Exploiting User Delay-Tolerance to Save Energy in Cellular Network: an Analytical Approach

Samantha Gamboa, Alexander Pelov, Patrick Maillé, Nicolas Montavont

Exploiting User Delay-Tolerance to Save Energy in Cellular Network: an Analytical Approach

Samantha Gamboa, Alexander Pelov, Patrick Maille and Nicolas Montavont
Institut Mines-Telecom / Telecom Bretagne / IRISA / Université Européenne de Bretagne
Department RSM, 2 rue de la Châtaigneraie, 35576 Cesson Sevigné, France
firstname.lastname@telecom-bretagne.eu

***Abstract*—The network reconfiguration (NR) techniques are key strategies to reduce the energy consumption in cellular networks. In this paper we propose a mechanism to further increase the energy gains resulting of the application of these strategies. Our approach exploits the delay tolerance of the users to increase the periods when no reconfiguration is needed. We developed the analytical model of the mechanism dynamics and we evaluated it numerically for two NR strategies. The results show up to 13 percentage points of performance improvement regarding the traditional NR strategies for a maximal delay of $50s$.**

I. Introduction

Reducing the energy consumption of cellular networks has two beneficial consequences for cellular operators: it decreases the environmental impact (e.g. reducing CO_2 emissions) and it diminishes the operational costs. As the traffic demand is rapidly increasing, the operators have to expand their network infrastructure (e.g. more equipments) which translates into more energy consumption and larger operational costs.

The main energy consuming part of cellular networks is the access network. Traditionally, it operates in an Always-ON paradigm to ensure the ubiquitous nature of cellular networks and the almost-instantaneous connection time. Moreover, it is dimensioned to have enough capacity to support high traffic peaks, which rarely occur. However, the traffic varies in time and space and this traditional approach leads to resource underutilization resulting in energy waste. To operate the access network in a more energy-efficient way, a management paradigm consists in adapting the resources to the traffic variations. Recently, a large number of studies (including efforts from industry and academia) have been proposed toward this idea of resource adaptation. To only cite a few, Green Touch, ICT-Earth, Opera-Net and Ict-c2power are large worldwide consortia which deal with network energy efficiency. Some solutions coming from the research community are already being implemented by hardware manufacturers, while standardization entities are including them in the design of future access network technologies.

In this paper we study a new design approach based on a new type of trade-off – *energy savings* against *initial service delay*. We consider the Network Reconfiguration (NR) techniques which adapt the capacity of the access network to the mid-term load variations. When more capacity is needed, resources are added to the system increasing the power consumption. In order to maximize the low power consumption periods, we propose to offset the start of a requested service for a given, maximal delay.

The users willing to offset some of their services are called Delay Tolerant Users (*DTU*). We introduced this notion in [1] where we designed a DTU-based sleep mode strategy for cellular base stations. In this paper, we go a step further and we present a general DTU-aware mechanism that can be applied to the different NR techniques to improve their performance. For the sake of having specific energy consumption and resource allocation models, we evaluate the theoretic impact of such strategies in a Long Term Evolution (*LTE*) cellular system.

The rest of the paper is organized as follows. We overview the related work in Sec. II. Sec. III defines the DTU-aware enhancement mechanism, which we formally analyze with the model presented in Sec. IV. The system parameters and evaluation parameters are given in Sec. V, while the results are presented in Sec. VI. Sec. VII. concludes the paper.

II. Related Work

The energy saving techniques in the cellular access network can be classified in 3 main complementary categories: Hardware Upgrade (HU), Network Reconfiguration (NR) and Radio Resource Management (RRM) [2]. While only the hardware upgrades can represent considerable energy savings for the base stations in normal operation, some of these improvements are further exploited by the smart management in the NR and RRM algorithms.

a) Hardware Upgrades (HU): Ferling et al. [3] proposed an optimized transceiver architecture with reduced power consumption. One of the key factors of this architecture is the *adaptive power amplifier (APA)*. Results in [4] show the energy gains using the features of the APA, namely operating point adjustment (OPA) and component deactivation (CD). The first allows the APA to operate in optimal efficiency for the different levels of signal load. The second enables the fast activation/deactivation of the APA when there is no signal to transmit. Another promising track in the design of energy-optimized hardware for the cellular access network is the design of *Adaptive Antenna systems*. This system provides several adaptivity degrees. One is to adapt the number of transmitting antennas to the load variations (Antenna Muting [5]). Other studies propose to use *Beamforming*, i.e. to adapt the radiation patterns of the radio signal transmitted by active antennas. This can be done in a semi-static manner, adapting the patterns to long/mid term spatial traffic variations or dynamically per-user [6].

