

- The configuration with $M^{\mathtt{AP}} = 2$ antennas per AP has the largest number of APs in the network, namely 80 APs. In the case of $M^{\mathtt{AP}} = 8$, there are only 20 APs. In the former case the overall AP deployment density is higher, meaning that the average distance between a UE and the APs in its vicinity decreases, i.e. there will be more APs close to a UE. For a given clustering threshold ψ, a higher AP density will likely cause UEs to have larger AP clusters, meaning more APs will contribute to a UE's received signal. Additionally, as the distance is decreased, this also means that the average path loss is reduced. This means a UE will experience a better propagation gain towards APs in its vicinity.
- Another consequence, however, is that the maximum beamforming gain for the $M^{\mathtt{AP}} = 2$ antenna case is only 3 dB, whereas the $M^{\mathtt{AP}} = 8$ antenna case has a maximum beamforming gain of 9 dB. This negatively affects the received signal strength and therefore SINR of a UE, lowering the attainable throughput.

As evidenced by the UE throughput graphs, the AP densification effects are stronger than the loss in beamforming gain. It must be noted that the impact of the increased densification on the resulting increase in cluster size is relatively weak. Figure 5.8 shows the average number of APs per UE cluster versus the expected number of active UEs $\mathbb{E}[U^{\mathtt{active}}]$ for the distinct cases of $M^{\mathtt{AP}}$ and for each clustering threshold ψ. The graphs show that the cluster sizes for a given clustering threshold are insensitive to the traffic load. This is because the clustering procedure selects a set of serving APs independently for each UE without regard for the other UEs that those APs might be serving. More importantly, Figure 5.8 shows that for a given clustering threshold there is only a minor increase in the cluster size as the number of antennas per AP $M^{\mathtt{AP}}$ decreases. The fact that there is only a minor increase is caused by the coverage requirement used in the clustering procedure. Recall that an AP a can only be included in a UE u's cluster if that AP provides coverage at u's location in addition to satisfying the clustering threshold requirement. To satisfy the coverage requirement, the average channel gain $G^{\mathtt{ul}}_{a,u}$ between AP a and UE u must exceed the coverage threshold $G^{\mathtt{ul,min}}$. The UL superscript indicates that the threshold is derived from an uplink throughput requirement as the coverage of a network is primarily determined by the limited transmit power available at the UE-side. The average channel gain $G^{\mathtt{ul}}_{a,u}$ includes an estimate for the combiner gain at the AP-side and is based on the number of AP antennas. This means that as the number of antennas per AP decreases, the likelihood that an AP provides coverage to a given UE decreases. Consequently, some APs that would be included in a UE's cluster based on the clustering threshold are actually not included due to the coverage requirement. Hence, the increased densification only marginally improves the cluster size.

The average AP PRB utilization, depicted in Figure 5.3, increases with the number of antennas at each

AP. This is related to the fact that the cluster sizes for a given clustering threshold only increases marginally as the number of antennas per AP decreases. This means that roughly the same number of APs will be involved in serving the active UEs in a given TTI across the different antenna cases. While the number of APs per cluster stay relatively similar, the number of deployed APs doubles each time when going from $M^{\mathtt{AP}} = 8$ via $M^{\mathtt{AP}} = 4$ to $M^{\mathtt{AP}} = 2$. This means that an ever smaller fraction of APs will be involved in serving the UEs in a given downlink TTI. With a larger fraction of APs idle in a given TTI, the PRB utilization will therefore be lower as the number of antennas per AP decreases.

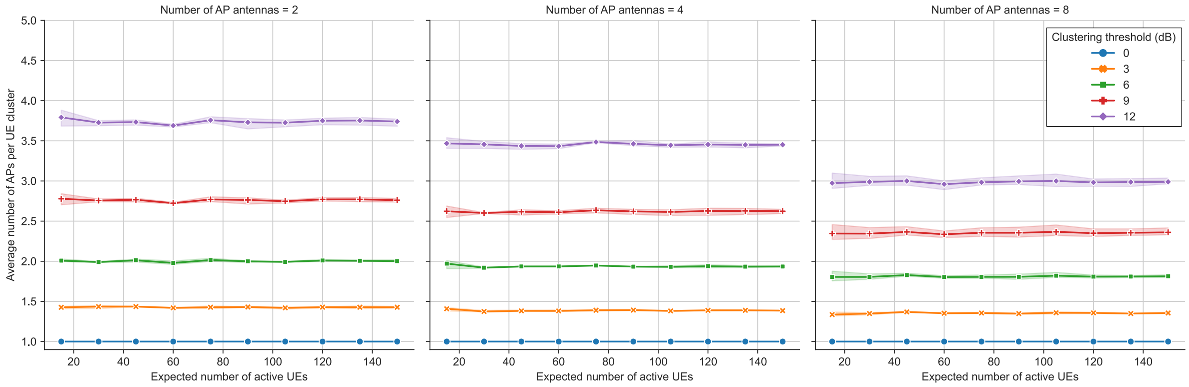


Figure 5.8: Average UE cluster size versus the traffic loads for the $M^{\mathtt{AP}} = 2$, $M^{\mathtt{AP}} = 4$ and $M^{\mathtt{AP}} = 8$ antenna configurations.

The average network required power becomes significantly higher as the number of AP antennas decreases (and hence the number of APs increases), as shown in Figure 5.6. This effect can be explained by the fact that the circuit power dominates the total required power of an AP. For the $M^{\mathtt{AP}} = 2$ antenna case, for example, there are 80 total APs, each with 8 W of fixed power. Even though the power required for the RF chain is lower compared to the $M^{\mathtt{AP}} = 8$ antenna case, the difference is relatively minor. For $M^{\mathtt{AP}} = 8$, the required power for the RF chain is 1.6 W and for $M^{\mathtt{AP}} = 2$ it is 0.4 W.

While more bits are transferred for the $M^{\mathtt{AP}} = 2$ antenna case in comparison with the $M^{\mathtt{AP}} = 8$ antenna case, the energy required for this is significantly higher. Figure 5.7 illustrates that the latter effect dominates: the average network energy efficiency indeed increases as the number of AP antenna decreases.

Recall that the auxiliary research question asks which AP deployment strategy yields better QoS and energy efficiency in a CF-mMIMO network: deploying a larger number of APs with fewer antennas each, or a smaller number of APs with more antennas each. Based on the simulations, *deploying a smaller number of APs with more antennas each provides better network energy efficiency, but worse QoS as quantified by the 10th UE throughput percentile*. To attain more insight into this tradeoff, Table 5.2 shows the average network energy efficiency and 10th UE throughput percentile of the best performing clustering threshold for three traffic loads for the $M^{\mathtt{AP}} = 2$ and $M^{\mathtt{AP}} = 8$ antenna cases. The three selected traffic loads are: low ($\mathbb{E}[U^{\mathtt{active}}] = 30$), medium ($\mathbb{E}[U^{\mathtt{active}}] = 75$) and high ($\mathbb{E}[U^{\mathtt{active}}] = 135$). If each traffic load is valued equally, the average increase in the 10th UE throughput percentile when going from $M^{\mathtt{AP}} = 8$ antennas to $M^{\mathtt{AP}} = 2$ antennas is 49.60%. However, this improvement comes at the cost of a 38.87% decrease in the average network energy efficiency. Whether this trade-off is worth it is a question that each MNO might answer differently. Two more factors that play a role in this decision that should be mentioned are the average PRB utilization and the financial considerations in terms of deployment and operating costs. An advantage of deploying more APs, in the case of $M^{\mathtt{AP}} = 2$ antennas, is that more UEs can be supported simultaneously as evidenced by the lower average PRB utilization at the highest considered traffic load. A disadvantage of this deployment scenario is that it will cost more money both to deploy and to maintain more APs. Not only will the energy bills be higher, but there are also higher maintenance costs given that more APs need to be kept operational.

5.2. SMM algorithm

Based on the results presented in Section 5.1, the $M^{\mathtt{AP}} = 2$ antenna configuration is selected as the deployment scenario to evaluate the SMM algorithm as it provides the highest UE throughputs.

Table 5.2: Comparison of the $M^{\mathtt{AP}} = 8$ and $M^{\mathtt{AP}} = 2$ antenna cases. The values are taken from the best performing clustering threshold ψ for each individual case.

10th UE throughput percentile

Traffic load ($\mathbb{E}[\mathbf{U}^{\mathtt{active}}]$)	$\mathbf{M}^{\mathtt{AP}}=\mathbf{8}$ **(Mb/s)**	$\mathbf{M}^{\mathtt{AP}}=\mathbf{2}$ **(Mb/s)**	**Relative increase**
Low (30)	32.41	40.28	24.29%
Medium (75)	18.73	28.68	53.12%
High (135)	11.98	20.54	71.40%

Average network energy efficiency

Traffic load ($\mathbb{E}[\mathbf{U}^{\mathtt{active}}]$)	$\mathbf{M}^{\mathtt{AP}}=\mathbf{8}$ **(Mb/J)**	$\mathbf{M}^{\mathtt{AP}}=\mathbf{2}$ **(Mb/J)**	**Relative decrease**
Low (30)	13.42	7.26	45.90%
Medium (75)	16.37	10.29	37.16%
High (135)	17.05	11.33	33.55%

Additionally, with its associated deployment of $A = 80$ APs, it provides the most opportunities for the SMM algorithm to conserve energy. Given that optimizing for the 10th UE throughput percentile is a typical MNO objective, a clustering threshold of $\psi = 6$ dB is adopted as it provides the highest 10th UE throughput percentile for the $M^{\mathtt{AP}} = 2$ antenna case. To evaluate the performance of the SMM algorithm, different parameter values are tested and compared to the scenario analyzed in Section 5.1 where the deep sleep mode is not available. This scenario is referred to as the baseline scenario. Recall that the baseline scenario *does* include the use of light sleep. The parameter values can be found in Table 5.3. In total, 144 SMM configurations are tested for a range of six traffic loads. Each of these 864 combinations is run on the simulator with three different seeds.

