Wake-Up Radio based Access in 5G under Delay Constraints: Modeling and Optimization

Soheil Rostami, Sandra Lagen, Mário Costa, *Member, IEEE*,
Mikko Valkama, *Senior Member, IEEE*, and Paolo Dini

***Abstract*—Recently, the concept of wake-up radio based access has been considered as an effective power saving mechanism for 5G mobile devices. In this article, the average power consumption of a wake-up radio enabled mobile device is analyzed and modeled by using a semi-Markov process. Building on this, a delay-constrained optimization problem is then formulated, to maximize the device energy-efficiency under given latency requirements, allowing the optimal parameters of the wake-up scheme to be obtained in closed form. The provided numerical results show that, for a given delay requirement, the proposed solution is able to reduce the power consumption by up to 40% compared with an optimized discontinuous reception (DRX) based reference scheme.**

***Index Terms*—Energy efficiency, mobile device, DRX, wake-up radio, 5G, optimization, delay constraint.**

I. Introduction

In order for the emerging fifth generation (5G) mobile networks to satisfy the ever-growing needs for higher data-rates and network capacities, while simultaneously facilitating other quality of service (QoS) improvements, computationally-intensive physical layer techniques and high bandwidth communication are essential [2], [3]. At the same time, however, the device power consumption tends to increase which, in turn, can deplete the mobile device's battery very quickly. Moreover, it is estimated that feature phones and smartphones consume 2 kWh/year and 7 kWh/year, respectively, based on charging every 60^{th} hour equal to 40% of battery capacity every day and a standby scenario of 50% of the remaining time [4]. Also, the carbon footprints of production of feature phones and smartphones are estimated to be 18 kg and 30 kg CO2e per device, respectively, which still is the major contributor of CO2 emission of mobile communication systems [4].

In general, battery lifetime is one of the main issues that mobile device consumers consider important from device usability point of view [5]. However, since the evolution of battery technologies tends to be slow [6], the energy efficiency of the mobile device's main functionalities, such as the cellular subsystem, needs to be improved [5], [7], [8]. Furthermore, since the data traffic has been largely downlink-dominated [9], the power saving mechanisms for cellular subsystems in receive mode are of great importance.

The 3rd generation partnership project (3GPP) has specified discontinuous reception (DRX) as one of the *de facto* energy saving mechanism for long-term evolution (LTE), LTE-Advanced and 5G New Radio (NR) networks [10]–[13]. DRX allows the mobile device to reduce its energy consumption by switching off some radio modules for long periods of time, activating them only for short intervals. To this end, the modeling and optimization of DRX mechanisms have attracted a large amount of research interest in recent years. The authors in [14] proposed an adaptive approach to configure DRX parameters according to users' activities, aiming to balance power saving and packet delivery latency. Koc *et al.* formulated the DRX mechanism as a multi-objective optimization problem in [15], satisfying the latency requirements of active traffic flows and the corresponding preferences for power saving. In [16], DRX is modeled as a semi-Markov process with three states (active, light-sleep, and deep-sleep), and the average power consumption as well as the average delay are calculated and optimized. Additionally, the authors in [17] utilized exhaustive search over a large parameter set to configure all DRX parameters. Such a method may not be attractive from a computational complexity perspective for real-time/practical applications, however, it provides the optimal DRX-based power consumption and is thus used in this article as a benchmark.

To improve the device's energy-efficiency beyond the capabilities of ordinary DRX, the concept of wake-up radio based access has been discussed, e.g., in [6], [18], [19]. Specifically, in the cellular communications context, the wake-up radio based approaches have been recently discussed and described, e.g., in [20] and [21]. In such concept, the mobile device monitors only a narrowband control channel signaling referred to as wake-up signaling at specific time instants and subcarriers, in an OFDMA-based radio access systems such as LTE or NR, in order to decide whether to process the actual upcoming physical downlink control channel (PDCCH) or discard it. Compared to DRX-based systems, this directly reduces the buffering requirements and processing of empty subframes as well as the corresponding power consumption. Furthermore, in [20], the concept of a low-complexity wake-up receiver (WRx) was developed to decode the corresponding wake-up signaling, and to acquire the necessary time and frequency synchronization. A wake-up scheme that enhances

S. Rostami is with Huawei Technologies Oy (Finland) Co. Ltd, Helsinki, Finland, and also with Tampere University, Finland. E-mail: soheil.rostami1@huawei.com, soheil.rostami@tuni.fi.

S. Lagen and P. Dini are with Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA), Barcelona, Spain. E-mails: {sandra.lagen, paolo.dini}@cttc.es.

M. Costa is with Huawei Technologies Oy (Finland) Co. Ltd, Helsinki, Finland. E-mail: mariocosta@huawei.com.

M. Valkama is with the Department of Electrical Engineering, Tampere University, Finland. E-mail: mikko.valkama@tuni.fi.

Limited subset of early stage results presented at IEEE GLOBECOM, Dec. 2019 [1].

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 675891 (SCAVENGE), Tekes TAKE-5 project, Spanish MINECO grant TEC2017-88373-R (5G-REFINE), and Generalitat de Catalunya grant 2017 SGR 1195.

the power consumption of machine type communications (MTC) is introduced in 3GPP LTE Release-15 [22], which is based on a narrowband signal, transmitted over the available symbols of configured subframes. It is also considered as the starting point of NR power saving study item in 3GPP NR Release-17 [23].

In general, the existing wake-up concepts and algorithms, such as those described in [6], [18]–[21], [24]–[32], build on static operational parameters that are determined by the radio access network at the start of the user's session, and kept invariant, even if traffic patterns change. Accordingly, methods to optimize such parameters that characterize the employed wake-up scheme are needed, to further reduce energy consumption according to the traffic conditions. This is one of the main objectives of this paper. Specifically, the main contributions of this paper are as follows. Firstly, the wake-up radio based access scheme is modeled by means of a semi-Markov process. In the model, we consider realistic WRx operation by introducing start-up and power-down periods of the baseband unit (BBU), false alarm and misdetection probabilities of the wake-up signaling, as well as the packet service time. With such a model, the average power consumption and buffering delay can be accurately quantified and estimated for a given set of wake-up related parameters. Secondly, by utilizing such a mathematical model, the minimization of terminal's power consumption under Poisson traffic model is addressed for a given delay constraint. As a result, a closed-form optimal solution for the operational parameters is obtained. Furthermore, the range of packet arrival rates, for which the wake-up scheme is suitable and energy-efficient, is determined. Finally, simulation-based numerical results are provided in order to validate the proposed model and methods as well as to investigate the power consumption of our proposed solution compared to the optimized DRX-based reference mechanism proposed in [17]. The approach described in [17] is selected as the benchmark since it provides the optimal power consumption in DRX-based reference systems. Furthermore, to the best of our knowledge, virtually all of the DRX-based literature ignores the start-up and power-down energy consumption, and [17] also neglects the packet service time. Therefore, we have modified the approach in [17] slightly in order to consider such additional energy consumption in the optimization.

The rest of this paper is organized as follows. Section II presents a brief review of the considered wake-up scheme and its corresponding parameters, and defines basic system assumptions. In Section III, we model the wake-up based access scheme by means of a semi-Markov process and derive the power consumption as well as the buffering delay. Then, building on these mathematical models, in Section IV, the optimization problem is formulated and the optimal solution for minimum power consumption is found in closed-form. These are followed by numerical results, remarks, and conclusions in Sections V, VI, and VII, respectively. Some details on the analysis related to the modeling of power consumption are reported in the Appendix. For readers' convenience, the most relevant variables used throughout this paper are listed in Table I. Terminology-wise, we use gNB to refer to the base-station unit and UE to denote the mobile device, according to 3GPP NR specifications [10].

Table I: Most important variables used throughout the article

Variable	Definition
PW_1	power consumption of cellular module while WRx is active
PW_2	power consumption of cellular module at active-decoding state (BBU is active)
PW_3	power consumption of cellular module at active-inactivity timer state (BBU is active)
PW_4	power consumption of cellular module at sleep state
S_k	UE state in the state machine modeling (S_1, …, S_4)
$S(\tau_n)$	UE state at the τ_n jump time
P_{kl}	transition probability from state S_k to state S_l
ω_k	holding time for state S_k
t_p	inter-packet arrival time
λ	packet arrival rate
t_{su}	start-up time of cellular module
t_{pd}	power-down time of cellular module
t_w	wake-up cycle
t_i	length of inactivity timer
t_s	packet service time
t_{on}	on-duration time
$\overline{\mathrm{P}}_c$	average power consumption
$\overline{\mathrm{P}}_b$	average power consumption over boundary constraint
$\overline{\mathrm{D}}$	average buffering delay
$\overline{\mathrm{D}}_{\max}$	maximum tolerable delay or delay bound
t_{w_b}	minimum feasible wake-up cycle over boundary constraint
t_w^*	optimal value of wake-up cycle
t_i^*	optimal value of inactivity timer
λ_t	turnoff packet arrival rate
η	relative power saving factor
ϕ	power consumption ratio of UE at S_2 and S_3

II. Basic Wake-Up Radio Concept and Assumptions

In the considered wake-up radio based scheme, or wake-up scheme (WuS) for short, as presented in [20], the mobile device is configured to monitor a narrowband wake-up signaling channel in order to enhance its battery lifetime. Specifically, in every wake-up cycle (denoted by t_w), the WRx monitors the so-called physical downlink wake-up channel (PDWCH) for a specific on-duration time (t_{on}) in order to determine whether data has been scheduled or not. Occasionally, based on the interrupt signal from WRx, the BBU switches on, decodes both PDCCH and physical downlink shared channel (PDSCH), and performs normal connected-mode procedures. The WuS can be adopted for both connected and idle states of radio resource control [20], and can be configured based on maximum tolerable paging delay that idle users may experience, or alternatively, based on the delay requirements of a specific traffic type at connected state.

The wake-up signaling per each WRx contains a single-bit control information, referred to as the wake-up indicator (WI), where a WI of 1 indicates the WRx to wake up the BBU, because there is one or multiple packets to be received, while a WI of 0 signals the opposite. Each WI is code multiplexed with a user-specific signature to selected time-frequency resources, as described in [20]. When a WI of 1 is sent to WRx, the network expects the target mobile device

to decode the PDCCH with a time offset identical to that of start-up time (t_{su}). Fig. 1 (a) and (b) depict the basic operation and representative power consumption behavior of the conventional DRX-enabled cellular module and that of the cellular module with WRx, respectively, at a conceptual level. As illustrated, the WuS eliminates the unnecessarily wasted energy in the first and second DRX cycles, while also reducing the buffering delay compared to DRX. Due to the specifically-designed narrowband signal structure of WuS, the WRx power consumption (PW_1) is much lower than that of BBU active, either due to packet decoding (PW_2) or when inactivity timer is running (PW_3) [20]. The lowest power consumption is obtained in sleep state (PW_4). We consider that during the BBU active states, the power consumption due to packet decoding is larger than or equal to that of running inactivity timer (i.e., $\text{PW}_2 \geq \text{PW}_3$), but both are fixed. For presentation purposes, we denote the ratio of such power consumption at BBU active states as $\phi = \frac{\text{PW}_2}{\text{PW}_3}$, where $\phi \geq 1$.

