On Optimizing Time-, Space- and Power-Domain Energy-Saving Techniques for Sub-6 GHz Base Stations

Emanuele Peschiera, *Graduate Student Member, IEEE*, Youssef Agram, François Quitin, *Member, IEEE*,
Liesbet Van der Perre, *Senior Member, IEEE*, François Rottenberg, *Member, IEEE*

Abstract—**What is the optimal base station (BS) resource allocation strategy given a measurement-based power consumption model and a fixed target user rate? Rush-to-sleep in time, rush-to-mute in space, awake-but-whisper in power, or a combination of them? We propose in this paper an efficient solution to the problem of finding the optimal number of active time slots, active antennas, and transmit power at active antennas in a multiple-input multiple-output (MIMO) orthogonal frequency-division multiplexing (OFDM) system under per-user rate and per-antenna transmit power constraints. The use of a parametric power consumption model validated on operator measurements of 4G and 5G BSs enhances the interpretation of the results. We discuss the optimal energy-saving strategy at different network loads for three BS configurations. Using as few BS antennas as possible is close to optimal in BSs not implementing time-domain power savings such as micro-discontinuous transmission (µDTX). Energy-saving schemes that jointly operate in the three domains are instead optimal when the BS hardware can enter time-domain power-saving modes, with a tendency for rush-to-mute in massive MIMO and for rush-to-sleep in BS with fewer antennas. Median energy savings up to 30% are achieved at low network loads.**

Index Terms—**Resource allocation, energy saving, green communications, MIMO, 6G mobile communication.**

I. INTRODUCTION

A. Motivation

The reduction of energy consumption in wireless networks is imperative to reach the environmental sustainability objectives and control the expenses of mobile operators [1], [2], which foresee a mobile data traffic exceeding 400 EB/month in 2029 [3]. Most of the energy consumption in a wireless network comes from the radio access network (RAN) and in particular from the base stations (BSs) [1], [4], [5]. Measurements indicate that the power consumption of current wireless networks decreases only slightly at low traffic loads [6], hence network energy-saving techniques are becoming essential in wireless systems [7]. Energy-saving techniques implement load-aware resource adaptations and BS sleep modes to break down consumption at low and medium loads [8], where definitions of load include the fraction of active orthogonal frequency-division multiplexing (OFDM) time slots, the utilized fraction of the network peak capacity, and the ratio of transmit power to maximum transmit power.

B. State of the Art

Energy-saving techniques can generally operate in four domains, *i.e.*, time, frequency, space, and power [8], [9]. The adaptive shutdown of transmit symbols, (sub)carriers,

E. Peschiera, L. Van der Perre and F. Rottenberg are with ESAT-DRAMCO, Campus Ghent, KU Leuven, 9000 Ghent, Belgium (email: emanuele.peschiera@kuleuven.be).

Y. Agram and F. Quitin are with Brussels School of Engineering, ULB, 1000 Brussels, Belgium.

antennas, as well as the adaptation of the transmit power at the BS, are among the main drivers to achieve power consumption savings. The use of accurate power consumption models becomes a crucial aspect [10]. A review of the published models up to 2023 is given in [11]. These are classified in (i) linear models if they express the power consumption as linearly dependent on the BS output power, (ii) subcomponent models if they estimate the consumption of individual processing blocks, and (iii) massive multiple-input multiple-output (MIMO) models when they target 5G BSs with 32 or more antennas. A recent work has proposed a parametric power consumption model able to predict the power consumption of sub-6 GHz 4G and 5G BSs in Belgium [12]. By building upon measurements of deployed network and datasheets of manufacturers, the authors model the consumption of power amplifiers (PAs), analog front-end (AFE), digital baseband (DBB), and power supply and cooling systems. Another recent work introduced the concept of waste factor, which quantifies the proportion of wasted power versus power effectively used for any intended function, evaluating therefore the total BS power consumption [13].

As anticipated, energy savings can be achieved by optimizing the resource usage in different domains. In the time domain, micro-discontinuous transmission (µDTX) has been used to switch off PAs on a symbol scale [14], [15]. The 5G networks have been designed to be lean, requiring less frequent control signaling and therefore enabling longer sleep durations [16]. The impact of using sleep modes during both data transmission and control signaling phases has been evaluated at standard level [7] and via system-level simulations [17], [18]. The tradeoff between energy saving and latency has been discussed in [19]. In [20], the authors introduce a two-dimensional modeling of the energy efficiency (EE) that incorporates the time dimension. However, fundamental works in this area have often considered that the transmission takes place at a constant rate per time slot [21], [22], preventing the assessment of the optimal transmit powers and rates at different time slots. The work in [23] fills this theoretical gap in the case of a single-input single-output (SISO) system by considering an average rate constraint over multiple time slots. It shows that transmission at constant power among a number of active time slots is optimal and that a rush-to-sleep approach, using as few time slots as possible at maximum transmit power, is optimal at low signal-to-noise ratio (SNR).

In the space domain, many works have focused on load-aware activation of antennas in massive MIMO BSs [10]. In general, the need for these energy-saving schemes becomes evident when optimizing total consumed power instead of transmit power only. The authors of [24] find the optimal number of active antennas and users that maximize the EE by

TABLE I
MAIN QUANTITIES AND RELATIVE DEFINITIONS.

Variable	Description
M, N, K, Q	Number of antennas, time slots, users, and subcarriers
M_{a}, N_{a}	Number of active antennas and active time slots
$M_{\mathrm{a,min}}$, $N_{\mathrm{a,min}}$	Minimal number of active antennas and active time slots
P_{a}, $P_{\mathrm{a,min}}$	Transmit and minimal transmit power at active antennas
$P_{\max}$, α	Maximum PA output power, PA consumption exponent
$\eta_{\mathrm{PA}}^{\max}$, ξ	PA maximum efficiency and weight of static consumption
p_k, ρ_k	Power allocation and single-antenna equiv. SNR at user k
f_{c}, B	Carrier frequency and bandwidth
$R_{k,0}$	Baseline target rate of user k
$R_k = \kappa R_{k,0}$, κ	Target rate of user k, rate scaling
σ_k^2, β_k	Noise power and large-scale fading coefficient at user k
τ_{DL}, τ_{UL}	Time ratios of downlink and uplink to total frame
τ_{sig}	Time ratio of reference signaling to downlink
$\delta_{\mathrm{PA}}^{\mathrm{dtx}}$, $\delta_{\mathrm{PA}}^{\mathrm{idle}}$, $\delta_{\mathrm{PA}}^{\mathrm{sleep}}$	Reduction factors for PA μDTX, idle, and sleep modes
$\delta_{\mathrm{TRX}}^{\mathrm{idle}}$, $\delta_{\mathrm{TRX}}^{\mathrm{sleep}}$	Reduction factors for AFE idle and sleep modes
$\delta_{\mathrm{phy}}^{\mathrm{idle}}$	Reduction factor for DBB idle mode
$\bar{P}_{\mathrm{PA},1}$, $\bar{P}_{\mathrm{AFE},1}$	PA and AFE contributions to P_1
$\bar{P}_{\mathrm{AFE,sleep}}$, $\bar{P}_{\mathrm{DBB}}$	AFE contribution to P_{sleep}, DBB consumption
$\eta_{\mathrm{s/c}}^{\mathrm{PA}}$, $\eta_{\mathrm{s/c}}^{\mathrm{AFE}}$, $\eta_{\mathrm{s/c}}^{\mathrm{DBB}}$	Supply and cooling efficiencies of PA, AFE, and DBB
γ, P_0, P_1, P_{sleep}	Parameters of power consumption model

building on a detailed circuit consumption model. A similar analysis is done in [25] by considering a multi-cell network and traffic load variations. The work [26] addresses the problem of minimizing power consumption under per-user signal-to-interference-plus-noise ratio (SINR) constraints. The precoder that minimizes BS consumption under SINR constraints for a multicarrier massive MIMO system is derived in [27]. It is shown that conventional zero forcing (ZF) among a subset of antennas is optimal. The work [28] specifically targets energy consumption reduction by maximizing the number of muted antennas subject to quality of service (QoS) constraints and solves the problem using neural networks. Space-domain resource adaptation becomes even more important with the increasing interest in cell-free MIMO, where it is essential to keep the total consumption under control by turning on/off access points and their antennas [29]–[31].

Power-domain energy-saving techniques can instead optimize either the total BS transmit power as is the case in classic precoding designs [32], or the individual transmit powers at the BS antennas in order to improve the PAs efficiency [33]. Constant-envelope precoding has been proposed in [34]. The authors in [35] find the optimal point-to-point MIMO precoder that maximizes the rate under a PA consumption constraint. More recently, flat-ZF precoding has been proposed with the same objective of reducing PA consumption [36].

Recent works have also looked at multi-domain resource allocation. The joint optimization of spectral and spatial resources in a multi-cell network under ZF precoding is addressed in [37], while optimal precoders and scheduling of resources are found in [38] via convex solvers. In [39], the authors optimize the sleep mode configuration, number of active antennas, and user-BS associations in a multi-cell network using reinforcement learning.

C. Contributions

In this paper, we answer the question of finding the optimal combination of power-, space- and time-domain energy-saving techniques for different BS configurations at varying network loads. Three specific allocations are identified and used as benchmark: a rush-to-sleep that minimizes the number of active time slots, a rush-to-mute that minimizes the number of active antennas, and an awake-but-whisper that minimizes the transmit power at active antennas. The relative performance of these three allocations is not addressed by previous works, and therefore we do not consider frequency-domain energy savings to limit the scope. We employ a power consumption model applicable to deployed sub-6 GHz BSs and we further use on-site measurements to initialize our numerical simulations [12], [40]. Compared to the conference version [41], we provide here the full expressions of the power consumption model parameters, the proofs of convexity of the optimization problem, an approximation of the optimal solution and the analysis of 4G and 5G BSs with a lower number of antennas. Our specific contributions are the following:

- We first show that the initial problem can be simplified, casted into a two-dimensional convex differentiable problem (with a sufficient condition on the domain convexity) and efficiently solved. We propose an approximation of the final discrete solution and provide the error decay.
- We discuss the optimal energy-saving strategy for the 4G and 5G BSs under study in [12], not implementing time-domain power savings (PA μDTX and AFE idle modes). A rush-to-mute approach provides quasi-optimal performance at any network load for massive MIMO BSs and at low and medium network loads for BSs with fewer antennas.
- We consider the case when the analyzed 4G and 5G BSs utilize time-domain power savings in the form of PA μDTX and AFE idle modes. It turns out that a combination of the three energy-saving techniques is the optimal one. Massive MIMO BSs still tends more to a rush-to-mute at any network load, while the BSs with fewer antennas steer towards a rush-to-sleep at low and medium loads.

The remainder of the paper is organized as follows. Section II introduces the transmission model in frequency, time, and space. Section III explains the BS power consumption model. The solution to the considered optimization problem is discussed in Section IV. Numerical experiments are presented in Section V. Section VI concludes the paper. Table I summarizes the main quantities used throughout the paper.

Notations: The symbols $\mathbb{E}\{\cdot\}$ and $\mathrm{tr}[\cdot]$ denote the expectation of a random variable and the trace of a matrix. The operators $\lceil\cdot\rceil$, $\lfloor\cdot\rfloor$, and $[\cdot]$ are the ceil, floor, and round, respectively. The operator $\lfloor x\rceil$, which we refer to as the ceil-floor, selects among the upper and lower bounding integers of x the one that optimizes the cost function. In two dimensions, $\lfloor x, y\rceil$ selects among the four different combinations of (x,y) that use ceil or floor for both variables, the one that optimizes the cost function. We use the notation $f(x) = \mathcal{O}(g(x))$, as $x \to a$, if there exist positive numbers μ and λ such that $|f(x)| \le \lambda g(x)$ when $0 < |x - a| < \mu$.
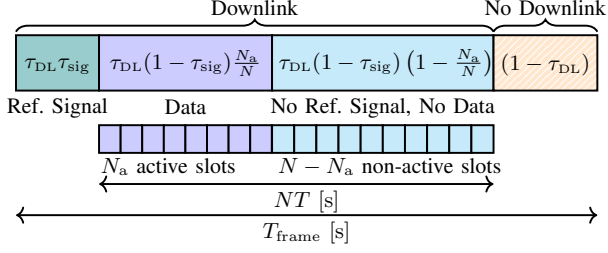
Fig. 1. Frame structure in time considered in this work, where the quantities inside each block are time ratios relative to the frame duration T_{frame}.

