Information-Theoretic Study of Time-Domain Energy-Saving Techniques in Radio Access

François Rottenberg

Abstract

Reduction of wireless network energy consumption is becoming increasingly important to reduce environmental footprint and operational costs. A key concept to achieve it is the use of lean transmission techniques that dynamically (de)activate hardware resources as a function of the load. In this paper, we propose a pioneering information-theoretic study of time-domain energy-saving techniques, relying on a practical hardware power consumption model of sleep and active modes. By minimizing the power consumption under a quality of service constraint (rate, latency), we propose simple yet powerful techniques to allocate power and choose which resources to activate or to put in sleep mode. Power consumption scaling regimes are identified. We show that a "rush-to-sleep" approach (maximal power in fewest symbols followed by sleep) is only optimal in a high noise regime. It is shown how consumption can be made linear with the load and achieve massive energy reduction (factor of 10) at low-to-medium load. The trade-off between energy efficiency (EE) and spectral efficiency (SE) is also characterized, followed by a multi-user study based on time division multiple access (TDMA).

Index Terms

Energy consumption, radio access technologies, physical layer, channel capacity.

I. INTRODUCTION

A. Motivation

As of 2022, yearly data volume has gone up to more than 3 Zettabytes (10²¹ bytes) and the traffic continues to rise at a rate of about 25%/year [1]. Energy efficiency has improved over the years but not fast enough, which results in an annual energy consumption growth of 2.5% for the ICT sector [1]. It is becoming increasingly important to reduce energy consumption of wireless communication networks to reach climate ambitions and reduce operational expenses, in other words, "break the energy curve" [2], [3]. Most energy of wireless networks is consumed by the radio access network (RAN) and more specifically at base stations [4], [5]. Moreover, the traffic load at a base station is highly varying across the day and most often lightly loaded, with

François Rottenberg is with ESAT-DRAMCO, Ghent Technology Campus, KU Leuven, 9000 Ghent, Belgium (e-mail: francois.rottenberg@kuleuven.be).

traffic at night being about 10 times lower than during the day [6]. This opens a big potential for energy reduction through the use of lean transmission techniques that dynamically activate or deactivate resources as a function of the load, letting the system dynamically switch from fully active to deep sleep mode.

B. State of the Art

Energy-saving techniques are a popular topic. A large effort has been made to integrate these techniques into industrial products and standards. The 5G standard was for instance designed with a lean paradigm in mind which resulted in, *e.g.*, less reference signaling to increase sleep duration [7]–[9]. At low-to-medium load, a popular scheduling technique is a "rush-to-sleep" approach which compacts transmission in as few symbols as possible. These symbols are transmitted at maximal power, leaving the remaining symbols in the frame free, so that the sleep duration is maximized. Many studies have been performed to evaluate the gains of such techniques based on standardized power models [5], system-level evaluations [10]–[14] and aided by actual measurements [15]. The use of machine learning was also identified as an interesting tool to predict traffic and/or to optimize energy-saving features [16], [17]. We refer to [8] for a review of these techniques.

Despite much work in the domain, there remains a fundamental gap to be filled by establishing an information-theoretic study of time-domain energy-saving techniques, even for basic systems such as single input single output (SISO) transceivers. Going back to the underlying physics of energy consumption of base stations and with a proper mathematical formulation of the communication link, a lot of additional understandings and improvements can be obtained: optimization of algorithms, finding optimal power scaling regimes as a function of load, guarantees of optimality for energy-saving features and/or finding the gap from it with existing techniques... As an example, it is not clear if or when a rush-to-sleep approach is optimal or not. We should mention that many recent works have performed this kind of energy-saving studies but they have focused on the spatial domain and more specifically the optimal operation of massive MIMO systems (number of active antennas, served users, power allocation) as a function of the load [18]–[22]. On the other hand, information-theoretic time-domain studies are lacking. The fundamental studies on energy-efficient communications have mainly focused on a stationary transmission of symbols in time at a constant rate and average transmit power $P_{\rm T}$ [23]–[27]. Considering an ideal consumption model and a quasi-static channel, this choice seems intuitive. To clarify it,

let us formalize the problem. The channel capacity of a complex discrete memoryless additive white gaussian noise (AWGN) channel, under an average transmit power constraint $P_{\rm T}$, is

$$R = \log_2\left(1 + \frac{P_{\rm T}}{L\sigma_n^2}\right) = \log_2\left(1 + \frac{P_{\rm T}}{\sigma^2}\right)$$
 [bits/channel use],

where $\sigma^2 = \sigma_n^2 L$ is the noise power at the receiver σ_n^2 normalized by the path loss L. If the transmission is divided in frames of N symbols, the average rate and transmit power are

$$R = \frac{1}{N} \sum_{n=0}^{N-1} \log_2 \left(1 + \frac{p_n}{\sigma^2} \right), \ P_T = \frac{1}{N} \sum_{n=0}^{N-1} p_n$$

where p_n is the transmit power of the n-th symbol. Considering a rate constraint R, let us find the power allocation that minimizes the consumed power P_{cons} . Under an ideal consumption model, we have $P_{\text{cons}}^{\text{ideal}} = P_{\text{T}}$ and the problem can be written as¹

$$\min_{p_0,\dots,p_{N-1}} \frac{1}{N} \sum_{n=0}^{N-1} p_n \text{ s.t. } \frac{1}{N} \sum_{n=0}^{N-1} \log_2 \left(1 + \frac{p_n}{\sigma^2} \right) = R.$$

Given the concavity of the $\log(.)$ function, we can write using the Jensen's inequality

$$R \le \log_2\left(1 + \frac{1}{\sigma^2} \frac{1}{N} \sum_{n=0}^{N-1} p_n\right) \leftrightarrow (2^R - 1)\sigma^2 \le \frac{1}{N} \sum_{n=0}^{N-1} p_n = P_T$$

and the bound is tight if uniform power allocation is used, i.e., $p_n = P_T = (2^R - 1)\sigma^2$, for n = 0, ..., N - 1. Intuitively, the log dependence of the rate implies diminishing returns. Starting from a non-uniform allocation, it can always be improved by reallocating some power from the time interval with the highest allocated power to the one with the lowest power.

In practice however, the consumed power $P_{\rm cons}$ is far from being equal or even linearly proportional to the transmit power $P_{\rm T}$. This is due to two main reasons, namely: i) as soon as a given time slot is active, a static load-independent power consumption is present due to activation of hardware components such as radio-frequency chains and baseband processing units; ii) the load-dependent power consumption, *i.e.*, the dependence of $P_{\rm cons}$ in p_n , is typically concave as power amplifiers (PAs) are more energy-efficient close to their saturation. Intuitively, this implies that the "cost" of using more power decreases when a large output power is transmitted. These two effects will counterbalance the log penalty and push towards using a reduced number of active time slots, especially in low-to-medium load scenarios.

¹This problem can be seen as a conventional waterfilling problem where water/noise levels are the same at each time slot.

C. Contributions

This paper presents a pioneering information-theoretic study of time-domain energy-saving techniques, using a realistic power consumption model. The transmission model considers singleantenna base stations and users. Even for such a basic system, a comprehensive study of energysaving features is lacking, which is the gap this paper is aiming to fill. The investigated techniques provide drastic energy reduction by dictating how to dynamically (de)activate hardware resources as a function of the load. The optimization problems are formalized as the minimization of $P_{\rm cons}$ for a given rate. More specifically, the structure of our paper and our contributions are structured as follows. Section II presents the hardware power consumption model used in this work, with two distinct contributions: active and sleep energy consumption. The active power consumption model is shown to address a large variety of PA classes. Section III considers the optimal allocation of time resources in a single-user scenario. The solution is approached step by step through lemmas to get more insight on its nature. Linear and exponential scaling regimes of $P_{\rm cons}$ as a function of the load R are identified. Asymptotic results for large N are provided that greatly simplify the analysis while having negligible performance penalty. We prove that a rush-to-sleep approach is optimal in a noise limited regime but not otherwise. The optimal trade-off EE-SE is also derived from previous results and we show that a maximal SE does not always provide a maximal EE. Section IV then extends previous results by considering successive sleep modes, resulting in drastic energy reductions. Section V considers the extension to a multi-user scenario where users are multiplexed using TDMA. The optimal allocation is provided for the most promising regime in terms of energy-savings, i.e., the low-to-medium-load scenario where P_{cons} linearly scales with the rate of each user and the system is not fully active. Finally, Section VI concludes the paper.

Notations: The operators $\lceil . \rceil$, $\lfloor . \rfloor$ and [.] are the ceil, floor and round operators, respectively. The operator $\lfloor x \rceil$ which we refer to as the ceil-floor operator selects among the upper and lower bounding integers of x the one that optimizes the cost function. The function W(z) is the Lambert W function, *i.e.*, the solution of $z = W(z)e^{W(z)}$. We use the notation f(x) = O(g(x)), as $x \to a$, if there exist positive numbers δ and λ such that $|f(x)| \le \lambda g(x)$ when $0 < |x-a| < \delta$.

II. POWER CONSUMPTION MODEL

As described in the introduction, we consider the transmission of N symbols, each of duration T [s]. The full frame has thus a duration NT [s]. Out of the N intervals, a number N_a are active

and actually transmitting information while the remaining $N-N_{\rm a}$ are inactive and in sleep mode. This implies that $0 \le N_{\rm a} \le N$. For minimizing latency and maximizing the sleep duration which allows entering a deeper sleep mode [28], the $N_{\rm a}$ active intervals are grouped together at the beginning of the transmission. We define as $E_{\rm active}$ the energy consumed during active time slots, which is assumed to depend on the transmit power at each time interval, i.e., $p_0, ..., p_{N_{\rm a}-1}$. On the other hand, $E_{\rm sleep}$ represents the energy consumed during sleep modes, which is non zero as all hardware components cannot be switched-off. It is assumed to depend on $(N-N_{\rm a})T$ as a longer sleep duration allows to enter a deeper sleep mode [28]. The average consumed power over the frame duration is thus given by

$$P_{\text{cons}} = \frac{E_{\text{active}}(p_0, \dots, p_{N_a-1}) + E_{\text{sleep}}((N - N_a)T)}{NT}.$$
(1)

In the following, we detail the models of the sleep and active energy consumption. As a benchmark, we also introduce the ideal consumption model

$$P_{\text{cons}}^{\text{ideal}} = \frac{1}{N} \sum_{n=0}^{N_{\text{a}}-1} p_n, \tag{2}$$

implying that the consumed power is equal to the transmit power. In other words, no losses are present.