In general, because NR supports wide bandwidth operation, packets can be served in a very short time duration. In addition, in case the user packet sizes are small, packet concatenation in NR is used, so that all buffered packets in a relatively short wake-up cycle can be served in a single transmission time interval (TTI). Accordingly, we assume that radio-link control entity (located at the gNB) concatenates all those packets arriving during the sleep state, and as soon as the BBU is triggered on, the device (UE) can receive and decode the concatenated packets for a service time of t_s, which equals to a single TTI. During the serving time, if there was a new packet arrival, the BBU starts serving the corresponding packet by the end of t_s. In case that there was no packet arrival by the end of t_s, the UE initiates its inactivity timer with a duration of t_i. After the inactivity timer is initiated, and if a new PDCCH message is received before the time expiration, the BBU enters the active-decoding state and serves the packet. However, if there is no PDCCH message received before the expiration of the inactivity timer, a sleep period starts, the WRx-enabled cellular module switches to sleep state, and WRx operates according to its wake-up cycle [20]. For reference, in case of DRX, the BBU sleeps according to its short and long DRX patterns [15], [33].

The introduction of a PDWCH has two fundamental consequences, namely misdetections and false alarms [20]. In the latter case, WRx wakes up in a predefined time instant, and erroneously decodes a WI of 0 as 1, leading to unnecessary BBU power consumption. The former, in turn, corresponds to the case where a WI of 1 is sent, but WRx decodes it incorrectly as 0. Such misdetection adds an extra delay and wastes radio resources. We denote the probability of misdetection and the probability of false alarm as P_{md} and P_{fa}, respectively. The requirements for the probability of misdetection of PDWCH are eventually stricter than those of the probability of false alarm [20].

One of the new features of 5G NR networks to reach their aggressive requirements is latency-optimized frame structure with flexible numerology, providing subcarrier spacings ranging from 15 kHz up to 240 kHz with a proportional change in cyclic prefix, symbol length, and slot duration [10]. Regardless

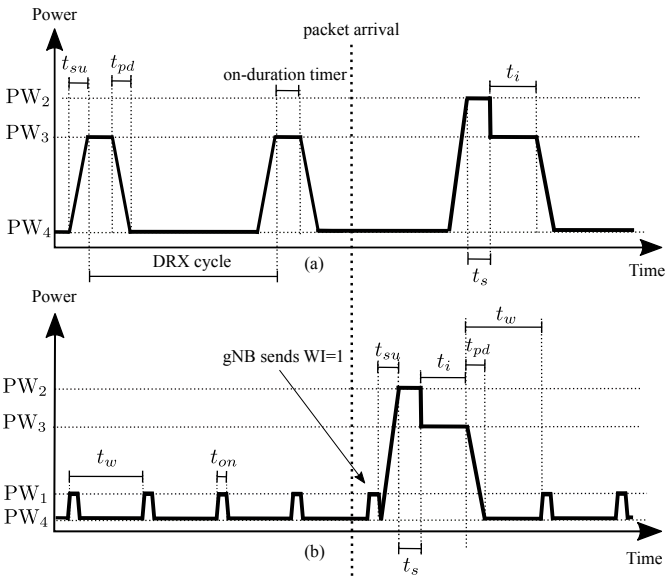

Figure 1: Power consumption profiles of (a) a typical DRX mechanism, and (b) a WRx-enabled cellular module.

of the numerology used, the length of one radio frame is fixed to 10 ms and the length of one subframe is fixed to 1 ms [2], as in 4G LTE/LTE-Advanced. However, in NR, the number of slots per subframe varies according to the numerology that is configured. Additionally, in order to support further reduced latencies, the concept of mini-slot transmission is introduced in NR, and hence the TTI varies depending on the service type ranging from one symbol, to one slot, and to multiple slots [2]. In this work, in order to provide consistent and exact timing definitions, different time intervals of the wake-up related procedures are defined as integer multiples of a TTI. Additionally, according to 3GPP, the packet service time (t_s) is one TTI, in which multiple packets can be concatenated. Furthermore, for the sake of clarity, a TTI duration of 1 ms is taken as the baseline system assumption for the WuS configurations, which then facilitates applying the proposed concepts also in future evolution of LTE-based systems.

Finally, it is important to note that from a system-level point of view, the configurable parameters of the WuS are the wake-up cycle (t_w) and the inactivity timer (t_i), whose values we will optimize in Section IV. The remaining parameters (t_{on}, t_{pd}, t_{su}, t_s) depend on physical constraints and signal design, and accordingly we assume them to be fixed, i.e., the optimization will be done for fixed (given) values of t_{on}, t_{pd}, t_{su} and t_s.

III. State Machine based Wake-up System Model

For mathematical convenience, the performance of the wake-up based system is studied and analyzed in the context of a Poisson arrival process with a packet arrival rate of λ packets per TTI. In the Poisson traffic model, each packet service session consists of a sequence of packets with exponentially distributed inter-packet arrival time (t_p) [33].

The power states of WuS are modeled as a semi-Markov process with four different states that correspond to WRx-ON

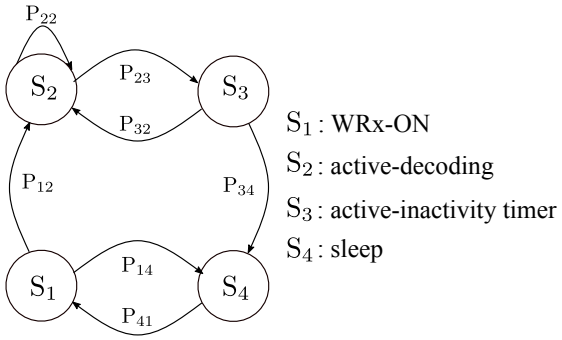

Figure 2: Semi-Markov process for the state transitions of the wake-up scheme, with states S_1, S_2, S_3 and S_4.

(state S_1), active-decoding (state S_2), active-inactivity timer (state S_3), and sleep (state S_4), as shown in Fig. 2. At S_1, WRx monitors PDWCH, and if WRx receives WI=0, UE transfers to S_4, otherwise (WI=1) it transfers to S_2. At S_2, UE decodes the packets for a fixed duration of t_s; if the device is scheduled before the end of t_s, it starts decoding the new packet, and remains at S_2, otherwise the device transfers to S_3. At S_3, t_i is running, and if the device is scheduled before the expiry of t_i, it enters in S_2, otherwise the device transfers to S_4. At S_4, the device is in sleep state, and cannot receive any signal, as opposed to being fully-functional at S_2 and S_3. Moreover, at the end of a wake-up cycle in sleep period, the UE moves to S_1. As noted already in the previous Section II, each state is associated with a different power consumption level, PW_k, $k \in \{1, 2, 3, 4\}$.

Transition probabilities: The transition probability from UE state S_k to S_l (P_{kl}) is defined as

$$P_{kl} = \lim_{n\to\infty} \Pr(S(\tau_n) = S_l | S(\tau_{n-1}) = S_k), \tag{1}$$

where $S(\tau_n)$ is the UE state at the τ_n jump time[1].

When the UE is at S_1, it moves to S_2 either because of false alarm or correct detection; otherwise, it moves to S_4. Accordingly, P_{12} and P_{14} can be expressed as

$$\begin{aligned} P_{12} &= \Pr[t_p > t_w | S_1] P_{fa} + \Pr[t_p \le t_w | S_1](1 - P_{md}) \\ &= e^{-\lambda t_w} P_{fa} + (1 - e^{-\lambda t_w})(1 - P_{md}), \end{aligned} \tag{2}$$

and

$$P_{14} = 1 - P_{12}. \tag{3}$$

When the UE is at S_2, it decodes the packet for a duration of t_s, and if the next packet is received before the end of the current service time, the UE starts decoding the new packet at the end of service time; otherwise, it moves to S_3. Therefore, P_{22} and P_{23} can be obtained as

$$P_{22} = \Pr[t_p \le t_s | S_2] = 1 - e^{-\lambda t_s}, \tag{4}$$

and

$$P_{23} = 1 - P_{22}. \tag{5}$$

[1]In a semi-Markov process, $S(t)$ is a stochastic process with a finite set of states ($S_1, \ldots, S_4$ in our case), having step-wise trajectories with jumps at times $0 < \tau_1 < \tau_2 ...$, and its values at the jump times ($S(\tau_n)$) form a Markov chain.

When the UE is at S_3, it moves to S_2 if the next packet is received before the expiry of t_i; otherwise, it moves to S_4. Therefore, P_{32} and P_{34} can be expressed as

$$P_{32} = \Pr[t_p \le t_i | S_3] = 1 - e^{-\lambda t_i}, \tag{6}$$

and

$$P_{34} = 1 - P_{32}. \tag{7}$$

Finally, at the end of every sleep cycle, the UE decodes PDWCH, and therefore,

$$P_{41} = 1. \tag{8}$$

Steady state probabilities: The steady state probability that the UE is at state S_k (P_k) is defined as

$$P_k = \lim_{n\to\infty} \Pr(S(\tau_n) = S_k). \tag{9}$$

By utilizing the set of balance equations ($P_k = \sum_{l=1}^{4} P_l P_{lk}$) and the basic sum of probabilities ($\sum_{k=1}^{4} P_k = 1$), the P_k's can be obtained as follows

$$P_1 = P_4 = \frac{P_{34} P_{23}}{2P_{34} P_{23} + P_{12}(1 + P_{23})}, \tag{10}$$

$$P_2 = P_1 \frac{P_{12}}{P_{23} P_{34}}, \tag{11}$$

$$P_3 = P_1 \frac{P_{12}}{P_{34}}. \tag{12}$$

Holding times: The corresponding holding time for state S_k is denoted by ω_k, $k \in \{1, 2, 3, 4\}$. The holding times for ω_1, ω_2 and ω_4 are constant and given by: $\mathbb{E}[\omega_1] = t_{on}$, $\mathbb{E}[\omega_2] = t_s$ and $\mathbb{E}[\omega_4] = t_w - t_{on}$. However, ω_3 is dependent on the inter-packet arrival time (t_p). If a packet arrives before t_i, ω_3 is equal to the inter-packet arrival time, otherwise ω_3 equals to t_i. Therefore, ω_3 can be calculated as a function of t_p as

$$\omega_3(t_p) = \begin{cases} t_p, & \text{for } \ t_p \le t_i \ , \\ t_i, & \text{for } \ t_p > t_i \ . \end{cases} \tag{13}$$

Hence, $\mathbb{E}[\omega_3]$ can be expressed as

$$\mathbb{E}[\omega_3] = \int_0^\infty \omega_3(t) f_p(t) dt = \frac{1 - e^{-\lambda t_i}}{\lambda}, \tag{14}$$

where $f_p(t) = \lambda e^{-\lambda t}$ is the probability density function of the exponentially distributed packet arrival time.