b) Network Reconfiguration (NR): In addition to the traditional *Network planning* where the parameters (BS types and position, maximal transmission power and cell sizes, bandwidth, etc.) are configured statically, several strategies are proposed to dynamically adapt the access network resources to save energy. To address long term traffic variations - e.g. monthly/yearly, the *Bandwidth Adaptation* strategy consists of changing the maximum usable bandwidth of the carriers depending on the load. With the introduction of the *Carrier Aggregation* in the standards [7], an evolved *Bandwidth Adaptation* technique consists in turning on/off complete carriers, changing the total bandwidth of the system [8]. A variant of this strategy is the *Capacity Adaptation* technique [4], where the maximal number of usable resource blocks is adapted which allows the OPA of the power amplifier. A more aggressive approach consists in turning on/off the entire transceiver chains, from individual sectors to complete base stations. The strategy called *Dynamic sectorization* turns on/off only certain sectors of a sectorized base station. Changing the beamform of the remaining sectors is used to maintain the coverage level while no dynamic adaptation of the transmission power is required [9]. When the transmission power is adapted, the terms *Cell breathing* [10] and *Cell on/off* [11] have been used in the literature. In most of the cases, the coverage is fulfilled by the increased transmission power of neighbouring BSs without performing beamforming.

c) Radio Resource Management (RRM): The RRM energy-aware strategies determine when and which resources to use (e.g. time, frequency, transmission power) to send the information over the radio interface. *Energy-aware schedulers* seek to exploit the characteristics of the hardware to rapidly adapt the radio interface to the instantaneous load while maximizing the energy efficiency of the transmission [12].

The delay tolerance in elastic data services has been studied in several scales, from deadline aware RRM schedulers [13] to delay tolerant bulk data offload [14]. In our previous work we designed a base station sleep strategy based on initial service delay tolerance and we evaluated it under delay tolerant voice calls [1]. To the best of our knowledge this was the first study considering the users delay tolerance in the design of NR strategies.

III. DTU-AWARE ENHANCEMENT MECHANISM

The approach proposed in this paper seeks to further increase the energy gains obtained by the Network Reconfiguration (NR) strategies described in Section II. These techniques adapt the capacity of the access network to the mid term load variations. In periods of low load, the access network can work using a subset of its components. All other components can be deactivated or turned to a low power consumption level. When more capacity is needed (due to load increasing), resources are incrementally added to the system increasing the power consumption. For example, when the load increases, one sector or an entire base station can be turned on (*Dynamic sectorization*, *Cell on/off*) or the usable resource blocks can be increased (*Capacity Adaptation*). We propose an initial delay for users to access their services. This contributes to extend the periods when the system operates at low power consumption (and no extra resources are added). By delaying some users, the system may avoid activating a new resource just to provide service to an instantaneous burst arrival.

The system in study is a cellular access network enabled to use the NR strategies. In particular, we consider the *Dynamic sectorization* and the *Capacity Adaptation* techniques. The system is modelled to be in one of two operational states, depending on the available resources. The system is in the state *all-On* when all components of the access network are operational. In this state the system has an available maximum capacity denoted by $C_{\max}$. The system is in the state *min-On* when only a fixed set of radio resources provides coverage and satisfies a minimum level of load. In this state the system has a capacity $C_{\min}$ which varies depending on the used strategy.

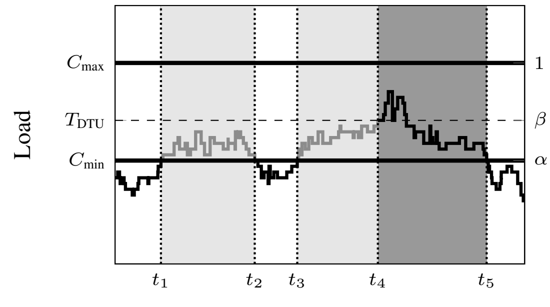


Fig. 1: **Load dynamic example of the DTU-aware network reconfiguration strategies**. White periods: system in *min-On* state – No delay. Light gray periods: system in *min-On* state – Delaying users. Dark gray periods: system in *all-On* state – No delay

In this paper we consider 3 types of system configurations. When the system is in normal operation (e.g. no NR strategy is applied), it remains in *all-On* state no matter the level of load. We denote this approach as the **Always On** strategy. When a NR strategy is applied, the system state changes depending on the load. This is denoted as **Traditional NR strategy**. When the delay tolerance of the users is considered in the state changing decision, the term **DTU-aware NR strategy** is employed.