A. Active Energy Consumption

Using the well-established model from [29], the active power consumption can be modelled as

$$P_{\text{active}} = \frac{P_{\text{PA}} + P_{\text{RF}} + P_{\text{BB}}}{(1 - \sigma_{\text{DC}})(1 - \sigma_{\text{MS}})(1 - \sigma_{\text{cool}})}$$

where $P_{\rm PA}$, $P_{\rm RF}$ and $P_{\rm BB}$ are the powers consumed by the PAs, the radio-frequency chains and the baseband unit respectively. The coefficients $\sigma_{\rm DC}$, $\sigma_{\rm MS}$ and $\sigma_{\rm cool}$ are the loss factors related to DC-DC power supply, mains supply and active cooling respectively.

We are interested in modelling the dependence of the active consumed power P_{active} in the output power at each active time slot $p_0, ..., p_{N_a-1}$. The base station (BS) power consumption analysis of [29] showed that mainly the PA consumed power P_{PA} scales with the output power. The other terms are thus considered load-independent. The active energy consumed across the frame duration can thus be written as

$$E_{\text{active}}(p_0, ..., p_{N_a-1}) = \frac{T}{\eta} \left(N_a \tilde{P}_0 + \sum_{n=0}^{N_{a-1}} P_{\text{PA}}(p_n) \right)$$

TABLE I

Values of loss factors and efficiency used in evaluations [29].

DC-DC	$\sigma_{ m DC}$	7.5%
Mains supply	$\sigma_{ m MS}$	9.0%
Cooling	$\sigma_{ m cool}$	10.0%
Efficiency	$\eta = (1 - \sigma_{\mathrm{DC}})(1 - \sigma_{\mathrm{MS}})(1 - \sigma_{\mathrm{cool}})$	75.8%

where $\tilde{P}_0 = P_{\rm BB} + P_{\rm RF}$ and $\eta = (1 - \sigma_{\rm DC})(1 - \sigma_{\rm MS})(1 - \sigma_{\rm cool})$. Values of loss factors and efficiencies used in evaluations are shown in Table I. To model the PA consumption, we use the following model

$$P_{\rm PA}(p) = P_{\rm PA,0} + \beta p^{\alpha}, \ 0 \le p \le P_{\rm max} \tag{3}$$

with $\alpha \in]0,1]$ and $\beta \geq 0$. The first term $P_{\mathrm{PA},0}$ represents the load-independent consumption while the second is load-dependent. This load dependency does not typically scale linearly with p. The fact that $\alpha \in]0,1]$ implies concavity of $P_{\mathrm{PA}}(p)$. This concavity comes from the fact that a typical PA efficiency is improved when moving closer to saturation [30]. The constant P_{max} denotes the maximal transmit power.In practice, P_{max} is (much) lower than the PA saturation power, that we denote by P_{sat} . The use of the so-called back-off $P_{\mathrm{max}}/P_{\mathrm{sat}}$ is required as recent technologies, e.g., orthogonal frequency division multiplexing (OFDM), have high peak-to-average power ratio (PAPR). A back-off (typically from -12 dB to -6 dB) prevents the PA to enter the saturation region, which would otherwise create nonlinear distortion impacting the signal quality and creating out-of-band emissions. The authors of [31] have justified in details the use of a similar model as (3) through their own measurements and a literature review [32], [33]. Model (3) is more general as it also includes a load-independent component, which is useful for particular PA architectures.

1) Ideal Power Amplifier: PA consumed power is linearly proportional to the output power giving $P_{\text{PA},0} = 0$, $\beta = \alpha = 1$ and

$$P_{\mathrm{PA}}^{\mathrm{ideal}}(p) = p.$$

2) Class A Power Amplifier: PA consumed power is independent of the load and has a maximal efficiency of 1/2 giving $P_{\rm PA,0}=2P_{\rm sat},~\beta=0$ and

$$P_{\rm PA}^{\rm A}(p) = 2P_{\rm sat}$$

where $P_{\rm sat}$ is the saturation power of the PA.

3) Class B Power Amplifier: The PA consumed power has a load dependence that scales with the square root of the output power giving $P_{\rm PA,0}=0$, $\beta=\frac{4}{\pi}\sqrt{P_{\rm sat}}$ and $\alpha=1/2$

$$P_{\rm PA}^{\rm B}(p) = \frac{4}{\pi} \sqrt{P_{\rm sat}} \sqrt{p}$$

This model has much relevance for typical base stations working with a significant back-off from saturation [31]. Some authors sometimes call it the "traditional" PA model [34]. Therefore, by default, we will use it in following evaluations, with a 8 dB back-off.

4) Envelope Tracking Power Amplifier: According to the curve fitted model proposed in [35], PA consumption was shown to be modelled as

$$P_{\mathrm{PA}}^{\mathrm{ET}}(p) \approx \frac{aP_{\mathrm{sat}}}{(1+a)\eta_{\mathrm{max}}} + \frac{1}{(1+a)\eta_{\mathrm{max}}}p,$$

where a = 0.0082. This model can again be seen as a special case of (3).

5) Doherty Power Amplifier: The ℓ-way Doherty PA consumed power is given by [25]

$$P_{\text{PA}}^{\text{Doherty}}(p) = \frac{4P_{\text{sat}}}{\ell\pi} \begin{cases} \sqrt{\xi} & 0 < \xi \le \frac{1}{\ell^2} \\ (\ell+1)\sqrt{\xi} - 1 & \frac{1}{\ell^2} < \xi \le 1 \end{cases}$$

where $\xi = p/P_{\rm sat}$. The class B model is obtained as a special case when $\ell = 1$. Except in such particular cases, the model proposed in (3) cannot exactly represent such PAs but can provide an approximation depending on the operating range of the Doherty amplifier.

Remark 1. We previously described how model (3) can address typical theoretical PA models. However, it can also be fitted for practical PAs based on measurements and/or datasheets. The PA was identified as the main load-dependent contribution. However, more generally, the model proposed in (3) can take into account other load-dependent terms.

In the light of this remark, we formalize the underlying assumption about the load-dependent active energy consumption model used throughout this work.

(As1): The active energy consumed across the frame is

$$E_{\text{active}}(p_0, ..., p_{N_a-1}) = T \left(N_a P_0 + \gamma \sum_{n=0}^{N_{a-1}} p_n^{\alpha} \right)$$
 (4)

where $\gamma=\frac{\beta}{\eta}\geq 0$ and $P_0=\frac{P_{\mathrm{BB}}+P_{\mathrm{RF}}+P_{\mathrm{PA},0}}{\eta}\geq 0,~\alpha\in]0,1],~0\leq p_n\leq P_{\mathrm{max}}$ for $n=0,\ldots,N_{\mathrm{a}}-1,$ $0\leq N_{\mathrm{a}}\leq N$ and $p_n=0$ for $n=N_{\mathrm{a}},\ldots,N-1$. The averaged consumed power is then

$$P_{\text{cons}} = \frac{N_{\text{a}}}{N} P_0 + \frac{\gamma}{N} \sum_{n=0}^{N_{\text{a}-1}} p_n^{\alpha} + \frac{E_{\text{sleep}}((N-N_{\text{a}})T)}{NT}.$$
 (5)

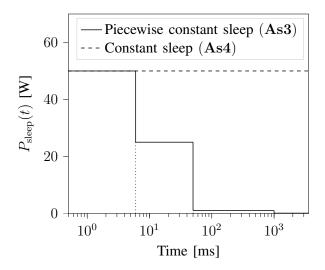


Fig. 1. Two classical sleep power models are shown: constant or successive sleep modes based on values in Table II.

B. Sleep Energy Consumption

In the ideal consumption model (2), the sleep energy consumption is exactly null. In practice, not all hardware can be switched off as always-on reference signals are required to allow users to access the network. Moreover, different hardware components have different activation/reactivation latencies. Hence, depending on the sleep duration, more or less components can be switched off. Therefore, power models have been proposed that consider successive sleep modes as a function of the sleep depth. We define the power consumption in sleep mode as $P_{\rm sleep}(t)$, where t=0 is used a reference for the system entering sleep and t is the sleep duration so that $E_{\rm sleep}(0)=0$. Given the fact that increasing sleep duration allows to switch off more hardware components, we introduce the following assumption.

(As2): $P_{\text{sleep}}(t)$ is monotonically non-increasing.

Proposition 1. Under (As2), the sleep energy consumption $E_{\text{sleep}}(t)$ is a concave function of the sleep duration t.

Proof. Directly follows from (As2).

Remark 2. (As2) does not imply continuity of $P_{\text{sleep}}(t)$, which can have jump discontinuities. Switching-off components might lead to a non-continuous drop of $P_{\text{sleep}}(t)$, as shown in Fig. 1.

Popular models for $P_{\text{sleep}}(t)$ include the use of different sleep modes. The model in [28] has been used as a qualitative and quantitative reference for several years by companies [36]. More

TABLE II

NUMERICAL VALUES OF POWER AND POWER MODELS USED IN EVALUATIONS.

Maximal transmit power	$P_{ m max}=20~{ m W}$			
Deep sleep power	$P_3 = 1 \text{ W}$			
Load-independent active power	$P_0 = 110P_3$			
	$P_{ m sleep}(t)$			
Successive sleep power model	Micro sleep	Light sleep	Deep sleep	Hibernating sleep
(As3) [5]	$T_1 = 0$	$T_2 = 6 \text{ ms}$	$T_3 = 50 \text{ ms}$	$T_4 = 1 \text{ s}$
	$P_1 = 50P_3$	$P_2 = 25P_3$	P_3	$P_4 = 0.1P_3$
Constant sleep power model (As4)	$P_{\text{sleep}}(t) = P_{\text{sleep}} = P_1$			

recently, 3GPP has introduced an improved model, expressed in relative units with respect to the deep sleep mode, that better reflects current trends [5], [37]. The model contains four sleep modes and numerical values are given in Table II for one configuration described in [5]. These values will be used as an example in evaluations.

We can formalize this model mathematically. Let us define as S the number of sleep modes, starting from mode 0 (no sleep) to mode S (deepest sleep mode). The start and end of each sleep mode are denoted by T_s and T_{s+1} , with $T_0 = 0$ and $T_{s+1} = +\infty$. The sleep power consumption during mode s is denoted by P_s . This notation is consistent with the definition of P_0 which denotes the load-independent active power consumption. This corresponds to "sleep mode zero", taking place until T_1 , where no hardware components are actually switched off.

(As3): the sleep power consumption is piecewise constant implying that $P_{\text{sleep}}(t) = P_s$ and

$$E_{\text{sleep}}(t) = \int_0^t P_{\text{sleep}}(t')dt' = \sum_{s'=0}^{s-1} P_{s'}(T_{s'+1} - T_{s'}) + (t - T_s)P_s$$

where s is the index such that $T_s < t \le T_{s+1}$ and $P_{s+1} \le P_s$ according to (As2).