A. Average Power Consumption

The average power consumption of the UE, denoted by $\overline{P}_c$, can be calculated as the ratio of the average energy consumption and the corresponding overall observation period. It is given by Eq. (15) at the top of the next page, where t_{su} and t_{pd} correspond to the length of the start-up and power-down stages (transition times), respectively. The corresponding average energy consumption of transitions between states are calculated as the areas under the power profiles of start-up and power-down stages, see Fig. 1, whose contribution to the average energy consumption is multiplied by its probability

$$\overline{\mathrm{P}}_c = \frac{0.5\mathrm{P}_1\mathrm{P}_{12}t_{su}(\mathrm{PW}_2-\mathrm{PW}_4)+0.5\mathrm{P}_3\mathrm{P}_{34}t_{pd}(\mathrm{PW}_3-\mathrm{PW}_4)+\sum_{n=1}^{4}\mathrm{P}_n\mathbb{E}[\omega_n]\mathrm{PW}_n}{\mathrm{P}_1\mathrm{P}_{12}t_{su}+\mathrm{P}_3\mathrm{P}_{34}t_{pd}+\sum_{n=1}^{4}\mathrm{P}_n\mathbb{E}[\omega_n]} \tag{15}$$

of occurrence ($\mathrm{P}_1\mathrm{P}_{12}$ and $\mathrm{P}_3\mathrm{P}_{34}$, respectively), thus leading to $0.5\mathrm{P}_1\mathrm{P}_{12}t_{su}(\mathrm{PW}_2-\mathrm{PW}_4)$ and $0.5\mathrm{P}_3\mathrm{P}_{34}t_{pd}(\mathrm{PW}_3-\mathrm{PW}_4)$, respectively.

For modeling simplicity, we assume that $t_{on}\approx 0$, $\mathrm{PW}_4\approx 0$, $P_{fa}\approx 0$, and $P_{md}\approx 0$. Therefore, Eq. (15) can be expanded as a multivariate function of t_w and t_i, denoted by $\overline{\mathrm{P}}_c(t_w,t_i)$, as follows

$$\overline{\mathrm{P}}_c(t_w,t_i) = \mathrm{PW}_3\frac{e^{\lambda t_i}(\phi t_s e^{\lambda t_s}+\frac{1}{\lambda})+\frac{1}{2}(\phi t_{su}+t_{pd})-\frac{1}{\lambda}}{e^{\lambda t_i}(t_s e^{\lambda t_s}+\frac{1}{\lambda})+\frac{t_w}{1-e^{-\lambda t_w}}+t_{su}+t_{pd}-\frac{1}{\lambda}}. \tag{16}$$

In order to provide more insight into $\overline{\mathrm{P}}_c(t_w,t_i)$, the instantaneous rate of change of the power consumption with respect to both t_w and t_i is calculated next. Assuming continuous variables, the partial derivatives of $\overline{\mathrm{P}}_c(t_w,t_i)$ with respect to t_w and t_i are given by

$$\frac{\partial\overline{\mathrm{P}}_c(t_w,t_i)}{\partial t_i} = \mathrm{PW}_3 e^{\lambda t_i}(\lambda\phi t_s e^{\lambda t_s}+1)\times \frac{\frac{1}{2}((2-\phi)t_{su}+t_{pd})+\frac{t_w}{1-e^{-\lambda t_w}}}{\left(e^{\lambda t_i}(t_s e^{\lambda t_s}+\frac{1}{\lambda})+\frac{t_w}{1-e^{-\lambda t_w}}+t_{su}+t_{pd}-\frac{1}{\lambda}\right)^2}, \tag{17}$$

and

$$\frac{\partial\overline{\mathrm{P}}_c(t_w,t_i)}{\partial t_w} = \mathrm{PW}_3\big((1+\lambda t_w)e^{-\lambda t_w}-1\big)\times \frac{\left(e^{\lambda t_i}(\phi t_s e^{\lambda t_s}+\frac{1}{\lambda})+\frac{1}{2}(\phi t_{su}+t_{pd})-\frac{1}{\lambda}\right)}{\left((1-e^{-\lambda t_w})(e^{\lambda t_i}(t_s e^{\lambda t_s}+\frac{1}{\lambda})+t_{su}+t_{pd}-\frac{1}{\lambda})+t_w\right)^2}. \tag{18}$$

It can be seen from Eq. (17) that $\frac{\partial\overline{\mathrm{P}}_c(t_w,t_i)}{\partial t_i} > 0$ for all feasible values of t_w and t_i. From Eq. (18), we can conclude that $\frac{\partial\overline{\mathrm{P}}_c(t_w,t_i)}{\partial t_w} < 0$ due to fact that $(1+\lambda t_w)e^{-\lambda t_w} < 1$.

Therefore, the average power consumption $\overline{\mathrm{P}}_c(t_w,t_i)$ is a strictly increasing function with respect to t_i at $t_i \geq 0$, and it is a strictly decreasing function with respect to t_w at $t_w \geq 0$. As expected, increasing the wake-up cycle t_w for a fixed t_i can reduce the power consumption. However, by increasing t_i for a fixed t_w, the power consumption increases.

B. Average Buffering Delay

We next assume that packets arriving during S_4 are buffered at the gNB until the UE enters S_2, thus causing buffering delay. Without loss of generality, we assume that the radio access network experiences unsaturated traffic conditions. Therefore, all packets that arrive are served without any further scheduling delay. Furthermore, to simplify the delay modeling, we omit the buffering delay caused by packets arriving on S_1 or at the start-up state of the modem. This is because the additional buffering delay of such packet arrivals is anyway very small ($t_{on}+t_{su}$). Additionally, thanks to the adoption of the WuS seeking to reduce unnecessary start-ups, the number of occurrences of such scenarios is low.

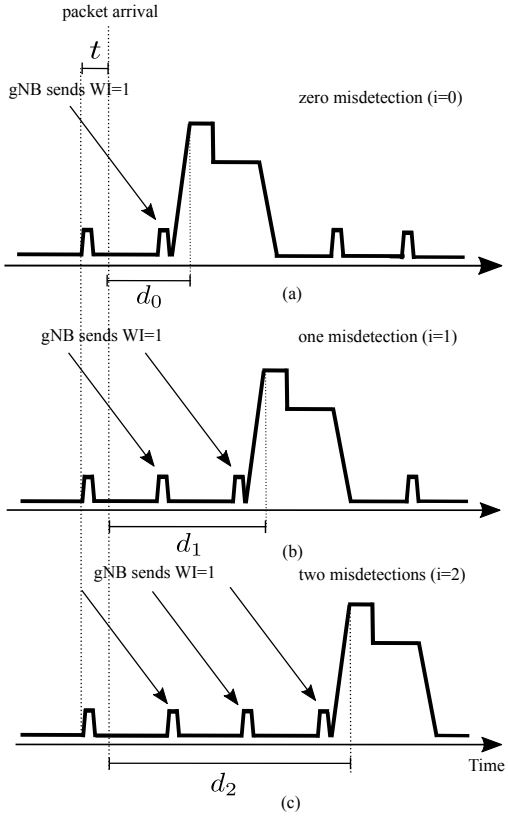

Figure 3: Buffering delay caused by wake-up scheme when (a) there is no misdetection, (b) there is a single misdetection, and (c) there are two consecutive misdetections.

Due to slot-based frame structure of NR, where PDCCH is sent at the beginning of the TTI, inherently, all packet arrivals (regardless WuS is utilized or not) suffer from small buffering delay. Since Poisson arrivals are independently and uniformly distributed on any short interval, we assume that the arrival instant of the packet is uniformly distributed within the TTI, and hence an average extra delay of $t_s/2$ will be introduced.

Now, as already briefly mentioned in Section II, misdetections can in general increase the buffering delay. For this purpose, Fig. 3 illustrates the buffering delay experienced by the UE with no misdetections, with a single misdetection, and with two consecutive misdetections. The number of consecutive misdetections and the corresponding buffering delay are referred to as i and d_i, respectively, and their dependence on t can be written as $d_i(t) = (i+1)t_w + t_{su} + t_{on} - t$, for $i \in \{0,1,...\}$. Therefore, the average buffering delay, denoted by $\overline{\mathrm{D}}$, can be expressed as

$$\overline{\mathrm{D}} = \mathrm{P}_4\sum_{i=0}^{\infty}(1-P_{md})(P_{md})^i\int_0^{t_w} f_p(t)d_i(t)dt + \frac{t_s}{2}. \tag{19}$$

$$\frac{\partial \overline{\mathrm{D}}(t_w,t_i)}{\partial t_w} = \frac{e^{\lambda t_i}(1+e^{\lambda t_s})\big(1-(1+\lambda t_w)e^{-\lambda t_w}\big)+2(1-e^{-\lambda t_w})+2\lambda t_{su}e^{-\lambda t_w}}{\Big(2+(1-e^{-\lambda t_w})(1+e^{\lambda t_s})e^{\lambda t_i}\Big)^2} \tag{21}$$

$$\frac{\partial \overline{\mathrm{D}}(t_w,t_i)}{\partial t_i} = \frac{-\lambda e^{\lambda t_i}(1+e^{\lambda t_s})(1-e^{-\lambda t_w})\big(t_w+(t_{su}-\frac{1}{\lambda})(1-e^{-\lambda t_w})\big)}{\Big(2+(1-e^{-\lambda t_w})(1+e^{\lambda t_s})e^{\lambda t_i}\Big)^2} \tag{22}$$

Furthermore, due to the small value of misdetection probability ($P_{md} \approx 0$), the contribution of multiple consecutive misdetections to average buffering delay is small. Thus, the average buffering delay for $P_{md} \approx 0$ can be expanded and solved as a multivariate function of t_w and t_i, $\overline{\mathrm{D}}(t_w,t_i)$, as follows

$$\overline{\mathrm{D}}(t_w,t_i) = \mathrm{P}_4 \int_0^{t_w} f_p(t) d_0(t) dt + \frac{t_s}{2} = \frac{t_w+(t_{su}-\frac{1}{\lambda})(1-e^{-\lambda t_w})}{2+(1-e^{-\lambda t_w})(1+e^{\lambda t_s})e^{\lambda t_i}} + \frac{t_s}{2}. \tag{20}$$

We note that $\overline{\mathrm{D}}(t_w,t_i)$ in Eq. (20) is strictly-speaking a lower bound of the delay expression in (19).

Similarly to $\overline{\mathrm{P}}_c(t_w,t_i)$, the partial derivatives of $\overline{\mathrm{D}}(t_w,t_i)$ with respect to continuous variables t_w and t_i are given by Eq. (21) and Eq. (22), respectively, at the top of the page.

From Eq. (21), it can be easily concluded that $\frac{\partial \overline{\mathrm{D}}(t_w,t_i)}{\partial t_w} > 0$, due to fact that $\big(1-(1+\lambda t_w)e^{-\lambda t_w}\big) > 0$. Moreover, it can be shown that $\frac{\partial \overline{\mathrm{D}}(t_w,t_i)}{\partial t_i} < 0$ as follows

$$\begin{aligned} &\lambda t_w > 0 \Longrightarrow \lambda t_w > 1-e^{-\lambda t_w} \Longrightarrow \\ &\frac{\lambda t_w}{1-e^{-\lambda t_w}} > 1 \Longrightarrow \frac{\lambda t_w}{1-e^{-\lambda t_w}} > 1-\lambda t_{su} \Longrightarrow \\ &t_w+(t_{su}-\frac{1}{\lambda})(1-e^{-\lambda t_w}) > 0 \Longrightarrow \frac{\partial \overline{\mathrm{D}}(t_w,t_i)}{\partial t_i} < 0. \end{aligned} \tag{23}$$

Therefore, the average buffering delay $\overline{\mathrm{D}}(t_w,t_i)$ is a strictly increasing function with respect to t_w at $t_w \geq 0$, and it is a strictly decreasing function with respect to t_i at $t_i \geq 0$. As expected, contrary to the behavior of $\overline{\mathrm{P}}_c(t_w,t_i)$, increasing the wake-up cycle t_w for a fixed t_i increases the buffering delay. On the other hand, by increasing t_i for a fixed t_w, the buffering delay can be reduced.