II. Transmission Model

We consider a BS with M antennas and K single-antenna users communicating in downlink using OFDM with Q active subcarriers carrying data symbols. The cyclic prefix is assumed to be larger than the channel excess delay. The OFDM symbol duration, including the cyclic prefix, is denoted by T. The users are multiplexed using space-division multiple access (SDMA) implying that they are being served at the same time and frequency, and we consider $M \geq K$. The transmission time is divided into frames lasting T_{frame} seconds. In every frame there are N OFDM symbols available to carry data symbols. Out of the N OFDM symbols, only N_{a} are active, i.e., used to communicate.[1] A frame is divided in four phases:

- Downlink and reference signal transmission, lasting $\tau_{\text{DL}}\tau_{\text{sig}}T_{\text{frame}}$ seconds, where $\tau_{\text{DL}} \in (0,1]$ is the time ratio of downlink transmission to the frame duration and $\tau_{\text{sig}} \in (0,1)$ is the time ratio of reference signal transmission to downlink transmission.
- Downlink and data transmission, with a duration of $\tau_{\text{DL}}(1 - \tau_{\text{sig}})\frac{N_{\text{a}}}{N}T_{\text{frame}}$ seconds.
- Downlink and no reference signal and no data transmission, lasting $\tau_{\text{DL}}(1 - \tau_{\text{sig}})\left(1 - \frac{N_{\text{a}}}{N}\right)T_{\text{frame}}$ seconds.
- No downlink that spans $(1 - \tau_{\text{DL}})T_{\text{frame}}$ seconds. This is either uplink transmission in time-division duplexing (TDD) or has zero duration in frequency-division duplexing (FDD) as uplink takes place on a different frequency.

The corresponding frame structure is given in Fig. 1. We illustrate the total duration of the four phases, but the reference signals as well as the active and non-active OFDM data symbols can be distributed across the frame. Without loss of generality, we consider one such frame in the following.

Throughout this work, we consider energy-saving techniques where only a subset of M_{a} antennas out of M are active, with $K \leq M_{\text{a}} \leq M$. The transmitted symbols at each subcarrier and OFDM symbol are considered zero mean and uncorrelated among users. There is one symbol per user at a given subcarrier and active OFDM symbol, with variance p_k that is the power allocation of user k. We denote the transmitted symbol vector as $\mathbf{s} \in \mathbb{C}^{K \times 1}$. The corresponding channel matrix between the active antennas and the users is denoted by $\mathbf{H} \in \mathbb{C}^{K \times M_{\text{a}}}$. We consider the use of a ZF precoder so that the precoded symbols are $\mathbf{H}^{\text{H}}(\mathbf{H}\mathbf{H}^{\text{H}})^{-1}\mathbf{s}$. The large-scale fading coefficient of user k is denoted by β_k. While the independent and identically distributed (i.i.d.)

[1]We interchangeably use the terms OFDM symbol and time slot.

Rayleigh fading is commonly used in the massive MIMO literature [42], several experimental studies have shown that it is often not verified in practice [43], [44]. Still, this assumption is made here as being useful for the purpose of finding tractable resource allocation strategies. Considering a unitary inverse fast Fourier transform (IFFT), the average power of the precoded symbols in the frequency domain is the same as in the time domain (after the IFFT). Under i.i.d. Rayleigh fading, the average total transmitted power is then given by

$$P_{\text{T}} = \mathbb{E}\left\{\text{tr}\left[\mathbf{H}^{\text{H}}(\mathbf{H}\mathbf{H}^{\text{H}})^{-1}\mathbf{s}\mathbf{s}^{\text{H}}(\mathbf{H}\mathbf{H}^{\text{H}})^{-1}\mathbf{H}\right]\right\}$$
$$\overset{(*)}{=} \text{tr}\left[\mathbb{E}\left\{(\mathbf{H}\mathbf{H}^{\text{H}})^{-1}\right\}\mathbb{E}\left\{\mathbf{s}\mathbf{s}^{\text{H}}\right\}\right] \overset{(**)}{=} \frac{1}{M_{\text{a}} - K}\sum_{k=1}^{K}\frac{p_k}{\beta_k} \quad (1)$$

where $(*)$ follows from the trace cyclic property and the fact that the symbols and channel coefficients are uncorrelated, while $(**)$ uses the expression of the mean of a complex inverse-Wishart distribution [45].

An array gain of a factor $M_{\text{a}} - K$ is present in (1). To minimize P_{T}, the maximal number of antennas should be active, i.e., $M_{\text{a}} = M$. In terms of consumed power, denoted by P_{cons}, this is often not optimal as P_{cons} is not linearly proportional to P_{T} because of the PA non-constant efficiency and the load-independent consumption of activating each antenna. This is further detailed in Section III. Given that all antennas are statistically identically distributed, the average transmit power is the same at each antenna and equal to

$$P_{\text{a}} = \frac{P_{\text{T}}}{M_{\text{a}}} = \frac{1}{M_{\text{a}}(M_{\text{a}} - K)}\sum_{k=1}^{K}\frac{p_k}{\beta_k}. \quad (2)$$

The selection of the M_{a} active antennas can be arbitrary given their identical distribution. In the following, we consider a maximal per-antenna transmit power constraint $P_{\text{a}} \leq P_{\text{max}}$. After going through the channel and OFDM demodulation at the receiver, the transmitted symbols are recovered with the addition of zero-mean circularly-symmetric complex Gaussian noise. The noise variance at user k is denoted by σ_k^2. Then, the total bit rate delivered to user k averaged over the frame duration is given by

$$R_{k,\text{deliv}} = \frac{Q}{T}\tau_{\text{DL}}(1 - \tau_{\text{sig}})\frac{N_{\text{a}}}{N}\log_2\left(1 + \frac{p_k}{\sigma_k^2}\right) \quad (3)$$

expressed in bits per second. To simplify notations, we define $R_k = R_{k,\text{deliv}}/\left(\frac{Q}{T}\tau_{\text{DL}}(1 - \tau_{\text{sig}})\right)$ as the normalized per-subcarrier average rate expressed in bits per OFDM symbol and per subcarrier. In the following, we consider that each user has a target average rate R_k. Given that we analyse one frame transmission, this rate constraint can also be seen as a latency constraint. If we consider that we first send all reference signals and then all the active OFDM symbols, we are constraining the delivery of $NTR_{k,\text{deliv}}$ bits to the k-th user in $\tau_{\text{DL}}\tau_{\text{sig}}T_{\text{frame}} + NT$ seconds at most.

III. Power Consumption Model

We adopt the model from [12] that expresses the BS power consumption averaged over the frame duration as

TABLE II
PARAMETERS OF POWER CONSUMPTION MODEL (5) FOR DIFFERENT BS CONFIGURATIONS.

Configuration	RU, duplex, protocol	M	K	f_c [GHz]	B [MHz]	$P_{\max}$ [W]	α	γ	P_0 [W][1]	P_1 [W][1]	P_{sleep} [W]
4T4R	RRU, FDD, 4G LTE	4	2	1.8	20	40	0.75	5.33	$\{0, 34.69\}$	$\{149.40, 114.71\}$	233.55
8T8R	RRU, TDD, 5G NR	8	4	3.5	100	40	0.75	5.38	$\{0, 69.98\}$	$\{229.47, 103.26\}$	363.78
64T64R	AAU, TDD, 5G NR	64	8	3.5	100	3.125	0.75	3.50	$\{0, 53.92\}$	$\{341.57, 161.95\}$	550.23

[1]Left value is obtained with $\delta_{\text{PA}}^{\text{dtx}} = 1$ and $\delta_{\text{TRX}}^{\text{idle}} = 1$ as in [12], right value is obtained with $\delta_{\text{PA}}^{\text{dtx}} = 0.25$ and $\delta_{\text{TRX}}^{\text{idle}} = 0.5$.

$$P_{\text{cons}} = \frac{\bar{P}_{\text{PA}}}{\eta_{\text{s/c}}^{\text{PA}}} + \frac{\bar{P}_{\text{AFE}}}{\eta_{\text{s/c}}^{\text{AFE}}} + \frac{\bar{P}_{\text{DBB}}}{\eta_{\text{s/c}}^{\text{DBB}}} \tag{4}$$

where $\bar{P}_{\text{PA}}$, $\bar{P}_{\text{AFE}}$, and $\bar{P}_{\text{DBB}}$ are the average power consumed by the PAs, AFE, and DBB across the frame duration (during both downlink and uplink), while $\eta_{\text{s/c}}^{\text{PA}}$, $\eta_{\text{s/c}}^{\text{AFE}}$, and $\eta_{\text{s/c}}^{\text{DBB}}$ are the supplying and cooling efficiencies of the PAs, AFE, and DBB. The above power consumption model predicts the consumption of sub-6 GHz 4G and 5G cellular BSs in Belgium within 10-20% uncertainty [12]. It can be applied to radio unit (RU) types including remote radio unit (RRU) and active antenna unit (AAU), working either in FDD or TDD, and running either 4G Long-Term Evolution (LTE) or 5G New Radio (NR). Also, it is parametric such that parameter values can be changed to reflect different implementations of BS components. The model in (4) can be adapted to the scenario that we target in this paper, and the average BS consumed power across the frame can be expressed as a function of N_a, M_a and P_a, giving

$$P_{\text{cons}} = \frac{N_a}{N} M_a \left(\frac{P_0}{M} + \gamma P_a^{\alpha} \right) + \frac{M_a}{M} P_1 + P_{\text{sleep}} \tag{5}$$

where P_0, P_1, P_{sleep}, γ are non-negative parameters, and $\alpha \in [0.5, 1]$. The term with P_0 depends on both the ratio of active time slots and active antennas. The term depending on γP_a^{α} models the power consumed by the PAs. The third term scales with the ratio of active antennas, while the last term is constant. Although (5) is kept compact to emphasize the dependence on the optimization variables, we will show in the following that the model parameters depend on system parameters such as K, $P_{\max}$, B and f_c, τ_{DL} and τ_{sig}, etc.

To map the model in (4) to (5), we recall that the average physical load in a frame is defined in [12] as the time-average on all the spatial layers of the instantaneous load, which is given by ratio of the instantaneous number of data subcarriers to the total number of data subcarriers. When a time slot is active, we assume that all its data subcarriers are used to serve all the users using SDMA. Hence, the average physical load in our scenario is given by N_a/N. Besides, we set the frame duration to T_{frame}, the number of active (resp. total) PAs and TX/RX chains to M_a (resp. M), and the number of active (resp. total) spatial layers to K_a (resp. K), where we consider $K_a = K$. We also consider one cell, which corresponds to one sector and one frequency band. The full-load PA output power and maximum PA output power are set to $P_{\max}$, which considers 8 dB output power back-off such that the PA is in the linear regime. We allow the PA to operate at an arbitrary output power P_a during data transmission. In the following, we provide the details on how this mapping is done and what

are the different contributions to the total P_{cons}.