We also introduce another popular sleep power model widely used in the literature and shown to be accurate to characterize 4G long term evolution (LTE) macro base stations [29].

(As4): the sleep power consumption is constant implying that $P_{\text{sleep}}(t) = P_{\text{sleep}}$ and $E_{\text{sleep}}(t) = P_{\text{sleep}}$ with $P_0 \ge P_{\text{sleep}}$. The averaged consumed power is then

$$P_{\text{cons}} = \frac{N_{\text{a}}}{N} P_0 + \frac{\gamma}{N} \sum_{n=0}^{N_{\text{a}-1}} p_n^{\alpha} + \frac{N - N_{\text{a}}}{N} P_{\text{sleep}}.$$
 (6)

Remark 3. (As4) is a particular case of (As3) with a single sleep mode: S = 1, $P_{\text{sleep}} = P_1$ and $T_1 = 0$. In the following, we use the term "(As3) – (As4)", when both assumptions hold.

III. OPTIMAL ALLOCATION OF TIME RESOURCES

This section considers the solution of minimizing the average consumed power under (As1) and a rate constraint

$$\min_{N_{\mathbf{a},p_0,\dots,p_{N_{\mathbf{a}}-1}}} P_{\text{cons}} \quad \text{s.t. } \frac{1}{N} \sum_{n=0}^{N_{\mathbf{a}}-1} \log_2\left(1 + \frac{p_n}{\sigma^2}\right) = R.$$
 (7)

We define a constant that will be useful throughout this section. Under $(\mathbf{As1}) - (\mathbf{As4})$, R_a is the constant that minimizes the convex problem

$$R_{\mathbf{a}} = \arg\min_{x \ge 0} \frac{P_0 - P_{\text{sleep}} + \gamma \sigma^{2\alpha} (2^x - 1)^{\alpha}}{x},$$

where P_{sleep} is the constant sleep power consumption defined in (As4). When $P_0 - P_{\text{sleep}} = 0$, it is given by

$$R_{\rm a} = (W(-\alpha^{-1}e^{-\alpha^{-1}}) + \alpha^{-1})/\log(2)$$
(8)

where W(z) is the Lambert W function, *i.e.*, the solution of $z = W(z)e^{W(z)}$. In the following, the general solution of (7) is approached step by step, introducing several lemmas, which provide insight on its form and scaling as a function of R. Finally, the trade-off SE versus EE will be characterized.

A. Optimal Allocation for Load-Dependent Consumed Power

The following lemma provides the power allocation that minimizes the load-dependent part of the average consumed power under a rate constraint and without a maximal per-time slot power constraint. Under (As1), the load-dependent part of P_{cons} can be identified as

$$P_{\rm ld}(p_0, ..., p_{N-1}) = \frac{\gamma}{N} \sum_{n=0}^{N-1} p_n^{\alpha}, \tag{9}$$

which equals P_{cons} for $P_0 = 0$ and $E_{\text{sleep}}(t) = 0$.

Lemma 1. Under (As1), for $P_0 = E_{\text{sleep}}(t) = 0$, $\gamma > 0$ and $P_{\text{max}} \to +\infty$, the minimum of (7), is achieved by uniformly allocating power among N_a time slots

$$N_{\mathbf{a}} = \lfloor \min(NR/R_{\mathbf{a}}, N) \rceil,$$

$$p_n = \begin{cases} \left(2^{R\frac{N}{N_{\mathbf{a}}}} - 1\right) \sigma^2 & \text{if } n = 0, ..., N_{\mathbf{a}} - 1\\ 0 & \text{otherwise} \end{cases}.$$

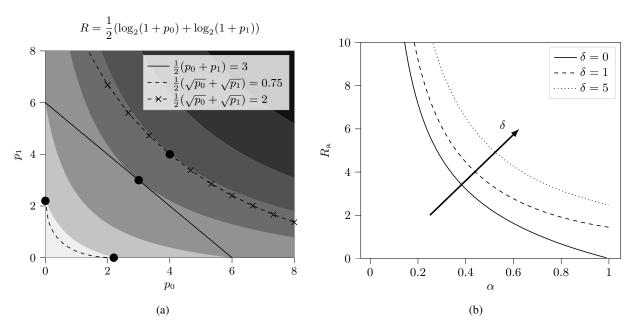


Fig. 2. (a) Contour plot of i) rate constraint and ii) load-dependent power consumption ($\gamma=1$) for a frame of N=2 time slots and two load-dependent power exponents α : 1 and 1/2. (b) Constant R_a as a function of the load-dependent power exponent α and the ratio $\delta=(P_0-P_{\rm sleep})/(\gamma\sigma^{2\alpha})$.

Proof. See Appendix VII-A.

Remark 4 (Frame of N=2 symbols). To illustrate Lemma 1, Fig. 2a considers the particular case $N=2, \ \gamma=1$. It shows contour plots of the constraint R and the cost function $P_{\rm ld}$, as a function of p_0 and p_1 . In the case $\alpha=1$, implying linearity of consumed and transmit power, the objective function curve $(p_0+p_1)/2=3$ is a straight line. As shown in the introduction, utilization of the two time slots is always optimal $(N_{\rm a}=2)$, with uniform power allocation. However, for $\alpha=1/2$, this is not anymore the case. For low values of R, it is better to only use $N_{\rm a}=1$ slot while for large R, $N_{\rm a}=2$ is optimal, again with uniform power.

Remark 5 (Scaling of N_a for general N and α). For an arbitrary value of N and $\alpha \in]0,1]$, the number of activated time slots N_a scales approximately linearly with the rate R up to the point where the maximal number N is allocated, *i.e.*, when $R > R_a$. When $R \leq R_a$, the rate per activated time slot is approximately equal to R_a . The "approximate" nature comes from the rounding operation. This error disappears for large N, as will be formalized properly in the following. The constant R_a is independent of R and is a function of α . As shown in Fig. 2b for the case $\delta = 0$ ($P_0 = P_{\text{sleep}} = 0$), R_a monotonically decreases as a function of α , implying that

more time slots are activated for a fixed value of R. In the asymptotic cases of α approaching 0 or 1, a single $(N_{\rm a}=1)$ or all time slots $(N_{\rm a}=N)$ are allocated, respectively.

B. Optimal Allocation with no Maximal Power Constraint

The following lemma gives the power allocation that minimizes the averaged consumed power under a rate constraint and without a maximal per-time slot power constraint. Removing this constraint provides the solution when it is not binding, *i.e.*, when the user experiences a good channel (low normalized noise variance $\sigma^2 = L\sigma_n^2$) and its target rate is not too high. The general case will be addressed in next subsection.

Lemma 2. Under (As1)-(As2) and for $P_{\text{max}} \to +\infty$, the minimum of the problem (7) is achieved by uniformly allocating power among N_{a} time slots

$$p_n = \begin{cases} \left(2^{R\frac{N}{N_a}} - 1\right)\sigma^2 & \text{if } n = 0, ..., N_a - 1\\ 0 & \text{otherwise} \end{cases}$$
 (10)

where $N_{\rm a}$ is the argument that minimizes

$$\min_{0 \le N_{\rm a} \le N} \frac{N_{\rm a}}{N} \left(P_0 + \gamma \sigma^{2\alpha} \left(2^{R \frac{N}{N_{\rm a}}} - 1 \right)^{\alpha} \right) + \frac{E_{\rm sleep}((N - N_{\rm a})T)}{NT}. \tag{11}$$

Under (As3) - (As4), the solution is

$$N_{\rm a} = \lfloor \min(NR/R_{\rm a}, N) \rceil.$$

Proof. See Appendix VII-A.

Remark 6 (Relation with Lemma 1). Lemma 2 extends the result of Lemma 1 by considering a non-zero sleep power P_{sleep} and load-independent active power consumption P_0 . Under a constant sleep power model (As4), the problem has a similar solution.

Remark 7 (Scaling of $R_{\rm a}$). As shown in Fig. 2b, $R_{\rm a}$ increases with $\delta=(P_0-P_{\rm sleep})/(\gamma\sigma^{2\alpha})$. This implies that less time slots should be activated if normalized noise power σ^2 , $P_{\rm sleep}$ and γ are small and if the active load-independent power consumption P_0 is high. A major difference with Lemma 1 is that $R_{\rm a}$ does not go to zero as α approaches 1 when $\delta>0$. This implies that, given nonzero static power consumption, not all time slots should always be activated. The particular load-independent case $\gamma=0$ implies that $R_{\rm a}=+\infty$ and $N_{\rm a}=1$. This makes sense as $P_{\rm cons}$ then only depends on $N_{\rm a}$. As shown in next subsection, considering a finite maximal transmit power per time slot will change this result.

Algorithm 1 Iterative resource allocation

```
Require: \sigma^2, R, N, P_{\text{max}}, R_{\text{max}}, R_{\text{a}}
   N_{\rm a} \leftarrow (11)
                                                                                                                              ▶ Init. by sol. of Theor. 2
   p_0, ..., p_{N-1} \leftarrow (10)
   N_{\text{max}} \leftarrow 0
   \hat{N} \leftarrow N
   while p_{N_a-1} > P_{\max} do
                                                                                                                       N_{\text{max}} \leftarrow N_{\text{max}} + 1
                                                                                                          ▶ Set one more time slot to max power
        R \leftarrow (NR - R_{\max})/(N-1)
                                                                                                                                  N \leftarrow N-1
        N_{\rm a} \leftarrow (11)
                                                                                                                       ▶ Update with sol. of Theor. 2
        p_0, ..., p_{N-1} \leftarrow (10)
   end while
   p_{N_{\mathbf{a}}}, ..., p_{N_{\mathbf{a}} + N_{\max} - 1} \leftarrow P_{\max}
   p_{N_{\mathbf{a}}+N_{\max}},...,p_{\hat{N}-1} \leftarrow 0
```

C. Optimal Allocation with Maximal Power Constraint

We now consider the additional constraint of a finite maximal per-time slot power P_{max} . This constraint may render the problem unfeasible. Therefore, we introduce the following assumption.

(As5): Problem (7) is feasible, i.e.,

$$R \le R_{\text{max}} = \log_2 \left(1 + \frac{P_{\text{max}}}{\sigma^2} \right).$$

The exact solution of problem (7) generally requires an iterative solution.

Proposition 2. Under (As1) - (As2), (As5), the solution of the problem (7) can be obtained by using Algorithm 1.

To avoid the need of an iterative solution and the ceil-floor operator, we use the fact that the problem greatly simplifies by considering the asymptotic case of a large N. Then, the ratio $N_{\rm a}/N$ can be considered asymptotically continuous instead of only taking discrete values.