The findings related to the impact of t_w and t_i on the average delay and power consumption are intuitive while are rigorously confirmed and quantified by the presented expressions.

IV. Optimization Problem Formulation and Solution

In this section, dual-parameter (t_w and t_i) constrained optimization problem is formulated with the objective of minimizing the UE power consumption under a buffering delay constraint. Specifically, the average buffering delay is constrained to be less than or equal to a predefined maximum tolerable delay or delay bound, denoted by $\overline{\mathrm{D}}_{\max}$, whose value is set based on the service type. To this end, building on the modeling results of the previous section, the optimization problem is now formulated as follows

$$\underset{t_w,t_i}{\text{minimize}} \quad \overline{\mathrm{P}}_c(t_w,t_i) \tag{24}$$

$$\text{subject to} \quad \overline{\mathrm{D}}(t_w,t_i) \leq \overline{\mathrm{D}}_{\max}, \tag{25}$$

$$t_w, t_i \in \{1,2,...\}, \tag{26}$$

where $\overline{\mathrm{P}}_c(t_w,t_i)$ and $\overline{\mathrm{D}}(t_w,t_i)$ are defined in Eq. (16) and Eq. (20), respectively.

The resulting optimization problem in (24)-(26) belongs to a class of intractable mixed-integer non-linear programming (MINLP) problems [34]. In this work, the corresponding MINLP is solved by using the equivalent non-linear programming problem with continuous variables, expressed below in (27)-(29), which is obtained by means of relaxing the second constraint (26) into a continuous constraint (see Eq. (29)), assuming that both parameters are positive real numbers larger than or equal to one (i.e., the minimum TTI unit). The relaxed optimization problem can be expressed as

$$\underset{t_w,t_i}{\text{minimize}} \quad \overline{\mathrm{P}}_c(t_w,t_i) \tag{27}$$

$$\text{subject to} \quad \overline{\mathrm{D}}(t_w,t_i) \leq \overline{\mathrm{D}}_{\max}, \tag{28}$$

$$t_w \geq 1, t_i \geq 1. \tag{29}$$

In general, the optimization problem in (27)-(29) is not jointly convex in t_w and t_i. Therefore, finding the global optimum is a challenging task. However, in the next subsections, we exploit the increasing/decreasing properties of the power consumption and delay expressions that we have derived in Section III, in order to derive additional properties of the problem that will allow us to find the optimal solution in closed form.

A. Unbounded Feasible Region

In this section, a schematic approach is used to illustrate the feasible region for the relaxed optimization problem in (27)-(29) and then the feasible region is narrowed down to the boundary of the delay constraint, whose points are proved to remain candidate solutions while the other feasible solutions are henceforth excluded.

Fig. 4 (a) and (b) show the increasing trend of the power consumption and the decreasing behaviour of the delay constraint as a function of t_i, while t_w is fixed at t_{w_0}, i.e. $\frac{\partial \overline{\mathrm{P}}_c(t_w,t_i)}{\partial t_i} > 0$ and $\frac{\partial \overline{\mathrm{D}}(t_w,t_i)}{\partial t_i} < 0$ (as proved in Section III). Let us consider an arbitrary point A in the interior of the feasible region ($t_{i_A} > t_{i_m}$, where $\overline{\mathrm{D}}(t_{w_0},t_{i_m}) = \overline{\mathrm{D}}_{\max}$). As it can be seen from Fig. 4 (a) and (b), there is always a point on the boundary of the delay constraint, referred to as B ($t_{i_B} = t_{i_m}$), where its power consumption $\overline{\mathrm{P}}_{c_B}$ is lower than that of A

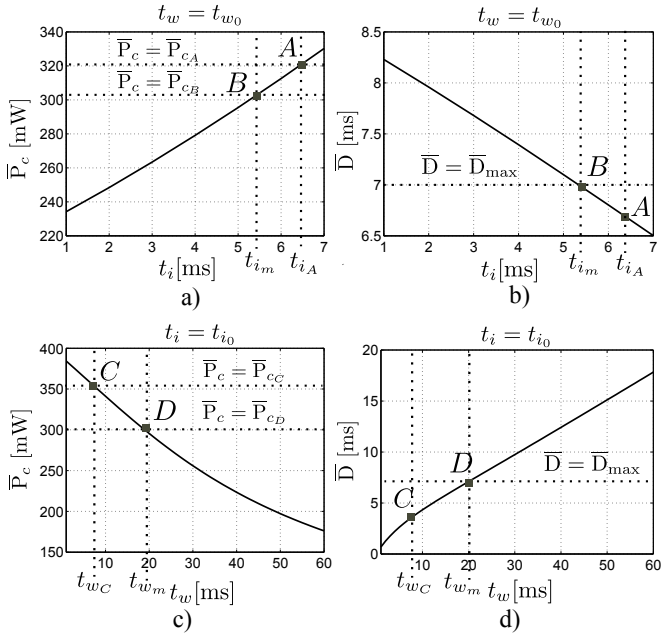

Figure 4: Schematic proof that optimal point lies over the boundary.

($\overline{\mathrm{P}}_{c_B} < \overline{\mathrm{P}}_{c_A}$). Hence, we can conclude that, for any fixed t_w, under a given delay constraint, there is a point on the boundary that attains the lowest power consumption.

Similarly, Fig. 4 (c) and (d) show the decreasing trend of the power consumption and increasing behaviour of the delay constraint as a function of t_w, while t_i is fixed at t_{i_0}, i.e. $\frac{\partial \overline{\mathrm{P}}_c(t_w,t_i)}{\partial t_w} < 0$ and $\frac{\partial \overline{\mathrm{D}}(t_w,t_i)}{\partial t_w} > 0$ (as proved in Section III). Consider an arbitrary point C in the interior of the feasible region ($t_{w_C} < t_{w_m}$ where $\overline{\mathrm{D}}(t_{w_m}, t_{i_0}) = \overline{\mathrm{D}}_{\max}$). As it can be seen from Fig. 4 (c) and (d), there is always a point on the boundary of the delay constraint, referred to as D ($t_{w_D} = t_{w_m}$), where its power consumption $\overline{\mathrm{P}}_{c_D}$ is lower than that of C ($\overline{\mathrm{P}}_{c_D} < \overline{\mathrm{P}}_{c_C}$). Then, we can conclude that, for any fixed t_i, under a given delay constraint, there is a point on the boundary that attains the lowest power consumption.

Therefore, because for both scenarios (fixed t_w and fixed t_i), the lowest power consumption occurs at the boundary of the delay constraint, we can conclude that the optimal point cannot be located in the interior of the feasible region, but rather it lies over the boundary. That is, any arbitrary point (t_w, t_i) in the feasible region of the relaxed optimization problem in (27)-(29) cannot be an optimal point, unless it lies on the boundary (rather than the interior) of the delay constraint, i.e., $\overline{\mathrm{D}}(t_w, t_i) = \overline{\mathrm{D}}_{\max}$.

B. Power Consumption over Boundary

Next, the equation of the boundary curve (expressed through t_i as a function of t_w) is derived, and then the power consumption profile of all points on the boundary is calculated as well as formulated as a function of t_w only. In particular, the boundary curve can be obtained by finding all the solutions for which the inequality constraint in (28) is satisfied with equality, while the constraint in (29) is met, i.e.,

$$\overline{\mathrm{D}}(t_w, t_i) = \overline{\mathrm{D}}_{\max} \text{ for all } t_w, t_i \geq 1. \tag{30}$$

By utilizing Eq. (20), we can isolate t_i, and the boundary curve can be formulated as follows

$$t_i(t_w) = \frac{1}{\lambda} \ln\Big(\frac{t_w + (t_{su} - \frac{1}{\lambda})(1 - e^{-\lambda t_w}) - 2(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})}{(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})(1 - e^{-\lambda t_w})(1 + e^{\lambda t_s})}\Big) \quad \text{for all } t_w \geq t_{w_b}, \tag{31}$$

where t_{w_b} (see Eq. (36)) is the minimum feasible value of t_w over the boundary.

By using Eq. (31), one can show that $t_i(t_w)$ is an increasing function with respect to t_w on any feasible t_w point over the boundary of the delay constraint (i.e., $\frac{dt_i(t_w)}{dt_w} \geq 0$), as follows. Let us use the composite function rule over (31), so that $\frac{dt_i(t_w)}{dt_w}$ can be calculated as follows

$$\frac{dt_i(t_w)}{dt_w} = \frac{1}{\lambda} \frac{d\ln(\mathrm{Arg})}{d\mathrm{Arg}} \frac{d\mathrm{Arg}(t_w)}{dt_w}, \tag{32}$$

where Arg refers to the argument of the logarithm in (31). Since the logarithmic function is monotonically increasing ($\frac{d\ln(\mathrm{Arg})}{d\mathrm{Arg}} \geq 0$), it is sufficient to prove that $\mathrm{Arg}(t_w)$ is increasing with respect to t_w,

$$\frac{d\mathrm{Arg}(t_w)}{dt_w} = \frac{1 - (1 - 2(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})\lambda + \lambda t_w)e^{-\lambda t_w}}{(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})(1 + e^{\lambda t_s})(1 - e^{-\lambda t_w})^2}, \tag{33}$$

from which we can write

$$\begin{aligned} 1 - (1 + \lambda t_w)e^{-\lambda t_w} > 0 \Longrightarrow \\ 1 - (1 - 2(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})\lambda + \lambda t_w)e^{-\lambda t_w} \geq 0 \Longrightarrow \\ \frac{d\mathrm{Arg}(t_w)}{dt_w} \geq 0. \end{aligned} \tag{34}$$

Therefore, we can conclude that $\frac{dt_i(t_w)}{dt_w} \geq 0$.

Additionally, one can prove that $t_i = 1$ and $t_w = t_{w_b}$ (in which t_{w_b} always exists, and is larger than or equal to one) is located over the boundary, as follows. Based on Eq. (31), and by induction on $t_i = 1$, we can write

$$\begin{aligned} e^{\lambda} = \frac{t_w + (t_{su} - \frac{1}{\lambda})(1 - e^{-\lambda t_w}) - 2(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})}{(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})(1 - e^{-\lambda t_w})(1 + e^{\lambda t_s})} \Longrightarrow \\ e^{-\lambda t_w} = -\frac{-\lambda t_w + \big((e^{\lambda}(1 + e^{\lambda t_s}) + 2)\overline{\mathrm{D}}_{\max} - t_{su}\big)\lambda + 1}{\lambda t_{su} - e^{\lambda}(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})(1 + e^{\lambda t_s})\lambda - 1} \Longrightarrow \\ e^{-\lambda t_w + F} = -e^{F}\frac{-\lambda t_w + F}{H} \Longrightarrow \\ -\lambda t_w + F = -\mathcal{W}\Big(\frac{H}{e^F}\Big) \Longrightarrow \\ t_w = \frac{1}{\lambda}\Big(F + \mathcal{W}\Big(\frac{H}{e^F}\Big)\Big), \end{aligned} \tag{35}$$

where $F = \big((e^{\lambda}(1 + e^{\lambda t_s}) + 2)(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2}) - t_{su}\big)\lambda + 1$, $H = \lambda t_{su} - e^{\lambda}(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})(1 + e^{\lambda t_s})\lambda - 1$ and $\mathcal{W}(x)$ is the Lambert W function [35]. For typical $\overline{\mathrm{D}}_{\max}$ and t_{su} values, $1 \ll \frac{F}{\lambda}$ and $H < 0$, therefore, the main branch of the Lambert W function ($\mathcal{W}_0$) can be considered as a solution for (35) that has a value greater than -1. Then, we can conclude that

$$t_{w_b} = \frac{1}{\lambda}\Big(F + \mathcal{W}_0\Big(\frac{H}{e^F}\Big)\Big) \geq 1, \tag{36}$$

and, as a result, all feasible points on the boundary curve can be specified and constrained by $t_i \geq 1$ and $t_w \geq t_{w_b}$.