A. Power Amplifier

The instantaneous power consumed by a PA is modeled as the sum of a static consumption and a dynamic consumption that depends on the instantaneous output power p [12]:

$$P_{\text{PA}}(p) = \xi \frac{P_{\max}^{\alpha}}{\eta_{\text{PA}}^{\max}} + (1 - \xi) \frac{P_{\max}^{1-\alpha} p^{\alpha}}{\eta_{\text{PA}}^{\max}}$$

where $\xi \in [0, 1]$ and $\eta_{\text{PA}}^{\max}$ is the maximum PA efficiency [12, eq. (5)] that depends on the PA technology, on the carrier frequency f_c and on $P_{\max}$. Parameters of common PA architectures are in [12, Table I]. Each of the M_a active PAs can be in one of four states (see Fig. 1):

- Downlink mode and reference signal transmission, where $p = \zeta_{\text{sig}} P_{\max}$, with $\zeta_{\text{sig}} \in (0, 1)$ being the ratio of reference signaling power to $P_{\max}$. This gives a power consumption equal to $\tau_{\text{DL}} \tau_{\text{sig}} P_{\text{PA}}(\zeta_{\text{sig}} P_{\max})$.
- Downlink mode and data transmission, with $p = P_a$, resulting in a power consumption of $\tau_{\text{DL}}(1 - \tau_{\text{sig}}) \frac{N_a}{N} P_{\text{PA}}(P_a)$.
- Downlink mode and no data and no reference signal transmission, with $p = 0$. This corresponds to discontinuous transmission (DTX) mode. The power consumption is $\tau_{\text{DL}}(1 - \tau_{\text{sig}}) \left(1 - \frac{N_a}{N}\right) P_{\text{PA}}(0) \delta_{\text{PA}}^{\text{dtx}}$, where $\delta_{\text{PA}}^{\text{dtx}} \in [0, 1]$ is a reduction factor when in DTX.
- Idle state when not in downlink mode, with $p = 0$, resulting in a power consumption given by $(1 - \tau_{\text{DL}}) P_{\text{PA}}(0) \delta_{\text{PA}}^{\text{idle}}$, where $\delta_{\text{PA}}^{\text{idle}} \in [0, 1]$ is a reduction factor when in idle mode.

The $M - M_a$ non-active PAs are instead in sleep mode and their consumed power is $(M - M_a) P_{\text{PA}}(0) \delta_{\text{PA}}^{\text{sleep}}$, where $\delta_{\text{PA}}^{\text{sleep}} \in [0, 1]$ is a reduction factor when in sleep mode. The reduction factors can represent opposite situations, e.g., $\delta_{\text{PA}}^{\text{dtx}} = 1$ (no µDTX) and $\delta_{\text{PA}}^{\text{dtx}} = 0$ (ideal µDTX). By grouping the above terms, the consumed power by the PAs averaged over the frame duration is given by [12, eq. (6)]

$$\bar{P}_{\text{PA}} = \frac{N_a}{N} M_a \tau_{\text{DL}} (1 - \tau_{\text{sig}}) \left(P_{\text{PA}} (P_a) - P_{\text{PA}}(0) \delta_{\text{PA}}^{\text{dtx}} \right) \\ + M_a \bar{P}_{\text{PA},1} + M P_{\text{PA}}(0) \delta_{\text{PA}}^{\text{sleep}}$$

where we defined

$$\bar{P}_{\text{PA},1} \triangleq \tau_{\text{DL}} \tau_{\text{sig}} P_{\text{PA}} (\zeta_{\text{sig}} P_{\max}) + \tau_{\text{DL}} (1 - \tau_{\text{sig}}) P_{\text{PA}}(0) \delta_{\text{PA}}^{\text{dtx}} \\ + (1 - \tau_{\text{DL}}) P_{\text{PA}}(0) \delta_{\text{PA}}^{\text{idle}} - P_{\text{PA}}(0) \delta_{\text{PA}}^{\text{sleep}}. \tag{6}$$

We can therefore identify in $\bar{P}_{\text{PA}}$ the terms composing γ and P_0/M, as well as the terms that are part of P_1/M and P_{sleep}.

B. Analog Front-End

The AFE performs digital-to-analog/analog-to-digital conversions, up-/down-conversion, pre-driving, low-noise ampli-

fication, frequency synthesis and general control. We consider the miscellaneous functions to be always active and the number of TX and RX chains to be the same. The M_a active TX/RX chains can be in (i) working mode, where we assume that the same power is consumed in downlink and uplink, and (ii) idle mode. The $M - M_\mathrm{a}$ non-active TX/RX chains are in sleep mode instead. By using [12, eq. (10)], the consumed power by the AFE averaged over the frame duration is expressed as

$$\bar{P}_\mathrm{AFE} = M_\mathrm{a}\bar{P}_\mathrm{AFE,1} + \bar{P}_\mathrm{AFE,sleep}$$

where we defined

$$\bar{P}_\mathrm{AFE,1} \triangleq P_\mathrm{TRX}\big((\tau_\mathrm{DL} + \tau_\mathrm{UL}) + (2 - \tau_\mathrm{DL} - \tau_\mathrm{UL})\,\delta_\mathrm{TRX}^\mathrm{idle}$$
$$- 2\delta_\mathrm{TRX}^\mathrm{sleep}\big) \tag{7}$$
$$\bar{P}_\mathrm{AFE,sleep} \triangleq P_\mathrm{misc} + 2MP_\mathrm{TRX}\delta_\mathrm{TRX}^\mathrm{sleep} \tag{8}$$

with P_misc scaling with the bandwidth B and MP_max, P_TRX scaling with B and P_max [12, eq. (9) and (10)], and $\delta_\mathrm{TRX}^\mathrm{idle}, \delta_\mathrm{TRX}^\mathrm{sleep} \in [0,1]$ representing the reduction factors when the AFE is in idle and sleep modes. The AFE contribution is thus (physical) load-independent and part of P_1/M and P_sleep.

C. Digital Baseband

The operations performed by the DBB are physical resource scheduling, (de)coding, (de)modulation, digital precoding and decoding, etc. The DBB consumption is modeled as the consumption at the data link layer $P_\mathrm{link}^\mathrm{ref}$, which is considered constant, plus the consumption at the physical layer, which comprises either working or idle mode. The average DBB consumed power over the frame is then given by [12, eq. (11)]

$$\bar{P}_\mathrm{DBB} = P_\mathrm{link}^\mathrm{ref} + \tilde{P}_\mathrm{phy}^\mathrm{ref}\big((\tau_\mathrm{DL} + \tau_\mathrm{UL}) + (2 - \tau_\mathrm{DL} - \tau_\mathrm{UL})\,\delta_\mathrm{phy}^\mathrm{idle}\big) \tag{9}$$

where $P_\mathrm{link}^\mathrm{ref}$ is a constant and $\tilde{P}_\mathrm{phy}^\mathrm{ref}$ is linearly proportional to the number of users K and scales with B. As for the AFE, the DBB consumption is (physical) load-independent and this time is only part of P_sleep.

D. Power Supply and Cooling Systems

Power converters (AC/DC and DC/DC) and active cooling modules in a BS also contribute to the BS consumption. We consider that the first introduce an overhead with respect to the total consumption with an efficiency η_supply, while the second introduce component-wise overheads with efficiencies $\eta_\mathrm{cool}^\mathrm{PA}, \eta_\mathrm{cool}^\mathrm{AFE}, \eta_\mathrm{cool}^\mathrm{DBB} \in (0,1]$ [12]. In case of no active cooling, $\eta_\mathrm{cool}^\mathrm{PA} = \eta_\mathrm{cool}^\mathrm{AFE} = \eta_\mathrm{cool}^\mathrm{DBB} = 1$. The supply and cooling efficiencies are eventually defined as $\eta_\mathrm{s/c}^\mathrm{PA} = \eta_\mathrm{supply}\eta_\mathrm{cool}^\mathrm{PA}$, $\eta_\mathrm{s/c}^\mathrm{AFE} = \eta_\mathrm{supply}\eta_\mathrm{cool}^\mathrm{AFE}$, and $\eta_\mathrm{s/c}^\mathrm{DBB} = \eta_\mathrm{supply}\eta_\mathrm{cool}^\mathrm{DBB}$. Now, using (6), (7), (8), (9) and linking them to (4) we can express the parameters of our power consumption model (5) as follows

$$\gamma = \frac{\tau_\mathrm{DL}(1 - \tau_\mathrm{sig})}{\eta_\mathrm{s/c}^\mathrm{PA}}(1 - \xi)\frac{P_\mathrm{max}^{1-\alpha}}{\eta_\mathrm{PA}^\mathrm{max}}$$

$$\frac{P_0}{M} = \frac{\tau_\mathrm{DL}(1 - \tau_\mathrm{sig})}{\eta_\mathrm{s/c}^\mathrm{PA}}\xi\frac{P_\mathrm{max}^\alpha}{\eta_\mathrm{PA}^\mathrm{max}}\big(1 - \delta_\mathrm{PA}^\mathrm{dtx}\big)$$

$$\frac{P_1}{M} = \frac{\bar{P}_\mathrm{PA,1}}{\eta_\mathrm{s/c}^\mathrm{PA}} + \frac{\bar{P}_\mathrm{AFE,1}}{\eta_\mathrm{s/c}^\mathrm{AFE}}$$

$$P_\mathrm{sleep} = \frac{M\delta_\mathrm{PA}^\mathrm{sleep}}{\eta_\mathrm{s/c}^\mathrm{PA}}\xi\frac{P_\mathrm{max}^\alpha}{\eta_\mathrm{PA}^\mathrm{max}} + \frac{\bar{P}_\mathrm{AFE,sleep}}{\eta_\mathrm{s/c}^\mathrm{AFE}} + \frac{\bar{P}_\mathrm{DBB}}{\eta_\mathrm{s/c}^\mathrm{DBB}}.$$

To compute the above parameters, we select three BS configurations (64T64R, 8T8R, and 4T4R) and utilize the values in [12, Table II]. We consider two pairs of values for the reduction factors $\delta_\mathrm{PA}^\mathrm{dtx}$ and $\delta_\mathrm{TRX}^\mathrm{idle}$, leading to two cases:

- Disabled time-domain hardware power-saving modes, with $\overline{\delta_\mathrm{PA}^\mathrm{dtx} = 1}$ and $\delta_\mathrm{TRX}^\mathrm{idle} = 1$ as considered in [12]. This corresponds to not using PA μDTX and AFE idle modes.
- Enabled time-domain hardware power-saving modes, with $\overline{\delta_\mathrm{PA}^\mathrm{dtx} = 0.25}$ and $\delta_\mathrm{TRX}^\mathrm{idle} = 0.5$. The obtained model parameters reflect the BSs under study in [12] but utilizing related to PA μDTX and AFE idle modes. The values of $\delta_\mathrm{PA}^\mathrm{dtx}$ and $\delta_\mathrm{TRX}^\mathrm{idle}$ are set to be equal to $\delta_\mathrm{PA}^\mathrm{idle}$ and $\delta_\mathrm{TRX}^\mathrm{sleep}$, respectively.

The obtained values of (5) are given in Table II.

IV. OPTIMAL ALLOCATION OF POWER, TIME AND SPATIAL RESOURCES

This section considers the problem

$$\min_{N_\mathrm{a},M_\mathrm{a},P_\mathrm{a}} P_\mathrm{cons} \text{ s.t. } R_k = \frac{N_\mathrm{a}}{N}\log_2\left(1 + \frac{p_k}{\sigma_k^2}\right), \; k = 1, \ldots, K. \tag{10}$$

In words, we consider the minimization of P_cons by optimizing the number of active time slots N_a and active antennas M_a, and the transmit power at active antennas P_a. The optimization is subject to a target rate constraint R_k for each user. Moreover, as detailed in Section II, the optimization variables have additional constraints (not directly shown in (10)) given their physical meaning: $0 \leq N_\mathrm{a} \leq N$, $K \leq M_\mathrm{a} \leq M$ and $0 \leq P_\mathrm{a} \leq P_\mathrm{max}$, N_a and M_a are integers while P_a is continuous. This section adopts a step-by-step approach to solve problem (10) and shows that it can be efficiently solved through a two-dimensional differentiable convex problem. More specifically, in Section IV-A, we first give the optimal transmit power P_a as a function of N_a and M_a. Then in Section IV-B, we find the optimal time and spatial resources N_a and M_a to allocate. Section IV-C proposes an approximation of the optimal solution.