Remark 8 (Large N assumption). The large N assumption is realistic in practice as frames are typically made of many symbols. Moreover, the assumption of having a sufficiently large N is central and implicit in this work. This article investigates the gain of activating only a share of transmission time slots. To be able to do this, some flexibility in the number of activated resources should be available, implying a sufficiently large N.

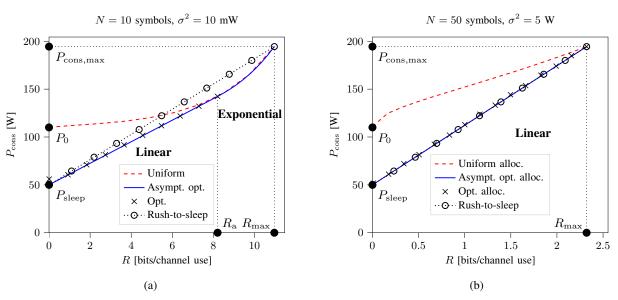


Fig. 3. Power consumption versus load using optimal/uniform power allocation with constant sleep mode (As4).

To provide the closed-form asymptotic solution, we define

$$\tilde{R} = \min(R_{\rm a}, R_{\rm max}), P_{\rm a} = (2^{R_{\rm a}} - 1)\sigma^2, \tilde{P} = \min(P_{\rm a}, P_{\rm max}).$$

Theorem 1. Under (As1) - (As5), the solution of problem (7) and two scaling regimes of P_{cons} as a function of R can be found:

 \square (Linear) If $R \leq \tilde{R}$, as $N \to +\infty$, the allocation

$$N_{\rm a} = \left[NR/\tilde{R}\right], \ p_n = \begin{cases} \tilde{P} & {
m if} \ n = 0,...,N_{
m a} - 1 \\ 0 & {
m otherwise} \end{cases}$$

is asymptotically optimal and achieves a consumed power

$$P_{\text{cons}} = P_{\text{sleep}} + R \frac{P_0 - P_{\text{sleep}} + \gamma \tilde{P}^{\alpha}}{\tilde{R}} + \epsilon$$

where ϵ is the gap from the optimum which asymptotically vanishes: $|\epsilon| = O(1/N)$.

 \square (Exponential) If $R > \tilde{R}$:

$$N_{\rm a}=N,\ p_n=\left(2^R-1\right)\sigma^2\quad {\rm for}\ n=0,...,N-1,$$

$$P_{\rm cons}=P_0+\gamma\sigma^{2\alpha}\left(2^R-1\right)^\alpha.$$

Proof. See Appendix VII-C.

Remark 9. For large N, the solution has a simple form: the ceil-floor operator is replaced by a rounding operator. The power and number of bits per-activated time slot $(\tilde{P} \text{ and } \tilde{R})$ are constant

in the linear regime. As shown in Fig. 3a and 3b, the approximation error can barely be seen and is already negligible for small values of N such as 10 or 50.

Remark 10 (Scaling regimes). Lemma 1 puts forward two scaling regimes: linear and exponential. They can easily be identified in Fig. 3a. In Fig. 3b, a higher noise regime is considered so that only the linear regime is present ($\tilde{R} = R_{\rm max}$).

Remark 11 (Gain with respect to uniform allocation). Fig. 3a and 3b also plot the gain with respect to a uniform allocation, *i.e.*, $p_n = (2^R - 1) \sigma^2$ for n = 0, ..., N - 1. As expected, the gain is larger at low load (rate) where the optimal allocation only activates few resources.

Remark 12 (Rush-to-sleep). If $R_{\rm max} \leq R_{\rm a}$ (high noise regime, Fig. 3b), a rush-to-sleep approach is optimal: active time slots transmitting at full power $P_{\rm max}$ and rate $R_{\rm max}$. This minimizes $N_{\rm a}$ and maximize sleep duration. If $R_{\rm a} < R_{\rm max}$ (low noise regime, Fig. 3a), reduced transmit power should be used instead and not even using sleep for $R > R_{\rm a}$. It is then better to fully activate the system with a uniform allocation.

Remark 13 (Converse problem). We here consider the minimization of $P_{\rm cons}$ for a fixed R. The maximization of R for a fixed $P_{\rm cons}$ can also be considered. It could occur if power is available and should directly be used, e.g., a solar panel or wind turbine without battery and/or not connected to the grid. The solution can be found using same methodology or directly by "reverting" the result of Theor. 1. Indeed, the allocation that minimizes $P_{\rm cons}$ for a fixed rate R is also the allocation that maximizes R for the minimal value of $P_{\rm cons}$ of the inital problem. Scaling regimes of R as a function of $P_{\rm cons}$ will be linear and logarithmic. It is here omitted due to space constraints.

D. Trade-Off: Spectral Efficiency versus Energy Efficiency

Considering that the transmission occupies a bandwidth $B = (1 + \alpha_{\rm rol})/T$, where $\alpha_{\rm rol} \in [0, 1]$ is the roll-off factor, so that the SE and the EE are

$$SE = \frac{R}{TB} = \frac{R}{1 + \alpha_{\text{rol}}} \quad \text{[bits/s/Hz]}, \ EE = \frac{BSE}{P_{\text{cons}}} = \frac{R}{TP_{\text{cons}}} \quad \text{[bits/Joule]}. \tag{12}$$

Given these relationships, the trade-off SE-EE can be easily identified from previous results.

Corollary 1. Under (As1) - (As5), the optimal EE for a given SE is:

 \square If $SE \leq \tilde{R}/(1+\alpha_{rol})$ (implying $R \leq \tilde{R}$), as $N \to +\infty$,

$$EE = \frac{BSE}{P_{\text{sleep}} + SE \frac{1 + \alpha_{\text{rol}}}{\tilde{R}} \left(P_0 - P_{\text{sleep}} + \gamma \tilde{P}^{\alpha} \right)} + O(1/N).$$

 \Box If ${\rm SE}>\tilde{R}/(1+\alpha_{\rm rol})$ (implying $R>\tilde{R}$):

$$EE = \frac{BSE}{P_0 + \gamma \sigma^{2\alpha} \left(2^{SE(1+\alpha_{rol})} - 1\right)^{\alpha}}.$$

Proof. From the definition of the EE in (12), it is clear that its maximization is equivalent to the minimization of P_{cons} . As a result, the optimal allocations and the results of Theor. 1 can directly be used, which lead to the above results. The two cases correspond to the linear and exponential scaling regimes of Theor. 1, respectively.

Corollary 2. Under (As1) - (As5), the SE that maximizes the optimal EE given in Corollary 1 is:

 \square If $R_{\rm a} \geq R_{\rm max}$:

$$\bar{SE} = SE_{max} = \frac{R_{max}}{1 + \alpha_{rol}}, \ EE_{max} = \frac{BSE_{max}}{P_0 + \gamma P_{max}^{\alpha}}.$$

 \square If $R_{\rm a} < R_{\rm max}$:

$$\bar{SE} = \frac{\bar{R}}{1 + \alpha_{rol}}, \ EE_{max} = \frac{B\bar{SE}}{P_0 + \gamma \sigma^{2\alpha} (2^{\bar{R}} - 1)^{\alpha}}$$

where \bar{R} is the constant that minimizes the convex problem

$$\min_{\bar{R} \in [R_a, R_{\text{max}}]} \frac{P_0 + \gamma \sigma^{2\alpha} \left(2^{\bar{R}} - 1\right)^{\alpha}}{\bar{R}}.$$

Proof. See Appendix VII-D.

Remark 14 (Scaling of EE-SE). Fig. 4a and 4b plot the EE as a function of the SE for the optimal and uniform allocation. Most gain in terms of EE (not absolute energy) is obtained at medium SE. The plots are obtained by varying the rate R, all other parameters being fixed.

Remark 15 (Optimal EE). As shown in Fig. 4a, for a low normalized noise, the maximal EE is not obtained at maximal SE while it is for a higher noise, as shown in Fig. 4b.

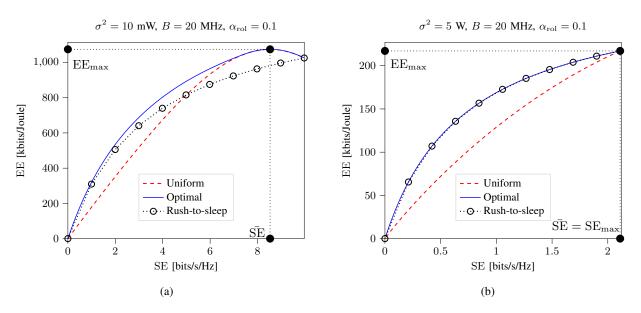


Fig. 4. Energy efficiency versus spectral efficiency for optimal/uniform allocation with constant sleep mode (As4).

IV. OPTIMAL ALLOCATION FOR PIECEWISE CONSTANT SUCCESSIVE SLEEP MODES

Fig. 3a and 3b have shown promising gains to reduce consumed power at low load. However, they are still limited by the relatively high sleep power consumption P_{sleep} . Similarly, Fig. 4a and 4b can seem limited as one would hope to obtain an EE that is approximately flat as a function of the SE. The reason is the same: a too high P_{sleep} .

To drastically reduce energy consumption at low-to-medium load, it is of paramount importance to implement successive sleep modes and use a frame duration long enough so that the system can enter these sleep modes. To find the optimal allocation, the iterative algorithm of Prop. 2 can be used, which requires to solve problem (11) at each iteration. This is an integer programming problem which can have a significant complexity. If the problem is relaxed by considering N_a continuous, it remains challenging to solve as it implies the minimization of the concave function $E_{\text{sleep}}(t)$ (Prop. 1).

