Trading off delay and energy saving through Advanced Sleep Modes in 5G RANs

Trading off delay and energy saving through Advanced Sleep Modes in 5G RANs

Daniela Renga, Zunera Umar, Michela Meo
Department of Electronics and Telecommunications
Politecnico di Torino
Turin, Italy
{firstname.lastname}@polito.it

***Abstract*—While designed for being energy efficient, the deployment of 5G networks will further increase Radio Access Networks (RANs) energy consumption with the twofold effect to raise sustainability issues and increase operational costs for Mobile Network Operators (MNOs). However, the energy waste occurring during low traffic periods can be mitigated through Advanced Sleep Modes (ASMs) that make the BSs enter into progressively deeper and less consuming sleep modes. Deep sleep modes, unfortunately, have longer reactivation times, and may jeopardize service quality.**

In this paper, focusing on 5G latency requirements in low traffic periods, we propose a framework to dynamically adapt the ASM configuration settings to the actual traffic load so as to meet a desired constraint on the average BS reactivation delay.

***Index Terms*—Sleep Modes, 5G networks, Energy Efficiency**

I. Introduction

In recent years, we have witnessed a staggering growth of mobile traffic, posing significant challenges to Mobile Network Operators (MNOs) in terms of network energy demand and operational costs. This trend is bound to be further boosted in the next years. According to the Cisco forecast, almost 30 billion of networked devices are expected by 2023, and global mobile data traffic will reach almost 80 exabytes per month by the same year [1]. In this context, the wide spreading of 5G technology, which enables mobile networks to provide incredibly huge capacity and ubiquitous and fast network access, is further pushing the process of mobile network densification. The enhancements that 5G technology is introducing in terms of higher bandwidth availability and ultra-low latency are facilitating the deployment of massive Internet of Things (IoT) applications, from IoT services focused on environmental monitoring, to smart mobility and autonomous vehicles, machine-to-machine (M2M) communications for factory automation in Industry 4.0, Smart Grid management, Smart Farming, just to cite some examples. By 2023, IoT devices will represent half of all networked devices, and over 10% of devices and connections will have 5G capability [1].

Considering that up to 80% of the total network consumption is accounted for the access segment [2], a rapid growth of energy demand to operate cellular networks is entailed by the boost of traffic volumes. Indeed, MNOs are already facing a substantial rise in the energy bill due to power supply [3]. Among the possible solutions to cope with these challenges, Base Station (BS) Sleep Modes (SMs) can be used to reduce mobile network consumption by deactivating unneeded radio resources during periods in which the traffic demand is low. However, the reactivation time from sleep modes introduces additional undesirable set-up delays and, consequently, the risk to deteriorate quality of service. This has slowed down the adoption of sleep modes. Recently, the introduction of Advanced Sleep Modes (ASMs), which allow to progressively deactivate the BS to deeper sleep modes that correspond to lower power levels but longer reactivation times [4], are making it possible to find convenient trade-offs between energy saving and reactivation delay. Clearly, ASMs can be applied in scenarios, like those expected in 5G deployments, in which the cell layout is dense and a few macro-cells provide full coverage and are always active, while several small cells can be activated on demand, only when needed.

While ASMs are very promising, they are defined through a large set of parameters whose setting is not straightforward and might compromise the effectiveness of the overall approach. To the best of our knowledge, no other work in the literature analyses the effect on the system performance of varying the ASM configuration settings and its potential in energy efficient 5G networks. In this paper, by focusing on BS operation with ASMs, we propose a simple method to optimally set the parameters that govern ASM so that a desired average BS reactivation delay is guaranteed while energy saving is maximized.

A. Related work

Significant potential for enhancing energy efficiency in Base Station communication is found through vastly studied sleep mode based techniques that utilize periods with low traffic load and effectively manage resources for achieving green communication in mobile networking. [5]–[7]. Deeper sleep states monotonically decrease average power consumption. Hence, various sleep policies suggest an effective utilization of low power modes to foster the need for energy efficient communication in mobile base stations [8]. Performance evaluation of strategic and random sleeping policies on a stochastic model is presented for Heterogeneous Cellular Networks (HCN) for QoS considerate optimization of sleep mode strategy with low computational complexity [9]. Furthermore, optimal sleep policy-based wake-up schemes suggest improvement in user-associated base station parameters [10], [11]. Among several

techniques for traffic knowledge aware base station sleep mode optimization [12]–[16], an access algorithm, mobile station round-robin base stations (MSRRBS), is compared with existing search methods in terms of complexity and performance [17].

Mobile cellular technology has been developed to achieve networking goals with increasing data rate requirements, which consequently increase the energy demand. As Renewable Energy sources are penetrated to power BSs, SM-based solutions allow for more efficient management of Renewable Energy generation, which is intermittent and unpredictable. [5], [7], [18]–[21]. The interest in SM based strategies expect to further increase with the widespread diffusion of 5G technologies [6], [22], [23]. Indeed, the 5G paradigm implies a gradual increase in network density, resulting in higher energy bills [23], [24]. Diffusion of sleep modes in 5G base station leads to raise energy gains through increasing SS burst period periodicity [25]. Moreover, the concept of femtocells or micro base stations in 5G brings forth the idea of efficient resource utilization by turning OFF few cells with a low load while distributing the traffic to the neighbouring cells [15], [26], [27]. As a result of the deactivation of some base stations, subscribers benefit from higher data rates due to less interference in the area, improving the overall capacity provided by the femtocell cluster [28].

Considering traffic prediction as a crucial factor that influences the decision of using sleep modes effectively [19], an approach towards performance specified Machine learning (ML) is discussed for base station ON-OFF switching using traffic prediction [12]. Device-to-Device (D2D) incorporation to prevent traffic loss is proposed for Q-learning based sleep-wakeup mechanism, in which QoS is offered through deactivating redundant base stations corresponding to the obtained local traffic profile [13]. Sleep mode techniques show promising outcomes by decreasing the overall consumption in base stations. However, they may raise concerns about the possibility of affecting the Quality of Service due to the time required to turn on BS components in the event of a new service request. This additional delay may limit the use of SMs due to the strict latency constraints for future application requirements, such as massive IoTs, besides the limitations defined in the 5G standard for mobile communication. In [29], authors suggest switching the BS to different sleep levels according to the traffic location and movement would maximize the performance by trading off energy and delays. Similarly, determining the delay distribution under the optimal policy study defines the linear relationship between the two parameters and hence the possibility to trade-off delay and energy efficiency [30], [31]. However, there remains a need for an efficient method to optimize sleep mode policy in BS according to delay constrained application requirement in 5G and imminent 6G technology.

Besides standard ASM operation presented in [4], other existing works from the literature focus on trading-off delay and energy consumption for advanced sleep modes. A distributed Q-learning approach is used to optimize the duration of sleep modes based on traffic prediction, and a weight is assigned to prioritize between latency and energy gains [32], [33]. However, this approach lacks a consistent performance for delay sensitive applications under varying traffic load. A similar approach based on reinforcement learning is used to trade-off energy savings and delays for ASMs [34]. Differently from other works, our approach is characterized by the peculiar capability to achieve the desired performance for delay sensitive 5G applications, still maintaining similar gains in terms of power saving.

In our previous work [35], we suggest an effective utilization of Advanced Sleep Modes (ASMs) to trade-off the delay and power consumption for different 5G scenarios. In particular, we propose Tailored SM operation (TO) to trade-off delays and power consumption by setting the duration of SMs according to scenarios and expected load. The method tests on various 5G scenarios and the results show that the application of ASMs based on TO can be used to dynamically trade-off delays and power consumption. However, what was missing is the integration of an automatic procedure in which the system can tune its parameters according to defined delay specifications for different 5G verticals. This paper extends our previous work to overcome this mentioned gap by proposing the closed-form solution for dynamically adjusting the duration of SMs under a given arrival rate. The main contributions are detailed below.

B. Contributions

In this paper, which is an extension of our previous work [35], we study an optimization method for sleep mode operated 5G BS with delay sensitive applications. Although standard ASMs setting shows good performance in terms of energy saving, it triggers up to almost 5.5 ms delay. According to the 5G standardization and ITU report on 5G performance requirements, the minimum requirement for user plane latency is 1 ms for URLLC and 4 ms for eMBB communication [36], [37]. To overcome the complexity of deciding the best parameter setting for SM operation, we propose a Delay Conservative Advanced Sleep Modes (DCASM) approach for BS operation, a method based on an analytic formulation to derive the proper SM settings and to better trade off energy saving and delay. Therefore:

- We develop a closed-form expression to derive the proper parameter settings for ASM operation under different scenarios and arrival rates, to make the system performance compliant with the required average delay constraints. By being based on analytic formulation, the approach can be easily integrated into 5G BS design while incorporating ASMs.
- Through the derived closed-form mathematical expression, we obtain the BS power consumption under the predefined DCASM settings, that allow to meet the scenario specific delay requirements. In this way, power consumption can be estimated ahead of time based on traffic prediction as well as in real time, and ASM parameters can be tailored to the specific needs of MNOs and vertical industries.
- We analyze the impact on delay due to the errors in real-time traffic predictions, which are needed to adapt ASM parameter setting to the actual traffic variations.