A. Optimal Transmit Power at Active Antennas

As a first step to solve problem (10), we can find that the active antenna transmit power P_a can be expressed in terms of N_a and M_a through the user rate constraint. Indeed, we can solve for p_k

$$R_k = \frac{N_\mathrm{a}}{N}\log_2\left(1 + \frac{p_k}{\sigma_k^2}\right) \iff p_k = \sigma_k^2\left(2^{R_k\frac{N}{N_\mathrm{a}}} - 1\right) \tag{12}$$

and substitute the expression of p_k into (2) giving

$$P_\mathrm{a} = \frac{1}{M_\mathrm{a}(M_\mathrm{a} - K)}\sum_{k=1}^{K}\frac{\sigma_k^2}{\beta_k}\left(2^{R_k\frac{N}{N_\mathrm{a}}} - 1\right). \tag{13}$$

The term $R_k\frac{N}{N_\mathrm{a}}$ can be interpreted as the user rate per active time slot, where R_k gives the frame-averaged number of bits per subcarrier and per time slot. Substituting (13) into problem (10), it simplifies to the two-dimensional problem (11) in

$$\min_{\substack{0 \le N_a \le N \\ K \le M_a \le M}} P_{\text{cons}} = \frac{N_a}{N} \frac{M_a}{M} P_0 + \frac{N_a}{N} M_a \gamma \left(\frac{1}{M_a(M_a - K)} \sum_{k=1}^{K} \frac{\sigma_k^2}{\beta_k} \left(2^{R_k \frac{N}{N_a}} - 1 \right) \right)^{\alpha} + \frac{M_a}{M} P_1 + P_{\text{sleep}} \quad \text{s.t. } (\mathcal{C}_{P_{\max}}) \quad (11)$$

M_a and N_a. Note that the user rate constraints are implicitly taken into account and thus satisfied. Moreover, the maximal power constraint $P_a \le P_{\max}$ can be converted in terms of M_a and N_a to

$$(\mathcal{C}_{P_{\max}}): \frac{1}{M_a(M_a - K)} \sum_{k=1}^{K} \frac{\sigma_k^2}{\beta_k} \left(2^{R_k \frac{N}{N_a}} - 1 \right) \le P_{\max}. \quad (14)$$

Clearly, P_a is minimized when $M_a = M$ and $N_a = N$ leading to the definition

$$P_{a,\min} = \frac{1}{M(M - K)} \sum_{k=1}^{K} \frac{\sigma_k^2}{\beta_k} \left(2^{R_k} - 1 \right). \quad (15)$$

Hence, problem (10) is feasible if $P_{a,\min} \le P_{\max}$. In other words, activating all resources ($M_a = M$, $N_a = N$) allows the system to meet the target users' rates within the power budget. As shown in Appendix VII-A, the inequality (14) can be further simplified as

$$M_a \ge \frac{K}{2} + \frac{1}{2} \sqrt{K^2 + 4 \sum_{k=1}^{K} \rho_k^{-1} \left(2^{R_k \frac{N}{N_a}} - 1 \right)} \quad (16)$$

where $\rho_k = P_{\max} \beta_k / \sigma_k^2$ and which implies $M_a \ge K$. Two useful quantities can be derived from this inequality. Firstly, the minimum amount of active antennas, defined as $M_{a,\min}$, is obtained by finding the minimum value of M_a that satisfies (16) when $N_a = N$

$$M_{a,\min} = \left\lceil \frac{K}{2} + \frac{1}{2} \sqrt{K^2 + 4 \sum_{k=1}^{K} \rho_k^{-1} \left(2^{R_k} - 1 \right)} \right\rceil. \quad (17)$$

Secondly, the minimum amount of active slots, defined as $N_{a,\min}$, is obtained by finding the minimum value of N_a that satisfies (16) when $M_a = M$

$$M \ge \frac{K}{2} + \frac{1}{2} \sqrt{K^2 + 4 \sum_{k=1}^{K} \rho_k^{-1} \left(2^{R_k \frac{N}{N_{a,\min}}} - 1 \right)}. \quad (18)$$

In general, $N_{a,\min}$ does not have a closed-form solution.

B. Optimal Space and Time Resources

Let us consider a continuous relaxation of problem (11). Making the change of variable $x = N/N_a$, $y = M_a$ and removing the constant term P_{sleep} not impacting the optimization, the cost function becomes

$$f(x, y) = \frac{y}{x} \frac{P_0}{M} + \frac{y}{x} \gamma \left(\frac{1}{y(y - K)} \sum_{k=1}^{K} \frac{\sigma_k^2}{\beta_k} \left(2^{R_k x} - 1 \right) \right)^{\alpha}$$
$$+ \frac{y}{M} P_1. \quad (19)$$

In Appendix VII-B, we show that $f(x, y)$ is convex in (x, y).[2]

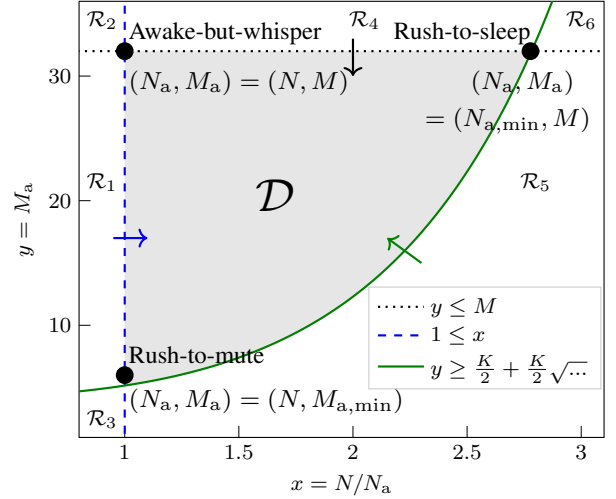[2] As a note, defining x as N_a or N_a/N instead of N/N_a would not have led to a convex function $f(x, y)$.

Fig. 2. Example of feasibility domain $\mathcal{D}$ for $M = 32$, $K = 4$, $P_{\max} = 1$, $\beta_k = 1$, $\sigma_k^2 = 10^{-1}$, $R_k = 4$, for $k = 1, \ldots, K$. The point $(N, M_{a,\min})$ has a small offset from the minimum y due to the ceiling operation in (17).

Moreover, the constraints $0 \le N_a \le N$, $K \le M_a \le M$ and (16) imply that x and y must belong to the feasibility domain $\mathcal{D}$ defined as

$$\mathcal{D} = \Big\{ 1 \le x, y \le M,$$
$$y \ge \frac{K}{2} + \frac{1}{2} \sqrt{K^2 + 4 \sum_{k=1}^{K} \rho_k^{-1} \left(2^{R_k x} - 1 \right)} \Big\}. \quad (20)$$

In a convex problem, both the objective function and the constraints need to be convex. In Appendix VII-C, we show that a sufficient condition for the convexity of the set $\mathcal{D}$ is

$$\rho_k = \frac{P_{\max} \beta_k}{\sigma_k^2} \ge \frac{2}{K} \quad \text{for } k = 1, \ldots, K. \quad (21)$$

This can be seen as a minimal average SNR constraint at each user when using a single transmit antenna at maximal power. This is a sufficient condition, which is required to make a strong general statement but can be neglected in certain regimes. For instance, for a small number of users and/or small target data rates, the $P_{\max}$ constraint is not binding and (21) can be left out. In the extreme opposite case, for a large number of users and/or large target data rates, a sufficiently large SNR at each user will be necessary to make the problem feasible, implying that (21) will be satisfied.

An illustration of $\mathcal{D}$ is shown in Fig. 2, which resembles a triangle with a curved bottom right side. Each of the three vertices corresponds to a specific energy-saving regime:

- **Rush-to-sleep** (time-domain only): all spatial resources are active with full transmit power giving $M_a = M$, $P_a = P_{\max}$ with minimum active time slots $N_a = N_{a,\min}$.
- **Rush-to-mute** (spatial-domain only): all time slots are active with full transmit power giving $N_a = N$, $P_a = P_{\max}$ with minimum number of active antennas $M_a = M_{a,\min}$.

- **Awake-but-whisper** (power-domain only): all spatial and temporal resources are activated giving $x = 1$, $N_a = N$ and $M_a = M$ with minimum transmit power $P_{a,\min}$.

The awake-but-whisper and rush-to-mute approaches have a zero PA idle-mode duration while the rush-to-sleep has the largest one, which also minimizes latency. As a result of the convexity of $f(x, y)$ and $\mathcal{D}$, the problem

$$(\bar{x}, \bar{y}) = \arg \min_{(x,y) \in \mathcal{D}} f(x, y) \tag{22}$$

is convex and differentiable. We can thus use simple and efficient techniques to perform the minimization, with guarantee to find the global minimum. Another favourable consequence of the convexity of (22) is that the solution of the original (discrete) problem can be found through the ceil-floor operator

$$(N_a, M_a) = \lfloor N/\bar{x}, \bar{y} \rceil. \tag{23}$$

Indeed, for a convex problem, the optimal discrete point should be one of the nearest (feasible) neighbours.

C. Approximation of Optimal Solution

The ceil-floor operation requires to evaluate and compare P_{cons} for four possible values as N_a is either $\lceil N/\bar{x} \rceil$ or $\lfloor N/\bar{x} \rfloor$ while M_a is either $\lfloor \bar{y} \rfloor$ or $\lceil \bar{y} \rceil$. In practice, if N is large, the ceil-floor operator can be approximated by a simpler rounding operation $N_a = [N/\bar{x}]$, not requiring any comparison (but just checking the feasibility). The resulting deviation from the optimal value of the cost function vanishes as $\mathcal{O}(1/N)$. The intuition is that P_{cons} depends on N_a through the ratio N_a/N, implying that the discretization error is at most $1/N$ and that asymptotically, the ratio N_a/N can take an infinite number (not only discrete) number of points between 0 and 1. A similar reasoning can be applied for the number of antennas: replacing the ceil-floor operation by a rounding operator $M_a = [\bar{y}]$. The resulting approximation error decays as $\mathcal{O}(1/M_a^{2\alpha})$ as M_a grows large. In summary, as shown in Appendix VII-D, using

$$N_a = [N/\bar{x}], \ M_a = [\bar{y}] \tag{24}$$

provides the minimal P_{cons} up to an error term that decreases as $\mathcal{O}(1/N + 1/M + 1/M_a^{2\alpha})$ when the number of time slots $N \to +\infty$ and the number of active antennas $M_a \to +\infty$. In practice, the large N assumption is realistic as frames are usually made of many symbols. The impact of the second and third contributions to the error term depends on which BS configuration is selected because that fixes the number of antennas M and users K, and imposes a constraint $M_a \geq K$.

As a final remark, problem (11) only depends on the large-scale fading coefficients of the user-to-BS channels, not on the instantaneous channels. These coefficients are simpler to estimate than the instantaneous channels as they vary more slowly in time. Moreover, the optimal allocation of time-space-power resources should only be done once per channel stationary time, *i.e.*, when the large-scale fading coefficients change, or when a different scheduling policy should be found. The optimization is thus already light given the problem convexity, but on top, it has to be updated on a relatively slow time scale.