Table 5.3: Test configurations for the SMM algorithm.

Parameter	**Symbol**	**Value**	**Unit**
Number of AP antennas	$M^{\mathtt{AP}}$	2	-
Number of APs	A	80	-
Clustering threshold	ψ	6	dB
To-deep-sleep-threshold	$R^{\mathtt{min,sleep}}$	$R^{\mathtt{min,UE}} \cdot \{1, 1.5, 2.0, 2.5\}$	Mb/s
To-active-threshold	$R^{\mathtt{max,wake}}$	$R^{\mathtt{min,sleep}} \cdot \{0.5, 0.75, 1.0\}$	Mb/s
AP neighborhood threshold	$G^{\mathtt{nh,min}}$	$-130, -115, -100$	dB
Window size	k	$22, 44, 87, 131$	Coherence blocks (in the time domain)
Expected number of active UEs	$\mathbb{E}[U^{\mathtt{active}}]$	$20, 40, 60, 80, 100, 120$	-

To compare the effectiveness of the different SMM algorithm configurations, both the throughput performance and the energy consumption need to be considered. For the former the generally most relevant metric from an MNO's perspective is the 10th UE throughput percentile as explained in Section 5.1. However, it is harder to quantify the energy consumption. Given that the simulator will only test one traffic load at a time, some method is needed to combine the average network required power for each individual traffic load into one number that can assess the performance of a particular SMM algorithm configuration. Taking a simple average would be misleading, as the traffic load fluctuates over the course of a day with unequal durations across the load levels. Instead, a better method would be to estimate the total network energy consumption over a day. Based on the traffic data from Ericsson [70], shown in Figure 3.5, the individual results for the average network required power at each traffic load can be multiplied by the number of hours that each traffic load occurs in order to derive the daily network energy consumption in kWh. Table 3.1 shows the number of hours spent within each traffic level.

Naturally, the SMM algorithm should only shut off APs that are not necessary to maintain a certain minimum QoS level for the UEs in the network. A minimum 10th UE throughput percentile target $R^{\mathtt{min,UE}}$ is imposed as an embodiment of this QoS requirement. This is in a line with a typical methodology applied by MNOs where the network is generally planned and optimized such that a 10th UE throughput percentile target is achieved during peak hours. It is not necessary to devise a method to obtain a daily 10th UE throughput percentile, such as for the energy consumption, as the goal is not to maximize or minimize this metric. Instead, the underlying requirement is that this minimum 10th UE throughput percentile target should be satisfied during each hour of the day. Therefore, any configuration of the

SMM algorithm that does not meet this minimum target in any hour of the day, hence for any of the considered loads, is disregarded. In this work, a minimum 10th UE throughput percentile of $R^{\texttt{min,UE}} = 20$ Mb/s is assumed to be a reasonable target. Given a peak hour traffic load of $\mathbb{E}[U^{\texttt{active}}] = 120$, a network deployment was derived such that this target is satisfiable when no AP is in the deep sleep mode. Figure 5.1 shows that a 10th UE throughput percentile of 20 Mb/s is attainable at the $\mathbb{E}[U^{\texttt{active}}] = 120$ traffic load for the case where $M^{\texttt{AP}} = 2$ and $\psi = 6$.

As shown in Table 5.3, the threshold used by the SMM algorithm to determine whether an AP can go to sleep, the to-deep-sleep-threshold $R^{\texttt{min,sleep}}$, is defined as a certain fraction of this target $R^{\texttt{min,UE}} = 20$ Mb/s. Even though the SMM algorithm also uses a 10th percentile, $R^{\texttt{min,UE}}$ is not used directly as the to-deep-sleep-threshold, as the algorithm operates on a much smaller timescale than the hourly level for which the 10th UE throughput percentile target is typically pursued. More specifically, the SMM algorithm operates on a timescale of approximately 250-1500 ms as determined by the window size k. Given that channel conditions can vary rapidly on such a small timescale, there are many opportunities for the to-deep-sleep-metric to have a high enough value for an AP to be put to sleep while its availability might actually be crucial to the UEs in its vicinity shortly after. Given that a full active-to-deep-sleep-to-active cycle takes at least 1 second, this means that putting an AP into deep sleep is a consequential decision that should not be taken haphazardly. Therefore, it might be beneficial to use a higher value for the to-deep-sleep-threshold than the 10th UE throughput percentile target to prevent premature switch-off decisions. The threshold used by the SMM algorithm to determine whether an AP should be woken up, the to-active-threshold $R^{\texttt{max,wake}}$, is defined as a fraction of the to-deep-sleep-threshold $R^{\texttt{min,sleep}}$ as it would not make sense to have higher to-active-threshold than a to-deep-sleep-threshold. If this were the case, there could be a ping-pong effect where APs that are put into the deep sleep mode are then immediately woken up again. Note that this ‘fraction notation’ is only used in this section. During operation, the $R^{\texttt{min,sleep}}$ and $R^{\texttt{max,wake}}$ thresholds are expressed in Mb/s as described in Section 4.1.

Figure 5.9 shows the performance of all SMM algorithm configurations for three selected traffic loads: low ($\mathbb{E}[U^{\texttt{active}}] = 20$), medium ($\mathbb{E}[U^{\texttt{active}}] = 60$) and high ($\mathbb{E}[U^{\texttt{active}}] = 120$). The graph showing the performance of the other three traffic loads can be found in Figure B.1 in Appendix B. These figures consist of a matrix of graphs, where the graphs in each column have the same traffic load and the graphs in each row have the same to-deep-sleep-threshold $R^{\texttt{min,sleep}}$. The graphs display the average network required power on the horizontal axis and the 10th UE throughput percentile on the vertical axis. Within each graph, the color of the marker represents the window size k, the marker type represents the to-active-threshold $R^{\texttt{max,wake}}$ and the marker size represents the AP neighborhood threshold $G^{\texttt{nh,min}}$.

Before the optimal parameter settings are derived, the trends with regard to each parameter are first observed and explained in the sections below. Section 5.2.5 presents the derivation of the best configuration.

5.2.1. Impact of the to-deep-sleep-threshold

The to-deep-sleep-threshold $R^{\texttt{min,sleep}}$ has a clear impact on both the resulting average network required power and the 10th UE throughput percentile. Figure 5.9 shows that, in general, the lower the to-deep-sleep-threshold $R^{\texttt{min,sleep}}$, the lower the average network power required and 10th UE throughput percentile are. This makes intuitive sense as lowering the bar for an AP to be put to sleep means it will be put to sleep more often and therefore conserve more energy while thereby simultaneously adversely affecting the throughput performance of UEs.

5.2.2. Impact of the to-active-threshold

The different settings of the to-active-threshold are indicated by the different marker types in Figure 5.9. When looking at groupings of the same colored and sized markers, i.e. when keeping all other settings equal, a pattern emerges. Looking in the direction starting at the bottom left and going diagonally towards the top right of the graphs, one generally first encounters the circle, then the cross and finally the square, corresponding to a to-active-threshold value of 0.5, 0.75 and 1.0, respectively. Put more formally, the lower the to-active-threshold, the lower the 10th UE throughput percentile and average network required power. This makes intuitive sense, as a lower value of the to-active-threshold raises the bar for APs to be woken up. This means deep sleeping APs will stay asleep for longer, thereby lowering the energy consumption of the network and simultaneously negatively impacting the 10th UE