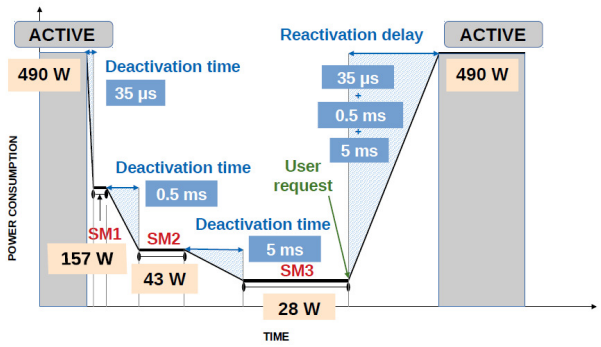


Fig. 1: Standard Advanced Sleep Mode operation.

II. Advanced Sleep Modes

Mobile access networks are often over-provisioned and prone to high intra-day load fluctuations, as well as to long low traffic periods during which a huge amount of energy is lost. The potential for ASMs to save energy is remarkable in such periods. The ASM paradigm was first developed by IMEC [38] within the Earth project [39], and is being integrated into the standardization process of 5G networks [40], [41]. In release 16 of 5G standardization, low power consumption and in-active mode is amalgamated for boost cells other than basic coverage providing cells, which strengthens the concept of sleep mode utilization in 5G base stations [42]. Here, we consider the ASM model stated in [4].

Fig. 1 shows the ASM operation as defined in the standard. A BS enters progressively into deeper SMs during periods of inactivity, with each SM featuring a specific amount of power consumption. Distinct subsets of BS components are gradually deactivated when no data has to be transmitted, starting from components characterized by the shortest activation delay as well as the lowest consumption, such as the Power Amplifier and some processing components. The power consumption of BS components is measured using the IMEC Power tool [43], and three SM levels are envisioned as a result. Therefore, the BS can be in different states that correspond to different power levels:

- **Active state**: the BS is transmitting. We denote the state by A. Transmission can also include signaling, i.e., the BS transmits synchronization and control signals (primary and secondary synchronization signals - PSS and SSS, respectively - and physical broadcast channel signals - PBSH) [44].
- **Idle state**: the BS is active and waiting for traffic to be transmitted.
- **Sleep mode states**: the BS can be in three possible sleep modes, denoted by SMi, with $i = 1, 2, 3$, that are progressively deeper, corresponding to lower consumption and higher reactivation times.
- **Deactivation states**: the BS is in the transition between two sleep modes, it is deactivating from a sleep mode SM$(i-1)$ to the deeper one, namely SMi; the state is denoted by Di. The deactivation from the active or idle state to state SM1 is denoted by $D1$.
- **Reactivation states**: the BS is reactivating from SMi to the active state A; reactivation states are denoted by Ri.

The adopted notation is reported in Table II, together with other notation and symbols used across the paper.

The transition times are the times required to move from state SMi to SM$(i+1)$ for the case of deactivation and to state SM$(i-1)$ for reactivation from current state SMi. The time needed to deactivate the BS from a sleep mode to the deeper one and the time to make the opposite transition from a deeper state to the considered one feature the same duration.[1] The standard defines the transition times and recommends the hold time to be spent in each SM before entering a deeper SM level [4], as reported in Table II. The Table also includes the power consumption values corresponding to each state. When the BS is in the idle state, it enters gradually into deeper SMs according to the pattern shown in Fig. 1. After entering a SM, the standard suggests that the BS spends a hold time in that SM before moving to the deeper SM. For SM3, BS could stay there for an arbitrary period until an request wakes the BS up. Note that in Fig. 1 both power levels and deactivation/reactivation times are not drawn to scale for sake of readability. While the transition times are defined by the internal switching operation and cannot be changed, the hold times spent in each SM can be changed and we act on these times to improve system performance.

When a new signalling or data transmission request arrives to the system and finds the BS in a SM, three different situations might occur:

- The BS is in SMi, and it immediately reverts back to active state to process the request. The request is temporarily buffered during the transition time required to move back to active state. The time to reactivate the BS is given by the sum of the transition times between all the SM states crossed during the transition from the current SMi to the active state.
- The request arrives while the BS is deactivating from SMi to SM$(i+1)$. In this case the deactivation process must be completed before starting reactivation from the state it has just been entered to the fully active state.
- The request arrives while the BS has already started the reactivation procedure due to a previous request. The remaining reactivation time period must elapse before starting the process of new request.

In all the three cases, requests are temporarily stored for the amount of time it takes to fully activate the BS.

The hold time that must be spent in state SMi before moving to next state SM$(i+1)$ is denoted by T_i. Let $T_{d_{i,i+1}}$ be the time required to make the BS enter SM$(i+1)$ from SMi, which coincides with the time needed to move back from SM$(i+1)$ to SMi. We denote by $T_{d_{0,1}}$ the time to move from the active state to SM1 and, similarly, $T_{r_{1,0}}$ is the time to move from SM1 to active. As already mentioned $T_{d_{1,0}} = T_{r_{0,1}}$. The time in the deactivation state Di is hence equal to $T_{d_{i,i+1}}$, while the total time spent in the reactivation state Ri is given by the sum of the times required to reactivate the BS from the current SMi through each intermediate state, i.e. SMi to SM$(i-1)$ up to the active state. For example, the time to reactivate the BS from

[1] There exists also a Sleep Mode 4 that is not considered here, as well as in other related work, since its reactivation time, equal to 0.5 s, is too long to be compliant with 5G latency constraints.

TABLE I: Notation.

Notations	Symbols
Active state	A
Sleep Mode state i	SMi
Deactivation state i, i.e. transition from state SM$i-1$) to state SMi	Di
Reactivation state i, i.e. transition from state SMi to state A	Ri
Hold time to be spent in SMi before entering state $SM(i+1)$	T_i
Time spent in deactivation state D$(i+1)$	$T_{d_{i,i+1}}$
Reactivation time from SMi to SM$(i-1)$	$T_{r_{i,i-1}}$
Time spent in reactivation state Ri	Tr_i
Remaining transient time	r
Average arrival rate	λ
Average inter-arrival time	T_{ia}
Average service rate	μ
Average delay	$\overline{d}$
Target delay	d^*
Base station Utilization	U
Average BS Power Consumption	PC
Average BS Power consumption during SM states and transition states	P_{SM}
BS Power consumption during Active state	P_A
Average Baseline Power consumption with BS always active	P_B
BS Power consumption in SMi	P_i
BS Power consumption in transient states Di	$P_{i,i+1}$
BS Power consumption while in reactivating from state SMi to SM$(i-1)$	$P_{i,i-1}$

TABLE II: Power levels, deactivation time to/reactivation time from a SM, hold time spent in a SM according to the standard ASM operation.

	Active	Idle	SM1	SM2	SM3
Power level	490 W	328 W	157 W	42.9 W	28.5 W
De- or re-activation	-	-	35.5 µs	0.5 ms	5 ms
Hold time	-	-	70 µs	1 ms	-

SM2 is given by $Tr_2 = T_{r_{2,1}} + T_{r_{1,0}}$. Note that no constraint is placed on the hold time to be spent in any SMi during the BS reactivation: as soon as an arrival occurs during any SMi, the BS can be immediately reactivated.

During an activation/deactivation slope, the power consumption corresponds to the power consumption of the level from which the BS is awaking or from which it is deactivating, respectively.

III. Delay Conservative ASM Operation

We propose a novel operational mode, called Delay Conservative ASM Operation (DCASM), that, by properly setting the parameters of the ASM operation, minimizes energy consumption while meeting delay constraints, hence targeting delay conservative 5G applications. The delay constraint is defined in terms of *the average delay experienced by a request that arrives when the BS is not active and has to wait for the BS to fully reactivate from a SM*.

To find the optimum configurations, a mathematical framework is developed. In particular, the setting concerns the value of T_2; indeed, T_1 is set to the same value adopted under the Standard ASM Operation, i.e., T_1=70 µs, since the reactivation delay from SM1 is negligible, and does not significantly affect the average reactivation delay.

In order to guarantee a desired average reactivation delay, we use T_2 as a control knob to tune the frequency with which, at its arrival, a request finds the BS in SM3, a state from which the reactivation delay is large. If T_2 is large, most of the arrivals find the BS in SM2 and a short average reactivation delay is experienced. Conversely, if T_2 is small, the BS can often enter SM3 from which reactivation delay is large.