In the traditional NR strategies the state changing decisions are taken based on the level of load (L). These load thresholds should be appropriately selected to ensure the service in the reconfiguration periods (e.g. no users blocked or dropped). For simplicity, in this paper we assume that the reconfiguration is done instantaneously. Thus, the reconfiguration threshold can be equal to the capacity in *min-On* state ($C_{\min} = \alpha C_{\max}$) without degrading the quality of service. However, the model can be further adapted to consider the reconfiguration periods and thresholds lower than $C_{\min}$. The parameter α is the traditional reconfiguration factor. It expresses the proportion of the system capacity that defines the reconfiguration threshold in the traditional NR strategy.

The objective of the DTU-aware NR strategy is to extend the periods when the system is in *min-On* state. We define a reconfiguration load threshold $T_{\text{DTU}} = \beta C_{\max}$. During the periods where the load is in-between the two thresholds ($C_{\min} < L < T_{\text{DTU}}$), the system will stay in *min-On* state. In other words, users will be accepted in the system even if (instantaneously) there are not enough resources available for them. Thus, they can experience a delay in the start of their service. The actual switching to *all-On* state will occur when T_{DTU} is reached, defining the parameter β as the DTU-aware reconfiguration factor. To ensure the service after the initial delay, the system should not accept more users than the capacity. Thus, T_{DTU} is constrained by $C_{\min}$. The system will switch to *min-On* state when no extra capacity is needed, i.e. when the C_{min} is reached (Fig. 1).

We define the maximum time the users are willing to wait as the *maximum tolerable delay (D)*. We assume throughout this paper that when the DTU-aware strategy is applied, all users are delay tolerant. The service of the impatient users should be prioritized (e.g. using access control or scheduling algorithms). However, this is out of the scope of this paper.

Delay tolerant users waiting in the system will be served either because some fixed resources are now available (user departures, $L < C_{\text{min}}$) or because the system increased its capacity to C_{max} (T_{DTU} was reached). Thus, the selection of T_{DTU} is critical to ensure the service before D occurs. In Section IV we present the formal analysis of the estimation of this threshold ensuring an average waiting time inferior to D.

Reducing the time in *all-On* state increase the energy-saving potential of the traditional NR strategies. In case of fast variation of the load, the DTU-aware approach can avoid the "ping-pong" effect switching between system states. This can contribute to the efficiency of the strategies as well as the lifetime of the equipments.

IV. THEORETICAL MODEL

We model the system traffic dynamic using the three strategies described in Section III. The Markov chains used for this purpose are explained in the first part of this section. The following part describes the threshold derivation depending of the maximum tolerable delay (D).

A. Markov chains

The number of active users in the system is given by a counting process, which we model as an ergodic and homogeneous discrete-state Markov Chain (*MC*). The parameters used to describe this process are the following: the interarrival time follows an exponential distribution with parameter λ. The service time is exponentially distributed with parameter μ. The system capacity C_{max} is fixed and the method of service is FIFO. We assume that during all the service time a user consumes a fixed number of resources (e.g. a given target downlink throughput) which is the same for all users. Thus, we can refer to the load and the system capacity in terms of number of users.

The Markov process $\{X_D(t) : t > 0\}$ models the user dynamic under the DTU-aware NR strategy. The corresponding Markov chain is depicted in Fig. 2. X_D can take its values in the state space $S_D = \{(i, j)\}$. The parameter i represents the number of active users in the system and j indicates the system state. If the system is in *all-On* state, $j = 1$ and if the state is *min-On* $j = 0$. The variation of S_D is limited by the strategy threshold (T_{DTU}) and the available capacity (C_{min} if the state is *min-On* and C_{max} if the state is *all-On*). For the states $(i, 0)$ where $C_{\text{min}} \le i \le T_{\text{DTU}}$, the departure rate is limited to $C_{\text{min}}\mu$ as the system is only capable of serving up to C_{min} users at the same time when in state *min-On*. These are the states of X_D for which some users are delayed.