If the sleep power consumption is assumed piecewise constant, according to (As3), a simple allocation can still be found. This sleep power model was detailed in Section II-B and Fig. 1. For a given $N_{\rm a}$, only sleep modes such that $T_s \leq (N-N_{\rm a})T$ can be entered. We define $N_{\rm a,s}^+ = N - \lfloor T_s/T \rfloor$ so that sleep mode s can be used if $N_{\rm a} \leq N_{\rm a,s}^+$. Depending on the target rate R, it might be unfeasible to use a given sleep mode because of the maximal power constraint per time slot. A minimum of active time slots $\frac{NR}{R_{\rm max}}$ is required to satisfy it. Otherwise, the power per time slot would have to be higher than $P_{\rm max}$ to satisfy the rate constraint. Mode s is thus

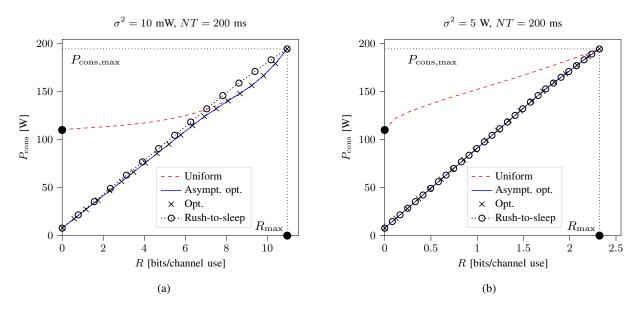


Fig. 5. Power consumption versus load with successive sleep power model (As3).

feasible only if $\frac{NR}{R_{\max}} \leq N_{\mathrm{a},s}^+$. As a result, deepest sleep modes are only possible for low values of R, which intuitively makes sense. For a target rate R, we define the set of feasible sleep modes as $\mathcal{S}_R = \left\{s \mid s \in \{0,\cdots,S-1\} \text{ and } \frac{NR}{R_{\max}} \leq N_{\mathrm{a},s}^+\right\}$. Moreover, we generalize the definition of R_{a} per sleep mode as the constant $R_{\mathrm{a}}(s)$ that minimizes the convex problem

$$R_{a}(s) = \arg\min_{x \ge 0} \frac{P_{0} - P_{s} + \gamma \sigma^{2\alpha} (2^{x} - 1)^{\alpha}}{x}.$$

We also define $\tilde{R}_s = \min(R_a(s), R_{\max})$.

Theorem 2. Under (As1) - (As3), (As5), as $N \to +\infty$, the solution of problem (7) is found by computing for all feasible sleep modes $s \in S_R$

$$\begin{split} N_{\mathrm{a},s} &= \left[\min \left(NR/\tilde{R}_{s}, N_{\mathrm{a},s}^{+} \right) \right], \ p_{n,s} = \begin{cases} (2^{\frac{RN}{N_{\mathrm{a},s}}} - 1)\sigma^{2} & \text{if } n = 0, ..., N_{\mathrm{a},s} - 1 \\ 0 & \text{otherwise} \end{cases} \\ P_{\mathrm{cons},s} &= \frac{N_{\mathrm{a},s}}{N} \left(P_{0} + \gamma \sigma^{2\alpha} \left(2^{\frac{RN}{N_{\mathrm{a},s}}} - 1 \right)^{\alpha} \right) + \frac{E_{\mathrm{sleep}}((N - N_{\mathrm{a},s})T)}{NT} + O(1/N) \end{split}$$

and choosing among these modes the one that has the minimal $P_{cons,s}$.

Remark 16. Fig. 5a and Fig. 5b are plotted based on the same simulation parameters as Fig. 3a and 3b, but with a different sleep model. Using successive power modes has a drastic impact on the consumed energy at low load. A key parameter to allow drastic savings is to have a large

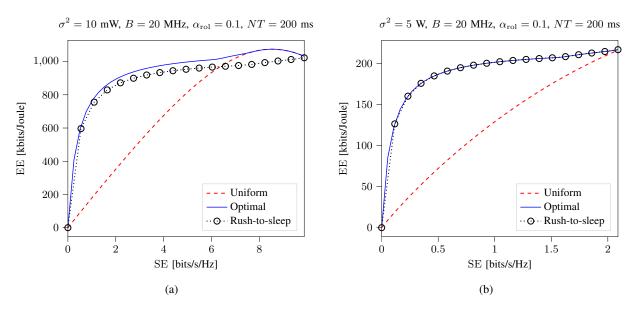


Fig. 6. Energy efficiency versus spectral efficiency with successive sleep power model $(\mathbf{As3})$.

enough frame duration so that deepest sleep modes can be used. For these figures, it was fixed to 200 ms. Hence, deep sleep power mode can be entered but not hibernating sleep power mode.

Remark 17. Similarly, Fig. 6a and 6b can be compared to Fig. 4a and 4b, where only the sleep model differs. As ideally expected, the EE quickly reaches a plateau. To still improve this behaviour, a longer frame duration can be used.

V. OPTIMAL ALLOCATION FOR TDMA SYSTEM

We now extend previous results by considering a downlink transmission from the BS to K users. We consider the constant power sleep model so that $(\mathbf{As1}) - (\mathbf{As4})$ hold. The users are multiplexed using TDMA. In the frame of N symbols, each symbol is allocated to the transmission towards at most one user so that no inter-user interference is present. Out of the N symbols, N_k symbols are allocated to user k, for k = 0, ..., K - 1. The power associated to the n-th symbol transmitted to user k is denoted by $p_{k,n}$, where $n = 0, ..., N_k - 1$ and k = 0, ..., K - 1. The normalized noise variance $\sigma_k^2 = \sigma_n^2 L_k$ is considered different at each user, as each can have a specific path loss L_k . The consumed power is then

$$P_{\text{cons}}^{\text{TDMA}} = \frac{N_{\text{a}}}{N} P_0 + \frac{\gamma}{N} \sum_{k=0}^{K-1} \sum_{n=0}^{N_{k-1}} p_{k,n}^{\alpha} + \frac{N - N_{\text{a}}}{N} P_{\text{sleep}}$$

where $N_{\rm a} = \sum_{k=0}^{K-1} N_k$ so that $0 \le N_{\rm a} \le N$. We consider the generalization of problem (7) of minimizing the power consumption, under $(\mathbf{As1}) - (\mathbf{As4})$ and per-user rate constraints R_k . The problem can be formulated as

$$\min_{\substack{N_k, p_{k,n} \\ n=0,\dots,N_k-1 \\ k=0,\dots,K-1}} P_{\text{cons}}^{\text{TDMA}} \text{ s.t. } \frac{1}{N} \sum_{n=0}^{N_k-1} \log_2 \left(1 + \frac{p_{k,n}}{\sigma_k^2} \right) = R_k \ \forall k, \ \sum_{k=0}^{K-1} N_k \le N. \tag{13}$$

We assume in the following that the problem has a feasible solution, which generalizes (As5). (As6): Problem (13) is feasible. Defining the maximal per-user rate $R_{k,\text{max}} = \log_2 \left(1 + \frac{P_{\text{max}}}{\sigma_k^2}\right)$, it implies that

$$\sum_{k=0}^{K-1} \lceil \frac{NR_k}{R_{k,\max}} \rceil \le N.$$

Moreover, we define the constant $R_{k,a}$ that minimizes the convex problem

$$R_{k,a} = \arg\min_{x>0} \frac{P_0 - P_{\text{sleep}} + \gamma \sigma_k^{2\alpha} (2^x - 1)^{\alpha}}{x}.$$

We also define $\hat{R}_k = \min(R_{k,\mathrm{a}}, R_{k,\mathrm{max}})$ and $\hat{P}_k = \min(P_{k,\mathrm{a}}, P_{\mathrm{max}})$. Two regimes can be considered for Problem (13), depending if the constraint $\sum_{k=0}^{K-1} N_k \leq N$ is binding or not.

In terms of power savings, the most promising case is the low-to-medium load regime where the constraint is not binding. In that case, the problem fully decouples per-user and the linear regime solution of Theor. 1 can directly be used on a per-user basis.

Theorem 3 (TDMA - linear regime). Under $(\mathbf{As1}) - (\mathbf{As4})$, $(\mathbf{As6})$, as $N \to +\infty$, if $\sum_{k=0}^{K-1} \frac{R_k}{R_k} \le 1 + O(1/N)$, the allocation

$$N_k = \left[NR_k / \hat{R}_k \right], \ p_{k,n} = \hat{P}_k + O(1/N) \text{ for } n = 0, ..., N_k - 1$$

for user k = 0, ..., K - 1 is an asymptotic solution of Problem (13) and

$$P_{\text{cons}} = P_{\text{sleep}} + \sum_{k=0}^{K-1} R_k \frac{P_0 - P_{\text{sleep}} + \gamma \hat{P}_k^{\alpha}}{\hat{R}_k} + O(1/N).$$

Proof. See Appendix VII-F.

Remark 18 (K=2 - TDMA). For K=2, Fig. 7a and 7b plot $P_{\rm cons}$ as a function of R_1 and R_2 . The optimal solution of Theor. 3 is plotted where it is valid, *i.e.*, the asymptotic linear regime $\frac{R_0}{\tilde{R}_0} + \frac{R_1}{\tilde{R}_1} \le 1$. As already observed in the single-user case (Fig. 3b), this regime can cover the whole feasible rate region, as shown in Fig. 7b, characterized by a relatively higher noise power

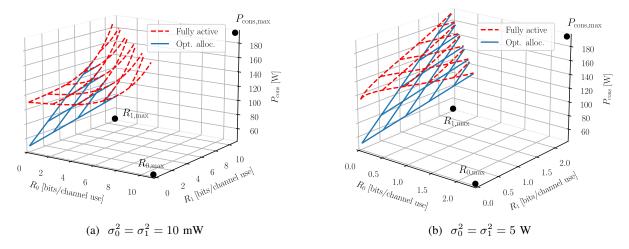


Fig. 7. Power consumption versus load in a K=2 user TDMA system using optimal/uniform power allocation with constant sleep mode (As4). The optimal allocation is valid in the asymptotic linear regime, *i.e.*, when $\frac{R_0}{\hat{R}_0} + \frac{R_1}{\hat{R}_1} \leq 1$.

implying that $\hat{R}_k = R_{k,\text{max}}$, $\forall k$. As a benchmark, a fully active uniform allocation was plotted where a share $N_k/N = R_k/(R_0 + R_1)$ of time slots was allocated to each user.

Remark 19 (rush-to-sleep - TDMA). If $\sum_{k=0}^{K-1} \frac{R_k}{\hat{R}_k} \le 1$, a rush-to-sleep allocation is asymptotically optimal for each user such that $R_{k,\max} \le R_{k,a}$. On the other hand, if $\sum_{k=0}^{K-1} \frac{R_k}{\hat{R}_k} > 1$, it is optimal to use a fully active system and no sleep.

If the constraint $\sum_{k=0}^{K-1} N_k = N$ is binding, Problem (13) is coupled between users and challenging to solve. As the sleep duration is zero, the power consumption becomes

$$P_{\text{cons}}^{\text{TDMA}} = P_0 + \frac{\gamma}{N} \sum_{k=0}^{K-1} \sum_{n=0}^{N_{k-1}} p_{k,n}^{\alpha}.$$

As this regime does not allow switching-off components, relatively small energy reduction potentials are expected. Given space constraints, we do not provide a more detailed solution. One possibility is to reduce the target rates of the users to make the constraint non-binding and then use Theor. 3. Another possibility is to use conventional scheduling policies, that are not aware of sleep capabilities, which makes sense as the system is fully active.