The complete process of DCASM is depicted in the flowchart diagram reported in Fig. 2. Depending on the considered scenario, a specific target is set for the maximum allowed average delay, let it be denoted by d^*. The arrival rate is estimated and fed as input to the framework which provides the optimal setting of T_2 such that the delay requirement is met while the power consumption is minimized, under the estimated value of the traffic.

We now present the formulation for setting the parameters in DCASM so as to meet a given delay requirement under given conditions of traffic.

A. Request Arrivals

We assume that new requests arrive according to a Poisson process. This assumption is reasonable due to the potentially large population of users which makes uncorrelated their behaviour. Since the process is Poisson, the inter-arrival time T_{ia} is a random variable with exponential distribution with parameter λ such that $\lambda = 1/\overline{T}_{ia}$ where $\overline{T}_{ia}$ denotes the mean inter-arrival time.

Arrivals can be categorized based on the activation delay that a new request would incur into and we consider only arrivals which trigger the BS reactivation, since at low load the probability to have an arrival in the short time of a reactivation is small and, in this case, the experienced delay is only a residual reactivation time. Arrivals occurring during sleep states SMi, with $i = 1,2,3$ experience a delay that is given by the reactivation time from SMi. Arrivals occurring during deactivation states Di have to wait for the deactivation to be completed followed by a reactivation.

Fig. 3a shows the probability density function of inter-arrival times. On the x-axis it is reported the sequence of the BS states visited during inactivity periods. In particular, the labels D_1, D_2, D_3 on the x-axis identify those ranges of inter-arrival times that lead to arrivals occurring during the transition

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3248291

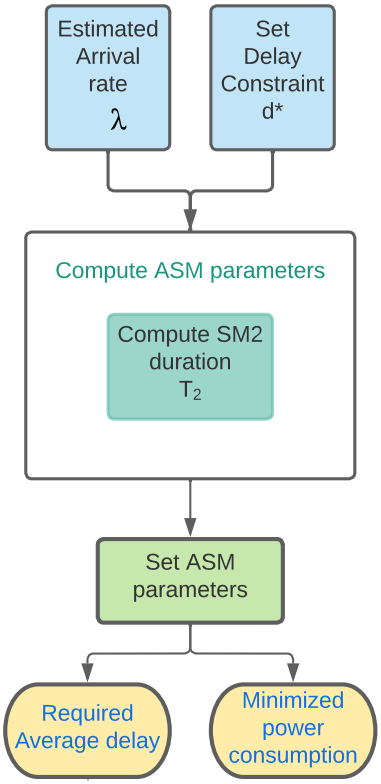


Fig. 2: DCASM operation flowchart.

towards a deeper SM, i.e., during the deactivation states Di. Conversely, the inter-arrival time corresponding to S_1, S_2, S_3 determine arrivals occurring during SMi. Depending on the state being visited when an arrival occurs, the experienced delay is different. Therefore, to define the probabilities of arrival in each state we present a Cumulative Density Function (CDF) of the inter-arrival times in Fig. 3b. The upper extremes of each interval is denoted as t_{D_1}, t_{S_1}, t_{D_2}, t_{S_2}, t_{D_3}, which can be found through accumulating the time from starting of deactivation from active state till the corresponding state. They are defined as follows:

$$\begin{aligned} t_{D_1} &= T_{d_{0,1}} \\ t_{S_1} &= T_{d_{0,1}} + T_1 \\ t_{D_2} &= T_{d_{0,1}} + T_1 + T_{d_{1,2}} \\ t_{S_2} &= T_{d_{0,1}} + T_1 + T_{d_{1,2}} + T_2 \\ t_{D_3} &= T_{d_{0,1}} + T_1 + T_{d_{1,2}} + T_2 + T_{d_{2,3}} \end{aligned} \tag{1}$$

Hence, the probabilities of arrivals occurring in each state can be derived as:

$$\begin{aligned} P(T_{ia} \in D_1) &= 1 - e^{-\lambda t_{D_1}} \\ P(T_{ia} \in S_1) &= -e^{-\lambda t_{S_1}} + e^{-\lambda t_{D_1}} \\ P(T_{ia} \in D_2) &= -e^{-\lambda t_{D_2}} + e^{-\lambda t_{S_1}} \\ P(T_{ia} \in S_2) &= -e^{-\lambda t_{S_2}} + e^{-\lambda t_{D_2}} \\ P(T_{ia} \in D_3) &= -e^{-\lambda t_{D_3}} + e^{-\lambda t_{S_2}} \\ P(T_{ia} \in S_3) &= e^{-\lambda t_{D_3}} \end{aligned} \tag{2}$$

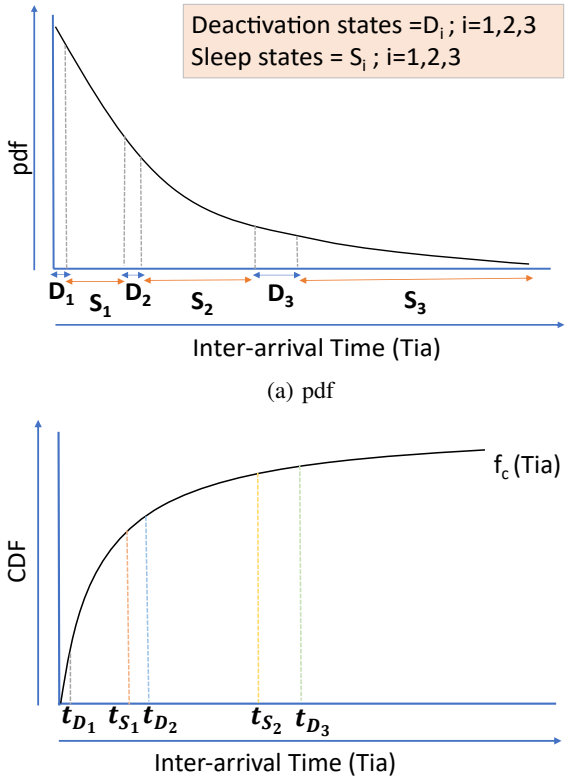


Fig. 3: Distribution of inter-arrival times comprising ASMs.

B. Delay Conservation

As described in previous section, an arrival in any sleep state SMi triggers a fixed reactivation delay whose value depends on the SM level; reactivation from SMi requires a time Tr_i. A request occurring during a deactivation phase experiences an additional delay, denoted by r, beside the reactivation time. This additional delay r represents the remaining time that must elapse from the moment in which the new request occurs (during a transition period) to complete the deactivation from the state SM$(i-1)$ to the state SMi.

The delay, denoted by d_j (with $j \in \mathcal{S} = \{D_1, S_1, D_2, S_2, D_3, S_3\}$), experienced by requests occurring during deactivation transient states (D_1, D_2, D_3) and sleep states (S_1, S_2, S_3) is given by:

$$\begin{aligned} &T_{ia} \in D_1 : d_{D_1} = r + Tr_1;\ r \le T_{d_{0,1}} \\ &T_{ia} \in S_1 : d_{S_1} = Tr_1 \\ &T_{ia} \in D_2 : d_{D_2} = r + Tr_2;\ r \le T_{d_{1,2}} \\ &T_{ia} \in S_2 : d_{S_2} = Tr_2 \\ &T_{ia} \in D_3 : d_{D_3} = r + Tr_3;\ r \le T_{d_{2,3}} \\ &T_{ia} \in S_3 : d_{S_3} = Tr_3 \end{aligned} \tag{3}$$

Where r represents the residual deactivation time. In our computations, we make the conservative assumption that the residual deactivation time required to reach the next state SMi is always equal to the upper bound of the remaining time r, i.e., the complete deactivation time $T_{d_{i-1,i}}$. Furthermore, as already mentioned in Sec. III, under DCASM operation the value of T_1 is fixed. The average delay, denoted by $\overline{d}$, must

be lower than the desired target d^*:

$$\overline{d} = \sum_{j \in \mathcal{S}} P(T_{ia} \in j) \cdot d_j < d^* \tag{4}$$

From (4), we can derive the hold time T_2 that allows to maintain the average reactivation delay below the desired target threshold, d^*. By replacing the values in (4) with formulas (2) and (3), we get:

$$(1-e^{-\lambda t_{D_1}})(T_{d_{0,1}}+Tr_1)+(e^{-\lambda t_{D_1}}-e^{-\lambda t_{S_1}})Tr_1+....+ (e^{-\lambda t_{D_3}})Tr_3 < d^* \tag{5}$$