When the system is configured with a traditional NR strategy, the state switching occurs when the load reaches C_{min}. This is the particular case of the process $X_D(t)$ having only one reconfiguration threshold ($T_{\text{DTU}} = C_{\text{min}}$).

Considering the Always-ON strategy, all resources are always available and the system is in *all-On* state all the time. This strategy is represented with X_D when $T_{\text{DTU}} = C_{\text{min}}$ and $j = 1$ for all states.

B. Waiting Times

For each user arriving in the system, the probability of waiting more than the *maximal tolerable delay* (D) should be controlled. Considering the characteristics of our model (inter arrival and service times exponentially distributed), the user waiting time W follows an *Erlang Distribution*. The shape and rate parameters of this distribution depend on the load characteristics (λ, μ) and the current system state, as explained below.

There are two possible ways of getting served for a waiting user. Consider a user arriving when there are already m users waiting in the system. This means that there are $n = C_{\text{min}} + m$ users already in the system. Regarding our model, and denoting by $p_{(i,j)}$ the steady-state probability of state (i, j), each user has a probability $p_{(n,0)}$ to find the process $X_D(t)$ in state $(n, 0)$ upon arrival. That user will be served either when m departures occur (no more users waiting before him, i.e. $X_D(t)$ is in the state $(C_{\text{min}}, 0)$), or the necessary arrivals occur and the system switches to the state *all-On* ($X_D(t)$ is in state $(T_{\text{DTU}} + 1, 1)$).

Thus, there are different sequences of events (user arrivals and departures) which can lead the waiting user to be served. Each of these events occurs after a time that is exponentially distributed with parameter $\gamma = C_{\text{min}}\mu + \lambda$. This event is an arrival or a departure with probability $p_A = \lambda/\gamma$ or $p_D = \mu/\gamma$ respectively. Consider a set of events S_a^d containing a arrivals and d departures. Consider there are n_a^d sequences of these events leading the waiting user to be served. Thus, the occurrence probability of S_a^d is given by:

$$p_{S_a^d} = n_a^d ({p_A}^a {p_D}^d) \tag{1}$$

where the values of a and d are finite and constrained by the maximum number of possible waiting users in the system ($h = T_{\text{DTU}} - C_{\text{min}}$) and the number of users already waiting when the arrival of the considered user occurs (m). The time after which a sequence of S_a^d occurs ($T_{S_a^d}$) follows an Erlang distribution with shape $k = a + d$ and rate γ. It should be guaranteed that in at least 95% of the cases a user arriving to the system will wait less than D to be served. Thus, the system will be dimensioned according to:

$$\begin{aligned} &\underset{T_{\text{DTU}}}{\text{maximize}} \sum_{i=0}^{T_{\text{DTU}}} p_{(i,0)} \\ &\text{subject to} \sum_{k=C_{\text{min}}}^{T_{\text{DTU}}-1} \sum_{a=0}^{m+h} \sum_{d=0}^{m} p_{(k,0)} p_{S_a^d} P(T_{S_a^d} > D) \le 0.05 \end{aligned} \tag{2}$$

V. EVALUATION PARAMETERS

For the evaluation of the DTU-aware network reconfiguration strategies, we consider a homogeneous LTE system composed of Macro base stations. The parameters concerning the deployment, the base station power consumption and the service characterization are summed up in Table I. We

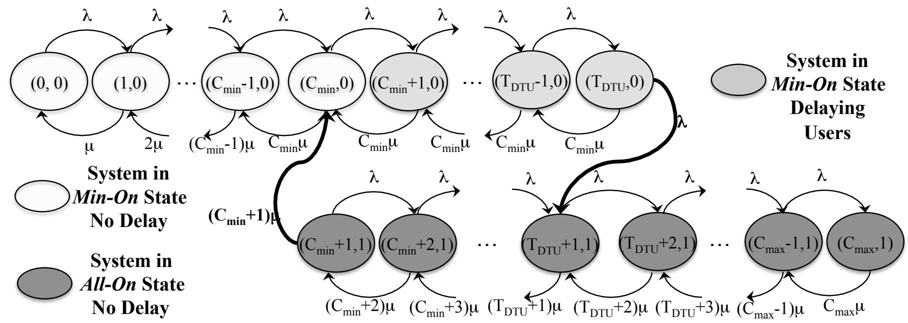


Fig. 2: Markov chain of the user dynamic in the DTU-aware network reconfiguration strategies. When T_{DTU} is equal to C_{min} (e.g. no light gray states), the Markov chain represents the dynamic of the traditional NR strategies