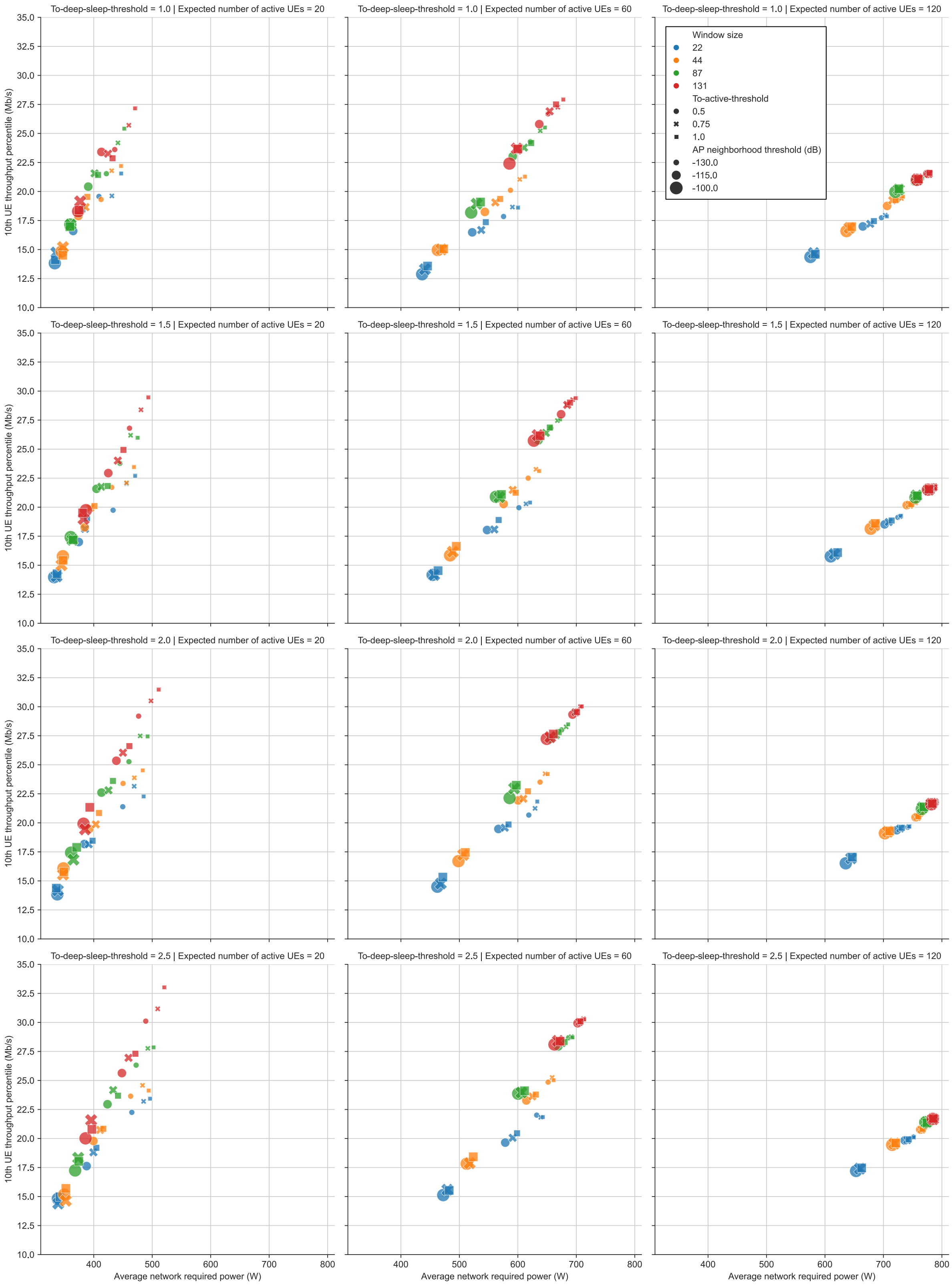


Figure 5.9: Performance of all SMM algorithm configurations for 3 selected traffic loads $\mathbb{E}[U^{\texttt{active}}] = \{20, 60, 120\}$. The graphs in each column have the same traffic load and the graphs in each row have the same to-deep-sleep-threshold $R^{\texttt{min,sleep}}$. Note that each graph contains 36 markers, however, it can be difficult to see all of them as values sometimes overlap.

throughput percentile.

5.2.3. Impact of the AP neighborhood threshold

By and large, Figure 5.9 shows that a higher AP neighborhood threshold, indicated by the larger marker sizes, results in a lower 10th UE throughput percentile and lower average network required power. Recall that the AP neighborhood threshold is used to construct a set of neighbors in the vicinity of a given AP a. The to-active-metric is based on the 10th percentile of the UE throughputs over that set of neighbors. The higher the AP neighborhood threshold, i.e. making the propagation 'radius' around an AP of UEs to consider in the calculation of the to-active-metric smaller, the more likely it is that the set of neighboring UEs was empty in the window consisting of the past k coherence blocks. In such cases, the to-active-metric cannot be calculated and therefore the associated AP is not allowed to wake up. This means that APs are more likely to stay in the deep sleep mode for longer, resulting in a lower average network required power and 10th UE throughput percentile.

5.2.4. Impact of the window size

In Figure 5.9, the window size is indicated by the marker colors. In general, smaller window sizes lead to a lower average network required power as well as a lower 10th UE throughput percentile. A shorter window size increases the likelihood that the $R_a^{\text{on}}(t) > R^{\texttt{min,sleep}}$ requirement is satisfied as the variations in channel conditions are typically less smoothed out compared to longer averaging windows. In other words, the SMM algorithm is less conservative about putting APs into deep sleep as the window size become smaller. However, this effect also applies in the opposite direction: a shorter window size also increases the likelihood that the $R_a^{\text{ds}}(t) < R^{\texttt{max,wake}}$ requirement is satisfied, meaning that deep sleeping APs are more likely to be woken up. The result is a higher likelihood of a ping-pong effect where APs may often go into and out of deep sleep. Consequently, a large fraction of APs is in a transition state at any given time. Figure 5.10 shows two graphs that depict what fraction of APs are in each state over the course of the simulation. Both graphs visualize a simulation with the same Random Number Generator (RNG) seed and SMM settings, except for the window size: the left graph corresponds to a window size of $k = 22$ coherence blocks and the right graph corresponds to a window size of $k = 87$ coherence blocks. It is clear from the graphs that there are significantly more transitioning APs for the shorter window size setting. As transitioning APs use the same amount of energy as light sleeping APs, the upside of a short window size is that it conserves energy as compared to leaving those APs in the active mode. However, the ping-pong effect is detrimental to the UE throughput performance, as transitioning APs cannot serve any UEs, and may furthermore cause hardware degradation.

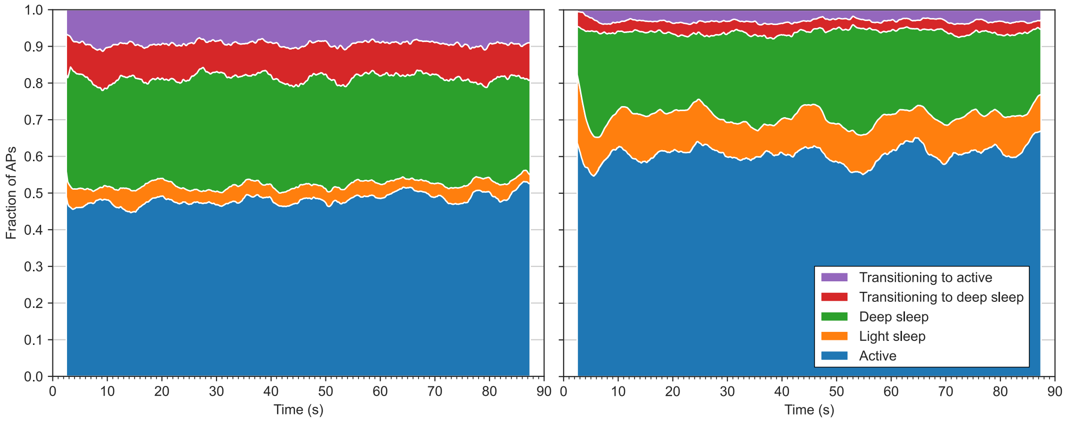


Figure 5.10: Comparison of the fraction of APs within each state over time. Identical SMM settings were used for both graphs, except the window size in the left graph is $k = 22$ coherence blocks, whereas the right graph has a window size of $k = 87$ coherence blocks. A symmetric rolling average of 5 seconds is used for the purpose of noise reduction.

5.2.5. Best configuration

The best configuration is defined as the particular parameter values for which the SMM algorithm uses the least amount of energy while still satisfying the minimum 10th UE throughput percentile target $R^{\texttt{min,UE}}$. As previously explained, there is a notion of 'best' per traffic load and a 'best' over all traffic loads, where the latter is measured by the daily network energy consumption metric.

A plot visualizing the performance of all configurations for each traffic load is shown in Figure 5.11. All configurations that do not satisfy the minimum 10th UE throughput percentile target $R^{\mathtt{min,UE}}$ for a particular traffic load are termed 'directly impermissible'. These 'directly impermissible' configurations are called 'indirectly impermissible' for all other traffic loads where they *do* satisfy $R^{\mathtt{min,UE}}$. All configurations that satisfy $R^{\mathtt{min,UE}}$ in all traffic loads are labeled 'permissible' configurations. To be considered the best overall configuration, it must not be impermissible for any traffic load.