Likewise, incorporating the expressions from (1) in (5), we obtain:

$$\left(1-e^{-\lambda T_{d_{0,1}}}\right)\left(T_{d_{0,1}}+T_{r_{1,0}}\right)+\left(e^{-\lambda T_{d_{0,1}}}-e^{-\lambda(T_{d_{0,1}}+T_1)}\right)T_{r_{1,0}} \left(e^{-\lambda(T_{d_{0,1}}+T_1+T_{d_{1,2}}+T_2+T_{d_{2,3}})}\right)\left(T_{r_{3,2}}+T_{r_{2,1}}+T_{r_{1,0}}\right) < d^* \tag{6}$$

Further simplifying (6) for a given target delay, the hold time, T_2, to be spent in SM2 before deactivating to SM3 is:

$$T_2 > -\lambda^{-1}\left\{\ln\left[d^*-T_{r_{1,0}}-T_{d_{0,1}}\left(1-e^{-\lambda T_{d_{0,1}}}\right) -e^{-\lambda\left(T_{d_{0,1}}+T_1\right)}\left(T_{r_{2,1}}+T_{d_{1,2}}\left(1-e^{-\lambda T_{d_{1,2}}}\right)\right)\right] -\ln\left[T_{d_{2,3}}\left(1-e^{-\lambda T_{d_{2,3}}}\right)+T_{r_{3,2}}\right]\right\}-\left(T_{d_{0,1}}+T_1+T_{d_{1,2}}\right) \tag{7}$$

Notice that the setting of ASM operation is decided based on desired average delay d^*. This parameter is set by the mobile operator based on its business models, provided verticals, expected breakdown of applications. Once it is defined, the setting of ASM operation follows, as described above.

C. Power Consumption Computation

Base station power consumption under DCASM operation is evaluated using the IMEC power model, which reports values of consumption at different sleep states and transient states [38]. From these values we compute the average power consumption, denoted by PC, under a given arrival rate and considering a specified target delay. To compute the average power consumption, we consider the base station utilization, U, defined as:

$$U = P(A)$$

where $P(A)$ is the probability that the BS is active. Assuming that the activity during low traffic periods can be modeled as a $M/M/\infty$ queue (since the few service requests do not saturate the BS capacity), we have:

$$P(A) = 1-e^{-\lambda/\mu}$$

where μ is the average service rate.

In order to compute the energy saving, we first compute the average energy consumed during sleeping periods, i.e., during the periods between the instant in which the deactivation starts and when the BS is back to full operation. For simplicity we neglect the transition through SM1, since the time spent in the transition from active state to SM1 is negligible. In a similar way to what was previously done, we distinguish different situations depending on the interval in which the arrival of a request, which triggers a reactivation, occurs; intervals are identified in Fig. 2 and defined in (1). The average energy consumed during a sleeping period can be computed from:

$$E_{SM} = \int_0^{t_{D_2}} 2P_1 T_{d_{1,2}} \lambda e^{-\lambda t} dt + \int_{t_{D_2}}^{t_{S_2}} \left(2P_1 T_{d_{1,2}} + P_2(t-T_{D_2})\right)\lambda e^{-\lambda t} dt + \int_{t_{S_2}}^{t_{D_3}} \left(2P_1 T_{d_{1,2}} + 2P_2 T_{d_{2,3}} + P_2 T_2\right)\lambda e^{-\lambda t} dt + \int_{t_{D_3}}^{\infty} \left(2P_1 T_{d_{1,2}} + 2P_2 T_{d_{2,3}} + P_2 T_2 + P_3(t-T_{D_3})\right)\lambda e^{-\lambda t} dt \tag{8}$$

The first integral refers to the case in which the arrival occurs during deactivation towards SM2: in this case, the power level is P_1 and the time spent with this level corresponds to full transition to SM2 plus immediate reactivation for a total time of $2T_{d_{1,2}}$. The second integral corresponds to the case in which the request arrives during SM2; in this case, in addition to the energy needed to deactivate and reactivate the BS, an amount of time equal to $(t-T_{D_2})$ is spent in SM2, i.e., at power level P_2. Similarly, the other integrals are derived. By solving (8), we obtain:

$$E_{SM} = 2P_1 T_{d_{1,2}} + 2P_2 T_{d_{2,3}} e^{-\lambda t_{S_2}} + \frac{P_2}{\lambda}\left(e^{-\lambda t_{D_2}} - e^{-\lambda t_{S_2}}\right) + \frac{P_3}{\lambda} e^{-\lambda t_{D_3}} \tag{9}$$

The average power during sleep modes is then:

$$P_{SM} = E_{SM} \cdot \lambda \tag{10}$$

since $1/\lambda$ is the average duration of a sleep period. Thus, the overall expression to compute the average power consumption, PC for delay constrained DCASM operation is:

$$PC = U \cdot P_A + (1-U) \cdot P_{SM} \tag{11}$$

where P_A is the power consumed when the BS is active and it is transmitting data.

D. Power saving

The power saving, denoted by PS, is computed as the fraction of power that is saved with respect to the baseline condition in which the BS is not performing ASMs operation, i.e., the BS is always active, even when there is no data to transmit. The value of PS can be derived through the following equation:

$$PS = \frac{P_B - PC}{P_B} \tag{12}$$

where P_B is the average power consumed in the baseline situation, which corresponds to a BS that is either fully active when traffic or signaling data are exchanged (hence consuming P_A) or idle when no service requests are received (in this case the consumption is equal to P_I, as reported in Table II).

IV. DCASM PERFORMANCE

To investigate the performance of DCASM, we consider a realistic traffic scenario. The paradigm associated with 5G is dense network deployment, with BSs coverage overlapping with macro-cells. Unneeded BSs are put in progressively deeper sleep mode during off peak periods to save energy,

and, if needed, macro-cell BSs trigger a BS activation when a new request arrives. Hence, an on demand activation upon signaling or service request could allow the BSs to operate with DCASMs.

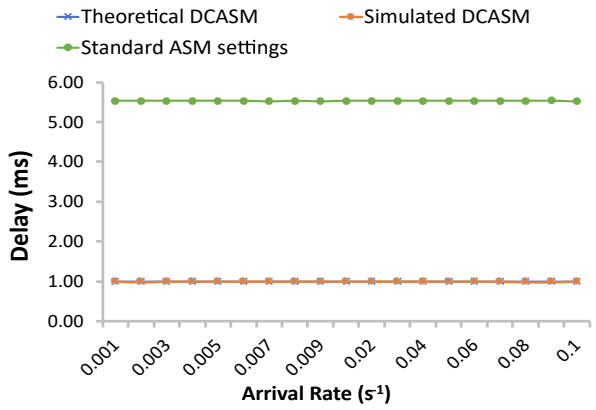
We consider a scenario, in which the BS handles (low) mobile user traffic. Requests inter-arrival times are exponentially distributed and the average service duration is equal to 30 s.

First, we validate the proposed mathematical framework that derives the optimized duration of T_2. To this purpose, a comparative analysis of theoretically derived T_2 settings and power consumption values is performed against simulation based results. Then, we discuss the impact of possible errors in predicting arrival rates on the evaluation of both reactivation delays and energy saving. Finally, we consider actual traffic traces as input to our proposed mathematical framework to analyze the performance of real-time optimization under different delay requirements.

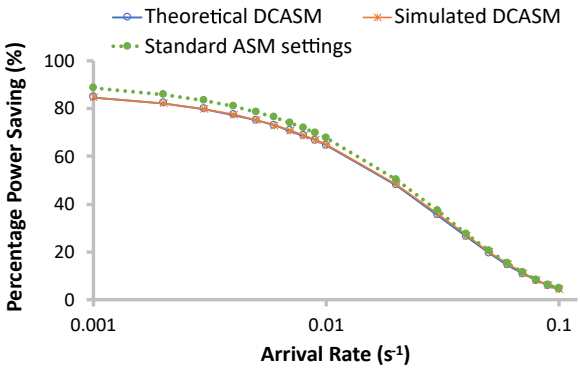
A. Validation

A python based simulator is developed using SimPy library for mimicking the behavior of a BS with ASMs under different scenarios. The tested value of the average arrival rate varies in [0.001,0.1] s^{-1}. The system performance is evaluated for a duration corresponding to approximately 500,000 arrivals for each arrival rate. While in the framework we consider an upper bound for the residual deactivation time, namely the remaining time r in (3), in the simulation the exact remaining time is used for the computation of the reactivation delay.