D. Practical Algorithm and Complexity

To solve problem (22), we utilize Algorithm 1 where every unconstrained problem is solved via Newton's method [46, Algorithm 9.5] with tolerance of 10^{-8}. After checking the problem feasibility, we solve the unconstrained two-dimensional problem $\min_{x,y} f(x, y)$. Through simulations, we observe that from 20 to 30 iterations are necessary to reach convergence. The gradient and Hessian of $f(x, y)$ depend on combinations of powers of $\phi(x) = \sum_{k=1}^{K} (\sigma_k^2/\beta_k) (2^{R_k x} - 1)$, $\phi'(x)$, $\phi''(x)$, x and y, where their values can be when reused among different terms in order to reduce complexity. If the solution of the problem lies in $\mathcal{D}$, the constraints are satisfied and the problem is solved. Otherwise, looking at the geometry of $\mathcal{D}$, either one or two constraints are not satisfied. We can identify six cases corresponding so six domains $\mathcal{R}_i$, $i = 1, \ldots, 6$, illustrated also in Fig. 2. Let us define

$$y_{\min}(x) = \frac{K}{2} + \frac{1}{2} \sqrt{K^2 + 4 \sum_{k=1}^{K} \rho_k^{-1} (2^{R_k x} - 1)}$$

as the value of y that satisfies the power constraint with equality. The feasibility domain can therefore be written as $\mathcal{D} = \{1 \leq x, y \leq M, y \geq y_{\min}(x)\}$. When the first, second and third constraints are not satisfied we obtain the domains $\mathcal{R}_1 = \{1 > x, y \leq M, y \geq y_{\min}(x)\}$, $\mathcal{R}_4 = \{1 \leq x, y > M, y \geq y_{\min}(x)\}$ and $\mathcal{R}_5 = \{1 \leq x, y \leq M, y < y_{\min}(x)\}$, respectively. When only the third, second and first constraints are satisfied we obtain the domains $\mathcal{R}_2 = \{1 > x, y > M, y \geq y_{\min}(x)\}$, $\mathcal{R}_3 = \{1 > x, y \leq M, y < y_{\min}(x)\}$ and $\mathcal{R}_6 = \{1 \leq x, y > M, y < y_{\min}(x)\}$, respectively. If the solution of $\min_{x,y} f(x, y)$ lies in $\mathcal{R}_1$, the first constraint is not satisfied. Making it binding sets the value of x to 1, and we can then solve the one-dimensional unconstrained problem $\min_y f(1, y)$. If the obtained solution is either larger than M or smaller than $y_{\min}(1)$, we make the unsatisfied constraint binding and this gives the final value of y. In case the solution of $\min_{x,y} f(x, y)$ lies in $\mathcal{R}_2$, we let the first constraint bind fixing x to 1, solve the problem $\min_y f(1, y)$ and further check the other two constraints in order to make the obtained y within bounds. We also let the second constraint bind fixing y to M, solve the problem $\min_x f(x, M)$ and further check the other two constraints fixing x either to 1 or to $x_{\max}$ if not satisfied, where $x_{\max}$ is the value solving $y_{\min}(x_{\max}) = M$. We then select among the two solutions, the one that minimizes $f(x, y)$. The same procedure is applied in the remaining four cases. Therefore, we might need to solve two of the three one-dimensional unconstrained problems $\min_y f(1, y)$, $\min_x f(x, M)$ and $\min_x f(x, y_{\min}(x))$. Through simulations, we observe that 5 to 20 iterations are necessary to solve any of the three problems. Computing the first and second derivatives of the three functions involves similar computations to the ones required for the gradient and Hessian of $f(x, y)$ but with fewer terms given that the problems are one-dimensional.

V. NUMERICAL EVALUATION

In this section, we numerically evaluate the proposed solution (24), which we refer to as optimized, and compare it with the awake-but-whisper, rush-to-sleep, and rush-to-sleep

---

Algorithm 1 Unconstrained optimization with inspection of the constraints to solve problem (22) and retrieve (24)

1: **Input:** $\{R_k\}$, $\{\beta_k\}$, $\{\sigma_k^2\}$, M, N, K, γ, α, P_0, P_1, $P_{\max}$
2: Compute $P_{a,\min}$ via (15). If $P_{a,\min} > P_{\max}$, quit
3: Set $(\bar{x}, \bar{y}) \leftarrow \arg\min_{(x,y)} f(x, y)$
4: **if** $(\bar{x}, \bar{y}) \in \mathcal{D}$ **then**
5:     Break, problem solved
6: **else if** $(\bar{x}, \bar{y}) \in \mathcal{R}_1$ **then**
7:     Set $\bar{x} \leftarrow 1$, $\bar{y} \leftarrow \arg\min_y f(1, y)$
        and make $\bar{y}$ within bounds
8: **else if** $(\bar{x}, \bar{y}) \in \mathcal{R}_2$ **then**
9:     Set $\bar{x}_1 \leftarrow 1$, $\bar{y}_1 \leftarrow \arg\min_y f(1, y)$
        and make $\bar{y}_1$ within bounds
10:     Set $\bar{y}_2 \leftarrow M$, $\bar{x}_2 \leftarrow \arg\min_x f(x, M)$
        and make $\bar{x}_2$ within bounds
11:     Set $(\bar{x}, \bar{y}) \leftarrow$ best* among $(\bar{x}_1, \bar{y}_1)$ and $(\bar{x}_2, \bar{y}_2)$
12: **else if** $(\bar{x}, \bar{y}) \in \mathcal{R}_3$ **then**
13:     Set $\bar{x}_1 \leftarrow 1$, $\bar{y}_1 \leftarrow \arg\min_y f(1, y)$
        and make $\bar{y}_1$ within bounds
14:     Set $\bar{x}_2 \leftarrow \arg\min_x f(x, y_{\min}(x))$, make $\bar{x}_2$ within bounds
        and set $\bar{y}_2 \leftarrow y_{\min}(\bar{x}_2)$
15:     Set $(\bar{x}, \bar{y}) \leftarrow$ best* among $(\bar{x}_1, \bar{y}_1)$ and $(\bar{x}_2, \bar{y}_2)$
16: **else if** $(\bar{x}, \bar{y}) \in \mathcal{R}_4$ **then**
17:     Set $\bar{y} \leftarrow M$, $\bar{x} \leftarrow \arg\min_x f(x, M)$
        and make $\bar{x}$ within bounds
18: **else if** $(\bar{x}, \bar{y}) \in \mathcal{R}_5$ **then**
19:     Set $\bar{x} \leftarrow \arg\min_x f(x, y_{\min}(x))$, make $\bar{x}$ within bounds
        and set $\bar{y} \leftarrow y_{\min}(\bar{x})$
20: **else if** $(\bar{x}, \bar{y}) \in \mathcal{R}_6$ **then**
21:     Set $\bar{y}_1 \leftarrow M$, $\bar{x}_1 \leftarrow \arg\min_x f(x, M)$
        and make $\bar{x}_1$ within bounds
22:     Set $\bar{x}_2 \leftarrow \arg\min_x f(x, y_{\min}(x))$, make $\bar{x}_2$ within bounds
        and compute $\bar{y}_2 \leftarrow y_{\min}(\bar{x}_2)$
23:     Set $(\bar{x}, \bar{y}) \leftarrow$ best* among $(\bar{x}_1, \bar{y}_1)$ and $(\bar{x}_2, \bar{y}_2)$
24: **end if**
25: **Output:** $(N_a, M_a) \leftarrow ([N/\bar{x}], [\bar{y}])$
*Best is the one that minimizes $f(x, y)$

---

strategies. The statistical distribution of the large-scale fading coefficients is derived from measurements of macro sub-6 GHz 4G and 5G BSs in Belgium [40], while the considered BS configurations and related parameters are given in Table II.

A. Large-Scale Fading and Network Load Computation

The evaluation scenario comprises the set of large-scale fading coefficients $\{\beta_k\}$ and users' target rates $\{R_k\}$. To compute the large-scale fading coefficients, we use the channel quality indicator (CQI) distribution obtained from on-site measurements of macro sub-6 GHz BSs in Belgium [40]. We selected the data from three sectors of two BSs:

- A 4G LTE RRU with 4 TX/RX chains and antennas, operating at $f_c = 1.8$ GHz with $B = 20$ MHz.
- A 5G NR AAU with 8 TX/RX chains and antennas, operating at $f_c = 3.5$ GHz with $B = 100$ MHz.

The raw data give the number of times each CQI index (from 0 to 15) is reported every hour over six days, and from those we compute the two CQI statistical distributions (one for 4G BSs and one for 5G BSs). Given the use of different modulation orders, 4G and 5G systems use two different CQI tables [47,

Table 5.2.2.1-2 and Table 5.2.2.1-3]. We project all the CQIs on the second table to perform a fair comparison. The CQI can then be mapped to the SNR by using the appropriate relations [48]. Once the SNR of user k is known, denoted as SNR_k, our rate model (3) shows that $\text{SNR}_k = p_k/\sigma_k^2$. Hence, the power allocation at user k is given by $\sigma_k^2 \text{SNR}_k$. The noise power is computed as $\sigma_k^2 = T_\sigma k_{\text{B}} B F_\sigma$, where $T_\sigma = 290$ K and $F_\sigma = 9$ dB are the noise temperature and the noise figure, and k_{B} is the Boltzmann constant [49]. Last, we use the expression of the total transmit power (1) to compute β_k by approximating $K = 1$, yielding

$$\beta_k = \frac{\sigma_k^2 \text{SNR}_k}{P_{\text{T}}(M - 1)}.$$

Based on the measurements and on the certificates of conformity that are publicly available upon request for deployed BSs [50], we set P_{T} to 160 W, 32 W, and 20 W for the 4T4T, 8T8R, and 64T64R configurations, respectively.

Concerning the users' target rates, let us consider some baseline rates $\{R_{k,0}\}$ normalized such that $\sum_{k=1}^K R_{k,0} = 1$, so $\{R_{k,0}\}$ are the users' shares of the sum rate. The baseline rates are then scaled by κ giving $R_k = \kappa R_{k,0}$. The maximum rate scaling $\kappa_{\max}$ is therefore the value of κ that solves (16) with equality when $M_a = M$ and $N_a = N$

$$M = \frac{K}{2} + \frac{1}{2}\sqrt{K^2 + 4\sum_{k=1}^K \rho_k^{-1}\left(2^{\kappa_{\max}R_{k,0}} - 1\right)}$$

for a fixed set of $\{\beta_k\}$ and $\{R_{k,0}\}$. With this choice and given the definition of $N_{a,\min}$ in (18), $N_{a,\min} = \lceil(\kappa/\kappa_{\max})N\rceil$. In the following, we generate $\{R_{k,0}\}$ as K realizations of a standard uniform distribution that are normalized to sum to 1, and we define the network load as the ratio $\kappa/\kappa_{\max}$. Larger κ indicate more loaded systems in terms of larger user rates.

B. Power Consumption and Optimal Number of Resources as a Function of Network Load

We consider a fixed set of $\{\beta_k\}$ and $\{R_{k,0}\}$, giving a unique value of $\kappa_{\max}$, and vary κ to observe the evolution of P_{cons} and (N_a, M_a) as the network load changes. We focus here on the massive MIMO configuration 64T64R. We notice for this and other configurations that with the computed values of β_k, the domain $\mathcal{D}$ is often not convex. This and the use of the rounding operator make the proposed solution an approximation of the optimal one. However, in many cases the solution lies on the left-edge of $\mathcal{D}$ (see Fig. 2), hence the non-convexity of the domain has no effect on the approximated solution.