VI. CONCLUSION

In this work, we have proposed a fundamental study of time-domain energy-saving techniques in radio access. The results provide key novel insights from an information-theoretic perspective.

Considering equal gain parallel communication channels, conventional information-theoretic results state that all channels (time slots in this study) should be equally used to minimize transmit (not consumed) power under a rate constraint. On the contrary, popular energy-saving techniques steer towards an extreme opposite "rush-to-sleep" approach: compact transmission in as few time slots as possible, at maximal transmit power, to maximize sleep duration.

Using a realistic power consumption model, our information-theoretic study bridges the gap between these two extremes. Simple allocations are provided that allow drastic energy savings reaching factors of 10 at low load. At low-to-medium load, the optimal number of active time slots is linearly proportional to the rate, resulting in a power consumption which linearly scales with the rate. At a higher load, all time slots become allocated. The rush-to-sleep approach is shown to be optimal in a high-noise regime but not otherwise. In a low-noise regime, it might be better to use a fully active system. Moreover, the fundamental trade-off between EE and SE is revisited leveraging the time-domain hardware sleep capabilities. Transmitting at maximal SE maximizes the SE in a high-noise regime while, in a low-noise regime, a reduced SE maximizes the EE. Considering a sleep model with increasing depth complicates the study but also greatly increases the energy-saving gains. For a piecewise constant model, simple allocations can still be found. Finally, for a multi-user TDMA system, single-user results are applicable on a per-user basis in the low-to-medium load regime, where the system should not be fully active and where sleep-aware energy-saving gains can be achieved.

VII. APPENDIX

We start by introducing two lemmas that will be useful in the following.

Lemma 3. An optimal solution of the following problem

$$\min_{p_0, \dots, p_{N-1}} \frac{\gamma}{N} \sum_{n=0}^{N-1} p_n^{\alpha} \quad \text{s.t. } \frac{1}{N} \sum_{n=0}^{N-1} \log_2 \left(1 + \frac{p_n}{\sigma^2} \right) = R$$
(14)

must have a uniform allocation among active time slots: $\forall n, n'$, if $p_n > 0$, $p_{n'} > 0$ then $p_n = p_{n'}$.

Proof. The case $\alpha=1$ was already treated in the introduction. For $\alpha=0$, the cost function only depends on the number of active time slots so that a single time slot must be allocated power. In the following, we will use a proof by contradiction for the case $0<\alpha<1$. Any non-uniform power allocation has at least two time slots, say n=0 and n=1 (potentially using a re-indexing), that are such that $p_0>0$, $p_1>0$ and $p_0\neq p_1$. We consider the power allocated

to the other time slots as fixed and optimize the cost function with respect to p_0 and p_1 only. For the sake of clarity, we define $\rho_n = p_n/\sigma^2$. The reduced problem is

$$\min_{\rho_0, \rho_1} \tilde{f}(\rho_0, \rho_1) = \sum_{n=0}^{1} \rho_n^{\alpha} \text{ s.t. } \sum_{n=0}^{1} \log(1 + \rho_n) = Z$$

where $Z = NR \log 2 - \sum_{n=2}^{N-1} \log (1 + \rho_n)$. The constraint implies that

$$\rho_1 = \frac{e^Z}{1 + \rho_0} - 1 \tag{15}$$

so that ρ_0 and ρ_1 take values only in the domain $[0, e^Z - 1]$ and have a one-to-one relationship. As $\rho_0 \to 0$, $\rho_1 \to e^Z - 1$ and vice versa. Moreover, the problem is symmetrical so that $\tilde{f}(\rho_0, \rho_1) = \tilde{f}(\rho_1, \rho_0)$. Using (15), the problem can be rewritten as a monovariable unconstrained problem

$$\min_{\rho_0} \tilde{f}(\rho_0) = \rho_0^{\alpha} + \rho_1^{\alpha} = \rho_0^{\alpha} + \left(\frac{e^Z}{1 + \rho_0} - 1\right)^{\alpha}.$$

The derivative of $\tilde{f}(\rho_0)$ and its limit at the bounds of its domain are given by

$$\tilde{f}'(\rho_0) = \frac{\alpha}{\rho_0^{1-\alpha}} - \frac{e^Z}{(1+\rho_0)^2} \frac{\alpha}{\rho_1^{1-\alpha}}, \lim_{\rho_0 \to 0} \tilde{f}'(\rho_0) = +\infty, \lim_{\rho_0 \to e^Z - 1} \tilde{f}'(\rho_0) = -\infty.$$
 (16)

To find the critical points, we set $f'(\rho_0) = 0$ and combining with (15), we find the condition

$$0 = \rho_1^{1-\alpha}(1+\rho_0) - \rho_0^{1-\alpha}(1+\rho_1)$$
(17)

which shows that there is always one critical point in $\rho_0 = \rho_1 = e^{Z/2} - 1$. Moreover, the fact that the problem is symmetric implies an odd number of critical points: if there is a critical point in $\tilde{\rho}_0$, there is one in $\frac{e^Z}{1+\tilde{\rho}_0}-1$. Using again (15), we can rewrite the condition (17) as

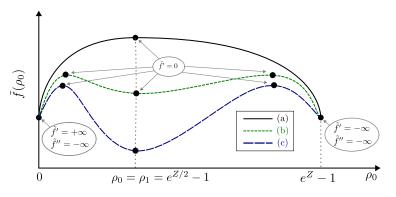
$$0 = (e^{Z} - 1 - \rho_0)^{1-\alpha} (1 + \rho_0)^{1+\alpha} - \rho_0^{1-\alpha} e^{Z}.$$

The point $\rho_0 = 0$ is not a critical point. Hence, we can restrict to $\rho_0 > 0$ and divide by $\rho_0^{1-\alpha}$

$$0 = \underbrace{(e^Z - 1 - \rho_0)^{1-\alpha} (1 + \rho_0)^{1+\alpha} \rho_0^{\alpha - 1} - e^Z}_{\tilde{g}(\rho_0)}.$$

The function $\tilde{g}(\rho_0)$ is infinite in $\rho_0=0$ and $-e^Z$ for $\rho_0=e^Z-1$. The number of roots of $\tilde{g}(\rho_0)$ and thus critical points of $\tilde{f}(\rho_0)$ is at most equal to one plus the number of critical points/alternations of $\tilde{g}(\rho_0)$. Setting $\tilde{g}'(\rho_0)=0$ gives the condition

$$0 = -(1 - \alpha)(1 + \rho_0)\rho_0 + (e^Z - 1 - \rho_0)(1 + \alpha)\rho_0 + (e^Z - 1 - \rho_0)(1 + \rho_0)(\alpha - 1)$$



Domain	$[0, e^Z - 1]$
Intercept	$\tilde{f}(0) = (e^Z - 1)^{\alpha}$
Symmetry	$\tilde{f}(\rho_0) = \tilde{f}\left(\frac{e^z}{1+\rho_0}-1\right)$
Derivatives	$\tilde{f}'(0) = +\infty, \ \tilde{f}'(e^Z - 1) = -\infty$
Concavity	$\tilde{f}''(0) = \tilde{f}''(e^Z - 1) = -\infty$
Critical points	Always 1 in $\rho_0 = \rho_1 = e^{Z/2} - 1$
	Potentially 2 others symmetrical

Fig. 8. Sketch of function $\tilde{f}(\rho_0)$.

which is a quadratic equation in ρ_0 . It has thus at most two solutions. As a result, $\tilde{g}(\rho_0)$ has max two alternations and $\tilde{f}(\rho_0)$ has at most 3 critical points. The second order derivative of $\tilde{f}(\rho_0)$ and its limit at the bounds of its domain are given by

$$\tilde{f}''(\rho_0) = -\alpha (1 - \alpha) \frac{1}{\rho_0^{2-\alpha}} + \frac{1}{\rho_1^{2-\alpha}} \frac{\alpha e^Z}{(1 + \rho_0)^3} \left(-(1 - \alpha) e^Z \frac{1}{1 + \rho_0} + 2\rho_1 \right)$$

$$\lim_{\rho_0 \to 0} \tilde{f}''(\rho_0) = -\infty, \quad \lim_{\rho_0 \to e^Z - 1} \tilde{f}''(\rho_0) = -\infty. \tag{18}$$

As shown in Fig. 8, three cases can be distinguished. In case (a), there is a single critical point in $\rho_0 = \rho_1 = e^{Z/2} - 1$ which is a maximum. In cases (b) and (c), there are three critical points: the middle one in $\rho_0 = \rho_1 = e^{Z/2} - 1$ will now be a minimum (local in (b), global in (c)) while the two on its sides are maxima. Hence, global minima can only be obtained for either $\rho_0 = 0$, $\rho_1 = e^Z - 1$ or $\rho_1 = 0$, $\rho_0 = e^Z - 1$ or $\rho_0 = \rho_1 = e^{Z/2} - 1$. Hence, it is impossible to find an optimal allocation such that $p_0 > 0$, $p_1 > 0$ and $p_0 \neq p_1$.

Lemma 4. Under (As1) - (As4), the following function is convex for x > 0

$$f(x) = \frac{P_0 - P_{\text{sleep}} + \gamma \sigma^{2\alpha} (2^x - 1)^{\alpha}}{x}.$$

Proof. Under (As4), $P_0 - P_{\text{sleep}} \ge 0$ and thus $(P_0 - P_{\text{sleep}})/x$ is convex. Given that the sum of two convex functions is convex, it is sufficient to show that

$$\frac{\gamma\sigma^{2\alpha}\left(2^x-1\right)^{\alpha}}{x} \text{ or equivalently } g(y)=\frac{\left(e^y-1\right)^{\alpha}}{y}$$

is convex for y > 0 (using $y = x \log 2$). Its second derivative is

$$g''(y) = \frac{(e^y - 1)^{\alpha - 2}(e^{2y}(\alpha^2 y^2 - 2\alpha y + 2) - e^y(\alpha y^2 - 2\alpha y + 4) + 2)}{y^3}.$$

Given that y>0, we have directly that $y^3>0$, $(e^y-1)^{\alpha-2}\geq 0$ and it is sufficient to show that

$$h(y) = e^{2y}(\alpha^2 y^2 - 2\alpha y + 2) - e^y(\alpha y^2 - 2\alpha y + 4) + 2 \ge 0.$$

Using the Taylor series expansion $e^y = \sum_{r=0}^{+\infty} \frac{y^r}{r!}$, which converges for all y, we find

$$h(y) = \sum_{r=0}^{+\infty} \frac{(2x)^r}{r!} (\alpha^2 y^2 - 2\alpha y + 2) - \sum_{r=0}^{+\infty} \frac{y^r}{r!} (\alpha y^2 - 2\alpha y + 4) + 2$$
$$= \sum_{r=2}^{+\infty} y^r \left(\alpha \frac{\alpha 2^{r-2} - 1}{(r-2)!} + \alpha \frac{2 - 2^r}{(r-1)!} + \frac{2^{r+1} - 4}{r!} \right).$$