Fig. 4a reports the average reactivation delay, d_a, and Fig. 4b the power saving, PS, observed at various values of the arrival rate when the average target delay d^* is 1 ms. Note that in our results the power saving is always computed with respect to the baseline case, in which ASM operation is not performed, i.e., the BS is always active. For each tested value of the arrival rate λ, the value of T_2 that is in accordance with the considered delay constraint is selected using analytical DCASM framework. The theoretical values of delay and power saving obtained according to (4) and (12) under these DCASM settings are represented by the blue markers in the graphs. Simulation values of the two performance indicators that are obtained under the same T_2 and λ DCASM settings are shown for comparison (orange symbols). Simulation values obtained under standard ASM settings are also reported (green symbols) to highlight the difference in reactivation delays and power saving between DCASM and the standard ASM approach. It can be observed that, using standard ASM settings, 5.5 ms average delay is observed for all the considered values of the arrival rate, whereas under DCASM settings the desired average reactivation delay, which is fixed to 1 ms, is guaranteed under any predetermined arrival rate. Moreover, simulation results validate the performance of DCASM, as reactivation delays under each arrival rate are almost in-lined with the desired 1 ms delay, further confirming the theoretical results. The difference observed between simulation and theoretical values of the average reactivation delay under DCASM is negligible, confirming that the conservative assumption on the residual deactivation time does not impact the derivation of the proper



(a) Delay



(b) Power Saving

Fig. 4: Reactivation delay and Power saving validation for different arrival rates under DCASM settings and standard ASM settings.

settings that allow to meet the delay constraint.

Notice that the target delay considered in our study is meant as a constraint on the reactivation delay that is experienced *on average* by the arrival of a request while the BS is in sleep mode. This is a different case with respect to a scenario in which a minimization of the worst case delay may be required, i.e., in which a maximum reactivation delay has to be guaranteed. A maximum reactivation delay can be simply guaranteed by considering the reactivation time from any state: if the desired maximum delay is smaller than the reactivation time from SMi, this sleep mode can simply never be entered. Considering the scenario analyzed in our work, the proposed DCASM approach allows to fully satisfy the requirement on the desired target average delay. Clearly, a fraction of arrivals occurs during deep sleep modes and undergoes a reactivation delay which is larger than the average, while a fraction of the arrivals experiences a lower delay than the average targeted one. For example, under DCASM operation with target delay d^*=1 ms, 9% of arrivals experience a reactivation delay larger than 1 ms. The fraction of arrivals for which the experienced reactivation delay under DCASM is larger than the delay constraint, denoted by f_c, can be easily derived under any setting of d^* from the following expression:

$$f_c = \sum_{j \in \mathcal{S}_v} P(T_{ia} \in j)$$

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3248291

where S_v is the subset of the states $\mathcal{S} = \{D_1, S_1, D_2, S_2, D_3, S_3\}$ that includes all sleep and transient states j for which $d_j > d^*$.

The power saving is shown in Fig. 4b for increasing values of the arrival rate. The blue and orange curves representing the DCASM (model and simulation) are in-lined and this further validates the proposed model. The BS tends to go to deeper sleep levels more often at lower rates, resulting in higher energy savings that reach 85% for low arrival rates. However, as expected, the power saving curve has a steep descent as the arrival rate grows larger. The BS under DCASM achieves a slightly smaller energy saving with respect to the standard ASM settings (shown in green in the figure) due to the smaller utilization of deep sleep modes that is needed to guarantee the delay constraint. The standard ASM setting achieves marginally better performance than DCASM at the expense of much higher reactivation delays. This is the "cost" to pay to guarantee the desired (lower) value of 1 ms for the average reactivation delay.

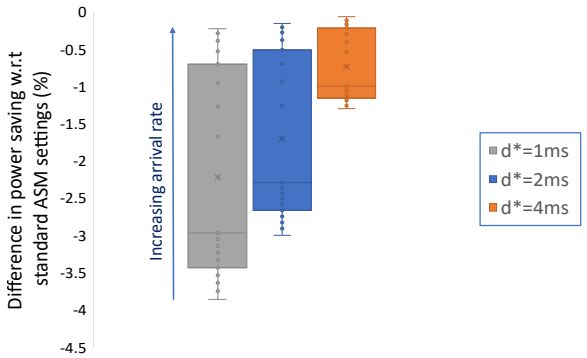


Fig. 5: Reduction of power saving under DCASM with respect to standard ASM operation, considerng different values of arrival rate and target delay.

To further investigate this, Fig. 5 shows the impact of DCASM delay guarantee on power saving with respect to standard ASM settings. The plot reports the reduction of power saving under DCASM in terms of difference between the power saving obtained under DCASM and the power saving obtained under standard ASM. Different color bars refer to different values of the delay constraint, while the points in the boxes refer to different values of the arrival rate λ, with increasing values indicated by the blue arrow. As already noted above, to guarantee the average delay constraint, DCASM achieves a slightly smaller power saving with respect to the standard ASM operation (hence, we have negative values), because DCASM reduces the frequency with which SM3 is entered. However, this is marginal: for low values of d^*, such as 1 ms, less than 4% is lost, while for $d^* = 4$ ms, the effect is further narrowed to no more than 1.5%.

Note that the Poisson arrival process adopted in our study to model 5G traffic may not be fully representative of the bursty traffic that characterizes Massive Machine-Type Communication, for which traffic modeling with a beta distribution is recommended by 3GPP [45]. Nevertheless, the higher arrival rate typically assumed during traffic bursts prevents the BS deactivation into any sleep modes. Conversely, our model may as well be representative of the system behavior during inter-

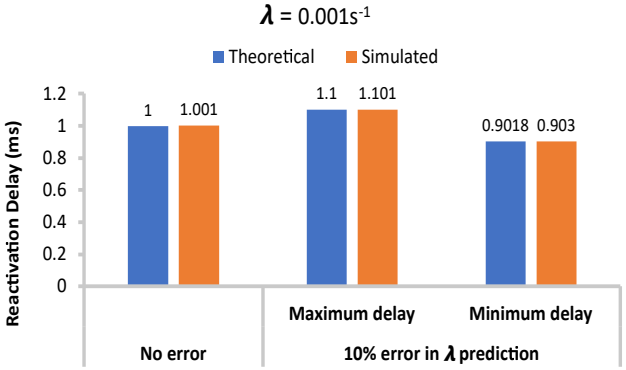


Fig. 6: Effect of error in arrival rate (λ) prediction on reactivation delay using optimized SM2 duration for 1 ms desired delay.

burst periods, during which the traffic load is low. While the overall energy saving may change depending on the prevalence of bursty traffic, we expect that the proper settings of the load dependent ASM configuration parameters between consecutive traffic bursts would not result in drastic change.

B. Effect of error in arrival rate prediction

Since the parameter optimization depends on the request arrival rate λ, and λ has to be predicted, we examine the impact of the error in the prediction on the actual delay and energy savings under DCASM. We consider a 10% error in the prediction of λ, either a negative error or a positive error, to evaluate both cases of delay underestimation and overestimation. In Fig. 6 reactivation delays are displayed when the BS operation is optimized based on DCASM and the prediction has no error, or the two cases of 10% error. An arrival rate of 0.001 s^{-1} is assumed. We keep the required average delay, d^*, fixed to 1 ms and observe the results obtained by both approaches, theoretical and simulated. With a negative 10% error (meaning that the predicted λ is underestimated by 10% with respect to the actual value), the estimated probability of going to SM3 is increased by some percentage points, consequently raising the average reactivation delay. On the contrary, if an overestimation error is assumed (such that the predicted value of λ is increased by 10% with respect to the actual value), the system is more frequently driven towards an active state, hence making it less likely to move to SM3, and thereby reducing the reactivation delay.

Fig. 7 shows the impact of error in prediction of λ on power saving of BS operated under DCASM operation for varying values of the arrival rate. Again, the desired average reactivation delay is equal of 1 ms. Positive or negative errors have no influence on percentage power savings for low values of λ and more than 80% saving is possible using the delay optimized DCASM process. However, for larger values of λ errors in the prediction produce a difference of power saving of up to 3%. In conclusion, under the proposed optimization technique, a 10% underestimation or overestimation error on λ prediction has a marginal impact on energy saving.

C. Real-time Traffic Prediction and Optimization

Real-time traffic prediction has been studied for quite some time, with the purpose of tuning the system parameters in ad-

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3248291

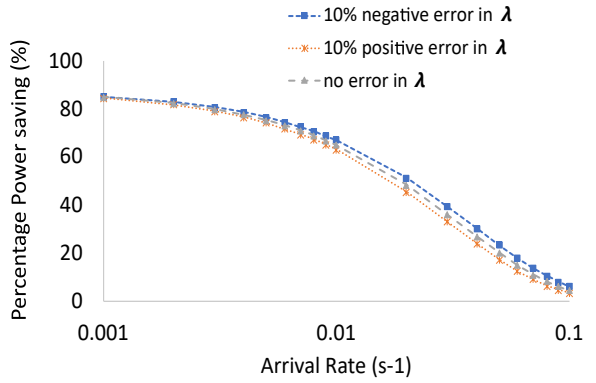


Fig. 7: Effect of error in arrival rate (λ) prediction on power saving using optimized SM2 duration for 1ms desired delay.