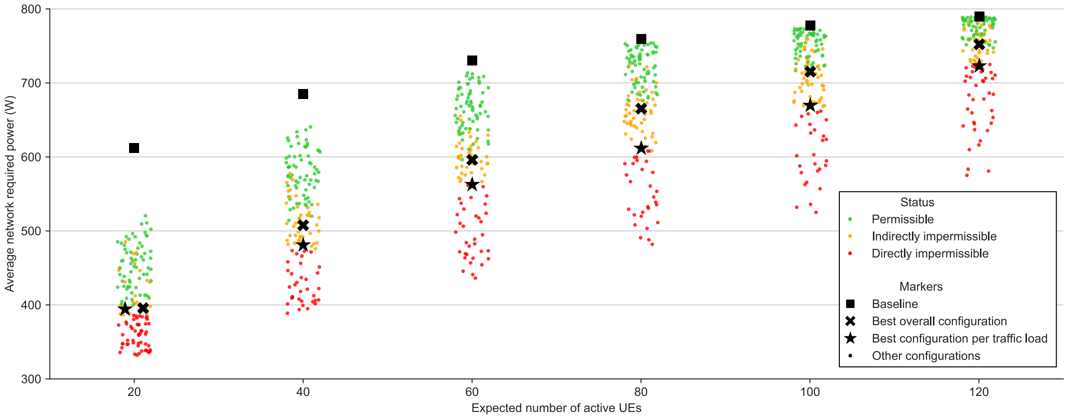


Figure 5.11: Performance of all SMM algorithm configurations for each traffic load $\mathbb{E}[U^{\mathtt{active}}]$. In this swarmplot showing the average network required power, the stars indicate the best performing configuration per traffic load, the crosses indicate the performance of the overall best configuration and the squares indicate the performance for the baseline scenario.

Of the 144 configurations tested, 68 are permissible and therefore remain contenders for the best overall configuration. For each permissible configuration, the daily network energy consumption is derived using the Ericsson traffic data. The top ten best-performing configurations are shown in Table 5.4. Based on the daily network energy consumption, the configuration with a

- to-deep-sleep-threshold $R^{\mathtt{min,sleep}} = 1 \cdot R^{\mathtt{min,UE}} = 20$ Mb/s;
- to-active-threshold $R^{\mathtt{max,wake}} = 0.5 \cdot R^{\mathtt{min,sleep}} = 10$ Mb/s;
- AP neighborhood threshold $G^{\mathtt{nh,min}} = -115$ dB; and
- window size $k = 87$ coherence blocks (≈ 1 s)

is determined to be the best overall configuration.

Table 5.4: Best performing configurations of the SMM algorithm.

To-deep-sleep-threshold $R^{\mathtt{min,sleep}}$ **(**$R^{\mathtt{min,UE}}$**)**	**To-active-threshold** $R^{\mathtt{max,wake}}$ **(**$R^{\mathtt{min,sleep}}$**)**	**AP neighborhood threshold** $G^{\mathtt{nh,min}}$ **(dB)**	**Window size** k **(coherence blocks)**	**Daily network energy consumption (kWh)**
1	0.5	-115	87	14.32
1.5	1.0	-115	44	14.39
1	0.75	-115	87	14.61
2	1	-115	44	14.74
1	1	-115	87	14.78
1.5	0.5	-130	44	14.78
1	0.5	-130	87	14.82
2.5	0.75	-115	44	14.88
2.5	1	-115	44	14.98
1.5	0.5	-115	87	15.00

Several interesting observations can be made from Table 5.4. First, no parameter has one best value across all ten configurations. Notably, the to-deep-sleep-threshold $R^{\mathtt{min,sleep}}$ and the to-active-threshold $R^{\mathtt{max,wake}}$ have each of their tested values represented in the table. However, for the AP neighborhood threshold $G^{\mathtt{nh,min}}$, the preferred value is more clear: 80% of the top ten best configuration use the value of -115 dB. Similary, for the windows size k it is clear that the lowest value (22) and highest value (131) are definitely not preferred. Second, the best configuration uses a value at the edge of the tested range for two of the parameters, the to-deep-sleep-threshold $R^{\mathtt{min,sleep}}$ and the to-active-threshold $R^{\mathtt{max,wake}}$, implying that a value beyond this range may perform even better. It must be noted, however, that the second-best configuration, which performs fairly similarly to the best configuration, only uses the most extreme value for one parameter: a value of 1 for the to-active-threshold $R^{\mathtt{max,wake}}$. Moreover, keep in mind that for that parameter specifically, going beyond 1 is probably ill-advised given the high likelihood of causing a ping-pong effect if the value of the to-active-threshold is higher than the to-deep-sleep-threshold. This suggests that the optimal value of the daily network energy consumption is at least likely close to the value found in this work.

To confirm the veracity of the results, seventeen extra seeds were run for the baseline scenario, the best overall configuration and the best configuration per traffic load. The results of these simulations are indicated by the special markers in Figure 5.11: the squares (baseline), crosses (best overall configuration) and stars (best configuration per traffic load). The results of the other configurations in this swarmplot, i.e. the non-best, non-baseline results indicated by the disks, are still based on the average value over three simulation seeds as it was not feasible to run twenty simulation seeds for every SMM configuration due to time constraints.

For the baseline scenario, i.e. when the use of deep sleep is disabled, the daily network energy consumption is 17.27 kWh, whereas the best overall SMM configuration has a value of 14.32 kWh. This is a 17.11% reduction in energy consumption. If one allows the SMM algorithm to be tuned to the traffic load, i.e. combining the best configurations for each traffic load, the best possible daily network energy consumption is 13.55 kWh, or 21.54% savings over the baseline. Allowing tuning improves the performance of the SMM algorithm by 5.34%. One should keep in mind that the baseline scenario already allows for light sleep. Without light sleep, the energy savings over the baseline scenario would likely be higher.

6 Concluding remarks

This thesis proposed a heuristic SMM algorithm for CF-mMIMO networks that conserves energy by putting APs to sleep that are not required to maintain coverage and an operator-targeted QoS level. The proposed low-complexity algorithm dynamically determines which APs should go to light or deep sleep, making its decisions based solely on practically available information while ensuring coverage and QoS is maintained. A model of a CF-mMIMO network was developed and implemented on a simulator to test the proposed SMM algorithm. Different from previous studies, the evaluation process utilized a realistic transition time to and from the deep sleep mode. Additionally, this work differs from the existing literature by addressing a realistic scenario based on the city center of Amsterdam, including lamppost-based AP deployments and a realistic basis for the spatial traffic distribution and daily traffic fluctuations. This chapter summarizes the main findings of the simulations and presents potential directions for future work.

6.1. Findings

This section presents the two key insights gained from the simulations. Each insight answers one of the research questions posed in Section 1.1:

- **Which AP deployment strategy yields better QoS and energy efficiency in a CF-mMIMO network: deploying a larger number of APs with fewer antennas each, or a smaller number of APs with more antennas each?**

  Based on the three evaluated options with APs equipped with 2, 4 or 8 antennas, when given the choice between deploying a larger number of APs with fewer antennas each or a smaller number of APs with more antennas each, with the same aggregate number of deployed antennas in both cases, the best choice is dependent on an MNO's goal. If QoS is most important, the simulation results show that deploying a larger number of APs with 2 antennas each provides the best 10th UE throughput percentile. However, if energy efficiency is deemed to be most critical, then the simulation results demonstrate that deploying a smaller number of APs with 8 antennas each provides the best average network energy efficiency. Other factors, such as the traffic capacity and financial considerations, might also influence an MNO's decision when selecting the best AP deployment strategy. An advantage of deploying a larger number of APs with fewer antennas each is that more UEs can be supported, as shown by the provided simulation results. A disadvantage of this deployment scenario is that both the deployment costs as well as the operational costs are higher.

- **How effective is a low-complexity heuristic SMM algorithm, making decisions based on practically available information, at reducing the total energy consumption in an urban user-centric CF-mMIMO network with APs characterized by a light and a deep sleep mode and a realistic non-zero transition time for the deep sleep mode, while preserving coverage and targeting a minimum QoS level?**

> The proposed SMM algorithm reduces the daily energy consumption of a CF-mMIMO network by up to 17.11% using the best overall configuration. If the parameters of the SMM algorithm are allowed to be adaptively tuned to the traffic load, the daily energy consumption can be improved by up to 21.54% over the situation where APs cannot use deep sleep. These savings are realized while simultaneously ensuring that 98% population coverage is maintained and that the 10th UE throughput percentile does not fall below 20 Mb/s. Furthermore, it relies on practically available information and incorporates a non-zero transition time with respect to entering and exiting the deep sleep mode.

These findings are valid for the considered scenarios which include: lamppost-based AP deployments in the city center of Amsterdam, a realistic spatial traffic distribution with full-buffer UEs going through successive periods of activity and inactivity, pedestrian mobility, Zero Forcing precoding and equal power sharing, Proportional-Fair scheduling with Semi-orthogonal User Selection and the use of perfect channel knowledge when pilots can be assigned such that UEs sharing the same orthogonal pilot have no overlap in their clusters.