consider static uniformly distributed users and homogeneous traffic demand. Thus, the system load variations as well as the state switching impact all base stations at the same time. In a real network the NR strategies should work on a per site basis rather than on a network region. This simplified evaluation approach is likely to underestimate the gains of the NR strategies and has no major impact in the analysis of the DTU-aware enhancement.

A. Deployment

The reference deployment is a homogeneous hexagonal deployment of 3-sectorised macro base stations with 10MHz of bandwidth and 2x2 MIMO antennas, each transmitting up to 20W of RF power. We consider a base station site density corresponding to dense urban scenario with 500 meters of inter site distance. We analyse a network region composed of five base station sites.

B. Traffic Characterization

The infrastructure of LTE allows the operators to provide data based services guaranteeing real-time quality of service constraints, e.g. video calls. The capacity of the reference system is obtained from system level simulations in [2] and defines a minimum quality of service required by high definition video transmission with 2Mbps even to the 5% of the users at cell edge. However, we consider a minimum of 500 kbps as the target for an acceptable video call quality [18]. When a video call is in progress, it uses a given number of allocated resource blocks in the LTE downlink. This number depends on the coding rate and modulation scheme which vary with the link quality reported by the user equipment. Considering uniformly distributed users and ideal channel conditions, we assume that each voice call uses in average approximately the same number of resource blocks. Therefore, we consider that the capacity and the offered load are linear functions of the number of simultaneous video calls.

TABLE I: System Parameters

Deployment [15]	
Deployment Type	Urban Macro
Inter Site Distance [m]	500
Site Area [Km^2]	0.2165
Number of BS sites	5
BS type	3-sector Reconfigurable
Transmission Bandwidth [MHz]	10
Antenna configuration	2x2 MIMO
Total Resource Blocks	50
Baseline BS power consumption [16]	
N_{TRX}	6, 4, 2
P_{max} [W]	20
P0 [W]	130
Δ_P	4.7
P_{sleep} [W]	75
Traffic characterization	
System capacity [$Mbps/Km^2$] [2]	115
Session target throughput [$kbps$]	500
System session capacity (C_{max})	235
Session duration [s] [17] (μ^{-1})	81

C. Base station power consumption

The baseline BS power consumption model used in this paper was introduced by Auer et al. [16]. This model relates P_{out} (the output power radiated at the antenna) and P_{in} (the total power needed by the base station to operate) for each type of LTE BS. The energy model is well approximated by the linear model given by:

$$P_{in} = \begin{cases} N_{TRX}(P_0 + \Delta_P P_{out}) & 0 < P_{out} < P_{max} \\ N_{TRX} P_{sleep} & P_{out} = 0 \end{cases} \quad (3)$$

where N_{TRX} is the number of transceiver chains (depending on the number of active sectors), P_0 represents the power consumption of an empty BS, Δ_P is the slope of the load dependent power consumption, P_{max} represents the maximum transmission power achievable by the base station and P_{sleep} represents the power consumption of the base station in sleep mode. In LTE systems the downlink transmission scheme uses orthogonal frequency-division multiplexing (OFDM). Thus, P_{out} depends on the BS physical resource allocation in the downlink.

The model considers light sleep modes – the transceiver deactivates a subset of its equipments and the power consumption drops to P_{sleep}. This feature is designed to adapt to short term load variations in the order of μs (e.g. DTX). The network reconfiguration techniques are designed to adapt to mid-term traffic variations (in terms of minutes). Thus, when applying the *Dynamic sectorization* strategy we consider deep sleep modes where the entire transceiver is shut down and the power consumption is near to zero Watt. The wake-up time from deep sleep ascends to some seconds [19]. As pointed in Section III our analysis is limited to the evaluation of the system in the stable states (*min-On* and *all-On*). Thus, we do not consider the sleep transitions in the evaluation presented in this paper.