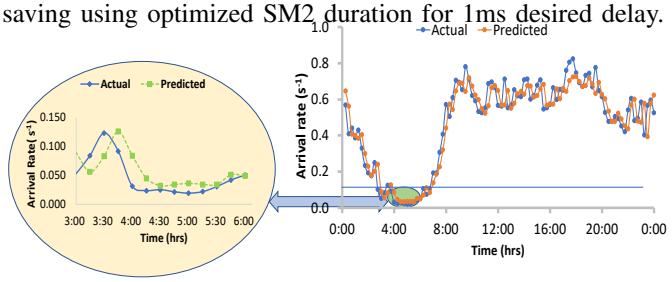


Fig. 8: Actual Vs Predicted BS traffic pattern of train station of city in Italy.

vance in real setups for efficient resource utilization in mobile BS communication. Hence, we investigate the BS performance under the presented optimization approach considering both real mobile traffic traces, collected in the city of Milan, Italy, by an Italian Mobile Operator, and their prediction, obtained through an Artificial Neural Network (ANN) based prediction algorithm, i.e. the approach named *1 ANN-4 outputs* in [19]. The adopted traffic forecast approach employs one ANN for each BS in the scenario, hence each ANN is specifically trained with the corresponding BS traffic profile. At the beginning of each time slot t, the ANN provides four outputs, predicting the BS traffic demand at time t (the time slot that is just beginning), t+1 (the following time slot), t+2 and t+3. The ANN receives five traffic samples as inputs: the traffic at the previous time slot, namely t-1, and four traffic samples corresponding to the time slots t-1 and t from the day ahead and from two days before.

Fig. 8 reports the actual average traffic pattern (normalized with respect to traffic peaks) observed during a sample day in a specific area of the city (blue curve) and the corresponding predicted values (orange curve). The relative error is reported in Fig. 9 (on the right) together with the actual and predicted traffic pattern (on the left). The largest relative error is observed around 4 a.m., at the end of a steep transition from a relatively higher load to a very low load. A similar steep descent before 3 a.m. justifies a minor peak of relative error at that hour, characterized by a higher load than 4 a.m. An intermediate peak in the relative error results evident at 5 a.m., when conversely the load is rather low.

We now study the effect of the error in predicting λ on delay and power consumption when the BS is working under DCASM operation in a real scenario. For our study, we

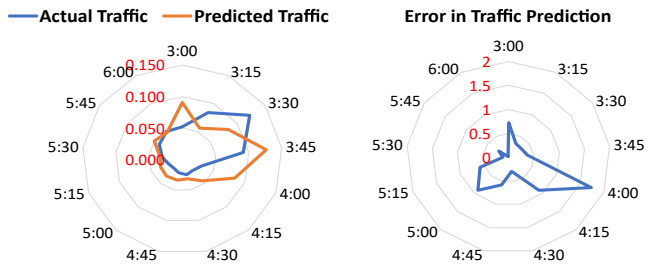


Fig. 9: Actual and Predicted BS traffic pattern with error in prediction [46].

focus on the day period between 3 a.m. and 6 a.m. (see the zoomed window in Fig. 8), when the arrival rate remains below 0.1 s^{-1} and sleep modes can be effectively utilized. We consider various 5G scenarios characterized by different requirements in terms of maximum allowed average delay (d^*). For brevity, we consider three average delay constraints, i.e. 1 ms, 2 ms, and 4 ms, considering the latest latency requirements in 5G standards where the user plane constraint for URLLC communication is 1 ms and 4 ms for eMBB communication [36], [37].

Fig. 10 shows the delays achieved under DCASM with the actual traffic pattern depicted in Fig. 8. In the ideal case (grey curves), BS operation is optimized based on the actual traffic; the orange curves report the delay obtained under the actual traffic pattern, when the operation is decided based on predicted traffic. Fig. 10a illustrates the case of DCASM operation assuming a target d^*=1 ms. In addition, the delay observed under standard ASM settings is also reported as reference (blue curve); the delay is about 5.5 ms at any time. Clearly, in the ideal case, when the optimal T_2 value is applied under the actual traffic profile, the resulting delay is equal to 1 ms at any time. When BS operation is set based on predicted traffic, although the actual delay remains far lower than the delay obtained under standard ASM setting, it results sometimes higher than the desired target. The worst performance is observed at 4 a.m., corresponding to the time when the largest traffic prediction error is registered, as shown in Fig. 9. Smaller delay deviations appear evident at 3 a.m. and 5 a.m., corresponding to minor peaks in the relative error of λ prediction. A correlation between the error in traffic prediction and the actual delay observed during real-time DCASM operation can hence be highlighted by these results, entailing that larger gaps between predicted and actual traffic values are responsible for the most relevant delay increase. Furthermore, during relatively higher arrival rate periods, i.e., from 3:15 a.m. to 3:30 a.m, the system shows more tolerance to traffic prediction error, with a compliant delay despite almost 35% prediction error, with respect to periods characterized by lower arrival rates, i.e. between 5:30 a.m. and 6:00 a.m.. Higher arrival rates are thus more prone to error in prediction in terms of reactivation delay as they bound the system towards the active state and hence less sleep mode utilization.

Fig. 10b and Fig. 10c show similar results for scenarios characterized by more relaxed delay constraints, i.e., d^* is equal to 2 ms and 4 ms. As the value of d^* increases, the

additional delay observed in a real system exceeding the maximum desired value tends to become less remarkable, exceeding by up to only 25% the target delay when d^* is set to 4 ms, against a value that results up to almost 3 fold higher than the desired maximum delay in the case of d^*=1 ms.

The proposed optimization technique shows some tolerance for traffic prediction error. In scenarios having stricter delay constraints, accurate prediction techniques must be adopted. Our findings emphasize that the DCASM optimization technique is robust under any delay sensitive 5G traffic scenario in terms of achieving required energy saving goals by means of ASM utilization.

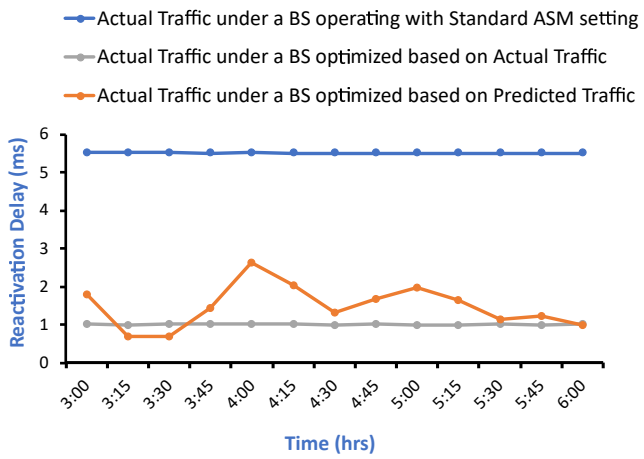
Finally, Fig. 11 reports the power saving obtained with DCASM BS operation based on traffic prediction. The same traffic trace depicted in Fig. 8 is considered. The power consumption is computed based on our proposed mathematical formulation. As a reference, the blue bars show the power saving achieved using the standard settings of ASM; i.e., without posing any delay constraint. Conversely, the remaining bars represent scenarios with three different delay requirements that are d^*=1 ms (orange bars), 2 ms (grey bars), 4 ms (yellow bars). BS operation under DCASM achieves energy consumption reduction of up to about 50% for the reported traffic profile which consist of low arrival rate periods. Furthermore, in terms of energy savings, the results reveal that the ASM operation under standard settings only slightly outperforms the recommended DCASM optimized settings, by just few percentage points, hence entailing the capability of DCASM operation to obtain virtually the same energy savings. From 3:00 a.m. to 4:00 a.m. the arrival rate exceeds 0.1 s^{-1} (see Fig. 8) and the BS is often switched on. No relevant effect due to different settings is observed in this period. Nevertheless, the presented optimized settings for delay conservation, DCASM, pose bound on average delays in real traffic scenarios with respect to standard settings, while minimally effecting the overall gains in energy saving.