The left plot of Fig. 3 shows that at any network load, the rush-to-mute approach achieves similar or equivalent performance to the optimized solution when there are no time-domain hardware power savings ($\delta_{\text{PA}}^{\text{dtx}} = 1$ and $\delta_{\text{TRX}}^{\text{idle}} = 1$). This is in line with the values in Table II, where P_0 is zero while P_1 assumes large values. All the four strategies consume P_{sleep} at zero network load, while they consume the same amount of power at maximum network load as that point corresponds to the activation of all the time slots and the antennas at $P_{\max}$. The difference between the proposed or rush-to-mute scheme and the awake-but-whisper or rush-to-sleep
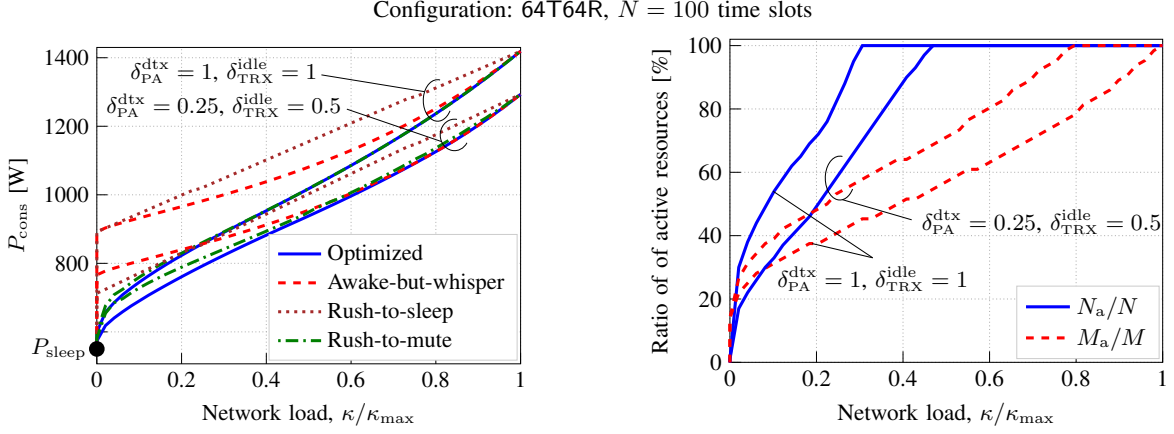
Fig. 3. For 64T64R configuration, (left) power consumption vs. network load for different energy-saving strategies, (right) optimal number of active spatial and time resources vs. network load. Disabled time-domain hardware power-saving modes corresponds to $\delta_{\mathrm{PA}}^{\mathrm{dtx}} = 1$ and $\delta_{\mathrm{TRX}}^{\mathrm{idle}} = 1$, while enabled time-domain hardware power-saving modes corresponds to $\delta_{\mathrm{PA}}^{\mathrm{dtx}} = 0.25$ and $\delta_{\mathrm{TRX}}^{\mathrm{idle}} = 0.5$.
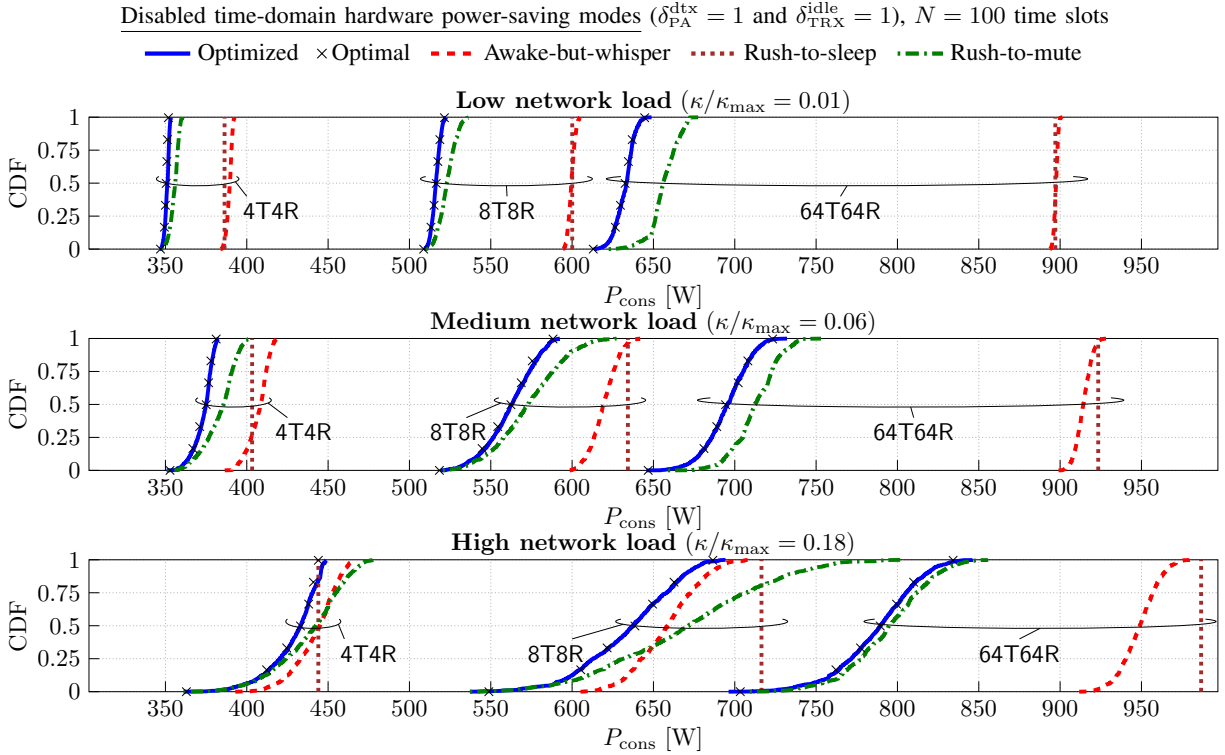

Fig. 4. For disabled PA μDTX and AFE idle-mode power savings, CDFs of power consumption for the three BS configurations at different network loads.

scheme becomes smaller as the network load increases. This is logical as most of the resources should be activated when there are large users' target rates. When time-domain power-saving modes are enabled ($\delta_{\mathrm{PA}}^{\mathrm{dtx}} = 0.25$ and $\delta_{\mathrm{TRX}}^{\mathrm{idle}} = 0.5$), the gap between the optimized and rush-to-mute solutions increases at low network loads because the active PAs and TR/RX chains consume less power. Also, the awake-but-whisper becomes optimal at very high network loads.

The right plot of Fig. 3 presents the change in number of optimal resources with the network load. The value of N_{a} first increases and then reaches its maximum value N at a certain value of network load. On the other hand, M_{a} keeps increasing until the maximum network load when there are

no time-domain power savings, while it reaches its maximum value M at a traffic load below 1 when time-domain power savings are enabled. We observe a similar trend for different $\{\beta_k\}$ and $\{R_{k,0}\}$. Therefore, when looking at the dependency on the network load in massive MIMO 64T64R configuration we find that:

- When there are no time-domain hardware power-saving modes, the optimal energy-saving strategy operates first in time, space and power, and then in space and power only.
- When there are time-domain hardware power-saving modes, the optimal energy-saving strategy operates first in time, space and power, then in space and power only, and last in power only.
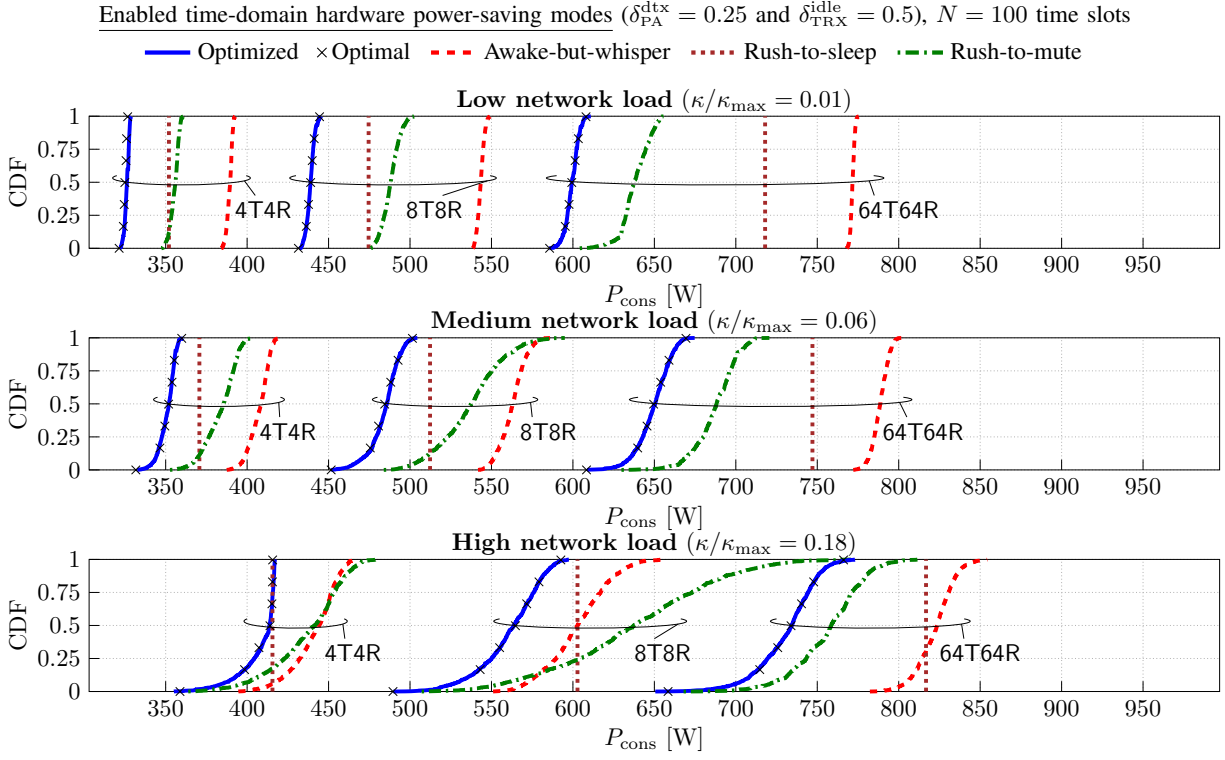
Enabled time-domain hardware power-saving modes ($\delta_{\text{PA}}^{\text{dtx}} = 0.25$ and $\delta_{\text{TRX}}^{\text{idle}} = 0.5$), $N = 100$ time slots

Fig. 5. For enabled PA µDTX and AFE idle-mode power savings, CDFs of power consumption for the three BS configurations at different network loads.

C. Average Power Consumption Performance

We now evaluate the CDF of the power consumed by the four energy-saving schemes for the three BS configurations at three network load levels: low, medium, and high network load, corresponding to $\kappa/\kappa_{\text{max}} = 0.01$, $\kappa/\kappa_{\text{max}} = 0.06$, and $\kappa/\kappa_{\text{max}} = 0.18$, respectively. We consider 10^3 realizations of $\{\beta_k\}$ and $\{R_{k,0}\}$. The CDFs of the rush-to-sleep approach assume values only in one point given that $N_{\text{a,min}}$ is fixed as $\lceil (\kappa/\kappa_{\text{max}}) N \rceil$. We observe that the CDFs obtained by using the optimized solution (solid blue) and the global optimum (black crosses) present a negligible difference.

In Fig. 4 we investigate the performance when PA µDTX and AFE idle-mode power savings are disabled. At low network load, the optimized and rush-to-mute schemes perform significantly better than the rush-to-mute and awake-but-whisper for any BS configuration. The optimized solution achieves 30% (64T64R), 14% (8T8R), and 10% (4T4R) median energy savings with respect to the rush-to-sleep and awake-but-whisper. At medium network load, the difference between the four energy-saving schemes decreases but the optimized and rush-to-mute remain well separated from the other two schemes. This gap further decreases at high network load, as highlighted also previously. Except for the 64T64R BS, the rush-to-mute approach becomes more distant to the proposed scheme while the awake-but-whisper (for the 8T8R BS) and both the awake-but-whisper and rush-to-mute (for the 4T4R BS) reduce their gap with the proposed scheme. We highlight that the three BS configurations cannot be directly compared as a larger M is associated with larger κ_{max}, K, and sum rate.

The performance of the rush-to-sleep technique compara-

tively improves when PA µDTX and AFE idle-mode power savings are enabled, as shown in Fig. 5. For the 64T64R configuration at any network load, the rush-to-mute scheme remains the closest to the optimal one, with a larger difference than in Fig. 4 indicating that the optimal scheme uses a combination of time- space-, and power-domain energy-saving techniques. The optimized scheme achieves 17%, 13%, and 6% median energy savings with respect to awake-but-whisper, rush-to-mute, and rush-to-sleep at medium network load. For the 8T8R configuration, the rush-to-sleep achieves the highest performance among the benchmarks at low and medium network loads, while at high network load its performance is similar to the awake-but-whisper. The rush-to-sleep is instead the closest approach to the proposed one for the 4T4R configuration at any network load. In general, even with active time-domain hardware power-saving modes, the cost of activating all the BS antennas in massive MIMO remains high and this still favors a rush-to-mute allocation. Different is the case for the two other configurations where the optimized solution exploits more the degrees of freedom in time.

VI. CONCLUSION

We studied the BS resource allocation problem in time, space, and power domains under QoS and power constraints using a state-of-the-art parametric power consumption model. The initially involved optimization problem can be simplified and solved via one, two or three unconstrained optimizations. Practical insights are drawn for 4G and 5G BSs. It turns out that a rush-to-mute approach, using as less antennas as possible at maximum output power, provides quasi-optimal performance for the analyzed BSs that are not implementing

time-domain power-saving modes. When the BS hardware can enter time-domain power-saving modes, the optimal energy-saving strategy utilizes combinations of intermediate number of active time slots, active antennas, and transmit power below the maximum one. The difference between the use of the three energy-saving domains decreases at high network loads. Median energy savings up to 25% are attained at medium network loads. Future work includes deriving scaling regimes and optimality conditions for the three one-dimensional energy-saving allocations, the inclusion of frequency-domain energy-saving techniques, and the analysis of cell-free scenarios. Further, evaluating the energy savings of the proposed scheme under actual traffic scenarios, using real hardware and considering instantaneous performance indicators over the average ones, is highly relevant.