The reduction of the transceiver power consumption when using *Capacity adaptation* depends on the selected operation point (OP) for each level of load [4]. In our evaluations we consider the OP2 (6% of power reduction - operating in loads bellow 79% of the capacity), OP3 (9% of power reduction - operating in loads bellow 63% of the capacity) and OP4 (13% reduction - operating in loads bellow 50% of the capacity).

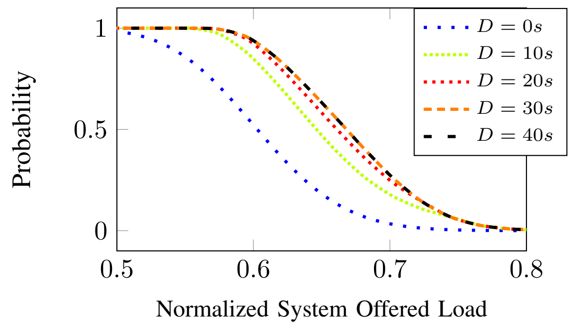
VI. RESULTS

We performed the numerical evaluation of the strategies for different offered load levels and maximal tolerable delays (D). For each combination, we performed an exhaustive search for the reconfiguration load threshold (T_{DTU}) satisfying (2).

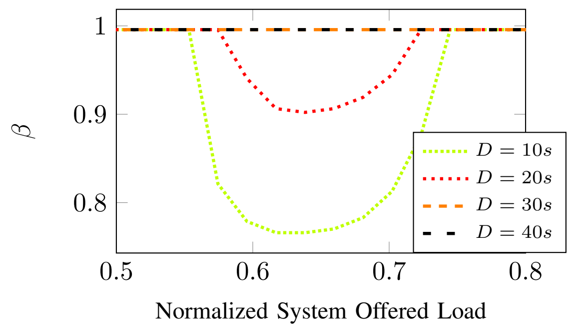
As presented in Section III, α is the traditional reconfiguration factor (e.g. the portion of the capacity where the traditional system reconfiguration is effective). The DTU-aware reconfiguration factor is the parameter β. By introducing delays in the service, we shift the moment of the reconfiguration. Our approach is efficient when the system is experimenting relative loads surrounding α. As shown in the example of Fig. 3(a), when the offered load is below 50% of its capacity, higher delay tolerance provides no benefits compared to traditional NR strategies. For these levels of load, the system operates in *min-On* state without reconfiguration needs. Similarly, for loads superior to 80% the system operates in *all-On* state no matter the strategy. We can see the benefits of delaying the services in loads between these two thresholds. For example, when the system is facing a offered load corresponding to the 65% of its capacity (0.65 in Fig. 3(a)), the probability of being in *min-On* state increases 35% when the users are able to tolerate up to 10 seconds of delay, compared to the traditional NR strategies. A further gain of 58% is observed with a maximum delay of 30 seconds.

The proposed strategy choose β satisfying the users delay constrains (Section IV-B). Thus, it is adjusted in the range of loads where the system is in *min-On* state and delaying users (Fig. 3(b)). It decreases when the probability of having waiting users increases (and the reconfiguration is needed to satisfy the delay constraint). And it increases when the probability to stay in *all-On* state dominates. The maximum tolerable delay (D) representing benefits using the DTU-aware strategy is limited. To ensure the service after the initial delay, the DTU-aware reconfiguration threshold (T_{DTU}) is constrained by the system capacity. For example in Fig. 3(b), we can see the system operates using the maximum reconfiguration factor ($\beta = 1$) for $D = 30s$. Users wiling to wait more time (e.g. $D = 40s$, Fig. 3(a)) do not represent further gains for the system, as a maximum delay of 30 seconds is ensured using the boundary parameters of the strategy for this α. This previous case should be differentiate from the case when no reconfiguration is needed. For example, in Fig. 3(b), when the offered load is inferior to 56%, waiting users are served by the system in *min-On* state. For loads superior to 78% the system will stay in *all-On* state. In these two cases $\beta = 1$, but no reconfiguration occurs.