The t_{w_b} in (36) is the smallest feasible t_w over the boundary curve because if we assume that there is a t_w smaller than t_{w_b}, based on Eq. (32) and (34), its corresponding t_i should become smaller than one, which belongs to the unfeasible region. Therefore, based on proof-by-contradiction, ($t_i = 1$, $t_w = t_{w_b}$) lies over the corner part of the boundary. Consequently, $t_i \geq 1$ and $t_w \geq t_{w_b}$ are equivalent constraints of the boundary of the delay constraint. Therefore, the point ($t_i = 1$, $t_w = t_{w_b}$) is an extreme point, and it is located over the boundary curve of the delay constraint, where t_{w_b} is larger than or equal to one.

Finally, by substituting the value of $e^{\lambda t_i}$ over the boundary (argument of logarithm in (31)) into (16), the average power consumption of all the points over the boundary, referred to as $\overline{\mathrm{P}}_b(t_w)$, can be obtained as follows

$$\overline{\mathrm{P}}_b(t_w) = \mathrm{PW}_3 \frac{u_1 + u_2 t_w + u_3 e^{-\lambda t_w}}{w_1 + w_2 t_w + w_3 e^{-\lambda t_w}} \quad \text{for all } t_w \geq t_{w_b}, \tag{37}$$

where

$$u_1 = \big(\frac{1}{2}(\phi t_{su} + t_{pd}) - \frac{1}{\lambda}\big)(1 + e^{\lambda t_s})(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2}) + (t_s e^{\lambda t_s} + \frac{1}{\lambda})\big(t_{su} - \frac{1}{\lambda} - 2(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})\big), \tag{38}$$

$$u_2 = t_s e^{\lambda t_s} + \frac{1}{\lambda}, \tag{39}$$

$$u_3 = -\big(\frac{1}{2}(\phi t_{su} + t_{pd}) - \frac{1}{\lambda}\big)(1 + e^{\lambda t_s})(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2}) - (t_s e^{\lambda t_s} + \frac{1}{\lambda})\big(t_{su} - \frac{1}{\lambda}\big), \tag{40}$$

$$w_1 = \big(\frac{1}{2}(t_{su} + t_{pd}) - \frac{1}{\lambda}\big)(1 + e^{\lambda t_s})(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2}) + (t_s e^{\lambda t_s} + \frac{1}{\lambda})\big(t_{su} - \frac{1}{\lambda} - 2(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})\big), \tag{41}$$

$$w_2 = t_s e^{\lambda t_s} + \frac{1}{\lambda} + (1 + e^{\lambda t_s})(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2}), \tag{42}$$

$$w_3 = -\big(\frac{1}{2}(t_{su} + t_{pd}) - \frac{1}{\lambda}\big)(1 + e^{\lambda t_s})(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2}) - (t_s e^{\lambda t_s} + \frac{1}{\lambda})\big(t_{su} - \frac{1}{\lambda}\big). \tag{43}$$

In the Appendix, we further analyze the expression in Eq. (37) in detail.

C. Optimal Parameter Values

The power consumption over the boundary curve in Eq. (37) depends on the packet arrival rate λ. Furthermore, as it is shown in the Appendix, $\overline{\mathrm{P}}_b(t_w)$ behaves differently for different ranges of λ. For this purpose, $\frac{d\overline{\mathrm{P}}_b(t_w)}{dt_w}$ is calculated (a detailed analysis is provided in the Appendix). Briefly, its sign for different ranges of λ within the feasible region of the wake-up cycle (i.e., $t_{w_b} \leq t_w$) can be expressed as follows

$$\mathrm{sgn}\left(\frac{d\overline{\mathrm{P}}_b(t_w)}{dt_w}\right) = \begin{cases} 1 & \text{for } 0 < \lambda \leq \lambda_t \ , \\ -1 & \text{for } \lambda_t < \lambda < 1 \ , \end{cases} \tag{44}$$

where sgn(.) refers to the sign function, and λ_t is referred to as the turnoff packet arrival rate. The turnoff packet arrival rate can be calculated using any typical root-finding algorithm that meets $F_1 = 0$ (see details in Appendix) where

$$F_1 = (1 + e^{\lambda t_s})(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})(t_s e^{\lambda t_s} + \frac{1}{\lambda}) \times \tag{45}$$

$$\big(\frac{1}{2}(t_{pd} - \phi t_{su}) + \frac{1}{\lambda} + 2(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})\big) - \tag{46}$$

$$\big(\frac{1}{2}(\phi t_{su} + t_{pd}) - \frac{1}{\lambda}\big)(1 + e^{\lambda t_s})^2(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})^2. \tag{47}$$

Theorem 1. *$t_w^* = t_{w_b}$ and $t_i^* = 1$ are the optimal parameter values of the optimization problem in (27)-(29) for the range $0 < \lambda \leq \lambda_t$.*

Proof. As it can be seen in (44), for all $0 < \lambda \leq \lambda_t$, the power consumption increases when increasing t_w over the boundary, so that the minimum power consumption is achieved at the minimum feasible t_w, i.e., $t_w = t_{w_b}$. Correspondingly, the optimal value of t_i can be calculated by substituting t_{w_b} into Eq. (31), which leads to $t_i = 1$. Therefore, for all $0 < \lambda \leq \lambda_t$, $t_w^* = t_{w_b}$ and $t_i^* = 1$ is the optimal solution of the optimization problem in (27)-(29). □

Theorem 2. *$t_w^* = +\infty$ and $t_i^* = +\infty$ are the optimal parameter values of the optimization problem in (27)-(29) for the range $\lambda_t < \lambda < 1$.*

Proof. As it can be seen in (44), for all $\lambda_t < \lambda < 1$, the power consumption decreases when increasing t_w over the boundary, so that the minimum power consumption is achieved at the maximum feasible t_w, i.e., $t_w = +\infty$. Correspondingly, the optimal value of t_i can be calculated by substituting t_w into Eq. (31), which is $t_i = +\infty$. Therefore, for all $\lambda_t < \lambda < 1$, $t_w^* = +\infty$ and $t_i^* = +\infty$ is the optimal solution of the optimization problem in (27)-(29). □

Corollary 1. *For the range $\lambda_t < \lambda < 1$, the optimal solution is equivalent to not utilizing WuS; the system is always at active-decoding and active-inactivity timer states (P_1+$P_4 \approx 0$). Hence, if the energy and delay overhead of switching on/off the BBU are taken into account, the WuS is not effective anymore for high λ values. Instead, other power saving mechanisms, such as DRX, microsleep, or pre-grant message could be used in this regime.*

Fig. 5 (a) and (b) illustrate how λ_t changes with $\overline{\mathrm{D}}_{\max}$ and $t_{su} + t_{pd}$, respectively. As it can be observed in Fig. 5 (a), the turnoff packet arrival rate is independent and insensitive to variations of the value of $\overline{\mathrm{D}}_{\max}$, however, it reduces, when $t_{su}+t_{pd}$ becomes larger (see Fig. 5 (b)). Therefore, in order to decide whether to enable WuS or not, regardless of the QoS requirement of the considered traffic, the network needs to compare the estimated packet arrival rate with pre-calculated and fixed λ_t.

Interestingly, for $\lambda_t < \lambda < 1$, the power consumption reduces by increasing t_w towards infinity (and correspondingly, t_i increases), while the delay constraint is satisfied. This can be interpreted in a way that for packet arrival rates higher than λ_t, the WuS is not effective anymore and only adds overhead energy consumption, thus implying that it is better not to switch off the BBU and to utilize short DRX cycles. As it is shown in Fig. 5 (b), when $t_{su} + t_{pd}$ becomes larger, turnoff

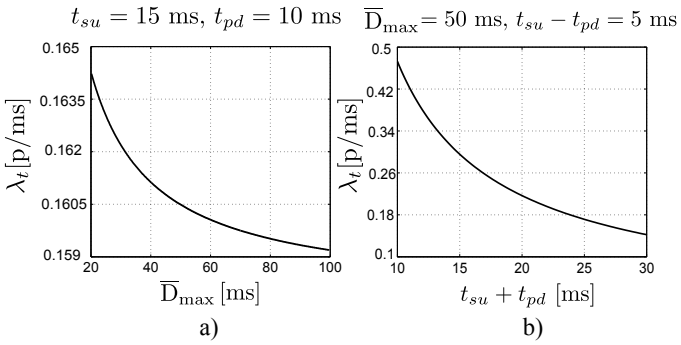

Figure 5: λ_t as function of delay bound and sum of transition times.

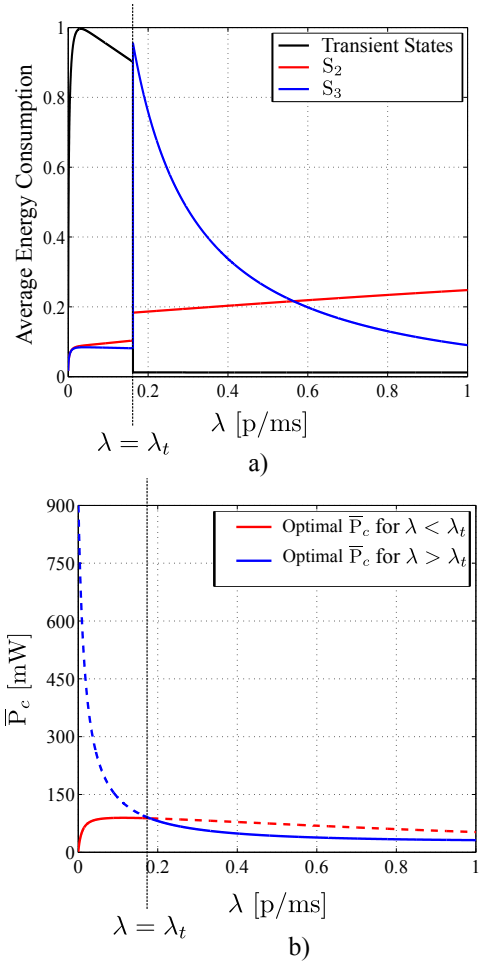

Figure 6: a) Normalized average energy consumption of each state, as result of optimal configuration of wake-up scheme parameters; b) average overall power consumption of optimized wake-up scheme.

packet arrival rates become smaller, which justifies the fact that for the higher packet arrival rates, the frequent start-up and power-down related energy consumption becomes larger.