V. DCASM with Periodic Signalling

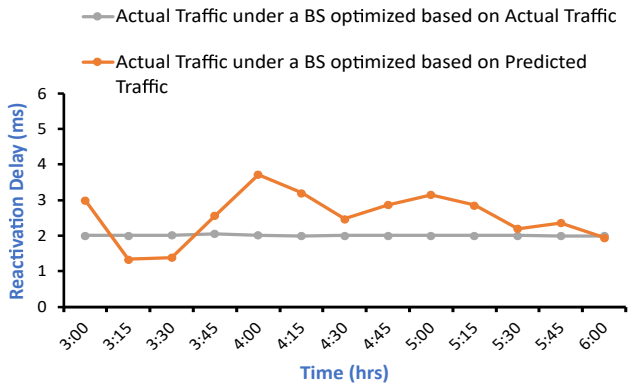
Base stations periodically send synchronization and control signals and the signalling state corresponds to the transmission of synchronization and control signals (primary and secondary synchronization signals - PSS and SSS, respectively - and physical broadcast channel signals - PBSH) [44]. In some configuration, there could be the need to guarantee some periodic signaling even if sleep modes are allowed. In this section, we adapt DCASM to operate in a scenario in which a periodic signalling has to be guaranteed.

Signaling data are assumed to be transmitted with periodicity Δ_t, meaning that after a time Δ_t the BS is activated to transmit signaling information. To derive the setting of T_2 that is required to guarantee a target delay d^* in this scenario, we adapt the derivation of T_2 by considering time windows of duration Δ_t. In a time window, requests arrive according to a truncated Poisson process whose inter-arrival times are limited to the interval $[0, \Delta_t]$, i.e., according to a random variable whose CDF is

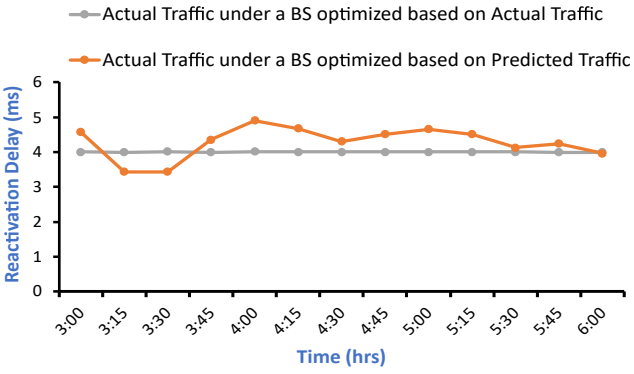
$$F_{ia}(t) = \frac{1 - e^{-\lambda t}}{1 - e^{-\lambda \Delta_t}} \quad t \in [0, \Delta_t] \tag{13}$$



(a) d^* = 1 ms



(b) d^* = 2 ms



(c) d^* = 4 ms

Fig. 10: Simulation analysis of reactivation delay for actual traffic pattern under standard and optimized settings for different average required delays (d^*).

The parameter T_2 that guarantees the delay constraint can be derived similar to previously, adapting (5) and (6) by using

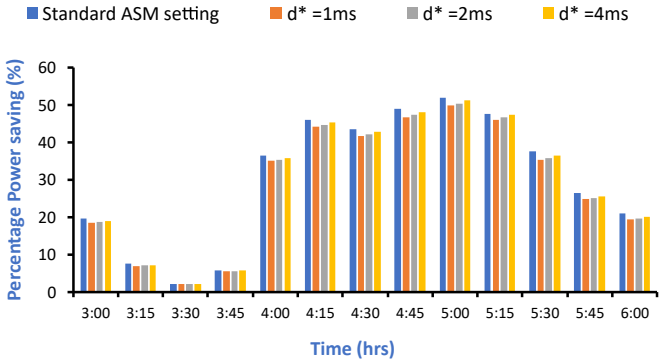
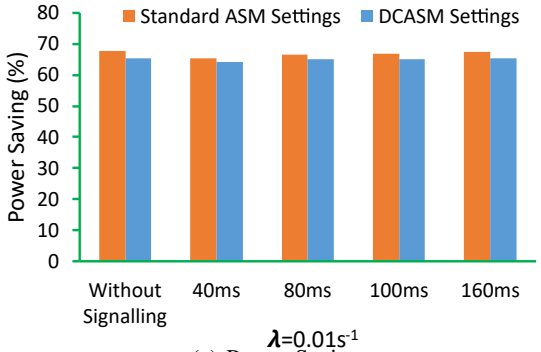


Fig. 11: Comparison analysis of power savings in daily actual traffic pattern with standard ASM and DCASM settings at different average required delays.

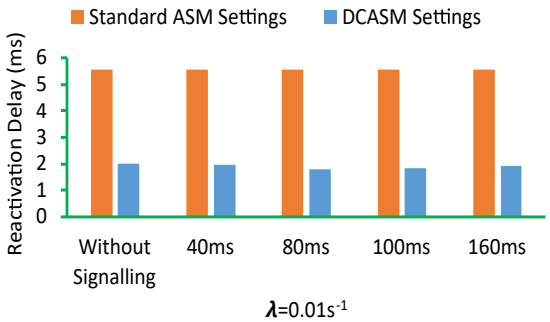
this new distribution for inter-arrival times. We obtain:

$$T_2 > -\lambda^{-1}\{\ln\left(d^*\left(1-e^{-\lambda\Delta_t}\right)+e^{-\lambda\Delta_t}\left(T_{r_{3,2}}+T_{r_{2,1}}+T_{r_{1,0}}\right)\right.(14)$$
$$-T_{r_{1,0}}-T_{d_{0,1}}\left(1-e^{-\lambda T_{d_{0,1}}}\right)-e^{-\lambda\left(T_{d_{0,1}}+T_1\right)}\left(T_{r_{2,1}}+T_{d_{1,2}}\right.$$
$$\left.\left.\left(1-e^{-\lambda T_{d_{1,2}}}\right)\right)\right)-\ln\left(T_{d_{2,3}}\left(1-e^{-\lambda T_{d_{2,3}}}\right)+T_{r_{3,2}}\right)\}-$$
$$\left(T_{d_{0,1}}+T_1+T_{d_{1,2}}\right)$$

Note that d^* represents the target average delay experienced by standard traffic arrivals, whereas signaling data do not experience delay. Indeed, given signaling periodicity, when the BS results in sleep mode, it is proactively switched on before signaling data transmission is needed at the scheduled time. Fig. 12a and Fig. 12b report the power saving and reactivation



(a) Power Saving



(b) Reactivation delay

Fig. 12: Power saving and reactivation delay under standard ASM and DCASM settings, assuming varying signalling periodicity and $\lambda = 0.01s^{-1}$.

delay achieved when the BS has to wake up from SM for periodic signalling considering different signaling periods. We compare the results for the standard setting and DCASM with delay constraint $d^* = 2$ ms; the case with no signaling is also reported for comparison. In the scenario with signalling, saving is smaller than in the case with no signaling but still above 65%, with DCASM saving a bit less due to the required average delay constraint. Whereas the standard setting leads always to 5.54 ms delay, DCASM meets the 2 ms average delay constraint considered in this case. The results prove that DCASM can properly operate even in presence of periodical signaling. In addition, they show that DCASM is a simple and flexible framework that can be adapted to different scenarios.

VI. Conclusion

The substantial raise of mobile traffic in 5G networks entails a huge growth of the energy demand and related operational cost for MNOs. Approaches based on Advanced Sleep Modes allow to save energy by gradually deactivating the BSs into progressively lower power levels when the traffic is low. However, the additional delay experienced by users due to the time required for BS reactivation may result critical for delay sensitive 5G applications.

In this work, we propose a closed-form delay optimization of ASM operation. The approach, called DCASM, is based on setting ASM parameters based on traffic so as to meet the strict delay requirements raised in 5G scenarios. Thus, by predicting traffic, it is possible to dynamically adapt BS operation to traffic so as to achieve energy saving while guaranteeing delay constraints.

Our results show that the presented method, unlike the ASM operation under standard settings, is effective in trading-off delay and energy saving. DCASM performance is not significantly affected by limited traffic prediction errors and, by using ML-based traffic prediction techniques which are already available, DCASM can be easily implemented to achieve energy saving while still guaranteeing delay constraints.

The proposed formulation of the ASMs parameter setting is simple and it enlightens the interplay between activation delay and energy saving, so as to alleviate one of the main concerns on the use of ASMs, which is the possible impact of activation delay. By being simple, the proposed setting can be easily incorporated in flexible and dynamic configurations of BSs, following a key principle of 5G networks. Therefore, DCASM can be effectively adopted in real 5G setups during low traffic periods to reduce the energy consumption without impairing Quality of Service, resulting suitable to fulfill the demanding 5G requirements.

References

[1] Cisco in *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2018-2023 White Paper*, February 2020.

[2] A. Gati, F. E. Salem, A. M. Galindo Serrano, D. Marquet, S. Le Masson, T. Rivera, D.-T. Phan-Huy, Z. Altman, J.-B. Landre, O. Simon, *et al.*, "Key technologies to accelerate the ICT Green evolution–An operator's point of view," *arXiv preprint arXiv:1903.09627*, 2019.