VII. APPENDIX

A. Simplification of $P_{\max}$ Constraint (14)

Let us use the definition $y = M_{\mathrm{a}}$ and consider it as continuous. Using $\rho_k = P_{\max}\beta_k/\sigma_k^2$, we can rewrite (14) as

$$0 \leq y^2 - Ky - \sum_{k=1}^{K} \rho_k^{-1} \left(2^{R_k \frac{N}{N_{\mathrm{a}}}} - 1 \right).$$

The right term is a quadratic expression in y with two roots

$$y = \frac{K}{2} \pm \frac{1}{2}\sqrt{K^2 + 4\sum_{k=1}^{K} \rho_k^{-1} \left(2^{R_k \frac{N}{N_{\mathrm{a}}}} - 1 \right)}.$$

Given that $M_{\mathrm{a}} \geq K$, also $y \geq K$ and we should only keep the one with positive sign. Given that the parabolic function is looking upward, we find the condition

$$y \geq \frac{K}{2} + \frac{1}{2}\sqrt{K^2 + 4\sum_{k=1}^{K} \rho_k^{-1} \left(2^{R_k \frac{N}{N_{\mathrm{a}}}} - 1 \right)}$$

which implies (16) going back to the discrete domain.

B. Convexity of the Function $f(x,y)$

We want to show that the function (19) is jointly convex, considering that $x \geq 1$ and $K \leq y \leq M$. Clearly, the term in y is convex. To show the convexity of the term in $\frac{y}{x}$, we check that its Hessian is positive semidefinite. A sufficient condition is that its determinant and trace are positive. The determinant is given by $\frac{1}{x^4} \geq 0$, while the trace is given by $\frac{y}{x^3} \geq 0$. Therefore, $\frac{y}{x}$ is convex in the domain of interest. We are left with the term $g(y)h(x)$, where

$$g(y) = \frac{y}{(y(y-K))^\alpha}, \quad h(x) = \frac{\left(\sum_{k=1}^{K} \zeta_k (e^{R_k' x} - 1) \right)^\alpha}{x}$$

with $R_k' = R_k \ln 2$ and $\zeta_k = \sigma_k^2/\beta_k$. To show convexity, let us show that the determinant and the trace of its Hessian are positive. The determinant condition requires that

$$h''(x)h(x)g''(y)g(y) - (h'(x)g'(y))^2 \geq 0$$

which is verified if the two following conditions hold

$$g''(y)g(y) - (g'(y))^2 \geq 0, \quad h''(x)h(x) - (h'(x))^2 \geq 0.$$

The trace condition is $h''(x)g(y) + h(x)g''(y) \geq 0$ and given that $h(x) \geq 0$ and $g(y) \geq 0$, it is verified by checking that both functions are convex, i.e., $h''(x) \geq 0, g''(y) \geq 0$. Let us first study $g(y)$, with derivatives

$$g'(y) = \frac{1}{y^\alpha(y-K)^{1+\alpha}} \left(y(1-2\alpha) - K(1-\alpha) \right)$$

$$g''(y) = \alpha\frac{2y^2(2\alpha-1) + 4Ky(1-\alpha) - K^2(1-\alpha)}{y^{1+\alpha}(y-K)^{2+\alpha}}.$$

The denominator of $g''(y)$ is positive and given that $\alpha \geq 0.5$, $2\alpha - 1 \geq 0$. Moreover, $y \geq K$ implies $4Ky \geq K^2$, hence $g''(y) \geq 0$. By developing $g''(y)g(y) - (g'(y))^2$, we obtain

$$\frac{y^2(2\alpha-1) + y2K(1-\alpha) - K^2(1-\alpha)}{y^{2\alpha}(y-K)^{2+2\alpha}}.$$

Again, the denominator is positive and given that $\alpha \geq 0.5$, we have $2\alpha - 1 \geq 0$. Moreover, $y \geq K$ implies $y2K \geq K^2$. Hence, $g''(y)g(y) - (g'(y))^2 \geq 0$. Let us now study $h(x)$ and let us define the function $k(x) = \ln h(x)$. Given that $h(x) = e^{k(x)}$, we have

$$h'(x) = e^{k(x)}k'(x), \quad h''(x) = e^{k(x)}(k'(x))^2 + e^{k(x)}k''(x)$$

$$h''(x)h(x) - (h'(x))^2 = e^{2k(x)}k''(x).$$

Therefore, it is sufficient to show that $k''(x) \geq 0$. After simplification and manipulation, $k''(x)$ can be expressed as

$$k''(x) = \alpha\frac{\sum_{k=1}^{K}\sum_{k'=1}^{K} \zeta_k\zeta_{k'} t_{k,k'}}{\left(\sum_{k=1}^{K} \zeta_k(e^{R_k' x} - 1) \right)^2} + (1-\alpha)\frac{1}{x^2}.$$

where

$$t_{k,k'} = \frac{1}{2}R_k'^2 e^{xR_k'}(e^{xR_{k'}'} - 1) + \frac{1}{2}(e^{xR_k'} - 1)e^{xR_{k'}'}R_{k'}'^2$$
$$- R_k' e^{R_k' x}e^{R_{k'}' x}R_{k'}' + \frac{1}{x^2}(e^{R_k' x} - 1)(e^{R_{k'}' x} - 1).$$

Given that $\alpha \in [0.5, 1]$, a sufficient condition for $k''(x) \geq 0$ is that $t_{k,k'} \geq 0 \; \forall k, k'$. Using $\sinh(x) \geq x$ for $x \geq 0$, we find the inequality

$$(e^{xR_k'} - 1) = 2e^{\frac{xR_k'}{2}} \sinh\left(xR_k'/2\right) \geq e^{\frac{xR_k'}{2}}xR_k'$$

and thus

$$t_{k,k'} \geq \frac{1}{2}R_k'^2 e^{xR_k'}(e^{xR_{k'}'} - 1) + \frac{1}{2}(e^{xR_k'} - 1)e^{xR_{k'}'}R_{k'}'^2$$
$$- R_k' e^{xR_k'}e^{xR_{k'}'}R_{k'}' + R_k' e^{xR_k'/2}e^{xR_{k'}'/2}R_{k'}'$$
$$= -\frac{1}{2}\left(R_k' e^{\frac{xR_k'}{2}} - e^{\frac{xR_{k'}'}{2}}R_{k'}' \right)^2 + \frac{1}{2}e^{xR_k' + xR_{k'}'}(R_k' - R_{k'}')^2.$$

We need to show that this term is greater or equal than zero. Multiplying by $x^2/2$ and defining $a = xR_k/2$ and $b = xR_{k'}/2$, it is equivalent to show that

$$e^{2a+2b}(a-b)^2 \geq \left(ae^a - be^b \right)^2.$$

Let us consider that $a > b$ (otherwise symmetric) and take the square root, we should then show that

$$e^{a+b}(a-b) \geq ae^a - be^b.$$

Subtracting $(a-b)e^a$ gives $(e^{a+b} - e^a)(a-b) \geq b(e^a - e^b)$,

and dividing by e^a, we obtain $(e^b - 1)(a - b) \geq b(1 - e^{b-a})$. We can then exploit $b \leq e^b - 1$ and $1 - e^{b-a} \leq a - b$ to arrive at the result.

C. Convexity of the Domain $\mathcal{D}$

The linear constraints in the definition of $\mathcal{D}$ in (20) are clearly convex. Let us look further at the $P_{\max}$ inequality, which can be rewritten as $-y + \frac{K}{2} + m(x) \leq 0$, where

$$m(x) = \sqrt{\frac{K^2}{4} + \sum_{k=1}^{K} \rho_k^{-1}(e^{R'_k x} - 1)}.$$

Given that $-y$ is convex, we need to check when $m''(x) \geq 0$. Let us further define

$$l(x) = \sum_{k=1}^{K} \rho_k^{-1}\left(e^{R'_k x} + \frac{K}{4}\rho_k - 1\right)$$

so that $m(x) = \sqrt{l(x)}$ with derivatives

$$m'(x) = \frac{1}{2}\frac{l'(x)}{\sqrt{l(x)}}, \ m''(x) = \frac{1}{4}\frac{2l''(x)l(x) - (l'(x))^2}{(l(x))^{3/2}}.$$

Given that $l(x) \geq 0$, we need to check when $2l''(x)l(x) - (l'(x))^2 \geq 0$. We have

$$2l''(x)l(x) - (l'(x))^2 =$$

$$\sum_{k=1}^{K}\sum_{k'=1}^{K}\frac{R_k 2^{R_k x}}{\rho_k \rho_{k'}}\left[2R_k\left(\frac{K}{4}\rho_{k'} - 1\right) + 2^{R_{k'} x}(2R_k - R_{k'})\right].$$

The double sum contains K^2 additive terms which can be partitioned in K terms where $k = k'$ and $K(K - 1)$ terms where $k \neq k'$. In the following we show that the condition (21) which can be rewritten as

$$\rho_k \geq \frac{2}{K} \iff \frac{K}{4}\rho_k - 1 \geq -\frac{1}{2} \tag{25}$$

for $k = 1, \ldots, K$ is sufficient to guarantee that $k''(x) \geq 0$. Let us first consider the K terms of the double sum where $k = k'$. These are positive if

$$2R_k\left(\frac{K}{4}\rho_k - 1\right) + 2^{R_k x}(2R_k - R_k) \geq 0$$

$$-2\left(\frac{K}{4}\rho_k - 1\right) \leq 2^{R_k x}$$

which is verified if (25) holds. Let us now consider the $K(K - 1)$ terms where $k \neq k'$. For each term k, k', there exists also a term k', k. The total set of $K(K - 1)$ terms can then be subdivided in $K(K - 1)/2$ pairs of distinct indices k, k'. For instance, if $K = 3$, there exists $3(3 - 1)/2 = 3$ pairs: 1, 2 and 2, 1; 1, 3 and 3, 1; 2, 3 and 3, 2. Let us consider the addition of one pair of terms k, k' and k', k. The contribution of these two terms is positive if $f_{k,k'} \geq 0$ where

$$f_{k,k'} = R_k 2^{R_k x}\left[2R_k\left(\frac{K}{4}\rho_{k'} - 1\right) + 2^{R_{k'} x}(2R_k - R_{k'})\right]$$
$$+ R_{k'} 2^{R_{k'} x}\left[2R_{k'}\left(\frac{K}{4}\rho_k - 1\right) + 2^{R_k x}(2R_{k'} - R_k)\right].$$

If (25) holds, we have

$$f_{k,k'} \geq R_k 2^{R_k x}\left[-R_k + 2^{R_{k'} x}(2R_k - R_{k'})\right]$$
$$+ R_{k'} 2^{R_{k'} x}\left[-R_{k'} + 2^{R_k x}(2R_{k'} - R_k)\right]$$
$$= 2^{R_k x + R_{k'} x}\left[2R_k^2 + 2R_{k'}^2 - 2R_{k'}R_k - \frac{R_k^2}{2^{R_{k'} x}} - \frac{R_{k'}^2}{2^{R_k x}}\right]$$

and given that $-1/2^{R_k x} \geq -1$ and $-1/2^{R_{k'} x} \geq -1$, we find

$$f_{k,k'} \geq 2^{R_k x + R_{k'} x}\left[2R_k^2 + 2R_{k'}^2 - 2R_{k'}R_k - R_k^2 - R_{k'}^2\right]$$
$$= 2^{R_k x + R_{k'} x}(R_k - R_{k'})^2$$

which is positive. Hence, the contribution from all pairs of indices where $k \neq k'$ gives a positive contribution to the double sum. The same holds for the terms where $k = k'$. The proof is thus complete.