Additionally, the main reason for such interpretation is due to the fact that for large wake-up cycles, P_{12} approaches to one, which is equivalent to the reduction of number of potential scheduled PDCCHs, and hence there is no gain by using the wake-up scheme over DRX anymore. Furthermore, for higher packet arrival rates, based on the optimal policy, most of the time the BBU is at either S_2 or S_3 ($\mathrm{P}_1 = \mathrm{P}_4 \approx 0$), to avoid wasted energy of start-up and power-down times, and to satisfy the delay constraint (illustrated in Fig. 6 (a)). As it can be seen in Fig. 6 (a), for small packet arrival rates, there is considerable energy consumption for transition between states, however once the packet arrival rate is higher than λ_t, this trend changes and the UE is mainly at S_2 or S_3, and it does not waste energy in start-up/power-down stages, due to the need for frequent start-up/power-down of the BBU. Such change in behaviour of the wake-up scheme can be explained by the objective of the system which is to reduce the overall power consumption, as it shown in Fig. 6 (b). Moreover, as it can be seen in Fig. 6 (a), for packet arrival rates higher than the turnoff packet arrival rate, most of the energy is consumed for decoding of the packets, and its energy consumption increases linearly with the packet arrival rate.

Finally, it can be shown that for λ less than the turnoff packet arrival rate ($0 < \lambda \leq \lambda_t$), the optimal parameter values (t_w^* and t_i^*) of the original MINLP (24)-(26) can be written based on the optimal values of the equivalent relaxed problem (27)-(29) as follows

$$t_w^* = \lfloor t_{w_b} \rfloor \text{ and } t_i^* = 1, \tag{48}$$

where $\lfloor\ \rfloor$ refers to the floor function.

Theorem 3. *$t_w^* = \lfloor t_{w_b} \rfloor$ and $t_i^* = 1$ are the optimal parameter values of the MINLP in* (24)-(26) *for the range $0 < \lambda \leq \lambda_t$.*

Proof. If f is a function of continuous variables x and y, it can easily be shown that

$$\begin{aligned} &\text{if } \frac{\partial f}{\partial x} > 0 \Longrightarrow \Delta f_x > 0,\ \forall x \in \mathbb{R}^+ \\ &\text{if } \frac{\partial f}{\partial x} < 0 \Longrightarrow \Delta f_x < 0,\ \forall x \in \mathbb{R}^+ \end{aligned} \tag{49}$$

where

$$\Delta f_x \triangleq f(\lfloor x \rfloor + 1) - f(\lfloor x \rfloor). \tag{50}$$

Therefore, by assuming f as representative of either $\overline{\mathrm{P}}_c(t_w, t_i)$ or $\overline{\mathrm{D}}(t_w, t_i)$ and x as of either t_i or t_w, one can prove, similarly to the case with continuous variables (proved in Section IV-A), that the optimal parameters of MINLP (24)-(26) are laid over the boundary. Therefore, the boundary of (24)-(26) consists of all combinations of $(\lfloor t_w \rfloor, t_i)$ for which $t_i \in \{1, 2, ...\}$ and $\overline{\mathrm{D}}(t_w, t_i) = \overline{\mathrm{D}}_{\max}$ (as formulated in (31)). Similarly, based on (49) and (32), as well as the properties of the floor function, we can state that increasing t_i over the boundary may increase $\lfloor t_w \rfloor$ ($\Delta \lfloor t_w \rfloor_{t_i} > 0$) or $\lfloor t_w \rfloor$ can remain in its previous value ($\Delta \lfloor t_w \rfloor_{t_i} = 0$). Furthermore, based on (44), for $0 < \lambda < \lambda_t$, we can conclude that $\Delta \overline{\mathrm{P}}_b(t_w, t_i)_{\ t_w} > 0$. Therefore, t_w^* is the smallest feasible value of the wake-up cycle over the boundary, i.e., $t_w^* = \lfloor t_{w_b} \rfloor$. However, $\lfloor t_{w_b} \rfloor$ may correspond to either $t_i = 1$ or larger values at the same time. Since the power consumption has the lowest value at the lowest t_i, for fixed t_w, we can conclude that $t_i^* = 1$. □

Corollary 2. *It turns out that the solution of the relaxed problem in* (27)-(29)*, after finding the integer part of one of*

Table II: Average transitional time and representative power consumption values of LTE-based cellular module during short and long DRX when carrier bandwidth is 20 MHz, and $\phi = 1.1$

DRX Cycle	PW_{sleep}	$\text{PW}_{\text{active}}$	$\text{PW}_{\text{decode}}$	t_{su}	t_{pd}
short	395 mW	850 mW	935 mW	1 ms	1 ms
long	$\approx$ 0 mW	850 mW	935 mW	15 ms	10 ms

Table III: Assumed power consumption parameters of the wake-up scheme with assumption of $\phi = 1.1$

PW_1	PW_2	PW_3	PW_4	t_{su}	t_{pd}	t_{on}
57mW	935 mW	850 mW	$\approx$ 0 mW	15 ms	10 ms	1/14 ms

the two optimization parameters' values, are actual optimal values for the original MINLP optimization problem in (24)-(26). Therefore, our relaxation approach yielded an equivalent reformulation.

V. Numerical Results

In this section, a set of numerical results are provided in order to validate our concept and the analytical results, as well as to show and compare the average power consumption of the optimized WuS over DRX for packet arrival rates less than the turnoff packet arrival rate. Power consumption of the mobile device in different operating states is highly dependent on the implementation, and also its operational configurations. Therefore, for the numerical results, the power consumption model used in [20], [33], [36], [37] is employed. Its parameters for DRX and WuS are shown in Table II and Table III, respectively. LTE-based power consumption values, shown in Table II, are considered as a practical example since those of the emerging NR modems are not publicly available yet. For simulations, we use $\phi = 1.1$ as an example numerical value, while methodology wise, numerical results can also be generated for any other value as well.

Two different sets of performance results, in terms of power consumption and delay, are presented based on the optimal configuration of the wake-up parameters (48). Namely, a) with simplified assumptions of zero false alarm/misdetection rates, and $t_{on} \approx 0$ ms, equivalent to analytical results (ana.), and b) with the realistic assumptions of $P_{fa} = 10\%$, $P_{md} = 1\%$, $t_{on} = 1/14$ ms obtained by simulations and [20], referred to as simulation results (sim.).

Table IV shows the optimal resulting values of t_w^* in (48) for different values of λ and $\overline{\text{D}}_{\max}$. As it can be observed, for tight delay requirements ($\overline{\text{D}}_{\max} = 30$ ms), t_w^* tends to be small, enabling the UE to reduce the duration of packet buffering. Interestingly, for mid range of packet arrival rates ($\lambda = 0.1$ p/ms), optimal wake-up cycle for a given delay bound is shorter than for both lower and higher packet arrival rates. The justification is as follows. For higher packet arrival rates, t_w^* becomes larger, the reason being that the inactivity timer is ON most of the time. Therefore, the need for smaller wake-up cycles decreases and correspondingly higher energy overhead is induced. For lower packet arrival rates, in turn, the value of t_w^* is higher due to the infrequent packet arrivals, hence achieving a smaller delay.

Fig. 7 illustrates the power consumption of the proposed WuS under ideal and realistic assumptions as a function of

Table IV: Optimal values of wake-up cycle under different delay requirements and packet arrival rates ($t_i^* = 1$ ms)

λ [p/ms]	0.01			0.08			0.15		
$\overline{\text{D}}_{\max}$[ms]	30	75	500	30	75	500	30	75	500
t_w^* [ms]	180	380	2099	124	315	2124	125	328	2246

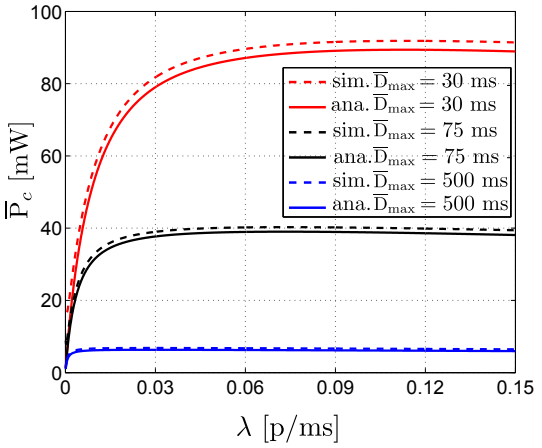

Figure 7: Average power consumption of the optimized wake-up scheme, under ideal and realistic assumptions as a function of packet arrival rate and delay bound.

the packet arrival rate ($\lambda < \lambda_t$), for different maximum tolerable delays. As it can be observed, for both analytical and simulation results, and for all delay bounds, the average power consumption initially increases, while then remains almost constant (especially for large delay bounds) as λ increases, due to the configuration of shorter wake-up cycles for mid packet arrival rates (see Table IV). Moreover, the UE consumes higher power in order to satisfy tighter delay requirements, which, as shown in Table IV, can be translated into shorter t_w^*. Furthermore, Fig. 7 shows that the simulation results closely follow the analytical results, the non-zero gap being due to the non-zero false alarm and misdetection rates. The relative gap between simulation-based and analytical results is somewhat larger for shorter delay bounds, which stems from the correspondingly higher number of wake-up instances.

Moreover, Fig. 8 depicts the average packet delay experienced by the WRx-enabled UE under ideal and realistic assumptions when packet arrival rates vary. As it can be observed, the analytical delay based on the optimal parameter configuration (48) is slightly shorter than the maximum tolerable delay. This is because of selecting the greatest integer less than or equal to the optimal wake-up cycle of the relaxed optimization problem. However, the actual average delay is slightly higher than the analytical average delay, especially for high delay bounds. The main reason for such negligible excess delay is the unavoidable misdetections, whose impact is more clear for large wake-up cycles corresponding to high delay bounds. In practice, to compensate for such small excess delay, the delay bound can be set slightly smaller than the actual average delay requirement.

Finally, for comparison purposes, the relative power saving of WuS over DRX representing the amount of power that can be saved with WuS as compared to the DRX-based reference

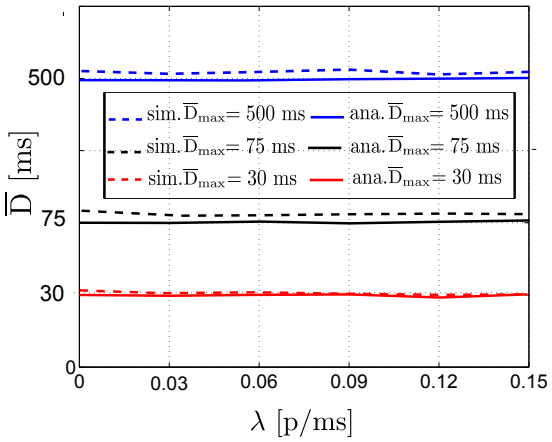

Figure 8: Average buffering delay of the optimized wake-up scheme, under ideal and realistic assumptions as a function of packet arrival rate and delay bound. For better visualization, the y-axis is deliberately not in linear scale.

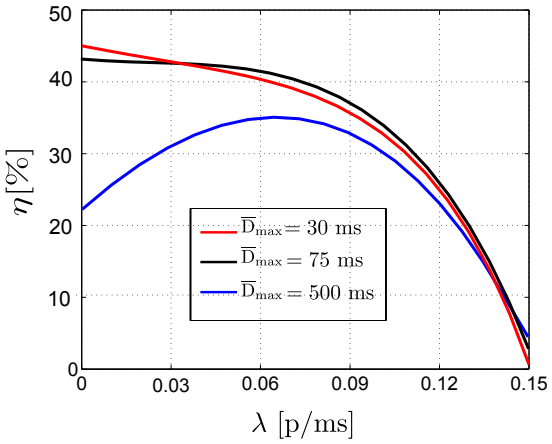

Figure 9: Achieved relative power saving values (η) of the proposed optimized wake-up scheme as a function of packet arrival rate and delay bound.

system is utilized, assuming the same delay constraints in both methods. The value of the relative power saving ranges from 0 to 100%, and a large value indicates that the WuS conserves energy better than the DRX. Formally, we express the relative power saving (η) as

$$\eta = \frac{\overline{\mathrm{P}}_{\mathrm{DRX}} - \overline{\mathrm{P}}_c}{\overline{\mathrm{P}}_{\mathrm{DRX}}} \times 100, \tag{51}$$

where $\overline{\mathrm{P}}_{\mathrm{DRX}}$ refers to average power consumption of DRX. Furthermore, for a fair comparison, we consider an exhaustive search over a large parameter set of DRX configuration, developed by authors in [17]. However, in order to take start-up and power-down power consumption into account, the solution in [17] is slightly modified to account for the transitory states.