6.2. Future work

This section presents the potential directions for future work. Four categories have been identified:

- The **integration of other RRM aspects** into a heuristic SMM algorithm. In this work, the proposed SMM algorithm is not integrated with the clustering, pilot assignment, scheduling or beamforming algorithms. An algorithm that considers one or more of these RRM aspects jointly with SMM could provide more energy savings.
- The **use of AI/ML techniques** as a basis for a standalone SMM or comprehensive RRM algorithm. As SMM/RRM tasks must be carried out on a small timescale to enable real-time operation, but obtaining an optimal solution is a highly computationally complex task, AI/ML techniques could be used to enable quick decision-making that is still data-driven. Given the recent advancements in AI, it is expected that such a solution could outperform a heuristic algorithm such as the one developed and assessed in this work. However, one should keep in mind that running an AI model can require significant amounts of energy which ought to be taken into account when calculating the energy efficiency of the network. It might therefore be advantageous to only use AI/ML techniques for a sub-component of an SMM algorithm, for example to enable traffic prediction.
- The **value of the simulation-based assessment can be further improved** by considering an **even more detailed traffic, energy consumption and/or channel estimation model**. The traffic model in this work is based on two exponential distributions with fixed scale values where active UEs are assumed to have a full buffer. While the spatial distribution aspect is already reasonably realistic as it is based on actual population data, the traffic model itself could potentially be improved by having an active time that is based on the experienced throughput. This would capture the effect that higher throughputs imply that a web page or file is downloaded more quickly, upon which the UE may go into idle mode again. Additionally, one could consider modeling several types of traffic such as calls, web browsing and video streaming with distinct arrival rates and packet sizes. Note, however, that such improvements may substantially increase the required simulation time and may prevent the assessment of a wide range of scenario or algorithm configurations within a reasonable timeframe. The accuracy of the energy consumption model could be improved by modeling more of the individual components separately. For example, the circuit power modeled in this thesis included several components in one number, while these components conceivably scale dissimilarly across APs designed to be equipped with different numbers of antennas or having different maximum transmit powers. It should be noted, however, that the power values for each of these components are incredibly hardware dependant, making it exceedingly difficult to craft a detailed yet generic model. Finally, this work considers channels to be perfectly known if UEs that share the same orthogonal pilot have no overlap in their clusters. In reality, some level of pilot contamination will be caused between any two UEs sharing the same pilot, depending on their respective channel conditions. The value of the simulations could be improved by explicitly modeling this channel estimation error.
- The **use of more advanced scheduling, beamforming and power control techniques** can

further improve the achievable spectral and energy efficiency of a CF-mMIMO network. Instead of wideband scheduling, PRB-level or subband-based scheduling could be adopted which would provide extra flexibility when scheduling UEs. Additionally, different values for the parameters of the SUS algorithm can be explored, e.g. for the orthogonality threshold $\gamma^{\mathtt{th}}$ and the exponential smoothing factor α. With regard to beamforming, other beamforming techniques such as MMSE precoding could be explored instead of ZF, as well as the impact of local versus network-wide beamforming. As for power control, this thesis adopted a straightforward equal power sharing scheme, but other power control schemes such as max-min fairness could also be considered as some works on CF-mMIMO networks [20], [22] have indicated that they could lead to even better spatial uniformity in terms of UE throughput.

References

[1] Delft High Performance Computing Centre (DHPC), *DelftBlue Supercomputer (Phase 2)*, 2024. [Online]. Available: https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2.

[2] ITU, “Measuring digital development: Facts and Figures 2024,” International Telecommunication Union, Tech. Rep., 2024, ISBN: 978-92-61-39861-3. [Online]. Available: https://www.itu.int/itu-d/reports/statistics/facts-figures-2024/.

[3] M. Tatipamula, *Past, Present, and Future of Wireless Networking*, Milpitas, California, United States, Oct. 2022. [Online]. Available: https://www.ieee-edps.com/archives/2022/c/1400tatipamula.pdf (visited on 04/02/2025).

[4] ITU-R, “Framework and overall objectives of the future development of IMT for 2030 and beyond,” International Telecommunication Union Radiocommunication Sector (ITU-R), Recommendation M.2160-0, Nov. 2023. [Online]. Available: https://www.itu.int/rec/R-REC-M.2160-0-202311-I.

[5] E. Björnson, C.-B. Chae, R. W. H. Jr, T. L. Marzetta, A. Mezghani, L. Sanguinetti, F. Rusek, M. R. Castellanos, D. Jun, and Ö. T. Demir, *Towards 6G MIMO: Massive Spatial Multiplexing, Dense Arrays, and Interplay Between Electromagnetics and Processing*, en, arXiv:2401.02844 [cs], Jan. 2024. DOI: 10.48550/arXiv.2401.02844. [Online]. Available: http://arxiv.org/abs/2401.02844.

[6] J. G. Andrews, X. Zhang, G. D. Durgin, and A. K. Gupta, “Are we approaching the fundamental limits of wireless network densification?” *IEEE Communications Magazine*, vol. 54, no. 10, pp. 184–190, Oct. 2016, ISSN: 1558-1896. DOI: 10.1109/MCOM.2016.7588290.

[7] National Institute of Standards and Technology, “The Spectrum Crunch,” en, Jun. 2016, Last Modified: 2022-04-05T16:18-04:00.

[8] J. Zhang, H. Miao, P. Tang, L. Tian, and G. Liu, “New Mid-Band for 6G: Several Considerations from the Channel Propagation Characteristics Perspective,” *IEEE Communications Magazine*, vol. 63, no. 1, pp. 175–180, Jan. 2025, ISSN: 1558-1896. DOI: 10.1109/MCOM.001.2300708.

[9] ITU-R, “World Radiocommunication Conference 2023 (WRC-23) - Final Acts,” International Telecommunication Union Radiocommunication Sector (ITU-R), Dubai, United Arab Emirates, Tech. Rep., Sep. 2024, pp. 499–501. [Online]. Available: https://www.itu.int/dms_pub/itu-r/opb/act/R-ACT-WRC.16-2024-PDF-E.pdf.

[10] Federal Communications Commission, *Expanding Use of the 12.7-13.25 GHz Band for Mobile Broadband or Other Expanded Use*, Volume: 88, Jul. 2023. [Online]. Available: https://www.federalregister.gov/documents/2023/07/10/2023-13500/expanding-use-of-the-127-1325-ghz-band-for-mobile-broadband-or-other-expanded-use.

[11] O. T. Demir, E. Björnson, and L. Sanguinetti, “Cell-Free Massive MIMO with Large-Scale Fading Decoding and Dynamic Cooperation Clustering,” in *WSA 2021; 25th International ITG Workshop on Smart Antennas*, 2021, pp. 1–6.

[12] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, “Cell-Free Massive MIMO: Uniformly great service for everyone,” in *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, ISSN: 1948-3252, Jun. 2015, pp. 201–205. DOI: 10.1109/SPAWC.2015.7227028.

[13] 3GPP, “Study on channel model for frequencies from 0.5 to 100 GHz,” 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.901, Mar. 2024. [Online]. Available: http://www.3gpp.org/DynaReport/38901.htm.

[14] H. Yang and T. L. Marzetta, “Capacity performance of multicell large-scale antenna systems,” in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct. 2013, pp. 668–675. DOI: 10.1109/Allerton.2013.6736589.

[15] A. Wyner, “Shannon-theoretic approach to a Gaussian cellular multiple-access channel,” *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1713–1727, 1994. DOI: 10.1109/18.340450.

[16] S. Shamai and B. Zaidel, “Enhancing the cellular downlink capacity via co-processing at the transmitting end,” in *IEEE VTS 53rd Vehicular Technology Conference, Spring 2001. Proceedings (Cat. No.01CH37202)*, vol. 3, 2001, 1745–1749 vol.3. DOI: 10.1109/VETECS.2001.944993.

[17] S. Zhou, M. Zhao, X. Xu, J. Wang, and Y. Yao, “Distributed wireless communication system: A new architecture for future public wireless access,” *IEEE Communications Magazine*, vol. 41, no. 3, pp. 108–113, 2003. DOI: 10.1109/MCOM.2003.1186553.

[18] 3GPP, “Handover procedures,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 23.009, Dec. 2000. [Online]. Available: http://www.3gpp.org/DynaReport/23009.htm.

[19] 3GPP, “Coordinated multi-point operation for LTE physical layer aspects,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.819, Sep. 2013. [Online]. Available: http://www.3gpp.org/DynaReport/36819.htm.

[20] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, “Cell-Free Massive MIMO Versus Small Cells,” en, *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017, ISSN: 1536-1276, 1558-2248. DOI: 10.1109/TWC.2017.2655515.

[21] Ö. T. Demir, E. Björnson, and L. Sanguinetti, *Foundations of User-Centric Cell-Free Massive MIMO*, en, arXiv:2108.02541 [cs, eess, math], Aug. 2021. DOI: 10.1561/2000000109. [Online]. Available: http://arxiv.org/abs/2108.02541.

[22] S. Chen, J. Zhang, J. Zhang, E. Björnson, and B. Ai, “A survey on user-centric cell-free massive MIMO systems,” en, *Digital Communications and Networks*, 2022. DOI: https://doi.org/10.1016/j.dcan.2021.12.005.

[23] E. Björnson and L. Sanguinetti, “Scalable Cell-Free Massive MIMO Systems,” en, *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020, ISSN: 0090-6778, 1558-0857. DOI: 10.1109/TCOMM.2020.2987311.

[24] S. Buzzi and C. D’Andrea, “Cell-Free Massive MIMO: User-Centric Approach,” en, *IEEE Wireless Communications Letters*, vol. 6, no. 6, pp. 706–709, Dec. 2017, ISSN: 2162-2337, 2162-2345. DOI: 10.1109/LWC.2017.2734893.

[25] 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,” en, *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 1, pp. 25–39, Mar. 2018, ISSN: 2473-2400. DOI: 10.1109/TGCN.2017.2770215.