D. Error Order of Using a Rounding Operation

Let us define the optimal allocation as $(N_{\mathrm{a}}, M_{\mathrm{a}}) = \lfloor N/\bar{x}, \bar{y}\rfloor$ and its approximation through the rounding operator as $\hat{N}_{\mathrm{a}} = [N/\bar{x}], \ \hat{M}_{\mathrm{a}} = [\bar{y}]$. We can also write

$$\hat{N}_{\mathrm{a}} = N_{\mathrm{a}} + \epsilon_1, \ \hat{M}_{\mathrm{a}} = M_{\mathrm{a}} + \epsilon_2$$

where $|\epsilon_1| \leq 1$, $|\epsilon_2| \leq 1$. The extra consumed power of the approximation with respect to the optimal allocation is defined

$$\epsilon_3 = P_{\mathrm{cons}}(\hat{N}_{\mathrm{a}}, \hat{M}_{\mathrm{a}}) - P_{\mathrm{cons}}(N_{\mathrm{a}}, M_{\mathrm{a}})$$

where the function $P_{\mathrm{cons}}(N_{\mathrm{a}}, M_{\mathrm{a}})$ is given in (11). Developing ϵ_3, we find

$$\epsilon_3 = \frac{\epsilon_2}{M}P_1 + (N_{\mathrm{a}}\epsilon_2 + \epsilon_1 M_{\mathrm{a}} + \epsilon_1\epsilon_2)\frac{P_0}{NM}$$
$$+ \frac{N_{\mathrm{a}} + \epsilon_1}{N}(M_{\mathrm{a}} + \epsilon_2)\gamma\left(\frac{\sum_{k=1}^{K}\frac{\sigma_k^2}{\beta_k}\left(2^{R_k \frac{N}{N_{\mathrm{a}} + \epsilon_1}} - 1\right)}{(M_{\mathrm{a}} + \epsilon_2)(M_{\mathrm{a}} + \epsilon_2 - K)}\right)^{\alpha}$$
$$- \frac{N_{\mathrm{a}}}{N}M_{\mathrm{a}}\gamma\left(\frac{\sum_{k=1}^{K}\frac{\sigma_k^2}{\beta_k}\left(2^{R_k \frac{N}{N_{\mathrm{a}}}} - 1\right)}{M_{\mathrm{a}}(M_{\mathrm{a}} - K)}\right)^{\alpha}.$$

As $M_{\mathrm{a}} \to +\infty$ and $N \to +\infty$, the first two terms can be bounded as $\mathcal{O}(1/N + 1/M)$. For the two last ones, after performing a first order Taylor expansion around $\epsilon_1 = 0$ and $\epsilon_2 = 0$, we find the bound $\mathcal{O}(1/N + 1/M_{\mathrm{a}}^{2\alpha})$ so that $|\epsilon_3| = \mathcal{O}(1/N + 1/M + 1/M_{\mathrm{a}}^{2\alpha})$.

REFERENCES

[1] GSMA, "5G energy efficiencies: Green is the new black," Nov. 2020. [Online]. Available: https://data.gsmaintelligence.com/api-web/v2/research-file-download?id=54165956&file=241120-5G-energy.pdf

[2] European Commission Joint Research Centre, G. Kamiya, and P. Bertoldi, "Energy Consumption in Data Centres and Broadband Communication Networks in the EU," Publications Office of the European Union, Luxembourg, Feb. 2024. [Online]. Available: https://data.europa.eu/doi/10.2760/706491

[3] Ericsson Mobility Report, Nov. 2023. [Online]. Available: https://www.ericsson.com/4ae12c/assets/local/reports-papers/mobility-report/documents/2023/ericsson-mobility-report-november-2023.pdf

[4] M. Gruber, O. Blume, D. Ferling, D. Zeller, M. A. Imran, and E. C. Strinati, "EARTH – Energy aware radio and network technologies," in *Proc. IEEE Int. Symp. Pers. Indoor Mob. Radio Commun.*, Tokyo, Japan, 2009, pp. 1–5.

[5] H. Viswanathan, S. Wesemann, J. Du, and H. Holma, "Energy efficiency in next-generation mobile networks," White paper, Nokia Bell Labs, Nov. 2022. [Online]. Available: https://www.nokia.com/asset/212810

[6] L. Golard, J. Louveaux, and D. Bol, "Evaluation and projection of 4G and 5G RAN energy footprints: the case of Belgium for 2020–2025," *Ann. Telecommun.*, Nov. 2022.

[7] 3GPP, "Study on network energy savings for NR (Release 18)," Tech. Rep. 38.864, Dec. 2022, version 18.0.0.

[8] T. Islam, D. Lee, and S. S. Lim, "Enabling network power savings in 5G-advanced and beyond," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1888–1899, Jun. 2023.

[9] D. Laselva, S. Hakimi, M. Lauridsen, B. Khan, D. Kumar, and P. Mogensen, "On the potential of radio adaptations for 6G network energy saving," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit*, Antwerp, Belgium, 2024, pp. 1157–1162.

[10] D. López-Pérez et al., "A survey on 5G radio access network energy efficiency: Massive MIMO, lean carrier design, sleep modes, and machine learning," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 653–697, Firstquarter 2022.

[11] A. M. Busch, K. Eger, and B. Richerzhagen, "Comparison of power consumption models for 5G cellular network base stations," in *Intelligent Distributed Computing XVI*, M. Köhler-Bußmeier, W. Renz, and J. Sudeikat, Eds. Cham, Switzerland: Springer Nature Switz., 2024, pp. 199–212.

[12] L. Golard, Y. Agram, F. Rottenberg, F. Quitin, D. Bol, and J. Louveaux, "A parametric power model of multi-band sub-6 GHz cellular base stations using on-site measurements," in *Proc. IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. (PIMRC)*, Valencia, Spain, 2024, pp. 1–7.

[13] T. S. Rappaport, M. Ying, N. Piovesan, A. De Domenico, and D. Shakya, "Waste factor and waste figure: A unified theory for modeling and analyzing wasted power in radio access networks for improved sustainability," *IEEE Open J. Commun. Soc*, vol. 5, pp. 4839–4867, 2024.

[14] P. Frenger, P. Moberg, J. Malmodin, Y. Jading, and I. Godor, "Reducing energy consumption in LTE with cell DTX," in *Proc. IEEE 73rd Veh. Technol. Conf. (VTC Spring)*, Budapest, Hungary, 2011, pp. 1–5.

[15] J.-F. Cheng, H. Koorapaty, P. Frenger, D. Larsson, and S. Falahati, "Energy efficiency performance of LTE dynamic base station downlink DTX operation," in *Proc. IEEE 79th Veh. Technol. Conf. (VTC Spring)*, Seoul, Republic of Korea, 2014, pp. 1–5.

[16] E. Dahlman, S. Parkvall, and J. Sköld, *5G NR: The Next Generation Wireless Access Technology*, 2nd ed. New York, NY, USA: Academic Press, 2020.

[17] F. E. Salem, A. Gati, Z. Altman, and T. Chahed, "Advanced sleep modes and their impact on flow-level performance of 5G networks," in *Proc. IEEE Veh. Technol. Conf. Fall*, Toronto, ON, Canada, 2017, pp. 1–7.

[18] P. Frenger and R. Tano, "More capacity and less power: How 5G NR can reduce network energy consumption," in *Proc. IEEE Veh. Technol. Conf. Spring*, Kuala Lumpur, Malaysia, 2019, pp. 1–5.

[19] D. Renga, Z. Umar, and M. Meo, "Trading off delay and energy saving through advanced sleep modes in 5G RANs," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 7172–7184, Nov. 2023.

[20] K. Liu et al., "Analysis of a new energy-efficient model for future wireless communication systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 6, pp. 5503–5514, Jun. 2024.

[21] Ł. Budzisz et al., "Dynamic resource provisioning for energy efficiency in wireless access networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2259–2285, Fourthquarter 2014.

[22] S. Zhang, Q. Wu, S. Xu, and G. Y. Li, "Fundamental green tradeoffs: Progresses, challenges, and impacts on 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 33–56, Firstquarter 2017.

[23] F. Rottenberg, "Information-theoretic study of time-domain energy-saving techniques in radio access," *IEEE Trans. Green Commun. Netw.*, pp. 1–1, 2024.

[24] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3059–3075, Jun. 2015.

[25] M. M. A. Hossain, C. Cavdar, E. Björnson, and R. Jantti, "Energy saving game for massive MIMO: Coping with daily load variation," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2301–2313, Mar. 2018.

[26] K. Senel, E. Björnson, and E. G. Larsson, "Joint transmit and circuit power minimization in massive MIMO with downlink SINR constraints: When to turn on massive MIMO?" *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1834–1846, Mar. 2019.

[27] E. Peschiera and F. Rottenberg, "Energy-saving precoder design for narrowband and wideband massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 4, pp. 1793–1806, Dec. 2023.

[28] N. Rajapaksha, J. Mohammadi, S. Wesemann, T. Wild, and N. Rajatheva, "Minimizing energy consumption in MU-MIMO via antenna muting by neural networks with asymmetric loss," *IEEE Trans. Veh. Technol.*, vol. 73, no. 5, pp. 6600–6613, May 2024.

[29] T. Van Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and load balancing optimization for energy-efficient cell-free massive MIMO networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6798–6812, Oct. 2020.

[30] N. Jayaweera, K. B. S. Manosha, N. Rajatheva, and M. Latva-aho, "Minimizing energy consumption in cell-free massive MIMO networks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 13 263–13 277, Sep. 2024.

[31] B. Yan, Z. Wang, J. Zhang, and Y. Huang, "Joint antenna activation and power allocation for energy-efficient cell-free massive MIMO systems," *IEEE Wirel. Commun. Lett.*, vol. 14, no. 1, pp. 243–247, Jan. 2025.

[32] M. B. Shenouda and T. N. Davidson, "Convex conic formulations of robust downlink precoder designs with quality of service constraints," *IEEE J. Sel. Top. Signal Process.*, vol. 1, no. 4, pp. 714–724, Dec. 2007.

[33] J. Joung, C. K. Ho, and S. Sun, "Spectral efficiency and energy efficiency of OFDM systems: Impact of power amplifiers and countermeasures," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 208–220, Feb. 2014.

[34] S. K. Mohammed and E. G. Larsson, "Per-antenna constant envelope precoding for large multi-user MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 3, pp. 1059–1071, Mar. 2013.

[35] H. V. Cheng, D. Persson, and E. G. Larsson, "Optimal MIMO precoding under a constraint on the amplifier power consumption," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 218–229, Jan. 2019.

[36] F. Sohrabi, C. Nuzman, J. Du, H. Yang, and H. Viswanathan, "Energy-efficient flat precoding for MIMO systems," *IEEE Trans. Signal Process.*, vol. 73, pp. 795–810, Feb. 2025.

[37] S. Marwaha et al., "Energy efficient operation of adaptive massive MIMO 5G HetNets," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 6889–6904, Jul. 2024.

[38] Z. Wei, P. Wang, Q. Shi, X. Zhu, and C. Masouros, "Towards structural sparse precoding: Dynamic time, frequency, space, and power multistage resource programming," *arXiv preprint arXiv:2310.09840*, 2023.

[39] S. Zhang, T. Cai, O. T. Demir, and C. Cavdar, "Multi-agent RL for sleep mode and antenna configuration with user offloading under dynamic traffic in massive MIMO networks," *IEEE Trans. Veh. Technol.*, pp. 1–16, 2025.

[40] Y. Agram, F. Rottenberg, and F. Quitin, "Measurement based time-domain power saving through radio equipment deactivation on sub-6GHz base station site," in *Proc. 9th Int. Conf. Green Commun., Comput. Technol. (GREEN 2024)*, Nice, France, 2024.

[41] E. Peschiera, Y. Agram, F. Quitin, L. Van der Perre, and F. Rottenberg, "On optimizing time-, space- and power-domain energy-saving techniques for sub-6 GHz massive MIMO base stations," submitted to 2025 IEEE 26th International Workshop on Signal Processing and Artificial Intelligence for Wireless Communications (SPAWC).

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

[43] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Massive MIMO performance evaluation based on measured propagation data," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3899–3911, Jul. 2015.

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

[45] D. Maiwald and D. Kraus, "On moments of complex Wishart and complex inverse Wishart distributed matrices," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 5, Munich, Germany, 1997, pp. 3817–3820.

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

[47] 3GPP, "NR; Physical layer procedures for data," Tech. Rep. 38.214, Oct. 2024, version 18.4.0.

[48] Y. Wang, W. Liu, and L. Fang, "Adaptive modulation and coding technology in 5G system," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Limassol, Cyprus, 2020, pp. 159–164.

[49] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[50] Environmental Dept., Flanders Government, "Certificate of conformity of transmitting antennas," Sep. 2021, certificate no. 00108929.