Fig. 9 shows the power saving results. It is observed that the proposed WuS, under realistic assumptions, outperforms DRX within the range $\lambda < \lambda_t$, especially for low packet arrival rates with tight delay requirements. The main reason is that in such scenarios, DRX-based device needs to decode the control channel very often, residing mainly in short DRX cycles, which causes extra power consumption. The WRx, in turn, needs to decode the wake-up signaling frequently, but with lower power overhead. Additionally, as expected, regardless of the delay requirements, for higher packet arrival rates, DRX infers relatively similar power consumption to the WuS. The reason is that in such cases, the DRX parameters can be configured in such a way that there is a low amount of unscheduled DRX cycles, either by utilizing short DRX cycles for very tight delay bounds or by employing long DRX cycles for large delay requirements. Overall, the results in Fig. 9 clearly demonstrate that WuS can provide substantial energy-efficiency improvements compared to DRX, with the maximum energy-savings being in the order of 40%.

VI. Discussions and Final Remarks

In this section, three interesting remarks are drawn and discussed.

Remark 1: The proposed WuS is fully independent of DRX, which means that both methods can co-exist, interact and be used together to reduce energy consumption of the UE even further. Based on the numerical results provided in this paper, our opinion regarding the power saving mechanisms for moderate and generic mobile users is that there is no 'One-Size-Fits-All Solution', unless the UE is well-defined and narrowed to a specific application and QoS requirement. For a broad range of applications and QoS requirements, there is a need for combining and utilizing different power saving mechanisms, and selecting the method that fits best for particular circumstances. For example, the WuS can be utilized for packet arrival rates lower than the turnoff packet arrival rate; for higher packet arrival rates or shorter delay bounds (e.g., smaller than 30 ms), DRX may eventually be the preferable method of choice; for ultra-low latency requirements, other power saving mechanisms may be needed, building on, e.g., microsleep [6] or pre-grant message [38], [39] concepts. Further, depending on whether an RRC context is established or not, the WuS is agnostic to the RRC states, and can be adopted for idle (delay bound is in range of some hundreds of milliseconds), inactive, and connected modes.

Remark 2: As mentioned in Section III, a latency-optimized frame structure with flexible numerology is adopted in 5G NR, for which the slot length scales down when the numerology increases [2]. In this work, different time intervals within the WuS are defined as multiples of a time unit of a TTI with a duration of 1 ms. As shown before, the minimum power consumption over the boundary is limited by the minimum feasible value of t_i. Therefore, if the TTI can be selected even smaller, i.e., with finer granularity, the optimal power consumption can be further reduced. In this line, Table V presents the power consumption with different TTI sizes (corresponding to NR numerologies 0, 1, 2, and 3 [10]), delay bounds, and packet arrival rates.

Interestingly, Table V shows how the 5G NR numerologies facilitate the use of WuS and improve the applicability and energy saving potential of WuS compared to longer TTI sizes. Besides the smaller t_i sizes, with smaller TTIs, the corresponding optimal wake-up cycles are more fine-grained. With shorter TTI sizes down to 125 μs, the proposed WuS can provide up to 12% additional energy savings compared to the baseline 1 ms TTI. Therefore, on average, the benefit

Table V: Minimum power consumption values [mW] as function of TTI size, delay bounds and packet arrival rates

λ [p/ms]	0.01			0.08			0.15		
$\overline{\mathrm{D}}_{\max}$[ms]	30	75	500	30	75	500	30	75	500
$\overline{\mathrm{P}}_c$ @ TTI $= 1$ ms	54.2	31.6	6.6	88.7	39	6.1	88.9	38.1	5.9
$\overline{\mathrm{P}}_c$ @ TTI $= 500$ μs	50.4	29.4	5.7	83.7	36.8	5.8	85.4	36.6	5.7
$\overline{\mathrm{P}}_c$ @ TTI $= 250$ μs	48.3	28.1	5.5	81.1	35.7	5.7	83.7	35.9	5.6
$\overline{\mathrm{P}}_c$ @ TTI $= 125$ μs	47.5	27.7	5.4	80.1	35.3	5.6	83.1	35.7	5.6

of the flexible NR frame structure is not only for low latency communication but it can also offer energy savings depending on the traffic arrival rates and delay constraints.

Remark 3: The traffic model assumed in this article is basically well-suited for the periodic nature of DRX. For instance, voice calls and video streaming have such periodic behaviour. However, in other application areas such as MTCs, where sensors can be aperiodically polled by either a user or a machine, the traffic will have more non-periodic patterns. In such case, the DRX may not fit well, while the WuS has more suitable characteristics, being more robust and agnostic to the traffic type.

VII. Conclusions and Future Work

In this article, wake-up based downlink access under delay constraints was studied in the context of 5G NR networks, with particular focus on energy-efficiency optimization. It was shown that the performance of the wake-up scheme is governed by a set of two parameters that interact with each other in an intricate manner. To find the optimal wake-up parameters configuration, and thus to take full advantage of the power saving capabilities of the wake-up scheme, a constrained optimization problem was formulated, together with the corresponding closed-form solution. Analytical and simulation results showed that the proposed scheme is an efficient approach to reduce the device energy consumption, while ensuring a predictable and consistent latency. The numerical results also showed that the optimized wake-up system outperforms the corresponding optimized DRX-based reference system in power efficiency. Furthermore, the range of packet arrival rates within which the WuS works efficiently was established, while outside that range other power saving mechanisms, such as DRX or microsleep, can be used.

Future work includes extending the proposed framework to bidirectional communication scenarios with the corresponding downlink and uplink traffic patterns and the associated QoS requirements, as well as to consider other realistic assumptions that impact the energy-delay trade-offs, such as the communication rate and scheduling delays. Additionally, an interesting aspect is to investigate how to configure the wake-up scheme parameters for application-specific traffic scenarios, such as virtual and augmented reality, when both uplink and downlink traffics are considered. Finally, focus can also be given to optimizing the wake-up scheme parameters based on the proposed framework, by utilizing not only traffic statistics but also short-term traffic pattern prediction by means of modern machine learning methods.

Appendix: Analysis of $\overline{\mathrm{P}}_b(t_w)$

In order to find the optimal value of t_w and correspondingly t_i, the derivation of $\overline{\mathrm{P}}_b(t_w)$ with respect to t_w is given as follows

$$\frac{d\overline{\mathrm{P}}_b(t_w)}{dt_w} = \mathrm{PW}_3 \frac{Y(t_w)}{\big(w_1 + w_2 t_w + w_3 e^{-\lambda t_w}\big)^2}, \tag{52}$$

where

$$Y(t_w) = F_1 + (F_2 - \lambda F_3 t_w) e^{-\lambda t_w}, \tag{53}$$

$$\begin{aligned} F_1 = (1 + e^{\lambda t_s})(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})(t_s e^{\lambda t_s} + \frac{1}{\lambda}) \times \\ \big(\frac{1}{2}(t_{pd} - \phi t_{su}) + \frac{1}{\lambda} + 2(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})\big) - \\ \big(\frac{1}{2}(\phi t_{su} + t_{pd}) - \frac{1}{\lambda}\big)(1 + e^{\lambda t_s})^2 (\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})^2, \end{aligned} \tag{54}$$

$$\begin{aligned} F_2 = -F_1 + (1 + e^{\lambda t_s})(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})^2 \times \\ (t_s e^{\lambda t_s} + \frac{1}{\lambda})\big(-2 + \lambda(2 - \phi) t_{su} + \lambda t_{pd}\big), \end{aligned} \tag{55}$$

$$F_3 = F_1 - 2(1 + e^{\lambda t_s})(\overline{\mathrm{D}}_{\max} - \frac{t_s}{2})^2 (t_s e^{\lambda t_s} + \frac{1}{\lambda}). \tag{56}$$

Based on typical values of $1 \leq \phi < 2$, the condition $0 \leq (2 - \phi) t_{su} + t_{pd}$ is met, and hence based on (54)-(56), we can conclude that $F_1 + F_2 > 0$, $F_2 + F_3 > 0$ and $F_1 > F_3$. Furthermore, it can be shown that F_1 is a decreasing function of λ, which has a single root. We refer to its root as λ_t; where for all $\lambda < \lambda_t$, then $F_1 > 0$ while, for all $\lambda > \lambda_t$, then $F_1 < 0$. Root-finding algorithms can be utilized to find λ_t as the λ value that meets $F_1 = 0$.

To determine whether $\frac{d\overline{\mathrm{P}}_b(t_w)}{dt_w}$ is positive or negative, $Y(t_w)$ needs to be analyzed. By differentiating $Y(t_w)$ with respect to t_w, we obtain

$$\frac{dY(t_w)}{dt_w} = -\lambda(F_2 + F_3 - \lambda F_3 t_w) e^{-\lambda t_w}. \tag{57}$$

Additionally, depending on whether F_3 and F_1 are positive or negative, the $\overline{\mathrm{P}}_b(t_w)$ behaves differently. In order to characterize the behaviour of $\overline{\mathrm{P}}_b(t_w)$, we define three mutually exclusive cases: Case A ($F_3 > 0$), Case B ($F_3 < 0$ and $F_1 > 0$), and Case C ($F_3 < 0$ and $F_1 < 0$). Due to fact that $F_1 > F_3$, in the former case, F_1 is always positive.

Case A ($F_3 > 0$): Based on (57), if F_3 is positive, $Y(t_w)$ is a decreasing function for the range of $t_w < \frac{F_2 + F_3}{\lambda F_3}$ and an increasing function for $t_w > \frac{F_2 + F_3}{\lambda F_3}$. Therefore, $Y(t_w)$ has a minimum point at $t_w = \frac{F_2 + F_3}{\lambda F_3}$, where the $Y(t_w)$ at $t_w = \frac{F_2 + F_3}{\lambda F_3}$ is positive, i.e., $Y(\frac{F_2 + F_3}{\lambda F_3}) = F_1 - F_3 e^{-\lambda t_w} > 0$ (due

to fact that $F_3 < F_1$). As a result, $Y(t_w)$ is always positive and hence $\overline{\mathrm{P}}_b(t_w)$ is a monotonous increasing function.

Based on (57), if F_3 is negative (Case B or Case C), we can conclude that $Y(t_w)$ is a monotonous decreasing function from $F_1 + F_2 > 0$ to F_1.

Case B ($F_3 < 0$ and $F_1 > 0$): In this case, for all values of wake-up cycle, $Y(t_w)$ is positive, and hence $\overline{\mathrm{P}}_b$ is a monotonous increasing function.