[26] S. Braam, R. Litjens, P. Smulders, and W. IJntema, “Assessment of Distributed Multi-User MIMO Transmission in 5G Networks,” in *Proceedings of the 18th ACM Symposium on Mobility Management and Wireless Access*, ser. MobiWac ’20, event-place: Alicante, Spain, New York, NY, USA: Association for Computing Machinery, 2020, pp. 125–132, ISBN: 978-1-4503-8119-2. DOI: 10.1145/3416012.3424629.

[27] M. Ito, I. Kanno, Y. Amano, Y. Kishi, W.-Y. Chen, T. Choi, and A. F. Molisch, “Joint AP On/Off and User-Centric Clustering for Energy-Efficient Cell-Free Massive MIMO Systems,” en, in *2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*, London, United Kingdom: IEEE, Sep. 2022, pp. 1–5, ISBN: 978-1-66545-468-1. DOI: 10.1109/VTC2022-Fall57202.2022.10013046.

[28] E. Björnson, J. Hoydis, and L. Sanguinetti, “Massive MIMO networks: Spectral, energy, and hardware efficiency,” en, *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017, Publisher: Now Publishers.

[29] Y. Zhang, H. Cao, P. Zhong, C. Qi, and L. Yang, “Location-Based Greedy Pilot Assignment for Cell-Free Massive MIMO Systems,” in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, 2018, pp. 392–396. DOI: 10.1109/CompComm.2018.8780756.

[30] H. Liu, J. Zhang, X. Zhang, A. Kurniawan, T. Juhana, and B. Ai, “Tabu-Search-Based Pilot Assignment for Cell-Free Massive MIMO Systems,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 2286–2290, 2020. DOI: 10.1109/TVT.2019.2956217.

[31] S. Buzzi, C. D’Andrea, M. Fresia, Y.-P. Zhang, and S. Feng, “Pilot Assignment in Cell-Free Massive MIMO Based on the Hungarian Algorithm,” *IEEE Wireless Communications Letters*, vol. 10, no. 1, pp. 34–37, 2021. DOI: 10.1109/LWC.2020.3020003.

[32] R. Sabbagh, C. Pan, and J. Wang, “Pilot Allocation and Sum-Rate Analysis in Cell-Free Massive MIMO Systems,” in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6. DOI: 10.1109/ICC.2018.8422575.

[33] H. Liu, J. Zhang, S. Jin, and B. Ai, “Graph Coloring Based Pilot Assignment for Cell-Free Massive MIMO Systems,” en, *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 9180–9184, Aug. 2020, ISSN: 0018-9545, 1939-9359. DOI: 10.1109/TVT.2020.3000496.

[34] D. Brélaz, “New methods to color the vertices of a graph,” *Commun. ACM*, vol. 22, no. 4, pp. 251–256, Apr. 1979, Place: New York, NY, USA Publisher: Association for Computing Machinery, ISSN: 0001-0782. DOI: 10.1145/359094.359101.

[35] S. Chen, J. Zhang, E. Bjornson, J. Zhang, and B. Ai, “Structured Massive Access for Scalable Cell-Free Massive MIMO Systems,” en, *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 4, pp. 1086–1100, Apr. 2021, ISSN: 0733-8716, 1558-0008. DOI: 10.1109/JSAC.2020.3018836.

[36] M. Attarifar, A. Abbasfar, and A. Lozano, “Random vs Structured Pilot Assignment in Cell-Free Massive MIMO Wireless Networks,” en, in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, Kansas City, MO, USA: IEEE, May 2018, pp. 1–6, ISBN: 978-1-5386-4328-0. DOI: 10.1109/ICCW.2018.8403508.

[37] F. Göttsch, N. Osawa, I. Kanno, T. Ohseki, and G. Caire, “Fairness Scheduling in User-Centric Cell-Free Massive MIMO Wireless Networks,” *IEEE Transactions on Wireless Communications*, pp. 1–1, 2024. DOI: 10.1109/TWC.2024.3386802.

[38] S. Mashdour, R. C. de Lamare, and J. P. Sales, *Study of Multiuser Scheduling with Enhanced Greedy Techniques for Multicell and Cell-Free Massive MIMO Networks*, en, arXiv:2303.06779 [cs, math], Mar. 2023. [Online]. Available: http://arxiv.org/abs/2303.06779 (visited on 07/09/2024).

[39] S. Mashdour and R. C. d. Lamare, *Study of Clustering Techniques and Scheduling Algorithms with Fairness for Cell-Free MIMO Networks*, en, arXiv:2404.18032 [cs], Apr. 2024. DOI: 10.48550/arXiv.2404.18032. [Online]. Available: http://arxiv.org/abs/2404.18032.

[40] K.-H. Shin, J.-W. Kim, S.-W. Park, J.-H. Yu, S.-G. Choi, H.-D. Kim, Y.-H. You, and H.-K. Song, “Dynamic Scheduling and Power Allocation with Random Arrival Rates in Dense User-Centric Scalable Cell-Free MIMO Networks,” *Mathematics*, vol. 12, no. 10, 2024, ISSN: 2227-7390. DOI: 10.3390/math12101515.

[41] T. Van Chien, E. Björnson, and E. G. Larsson, *Optimal Design of Energy-Efficient Cell-Free Massive MIMO: Joint Power Allocation and Load Balancing*, en, arXiv:1911.11375 [cs], Nov. 2019. DOI: 10.48550/arXiv.1911.11375. [Online]. Available: http://arxiv.org/abs/1911.11375.

[42] T. Van Chien, E. Björnson, and E. G. Larsson, *Joint Power Allocation and Load Balancing Optimization for Energy-Efficient Cell-Free Massive MIMO Networks*, en, arXiv:2002.01504 [cs, math], Jun. 2020. [Online]. Available: http://arxiv.org/abs/2002.01504.

[43] G. Femenias, N. Lassoued, and F. Riera-Palou, “Access Point Switch ON/OFF Strategies for Green Cell-Free Massive MIMO Networking,” *IEEE Access*, vol. 8, pp. 21 788–21 803, Jan. 2020. DOI: 10.1109/ACCESS.2020.2969815.

[44] J. García-Morales, G. Femenias, and F. Riera-Palou, “Energy-Efficient Access-Point Sleep-Mode Techniques for Cell-Free mmWave Massive MIMO Networks With Non-Uniform Spatial Traffic Density,” *IEEE Access*, vol. 8, pp. 137 587–137 605, 2020. DOI: 10.1109/ACCESS.2020.3012199.

[45] F. Riera-Palou, G. Femenias, J. Garcia-Morales, and H. Q. Ngo, “Selective Infrastructure Activation in Cell-free Massive MIMO: A Two Time-scale Approach,” en, in *2021 IEEE Globecom Workshops (GC Wkshps)*, Madrid, Spain: IEEE, Dec. 2021, pp. 1–7, ISBN: 978-1-66542-390-8. DOI: 10.1109/GCWkshps52748.2021.9682022.

[46] S. Jung and S.-E. Hong, “Performance analysis of Access Point Switch ON/OFF schemes for Cell-free mmWave massive MIMO UDN systems,” en, in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, Republic of: IEEE, Oct. 2021, pp. 644–647, ISBN: 978-1-66542-383-0. DOI: 10.1109/ICTC52510.2021.9620943.

[47] S.-E. Hong and J.-H. Na, “Joint Access Point Beamforming and Switch On/Off Scheme for Energy Efficient Cell-Free mmWave massive MIMO,” en, in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, Republic of: IEEE, Oct. 2022, pp. 730–732, ISBN: 978-1-66549-939-2. DOI: 10.1109/ICTC55196.2022.9952389.

[48] F. Riera-Palou, G. Femenias, D. Lopez-Perez, N. Piovesan, and A. De Domenico, “Sleep Mode Strategies for Energy Efficient Cell-Free Massive MIMO in 5G Deployments,” en, in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, Rome, Italy: IEEE, May 2023, pp. 618–624, ISBN: 9798350333077. DOI: 10.1109/ICCWorkshops57953.2023.10283562.

[49] Q. He, Ö. T. Demir, and C. Cavdar, “Dynamic AP Selection and Cluster Formation with Minimal Switching for Green Cell-Free Massive MIMO Networks,” en, in *2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, Gothenburg, Sweden: IEEE, Jun. 2023, pp. 234–239, ISBN: 9798350311020. DOI: 10.1109/EuCNC/6GSummit58263.2023.10188295.

[50] Q. Zheng, P. Zhu, J. Li, D. Wang, and X. You, “Energy Efficiency Enhancement in User-Centric and Cell-Free Millimeter-Wave Massive MIMO Systems With Hybrid Beamforming,” en, *IEEE Transactions on Vehicular Technology*, vol. 73, no. 4, pp. 5484–5499, Apr. 2024, ISSN: 0018-9545, 1939-9359. DOI: 10.1109/TVT.2023.3336071.

[51] R. Beerten, V. Ranjbar, A. P. Guevara, H. Sallouha, and S. Pollin, *Location-Based Load Balancing for Energy-Efficient Cell-Free Networks*, en, arXiv:2404.18799 [eess], Apr. 2024. [Online]. Available: http://arxiv.org/abs/2404.18799.