To show that $h(y) \ge 0$ for y > 0 and $\alpha \in [0, 1]$, it is sufficient to show that for all $r \ge 2$

$$\alpha \frac{2^{r-2}\alpha - 1}{(r-2)!} + \alpha \frac{2 - 2^r}{(r-1)!} + \frac{2^{r+1} - 4}{r!} \ge 0$$

$$\leftrightarrow \alpha^2 2^{r-2} r(r-1) + \alpha r(3 - 2^r - r) + 2^{r+1} - 4 \ge 0.$$

For r=2, this is verified as $2\alpha^2-6\alpha+4=2(\alpha-1)(\alpha-2)$ is always positive for $\alpha\in[0,1]$. For r=3, this is also verified as $\alpha^212-\alpha24+12=12(\alpha-1)^2$ is again positive for $\alpha\in[0,1]$. For $r\geq 4$, we have $r(r-1)\geq r^2/2$, $3-2^r-r\geq -2^{r+1}$ and $-2^{-r+4}\geq -1$ so that

$$\alpha^{2}2^{r-2}r(r-1) + \alpha r(3-2^{r}-r) + 2^{r+1} - 4 \ge 2^{r-2}(\alpha^{2}r^{2}/2 - \alpha r8 + 8 - 2^{-r+4})$$
$$\ge 2^{r-2}(\alpha^{2}r^{2}/2 - \alpha r8 + 7)$$

which roots are in $8r \pm \sqrt{50}r$. Given that $8r - \sqrt{50}r \ge 0.92r > 1$ for $r \ge 2$, both roots are strictly larger than 1 and the term is positive for $\alpha \in [0,1]$, which concludes the proof.

A. Proof of Lemmas 1 and 2

One can first note that Lemma 1 is a particularization of Lemma 2 when $P_0 = E_{\text{sleep}}(t) = 0$. Hence, the result will be found as a specific case in the following. Under $(\mathbf{As1})$ - $(\mathbf{As2})$, Problem (7) can be rewritten as

$$\min_{N_{\rm a}} \frac{N_{\rm a}}{N} P_0 + \left[\min_{p_0, \dots, p_{N_{\rm a}-1}} \frac{\gamma}{N} \sum_{n=0}^{N_{\rm a}-1} p_n^{\alpha} \right] + \frac{E_{\rm sleep}((N-N_{\rm a})T)}{NT} \tag{19}$$

which shows that only the second term depends on the power allocation, *i.e.*, the term defined as $P_{\rm ld}$. From Lemma 3, an optimal allocation needs to be uniform in the number of activated time slots. For a given $N_{\rm a}$, the rate constraint fixes the transmit power per active time slot

$$p_n = \left(2^{R\frac{N}{N_a}} - 1\right)\sigma^2 \text{ if } n = 0, ..., N_a - 1.$$

Hence, the problem can be reformulated as finding the optimal number of active slots $N_{\rm a}$ that minimizes the consumed power, *i.e.*, Problem (11) in Lemma 2. Moreover, under (As3)-(As4), the problem becomes

$$\min_{N_{\rm a}} P_{\rm sleep} + \frac{N_{\rm a}}{N} \left(P_0 - P_{\rm sleep} + \gamma \sigma^{2\alpha} \left(2^{\frac{NR}{N_{\rm a}}} - 1 \right)^{\alpha} \right).$$

We define $x = \frac{RN}{N_a}$ and relax the problem by considering x as continuous

$$\min_{x} f(x) = \frac{P_0 - P_{\text{sleep}} + \gamma \sigma^{2\alpha} (2^x - 1)^{\alpha}}{x}.$$

From Lemma 4, we know that f(x) is convex. Given the definition of x and the integer nature of $N_{\rm a}$, x can only take discrete values in practice. Given the fact that f(x) is convex, it is guaranteed that one of the neighboring possible values of $R_{\rm a} = \arg\min f(x)$ is optimal. As a result, the solution is given by either $N_{\rm a} = \lceil \frac{RN}{R_{\rm a}} \rceil$, $N_{\rm a} = \lfloor \frac{RN}{R_{\rm a}} \rfloor$ or N if $\frac{RN}{R_{\rm a}} > N$. Using the ceil-floor notation concludes the proof of Lemma 2. The above result can also be particularized to the problem of Lemma 1 by setting $P_0 = E_{\rm sleep}(t) = 0$ and the problem simplifies to

$$\min_{x} f(x) = \frac{(2^{x} - 1)^{\alpha}}{x} \leftrightarrow \min_{y} g(y) = \frac{(e^{y} - 1)^{\alpha}}{y}.$$

where $y = x \log 2$. Setting its derivative to zero, we find

$$\alpha e^{y} y = e^{y} - 1 \leftrightarrow y = W\left(-\alpha^{-1} e^{-\alpha^{-1}}\right) + \alpha^{-1} \leftrightarrow R_{a} = W\left(-\alpha^{-1} e^{-\alpha^{-1}}\right) + \alpha^{-1}/\log 2,$$

which concludes the proof of Lemma 1.

B. Proof of Proposition 2

The algorithm is initialized by the solution of the relaxed problem assuming that no max power constraints are binding. The solution is then given by Theor. 2. If the solution is such that $p_n \leq P_{\max}$, $\forall n$, the problem is solved. On the other hand, if, for at least one time slot $p_n > P_{\max}$, at least one of the constraints must be binding. Hence, the algorithm allocates the maximal power P_{\max} to one additional time slot. The rate constraint on the remaining time slots is then adapted. The power allocation is re-computed assuming that no max power constraint is binding on the remaining time slots not yet set to P_{\max} . Again, the solution is given by Theor. 2. Again, the algorithm checks if the max constraint is verified. If yes, the algorithm has converged. If not, it enters a novel iteration and so on until convergence.

C. Proof of Theorem 1

Let us consider one by one four different cases. On the one hand, the exponential regime mentioned in the theorem where $R > \tilde{R}$ and i) $\tilde{R} = R_{\rm max}$ or ii) $\tilde{R} = R_{\rm a}$. On the other hand, the linear regime mentioned in the theorem where $R \leq \tilde{R}$ and iii) $\tilde{R} = R_{\rm a}$ or iv) $\tilde{R} = R_{\rm max}$.

<u>Case i)</u>: This case is not applicable according to (As5) as it is unfeasible to have $R > R_{\text{max}}$.

<u>Case ii)</u>: The case together with (As5) implies $R_{\text{max}} \leq R > R_{\text{a}}$. From Lemma 2, we can find that $N_{\text{a}} = N$ and the corresponding power allocation, which is well feasible as it does not violate the P_{max} constraint. The exponential regime result of Theor. 1 is then found.

<u>Case iii)</u>: This case implies $R \leq R_{\rm a} \leq R_{\rm max}$. Let us first consider that the maximal power constraint per time slot is not active such that we can use the result of Lemma 2. If $R \leq R_{\rm a}$, the optimal number and ratio of active time slots are

$$N_{\rm a} = \lfloor RN/R_{\rm a} \rceil = \left[\frac{RN}{R_{\rm a}} \right] + \epsilon_1 = \frac{RN}{R_{\rm a}} + \epsilon_2$$
$$\frac{N_{\rm a}}{N} = \left[\frac{RN}{R_{\rm a}} \right] / N + \frac{\epsilon_1}{N} = \frac{R}{R_{\rm a}} + \frac{\epsilon_2}{N}.$$

where $|\epsilon_1| < 1$ and $|\epsilon_2| < 1$. The optimal ratio $N_{\rm a}/N$ asymptotically converges to $R/R_{\rm a}$ and the same occurs if the optimal number of time slots $N_{\rm a}$ is approximated using a rounding operator instead of the ceil-floor operator. Using this result, as $N \to +\infty$, the optimal power allocation per active time slot of Lemma 2 can be rewritten as

$$p_n = \left(2^{R_{\overline{N_a}}} - 1\right)\sigma^2 = P_a + O(1/N)$$
 (20)

where $P_{\rm a}=\left(2^{R_{\rm a}}-1\right)\sigma^2$ and $N_{\rm a}$ can be the ideal value $\lfloor RN/R_{\rm a} \rceil$ or its approximation using the rounding operator. Given that $R_{\rm a} \leq R_{\rm max}$, this allocation does not violate the $P_{\rm max}$ constraint and the result is feasible. The power consumption becomes

$$P_{\text{cons}} = P_{\text{sleep}} + \frac{N_{\text{a}}}{N} \left(P_0 - P_{\text{sleep}} + \gamma P_{\text{a}}^{\alpha} + O(1/N) \right) = P_{\text{sleep}} + R \frac{P_0 - P_{\text{sleep}} + \gamma P_{\text{a}}^{\alpha}}{R_{\text{a}}} + O(1/N).$$

Case iv): This case implies $R \leq R_{\rm max} \leq R_{\rm a}$ and thus $P_{\rm a} > P_{\rm max}$ such that allocation (20) is not feasible. As an alternative, the iterative Algorithm 1 can be used and simplified in the asymptotic regime. Indeed, as $N \to +\infty$, at each iteration, the algorithm allocates a constant power $P_{\rm a}$ (independent of R) to active time slots, not yet set to $P_{\rm max}$. At the convergence of the algorithm, the allocation will have approximately $RN/R_{\rm max}$ active time slots with maximal power $P_{\rm max}$ and rate $R_{\rm max}$. As a result, as $N \to +\infty$, at the optimum, $N_{\rm a}/N = [R/R_{\rm max}] + O(1/N) = R/R_{\rm max} + O(1/N)$ and the allocation

$$N_{\rm a} = \left[\frac{RN}{R_{\rm max}}\right], \ p_n = P_{\rm max} \ {\rm for} \ n = 0, ..., N_{\rm max}$$

is asymptotically optimal and achieves a consumed power

$$P_{\text{cons}} = P_{\text{sleep}} + R \frac{P_0 - P_{\text{sleep}} + \gamma P_{\text{a}}^{\alpha}}{R_{\text{max}}} + O(1/N).$$

Cases iii) and iv) can be written more compactly using the definitions of \tilde{R} and \tilde{P} , giving the linear regime result of Theor. 1.