Case C ($F_3 < 0$ and $F_1 < 0$): In this case, $Y(t_w)$ for $t_w < t_{ws}$ is positive, and it is negative for $t_w > t_{ws}$; where t_{ws} is a stationary point, i.e., $Y(t_{ws}) = 0$ or equivalently, $\frac{\partial \overline{\mathrm{P}}_b(t_w)}{\partial t_w}|_{t_w = t_{w_s}} = 0$. As a result, $\overline{\mathrm{P}}_b(t_w)$ is an increasing function within $t_w < t_{ws}$, and a decreasing function for $t_w > t_{ws}$. For the typical range of parameters, we have consistently observed through simulations that $t_{w_s} < t_{w_b}$, so that we can conclude that $\overline{\mathrm{P}}_b(t_w)$ is a decreasing function for the feasible range of the wake-up cycle (i.e., $t_w > t_{w_b}$).

To sum up, for $F_1 < 0$, or equivalently, $\lambda_t < \lambda$ (Case C), $\overline{\mathrm{P}}_b(t_w)$ is a monotonous decreasing function while, for $\lambda < \lambda_t$ (Case A or B), $\overline{\mathrm{P}}_b(t_w)$ is a monotonous increasing function. Due to the relevance of λ_t, we refer to it as the turnoff packet arrival rate.

REFERENCES

[1] S. Rostami, S. Lagen, M. Costa, P. Dini, and M. Valkama, "Optimized wake-up scheme with bounded delay for energy-efficient MTC," in *IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec. 2019.

[2] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "NR: The new 5G radio access technology," *IEEE Communications Standards Magazine*, vol. 1, pp. 24–30, Dec 2017.

[3] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, pp. 74–80, February 2014.

[4] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Communications Magazine*, vol. 49, pp. 55–62, August 2011.

[5] "Designing mobile devices for low power and thermal efficiency," tech. rep., Qualcomm Technologies, Inc., Oct. 2013.

[6] M. Lauridsen, *Studies on Mobile Terminal Energy Consumption for LTE and Future 5G*. PhD thesis, Aalborg University, Jan. 2015.

[7] A. Carroll and G. Heiser, "An analysis of power consumption in a smartphone," in *Proc. USENIXATC2010*, (Berkeley, CA, USA), pp. 21–21, USENIX Association, 2010.

[8] M. Lauridsen, P. Mogensen, and T. B. Sorensen, "Estimation of a 10 Gb/s 5G receiver's performance and power evolution towards 2030," in *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, pp. 1–5, Sept 2015.

[9] "IMT traffic estimates for the years 2020 to 2030," tech. rep., ITU-R, July 2015.

[10] "TS 38.300, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NR; NR and NG-RAN overall description," tech. rep., 3GPP, Jan. 2019.

[11] S. P. Erik Dahlman and J. Skold, *4G LTE/LTE-Advanced for Mobile Broadband*. Academic Press, 2011.

[12] "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," tech. rep., 3GPP TS 36.213 version 10.1.0 Release 10, Apr 2010.

[13] "5G;NR; User Equipment (UE) procedures in idle mode and in RRC inactive state," tech. rep., 3GPP TS 38.304 version 15.1.0 Release 15, Oct. 2018.

[14] E. Liu, J. Zhang, and W. Ren, "Adaptive and autonomous power-saving scheme for beyond 3G user equipment," *IET Communications*, vol. 7, pp. 602–610, May 2013.

[15] A. T. Koc, S. C. Jha, R. Vannithamby, and M. Torlak, "Device power saving and latency optimization in LTE-A networks through DRX configuration," *IEEE Transactions on Wireless Communications*, vol. 13, pp. 2614–2625, May 2014.

[16] Y. Y. Mihov, K. M. Kassev, and B. P. Tsankov, "Analysis and performance evaluation of the DRX mechanism for power saving in LTE," in *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, pp. 000520–000524, Nov 2010.

[17] H. Ramazanali and A. Vinel, "Tuning of LTE/LTE-A DRX parameters," in *Proc. IEEE CAMAD 2016*, pp. 95–100, Oct 2016.

[18] I. Demirkol, C. Ersoy, and E. Onur, "Wake-up receivers for wireless sensor networks: benefits and challenges," *IEEE Wireless Communications*, vol. 16, pp. 88–96, Aug 2009.

[19] N. S. Mazloum and O. Edfors, "Performance analysis and energy optimization of wake-up receiver schemes for wireless low-power applications," *IEEE Transactions on Wireless Communications*, vol. 13, pp. 7050–7061, Dec 2014.

[20] S. Rostami, K. Heiska, O. Puchko, J. Talvitie, K. Leppanen, and M. Valkama, "Novel wake-up signaling for enhanced energy-efficiency of 5G and beyond mobile devices," in *2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–7, Dec 2018.

[21] M. Lauridsen, G. Berardinelli, F. M. L. Tavares, F. Frederiksen, and P. Mogensen, "Sleep modes for enhanced battery life of 5G mobile terminals," in *Proc. IEEE VTC 2016 Spring*, pp. 1–6, May 2016.

[22] "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); overall description," tech. rep., 3GPP, TS 36.300, Mar. 2019.

[23] "New SID: Study on UE power saving in NR," tech. rep., 3GPP, RP-181463, June. 2018.

[24] S. Rostami, K. Heiska, O. Puchko, K. Leppanen, and M. Valkama, "Wireless powered wake-up receiver for ultra-low-power devices," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–5, April 2018.

[25] S. Tang, H. Yomo, Y. Kondo, and S. Obana, "Wakeup receiver for radio-on-demand wireless lans," in *2011 IEEE Global Telecommunications Conference (GLOBECOM)*, pp. 1–6, Dec 2011.

[26] L. R. Wilhelmsson, M. M. Lopez, S. Mattisson, and T. Nilsson, "Spectrum efficient support of wake-up receivers by using (O)FDMA," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, April 2018.

[27] N. Kouzayha, Z. Dawy, J. G. Andrews, and H. ElSawy, "Joint downlink/uplink RF wake-up solution for IoT over cellular networks," *IEEE Transactions on Wireless Communications*, vol. 17, pp. 1574–1588, March 2018.

[28] J. Oller, E. Garcia, E. Lopez, I. Demirkol, J. Casademont, J. Paradells, U. Gamm, and L. Reindl, "IEEE 802.11-enabled wake-up radio system: design and performance evaluation," *Electronics Letters*, vol. 50, pp. 1484–1486, Sep. 2014.

[29] F. Ait Aoudia, M. Magno, M. Gautier, O. Berder, and L. Benini, "Analytical and experimental evaluation of wake-up receivers based protocols," in *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–7, Dec 2016.

[30] R. P. K. Ponna and D. Ray, "Saving energy in cellular IoT using low-power wake-up radios," 2018. Student Paper, Lund University.

[31] "Low-power wake-up receiver for 802.11," tech. rep., IEEE 802.11-15/1307r1.

[32] "Proposal for wake-up receiver (WUR) study group," tech. rep., IEEE 802.11-16/0722r0.

[33] C. C. Tseng, H. C. Wang, F. C. Kuo, K. C. Ting, H. H. Chen, and G. Y. Chen, "Delay and power consumption in LTE/LTE-A DRX mechanism with mixed short and long cycles," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 1721–1734, March 2016.

[34] P. Belotti, C. Kirches, S. Leyffer, J. Linderoth, J. Luedtke, and A. Mahajan, "Mixed-integer nonlinear optimization," *Acta Numerica*, vol. 22, p. 1–131, 2013.

[35] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambertw function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, Dec 1996.

[36] T. Kolding, J. Wigard, and L. Dalsgaard, "Balancing power saving and single user experience with discontinuous reception in lte," in *2008 IEEE International Symposium on Wireless Communication Systems*, pp. 713–717, Oct 2008.

[37] M. Lauridsen, L. Noel, T. Sorensen, and P. Mogensen, "An empirical LTE smartphone power model with a view to energy efficiency evolution," *Intel Technology Journal*, vol. 18, pp. 172–193, 3 2014.

[38] S. Rostami, K. Heiska, O. Puchko, K. Leppanen, and M. Valkama, "Robust pre-grant signaling for energy-efficient 5G and beyond mobile devices," in *2018 IEEE International Conference on Communications (ICC)*, pp. 1–6, May 2018.

[39] S. Rostami, K. Heiska, O. Puchko, K. Leppanen, and M. Valkama, "Pre-grant signaling for energy-efficient 5G and beyond mobile devices: method and analysis," *IEEE Transactions on Green Communications and Networking*, vol. 3, pp. 418–432, June 2019.

Soheil Rostami received the M.Sc. degree in Mobile Communication Systems with distinction from the University of Surrey, UK, in 2011. During 2012–2017, he was with the Faculty of Engineering and Science, University of Greenwich, UK, and with Nokia Bell Labs, Belgium. Since 2017, he has been working as a Researcher with 5G Radio Network Technologies team in Huawei Technologies Oy (Finland) Co., Ltd. In 2018, he was a Visiting Researcher with Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Spain.

Sandra Lagen received the Telecommunications Engineering, M.S., and Ph.D. degrees from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2011, 2013, and 2016, respectively. Her dissertation was awarded by COIT the best national Ph.D. thesis on high-speed broadband mobile communications (2017) and received a Special Doctoral Award by UPC (2019). In 2015, she did a research appointment at Nokia Networks in Aalborg, Denmark. Since 2017, she is a researcher at CTTC. She is recipient of IEEE WCNC 2018 best paper award. Her research interests include wireless communications, spectrum and interference management, and optimization theory.

Mário Costa (S'08–M'13) received the M.Sc. degree (Hons.) in communications engineering from the Universidade do Minho, Portugal, in 2008, and the D.Sc. (Tech.) degree in electrical engineering from Aalto University, Finland, in 2013. In 2014, he was a Visiting Postdoctoral Research Associate with Princeton University. Since 2014, he has been with Huawei Technologies Oy (Finland) Co., Ltd., first as a Senior Researcher and currently as a Principal Data Scientist. His research interests include statistical signal processing, machine learning, and wireless communications.

Mikko Valkama received the M.Sc. (Tech.) and D.Sc. (Tech.) Degrees (both with honors) in electrical engineering (EE) from Tampere University of Technology (TUT), Finland, in 2000 and 2001, respectively. In 2002, he received the Best Doctoral Thesis -award by the Finnish Academy of Science and Letters for his dissertation entitled "Advanced I/Q signal processing for wideband receivers: Models and algorithms". In 2003, he was working as a visiting post-doc research fellow with the Communications Systems and Signal Processing Institute at SDSU, San Diego, CA. Currently, he is a Full Professor and Department Head of Electrical Engineering at newly formed Tampere University (TAU), Finland. His general research interests include radio communications, radio localization, and radio-based sensing, with particular emphasis on 5G and beyond mobile radio networks.

Paolo Dini is currently a Senior Researcher at the Centre Tecnologic de Telecomunicacions de Catalunya (CTTC). He received M.Sc. and Ph.D. degree from Università di Roma "La Sapienza", in 2001 and 2005, respectively. His research interests encompass system modeling and optimization by machine learning, data analytics, and optimal control with particular emphasis on sustainability and energy saving, including energy harvesting networks and smart power grids. He received two awards by the Cisco Silicon Valley Foundation for his research on heterogeneous mobile networks in 2008 and 2011. He is currently the coordinator of the EU H2020 MSCA SCAVENGE European Training Network on sustainable mobile networks with energy harvesting capabilities.