[52] F. Kooshki, A. G. Armada, M. M. Mowla, A. Flizikowski, and S. Pietrzyk, “Energy-Efficient Sleep Mode Schemes for Cell-Less RAN in 5G and Beyond 5G Networks,” en, *IEEE Access*, vol. 11, pp. 1432–1444, 2023, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3233430.

[53] C. F. Mendoza, S. Schwarz, and M. Rupp, “Deep Reinforcement Learning for Dynamic Access Point Activation in Cell-Free MIMO Networks,” in *WSA 2021; 25th International ITG Workshop on Smart Antennas*, Nov. 2021, pp. 1–6.

[54] H. Suh, J. Oh, S. Kang, and T. Hwang, “DRL-Based AP Switch On/Off Scheme for Cell-Free Massive MIMO MEC Networks,” in *2023 14th International Conference on Information and Communication Technology Convergence (ICTC)*, ISSN: 2162-1241, Oct. 2023, pp. 235–237. DOI: 10.1109/ICTC58733.2023.10392593.

[55] L. Sun, J. Hou, and R. Chapman, “Multi-Agent Deep Reinforcement Learning for Access Point Activation Strategy in Cell-Free Massive MIMO Networks,” en, in *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Hoboken, NJ, USA: IEEE, May 2023, pp. 1–6, ISBN: 978-1-66549-427-4. DOI: 10.1109/INFOCOMWKSHPS57453.2023.10225848.

[56] W. Li, Y. Jiang, Y. Huang, and F.-C. Zheng, “Energy-Efficient Access Point Sleep Control in User-Centric Cell-Free Massive MIMO Systems,” in *2024 16th International Conference on Wireless Communications and Signal Processing (WCSP)*, Oct. 2024, pp. 585–590. DOI: 10.1109/WCSP62071.2024.10827577.

[57] Mosek ApS, *MOSEK*, Apr. 2025. [Online]. Available: https://www.mosek.com/ (visited on 04/18/2025).

[58] B. Aslan and G. Zech, “Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 537, no. 3, pp. 626–636, 2005, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2004.08.071.

[59] 3GPP, *Overview of AI/ML related Work in 3GPP*, Feb. 2025. [Online]. Available: https://www.3gpp.org/news-events/3gpp-news/ai-ml-2025 (visited on 04/16/2025).

[60] Centraal Bureau voor de Statistiek, *Kaart van 100 meter bij 100 meter met statistieken 2023*, nl-NL, Last Modified: 25-07-2024T10:46:55, Jul. 2024. [Online]. Available: https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/kaart-van-100-meter-bij-100-meter-met-statistieken (visited on 01/14/2025).

[61] Gemeente Westerkwartier, *Beeldplan Licht in de Openbare Ruimte*, 2016. [Online]. Available: https://westerkwartier.verlichtinginbeeld.nl/factsheets/lantaarnpalen-en-duurzaamheid/ (visited on 01/14/2025).

[62] S. Chen, J. Zhang, E. Björnson, Ö. T. Demir, and B. Ai, *Energy-Efficient Cell-Free Massive MIMO Through Sparse Large-Scale Fading Processing*, en, arXiv:2208.13552 [eess], Apr. 2023. [Online]. Available: http://arxiv.org/abs/2208.13552.

[63] N. Jayaweera, K. B. S. Manosha, N. Rajatheva, and M. Latva-aho, "Minimizing Energy Consumption in Cell-free Massive MIMO Networks," en, *IEEE Transactions on Vehicular Technology*, pp. 1–15, 2024, ISSN: 0018-9545, 1939-9359. DOI: 10.1109/TVT.2024.3392790.

[64] Gemeente Amsterdam, *Openbare verlichting - dataset lichtpunten*, Sep. 2023. [Online]. Available: https://maps.amsterdam.nl/open_geodata/ (visited on 01/14/2025).

[65] Ministerie van Economische Zaken, *Dekkings- en snelheidsverplichting - Telecomaanbieders - Rijksinspectie Digitale Infrastructuur (RDI)*, nl-NL, Last Modified: 2025-01-07T14:10 Publisher: Ministerie van Economische Zaken, Jan. 2023. [Online]. Available: https://www.rdi.nl/onderwerpen/telecomaanbieders/dekkingseis-en-snelheidsverplichting (visited on 01/14/2025).

[66] 3GPP, "NR; Base Station (BS) radio transmission and reception," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.104, Dec. 2024. [Online]. Available: http://www.3gpp.org/DynaReport/38104.htm.

[67] 3GPP, "NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.101-1, Sep. 2024. [Online]. Available: http://www.3gpp.org/DynaReport/38101-1.htm.

[68] 3GPP, "NR; User Equipment (UE) radio transmission and reception; Part 2: Range 2 Standalone," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.101-2, Dec. 2024. [Online]. Available: http://www.3gpp.org/DynaReport/38101-2.htm.

[69] ITU-R, "Modelling and simulation of IMT networks and systems for use in sharing and compatibility studies," International Telecommunication Union Radiocommunication Sector (ITU-R), Recommendation M.2101-0, Feb. 2017. [Online]. Available: https://www.itu.int/rec/R-REC-M.2101-0-201702-I.

[70] Ericsson, *Exploring how traffic patterns drive network evolution*, en, Jun. 2023. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/articles/traffic-patterns-drive-network-evolution (visited on 05/10/2025).

[71] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005, ISBN: 978-0-521-84527-4.

[72] Koninkrijk der Nederlanden, "Verlening vergunningen veiling 3,5 GHz door Rijksinspectie Digitale Infrastructuur," *Staatscourant*, vol. 23967, Jul. 2024.

[73] Z. Lai, N. Bessis, G. de la Roche, P. Kuonen, J. Zhang, and G. Clapworthy, "The Characterisation of Human Body Influence on Indoor 3.5 GHz Path Loss Measurement," in *2010 IEEE Wireless Communication and Networking Conference Workshops*, Apr. 2010, pp. 1–6. DOI: 10.1109/WCNCW.2010.5487656.

[74] R. Fraile, J. F. Monserrat, J. Gozálvez, and N. Cardona, "Mobile radio bi-dimensional large-scale fading modelling with site-to-site cross-correlation," en, *European Transactions on Telecommunications*, vol. 19, no. 1, pp. 101–106, Jan. 2008, ISSN: 1124-318X, 1541-8251. DOI: 10.1002/ett.1179.

[75] S. Jaecke, L. Raschkowsk, L. Thiele, F. Burkhardt, and E. Eberlein, "QuaDRiGa - Quasi Deterministic Radio Channel Generator, User Manual and Documentation," Fraunhofer Heinrich Hertz Institute, Technical Report v2.8.1, Dec. 2023.

[76] M. Ajmal, A. Siddiqa, B. Jeong, J. Seo, and D. Kim, “Cell-free massive multiple-input multiple-output challenges and opportunities: A survey,” *ICT Express*, vol. 10, no. 1, pp. 194–212, 2024, ISSN: 2405-9595. DOI: https://doi.org/10.1016/j.icte.2023.10.007.

[77] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x. DOI: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[78] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, 2006, ISBN: 978-0-471-74881-6. [Online]. Available: https://books.google.nl/books?id=EuhBluW31hsC.

[79] A. Jalali, R. Padovani, and R. Pankaj, “Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system,” in *VTC2000-Spring. 2000 IEEE 51st Vehicular Technology Conference Proceedings (Cat. No.00CH37026)*, ISSN: 1090-3038, vol. 3, May 2000, 1854–1858 vol.3. DOI: 10.1109/VETECS.2000.851593.

[80] T. Yoo and A. Goldsmith, “On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming,” *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, Mar. 2006, ISSN: 1558-0008. DOI: 10.1109/JSAC.2005.862421.

[81] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, “Local Partial Zero-Forcing Precoding for Cell-Free Massive MIMO,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4758–4774, Jul. 2020, ISSN: 1558-2248. DOI: 10.1109/TWC.2020.2987027.

[82] K. Brueninghaus, D. Astely, T. Salzer, S. Visuri, A. Alexiou, S. Karger, and G.-A. Seraji, “Link performance models for system level simulations of broadband radio access systems,” in *2005 IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, ISSN: 2166-9589, vol. 4, Sep. 2005, 2306–2311 Vol. 4. DOI: 10.1109/PIMRC.2005.1651855.

[83] 3GPP, “Study on International Mobile Telecommunications (IMT) parameters for 6.425-7.025 GHz, 7.025-7.125 GHz and 10.0-10.5 GHz,” 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.921, Mar. 2024. [Online]. Available: http://www.3gpp.org/DynaReport/38921.htm.

[84] 3GPP, “NR; Physical layer procedures for data,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.214, Sep. 2024. [Online]. Available: http://www.3gpp.org/DynaReport/38214.htm.

[85] P. Mogensen, W. Na, I. Z. Kovacs, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, “LTE Capacity Compared to the Shannon Bound,” in *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, ISSN: 1550-2252, Apr. 2007, pp. 1234–1238. DOI: 10.1109/VETECS.2007.260.

[86] F. Tan, Q. Deng, and Q. Liu, “Energy-efficient access point clustering and power allocation in cell-free massive MIMO networks: A hierarchical deep reinforcement learning approach,” en, *EURASIP Journal on Advances in Signal Processing*, vol. 2024, no. 1, p. 18, Jan. 2024, ISSN: 1687-6180. DOI: 10.1186/s13634-024-01111-9.