[3] D. Renga and M. Meo, "Dimensioning Renewable Energy Systems to Power Mobile Networks," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 2, pp. 366–380, 2019.

[4] F. E. Salem, A. Gati, Z. Altman, and T. Chahed, "Advanced Sleep Modes and Their Impact on Flow-Level Performance of 5G Networks," in *2017*

IEEE 86th Vehicular Technology Conference (VTC-Fall), pp. 1–7, Sep. 2017.

[5] J. Wu, Y. Zhang, M. Zukerman, and E. K. Yung, "Energy-Efficient Base-Stations Sleep-Mode Techniques in Green Cellular Networks: A Survey," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 803–826, 2015.

[6] F. Han, S. Zhao, L. Zhang, and J. Wu, "Survey of Strategies for Switching Off Base Stations in Heterogeneous Networks for Greener 5G Systems," *IEEE Access*, vol. 4, pp. 4959–4973, 2016.

[7] Sriram Prasanth T, Sai Srujan Kumar M, and Shankar T, "A survey on techniques related to base station sleeping in Green communication and CoMP analysis," in *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, pp. 1059–1067, 2016.

[8] O. Onireti, A. Mohamed, H. Pervaiz, and M. Imran, "Analytical approach to base station sleep mode power consumption and sleep depth," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–7, 2017.

[9] C. Liu, B. Natarajan, and H. Xia, "Small Cell Base Station Sleep Strategies for Energy Efficiency," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 3, pp. 1652–1661, 2016.

[10] X. Guo, Z. Niu, S. Zhou, and P. R. Kumar, "Delay-Constrained Energy-Optimal Base Station Sleeping Control," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1073–1085, 2016.

[11] Z. Niu, X. Guo, S. Zhou, and P. R. Kumar, "Characterizing Energy–Delay Tradeoff in Hyper-Cellular Networks With Base Station Sleeping Control," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 4, pp. 641–650, 2015.

[12] I. Donevski, G. Vallero, and M. A. Marsan, "Neural Networks for Cellular Base Station Switching," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 738–743, 2019.

[13] F. H. Panahi, F. H. Panahi, G. Hattab, T. Ohtsuki, and D. Cabric, "Green Heterogeneous Networks via an Intelligent Sleep/Wake-Up Mechanism and D2D Communications," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 4, pp. 915–931, 2018.

[14] J. Wu, Y. Bao, G. Miao, S. Zhou, and Z. Niu, "Base-Station Sleeping Control and Power Matching for Energy–Delay Tradeoffs With Bursty Traffic," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 3657–3675, 2016.

[15] I. L. C. Araujo and A. Klautau, "Traffic-aware sleep mode algorithm for 5G networks," in *2015 International Workshop on Telecommunications (IWT)*, pp. 1–5, 2015.

[16] N. Mohammad Pour Nejatian, M. Nayebi, and F. Ashtiani, "Effect of different traffic patterns on power consumption of sleep mode in the IEEE 802.16e MAC," in *2007 IEEE International Conference on Telecommunications and Malaysia International Conference on Communications*, pp. 649–653, 2007.

[17] C. Meng, X. Li, X. Lu, T. Liang, Y. Jiang, and W. Heng, "A low complex energy saving access algorithm based on base station sleep mode," in *2013 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 491–495, 2013.

[18] D. Renga, H. Al Haj Hassan, M. Meo, and L. Nuaymi, "Energy Management and Base Station On/Off Switching in Green Mobile Networks for Offering Ancillary Services," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 3, pp. 868–880, 2018.

[19] G. Vallero, D. Renga, M. Meo, and M. A. Marsan, "Greener RAN Operation Through Machine Learning," *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 896–908, 2019.

[20] M. Meo, D. Renga, and M. T. Sarti, "Household users cooperation to reduce cost in green mobile networks," in *2018 IEEE International Telecommunications Energy Conference (INTELEC)*, pp. 1–8, 2018.

[21] H. Al Haj Hassan, D. Renga, M. Meo, and L. Nuaymi, "A Novel Energy Model for Renewable Energy-Enabled Cellular Networks Providing Ancillary Services to the Smart Grid," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 2, pp. 381–396, 2019.

[22] F. Elsherif, E. K. P. Chong, and J. Kim, "Energy-Efficient Base Station Control Framework for 5G Cellular Networks Based on Markov Decision Process," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 9267–9279, 2019.

[23] H. Ye, Z. Ju, B. Wu, J. Pei, and S. Fu, "Joint Base Station Cooperative Transmission and ON-OFF Mechanism in Internet of Things Networks," in *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pp. 336–339, 2017.

[24] Z. Jian, W. Muqing, and Z. Min, "Energy-Efficient Switching ON/OFF Strategies Analysis for Dense Cellular Networks With Partial Conventional Base-Stations," *IEEE Access*, vol. 8, pp. 9133–9145, 2020.

[25] P. Lähdekorpi, 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 (CSCN)*, pp. 163–168, 2017.

[26] J. Peng, P. Hong, and K. Xue, "Stochastic Analysis of Optimal Base Station Energy Saving in Cellular Networks with Sleep Mode," *IEEE Communications Letters*, vol. 18, no. 4, pp. 612–615, 2014.

[27] C. Bouras and G. Diles, "Energy efficiency in sleep mode for 5G femtocells," in *2017 Wireless Days*, pp. 143–145, 2017.

[28] C. Bouras and G. Diles, "Sleep mode performance gains in 5G femtocell clusters," in *2016 8th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pp. 141–146, 2016.

[29] A. El-Amine, H. A. Haj Hassan, M. Iturralde, and L. Nuaymi, "Location-Aware Sleep Strategy for Energy-Delay Tradeoffs in 5G with Reinforcement Learning," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–6, 2019.

[30] Z. Niu, J. Zhang, X. Guo, and S. Zhou, "On energy-delay tradeoff in base station sleep mode operation," in *2012 IEEE International Conference on Communication Systems (ICCS)*, pp. 235–239, 2012.

[31] N. Islam, A. Alazab, and M. Alazab, "A Reinforcement Learning Based Algorithm Towards Energy Efficient 5G Multi-Tier Network," in *2019 Cybersecurity and Cyberforensics Conference (CCC)*, pp. 96–101, 2019.

[32] A. El-Amine, M. Iturralde, H. A. Haj Hassan, and L. Nuaymi, "A Distributed Q-Learning Approach for Adaptive Sleep Modes in 5G Networks," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, 2019.

[33] 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)*, pp. 1–6, 2019.

[34] F. E. Salem, Z. Altman, A. Gati, T. Chahed, and E. Altman, "Reinforcement Learning Approach for Advanced Sleep Modes Management in 5G Networks," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, pp. 1–5, 2018.

[35] M. Meo, D. Renga, and Z. Umar, "Advanced Sleep Modes to comply with delay constraints in energy efficient 5G networks," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, pp. 1–7, 2021.

[36] "Minimum requirements related to technical performance for imt-2020 radio interface(s)." Available at https://www.itu.int/pub/R-REP-M.2410-2017 (2021/08/26).

[37] "5g and energy efficiency." Available at https://global5g.5g-ppp.eu/sites/default/files/BookletA4_EnergyEfficiency.pdf (2021/08/26).

[38] "Interuniversity microelectronics centre (imec)." Available at https://www.imec-int.com (2020/08/12).

[39] M. Gruber, O. Blume, D. Ferling, D. Zeller, M. A. Imran, and E. C. Strinati, "EARTH — Energy Aware Radio and Network Technologies," in *2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–5, 2009.

[40] "On Requirements and Design of SS Burst Set and SS Block Index Indication," in *3GPP TSG-RAN WG1 Meeting #88 R1-1703092*, (Athens, Greece), 2017.

[41] "NR Synchronization Complexity and Periodicity," in *3GPP TSG-RAN WG1 Meeting #88 R1-1702122*, (Athens, Greece), 2017.

[42] "Etsi ts 138 300 v16.7.0(2021-10)." Available at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3191 (2022/01/04).

[43] 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)*, pp. 1–7, 2015.

[44] E. Dahlman, S. Parkvall, and J. Sköld, "Chapter 5 - nr overview," in *5G NR: the Next Generation Wireless Access Technology* (E. Dahlman, S. Parkvall, and J. Sköld, eds.), pp. 57 – 71, Academic Press, 2018.

[45] "3GPP Generation Partnership Project; Technical Specification Group Radio Access Network; Study on RAN Improvements for Machine-type Communications (Release 11), Tech. Rep. 3GPP TR 37.868 V11.0.0." Available at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2630, 2011.

[46] G. Vallero, D. Renga, M. Meo, and M. A. Marsan, "Greener RAN Operation Through Machine Learning," *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 896–908, 2019.