D. Proof of Corollary 2

The results of Corol. 2 can be found by minimizing the EE expression in the two regimes of Corol. 1. In the regime where $SE \leq \tilde{R}/(1+\alpha_{\rm rol})$, it is clear that the EE is maximized for the largest SE, *i.e.*, when $SE = \tilde{R}/(1+\alpha_{\rm rol})$. Moreover, if $R_{\rm a} \geq R_{\rm max}$, we have $\tilde{R} = R_{\rm max}$ and the second regime of Corol. 1 is not feasible. The optimal SE corresponds to the maximal SE, $SE_{\rm max} = R_{\rm max}/(1+\alpha_{\rm rol})$. On the other hand, if $R_{\rm a} < R_{\rm max}$, $\tilde{R} = R_{\rm a}$ and the regime $SE > \tilde{R}/(1+\alpha_{\rm rol})$ can be entered. The optimization over \bar{R} then provides the optimum and can only improve the optimum as \bar{R} is allowed to take value $R_{\rm a}$.

E. Proof of Theorem 2

It is direct to see that one should choose the optimal allocation among feasible sleep modes. Under (As3), the sleep energy consumption at time t if sleep mode s is used is $E_{\text{sleep},s}(t) = E_{\text{sleep}}(T_s) + (t - T_s)P_s$ where $E_{\text{sleep}}(T_s) = \sum_{s'=0}^{s-1} P_{s'}(T_{s'+1} - T_{s'})$. The consumed power using sleep mode s can then be written as

$$P_{\text{cons},s} = \frac{\tilde{E}_s}{NT} + \frac{N_a}{N} P_0 + \frac{\gamma}{N} \sum_{n=0}^{N_{a-1}} p_n^{\alpha} + \frac{N - N_a}{N} P_s$$

where $\tilde{E}_s = E_{\rm sleep}(T_s) - T_s P_s$. This form is similar to the one given in (6), under (As4). The sole differences are the presence of P_s instead of $P_{\rm sleep}$ and the constant $\tilde{E}_s/(NT)$, which affects the cost function but does not impact the optimization. The result of Theorem 1 can then be used: uniform allocation among $N_{\rm a,s}$ active mode is optimal. The only difference is the fact that the maximal value of $N_{\rm a,s}$ is $N_{\rm a,s}^+$ instead of N, so that the sleep duration is sufficient to entermode s. As a result, we find $N_{\rm a,s} = \left[\min\left(NR/\tilde{R}_s,N_{\rm a,s}^+\right)\right]$.

F. Proof of Theorem 3

If the constraint $\sum_{k=0}^{K-1} N_k \le N$ is not binding, Problem (13) is fully decoupled between users and can be solved by solving for k = 0, ..., K - 1 an independent per-user problem

$$\min_{\substack{N_k, p_{k,n} \\ n=0, \dots, N_k-1}} \frac{N_k}{N} P_0 + \frac{\gamma}{N} \sum_{n=0}^{N_{k-1}} p_{k,n}^\alpha + \frac{N-N_k}{N} P_{\text{sleep}} \text{ s.t. } \frac{1}{N} \sum_{n=0}^{N_k-1} \log_2 \left(1 + \frac{p_{k,n}}{\sigma_k^2}\right) = R_k$$

so that the asymptotic solution of Theorem 1 can be used giving $N_k = \left[NR_k/\hat{R}_k\right]$. If the constraint $\sum_{k=0}^{K-1} N_k \leq N$ is not violated, this is the asymptotic solution of Problem (13). From

²No deeper sleep mode than s is considered even if $T_{s+1} < t$, which could decrease sleep energy consumption. Still, this does not affect the optimization result a deeper sleep mode will perform better and will be chosen instead.

Section VII-C, we know that, as $N \to +\infty$, $N_k/N = R_k/\hat{R}_k + O(1/N)$ and the constraint can thus be equivalently written as $\sum_{k=0}^{K-1} R_k/\hat{R}_k \le 1 + O(1/N)$.

ACKNOWLEDGMENT

The author would like to thank his colleagues from the DRAMCO-KU Leuven lab and Dr. Pål Frenger for many fruitful discussions and valuable comments.

REFERENCES

- [1] J. Malmodin, "The power consumption of mobile and fixed network data services The case of streaming video and downloading large files," *Electronics Goes Green 2020+*, p. 10, 2020.
- [2] E. Ekudden, "Breaking the energy curve," Ericsson, Tech. Rep., Mar. 2020.
- [3] C. Andersson, J. Bengtsson, G. Byström, P. Frenger, Y. Jading, and M. Nordenström, "Improving energy performance in 5G networks and beyond," *Ericsson Technology Review*, no. 8, pp. 2–11, August 2022.
- [4] M. Gruber, O. Blume *et al.*, "EARTH Energy Aware Radio and Network Technologies," in 2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, Sep. 2009, pp. 1–5.
- [5] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on network energy savings for NR (Release 18)," 3rd Generation Partnership Project (3GPP), Tech. Rep. 38.864, Dec. 2022, version 18.0.0.
- [6] "Green 5G White Paper," Huawei, Tech. Rep., Oct. 2021.
- [7] E. Dahlman, S. Parkvall, and J. Sköld, 5G NR: The next generation wireless access technology. Academic Press, 2020.
- [8] D. López-Pérez, A. De Domenico *et al.*, "A Survey on 5G Radio Access Network Energy Efficiency: Massive MIMO, Lean Carrier Design, Sleep Modes, and Machine Learning," *IEEE Comm. Surv. & Tut.*, vol. 24, no. 1, pp. 653–697, 2022.
- [9] S. Zhang, Q. Wu, S. Xu, and G. Y. Li, "Fundamental Green Tradeoffs: Progresses, Challenges, and Impacts on 5G Networks," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 33–56, 2017.
- [10] S. Tombaz, P. Frenger, F. Athley, E. Semaan, C. Tidestav, and A. Furuskar, "Energy Performance of 5G-NX Wireless Access Utilizing Massive Beamforming and an Ultra-Lean System Design," in 2015 IEEE Global Communications Conference (GLOBECOM), Dec. 2015, pp. 1–7.
- [11] P. Lahdekorpi, M. Hronec, P. Jolma, and J. Moilanen, "Energy efficiency of 5G mobile networks with base station sleep modes," in 2017 IEEE Conference on Standards for Communications and Networking, Helsinki, Sep. 2017, pp. 163–168.
- [12] M. Matalatala, M. Deruyck *et al.*, "Simulations of beamforming performance and energy efficiency for 5G mm-wave cellular networks," in 2018 IEEE Wireless Communications and Networking Conference (WCNC), Apr. 2018, pp. 1–6.
- [13] P. Frenger and K. W. Helmersson, "Energy Efficient 5G NR Street-Macro Deployment in a Dense Urban Scenario," in 2019 IEEE Global Communications Conference (GLOBECOM), Dec. 2019, pp. 1–6.
- [14] P. Frenger and R. Tano, "More Capacity and Less Power: How 5G NR Can Reduce Network Energy Consumption," in 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring), Apr. 2019, pp. 1–5.
- [15] L. Golard, J. Louveaux, and D. Bol, "Evaluation and projection of 4G and 5G RAN energy footprints: the case of Belgium for 2020–2025," *Annals of Telecommunications*, Nov. 2022.
- [16] F. E. Salem, T. Chahed, Z. Altman, and A. Gati, "Traffic-aware Advanced Sleep Modes management in 5G networks," in 2019 IEEE Wireless Communications and Networking Conference (WCNC), Marrakesh, Apr. 2019, pp. 1–6.
- [17] N. Piovesan, A. De Domenico, M. Bernabe et al., "Forecasting Mobile Traffic to Achieve Greener 5G Networks: When Machine Learning is Key," in 2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2021, pp. 276–280.

[18] H. V. Cheng, D. Persson, E. Bjornson, and E. G. Larsson, "Massive MIMO at night: On the operation of massive MIMO in low traffic scenarios," in 2015 IEEE International Conference on Communications (ICC), London, 2015, pp. 1697–1702.

- [19] 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.
- [20] E. Bjornson, 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.
- [21] J. C. Marinello, T. Abrao, A. Amiri, E. de Carvalho, and P. Popovski, "Antenna Selection for Improving Energy Efficiency in XL-MIMO Systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13305–13318, Nov. 2020.
- [22] E. Peschiera and F. Rottenberg, "Linear Precoder Design in Massive MIMO under Realistic Power Amplifier Consumption Constraint," in 2022 Joint WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux, 2022, p. 5.
- [23] Ł. Budzisz, F. Ganji, Rizzo *et al.*, "Dynamic Resource Provisioning for Energy Efficiency in Wireless Access Networks: A Survey and an Outlook," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 2259–2285, 2014.
- [24] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental trade-offs on green wireless networks," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 30–37, Jun. 2011.
- [25] Jingon Joung, Chin Keong Ho, and Sumei 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.
- [26] J. Wu, S. Rangan, and H. Zhang, Eds., Green Communications: Theoretical Fundamentals, Algorithms, and Applications. CRC Press, Apr. 2016.
- [27] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An Overview of Sustainable Green 5G Networks," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 72–80, Aug. 2017.
- [28] 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), May 2015, pp. 1–7.
- [29] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson *et al.*, "How much energy is needed to run a wireless network?" *IEEE Wireless Communications*, vol. 18, no. 5, pp. 40–49, Oct. 2011.
- [30] S. C. Cripps, *RF power amplifiers for wireless communications*, 2nd ed., ser. Artech House microwave library. Boston, Mass.: Artech House, 2006.
- [31] D. Persson, T. Eriksson, and E. G. Larsson, "Amplifier-Aware Multiple-Input Multiple-Output Power Allocation," *IEEE Communications Letters*, vol. 17, no. 6, pp. 1112–1115, Jun. 2013.
- [32] A. Grebennikov, RF and microwave power amplifier design. McGraw-Hill Education, 2015.
- [33] S. Mikami, T. Takeuchi et al., "An Efficiency Degradation Model of Power Amplifier and the Impact against Transmission Power Control for Wireless Sensor Networks," in 2007 IEEE Radio and Wireless Symposium, Long Beach, CA, USA, 2007, pp. 447–450.
- [34] M. M. A. Hossain, C. Cavdar, E. Bjornson, and R. Jantti, "Energy Saving Game for Massive MIMO: Coping With Daily Load Variation," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 3, pp. 2301–2313, Mar. 2018.
- [35] M. M. A. Hossain and R. Jantti, "Impact of efficient power amplifiers in wireless access," in 2011 IEEE Online Conference on Green Communications, Sep. 2011, pp. 36–40.
- [36] Ericsson, "Modeling and evaluation methodology for network energy saving," Tech. Rep. 3GPP TSG RAN WG1 #109-e, May 2022, R1-2204881.
- [37] T. Islam, D. Lee, and S. S. Lim, "Enabling Network Power Savings in 5G-Advanced and Beyond," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 6, pp. 1888–1899, Jun. 2023.