[87] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, “Optimal Design of Energy-Efficient Multi-User MIMO Systems: Is Massive MIMO the Answer?” *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3059–3075, 2015. DOI: 10.1109/TWC.2015.2400437.

[88] B. Debaillie, C. Desset, and F. Louagie, “A Flexible and Future-Proof Power Model for Cellular Base Stations,” in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, 2015, pp. 1–7. DOI: 10.1109/VTCSpring.2015.7145603.

[89] N. Piovesan, D. López-Pérez, A. De Domenico, X. Geng, H. Bao, and M. Debbah, “Machine Learning and Analytical Power Consumption Models for 5G Base Stations,” en, *IEEE Communications Magazine*, vol. 60, no. 10, pp. 56–62, Oct. 2022, arXiv:2209.11600 [cs], ISSN: 0163-6804, 1558-1896. DOI: 10.1109/MCOM.001.2200023.

[90] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, “How much energy is needed to run a wireless network?” *IEEE Wireless Communications*, vol. 18, no. 5, pp. 40–49, Oct. 2011, ISSN: 1558-0687. DOI: 10.1109/MWC.2011.6056691.

[91] E. Bobrov, B. Chinyaev, V. Kuznetsov, H. Lu, D. Minenkov, S. Troshin, D. Yudakov, and D. Zaev, *Adaptive Regularized Zero-Forcing Beamforming in Massive MIMO with Multi-Antenna Users*, en, arXiv:2107.00853 [cs], Jul. 2023. DOI: 10.48550/arXiv.2107.00853. [Online]. Available: http://arxiv.org/abs/2107.00853 (visited on 07/05/2025).

[92] E. Björnson and L. Sanguinetti, *Making Cell-Free Massive MIMO Competitive With MMSE Processing and Centralized Implementation*, en, arXiv:1903.10611 [cs], Sep. 2019. DOI: 10.48550/arXiv.1903.10611. [Online]. Available: http://arxiv.org/abs/1903.10611 (visited on 07/05/2025).

A

Simulation parameters

This appendix contains an overview of the simulation parameters that are fixed in all scenarios.

Table A.1: Overview of the simulation parameters that are fixed in all scenarios. Details and sources can be found in Chapter 3.

Parameter	Symbol	Value	Unit
Area sidelength	-	500	m
Frequency	f^{c}	7.8	GHz
Carrier bandwidth	B	100	MHz
Coherence bandwidth	B^{c}	1,440	kHz
Numerology	-	2	-
Subcarrier spacing	-	60	kHz
Number of PRBs	-	134	-
Coherence time	T^{c}	11.5	ms
TTI duration	-	0.25	ms
Number of pilot TTIs	-	1	-
Number of uplink TTIs	-	11	-
Number of downlink TTIs	-	34	-
SF decorrelation distance	-	13	m
SF Inter-AP correlation	$\sigma^{\texttt{sh}}$	0.5	-
SF pixel size	-	1	m
SF standard deviation	-	7.82	dB
Body loss	L^{b}	6	dB
Minimum coupling loss	$L^{\texttt{mc}}$	53	dB
Temperature	T	290	K
Coverage pixel size	-	10	m
Minimum gain threshold	$G^{\texttt{ul,min}}$	-132.55	dB
Population coverage threshold	-	98	%
AP antenna element spacing	-	3.84	cm
AP maximum antenna element gain	-	8	dBi
AP height	$h^{\texttt{AP}}$	6	m
AP elevation angle (downtilt)	-	0	°
AP noise figure	-	3	dB
AP maximum transmit power per antenna	$P^{\texttt{AP,ant,tx,max}}$	0.25	W
AP fixed power	$P^{\texttt{AP,fixed}}$	8	W
AP fixed power per RF chain	$P^{\texttt{AP,chain}}$	0.2	W
AP power amplifier efficiency	η	0.39	-
Total number of UEs	U	3,250	-
Number of UE antennas	$M^{\texttt{UE}}$	4	-
UE antenna element gain	-	0	dBi
UE height	$h^{\texttt{UE}}$	1.5	m
UE transmit power	$P^{\texttt{UE,tx,max}}$	23	dBm
UE noise figure	-	8	dB
UE mobility	v	3.0	km/h
Fronthaul fixed power	$P^{\texttt{FH,fixed}}$	0.825	W
Fronthaul traffic-dependent power	$\zeta^{\texttt{FH}}$	0.25	W/Gbps
CPU fixed power	$P^{\texttt{CPU,fixed}}$	5	W
CPU traffic-dependent power	$\zeta^{\texttt{CPU}}$	0.1	W/Gbps
SUS orthogonality threshold	$\gamma^{\texttt{th}}$	0.5	-
PF exponential smoothing factor	β	0.01	-
Shannon correction factor	C^{f}	0.75	-
Maximum downlink data rate	$R^{\texttt{max}}$	460.56	Mb/s

B

Extended simulation results

This appendix contains the simulation results that were excluded from the thesis.

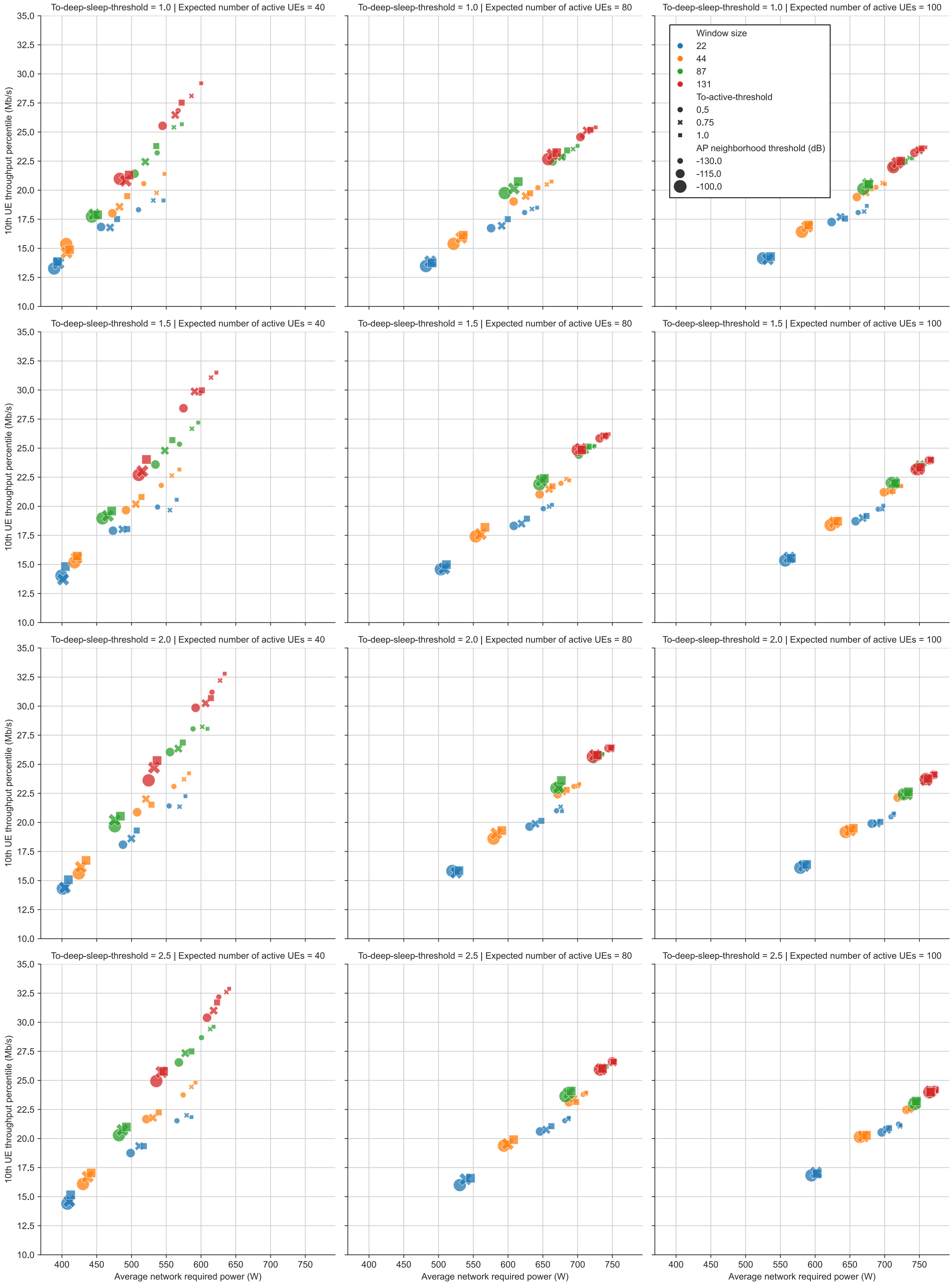


Figure B.1: Performance of all SMM algorithm configurations for 3 selected traffic loads $\mathbb{E}[U^{\texttt{active}}] = \{40, 80, 100\}$. The graphs in each column have the same traffic load and the graphs in each row have the same to-deep-sleep-threshold $R^{\texttt{min,sleep}}$. Note that each graph contains 36 markers, however, it can be difficult to see all of them as values sometimes overlap.