We evaluated each network reconfiguration strategy separately. The values of the parameter α were chosen considering the specification of the traditional strategies and the results from [4]. For each level of offered load, we calculated the average power consumption for the system using the traditional NR strategy, the DTU-aware NR strategy and the system in Always On operation. In Fig. 4 we present the results for three of the evaluated strategies (in the range of loads where each one is efficient). Considerable reductions in the average power consumption are observed when the users are able to tolerate some delay in the start of their services. Detailed graphics for the other strategies are omitted due to space restrictions. However, the results concerning the daily energy savings achieved by all the DTU-aware NR strategies are summed up in Table II. We considered three different deployment traffic profiles, each one characterized by a traffic peak. The load variations along the day are relative to the peak [15]. Each daily traffic profile was approximated using eight level of offered loads capturing the main variations during the day. We calculated the average energy consumption of the system depending on: the time it expends facing each level of load and the power consumption needed in average to serve it. Finally, we aggregated the results for a complete day.



(a) Probability the system is in low power consumption state *min-On*



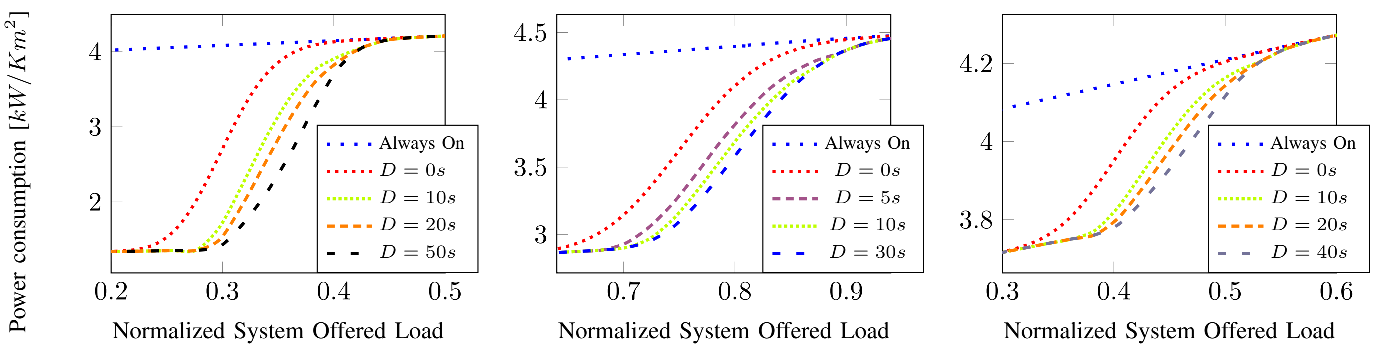
(b) DTU-aware reconfiguration factor β. Fraction of the capacity where the system changes to *all-On* state

Fig. 3: Example of the DTU-aware NR strategy performance. Traditional reconfiguration factor $\alpha = 0.6$ ($C_{\text{min}} = 141$ sessions)

The higher energy savings are observed using the dynamic sectorization strategy, switching between one and three base station sectors. The reference system is clearly overprovisioned for low traffic scenarios showing 65% of possible energy reductions using the strategy. For this scenario, the addition of the DTU-aware capability to the strategy represents a gain of two percentage points. In the medium traffic scenario the best performance of the DTU-aware strategies is achieved with 13 percentage points of improvement regarding the traditional dynamic sectorization strategy. This represents daily extra savings around 13 kWh if the users are able to tolerate up to 50 seconds of delay in the start of their services when required.

VII. CONCLUSION

In this paper, we proposed a mechanism to further increase the energy savings resulting of the application of network reconfiguration (NR) strategies in cellular access networks. Our mechanism considers the delay tolerance of the users to extend the periods of time when the network can remain in low power consumption. We developed a generic model of the mechanism dynamics and we analyzed it numerically for two NR strategies. We calculated the average power consumption



(a) Dynamic Sectorization: One ↔ Three sectors ($\alpha = 0.3$)

(b) *Dynamic Sectorization: Two ↔ Three sectors* ($\alpha = 0.75$)

(c) *Capacity Adaptation:* $OP4$ ($\alpha = 0.4$)

Fig. 4: Average system power consumption depending on the system load. Comparison between the system in normal operation (Always On), using the traditional NR strategies ($D = 0s$) and using the DTU-aware NR strategies ($D > 0s$)

			Deployment traffic profile					
			Low Traffic		**Medium Traffic**		**High Traffic**	
Traffic Peak			$28Mbps/Km^2$		$52Mbps/Km^2$		$92Mbps/Km^2$	
Equivalent of the system capacity			25%		46%		80%	
Strategy			**Energy reductions**					
Parameters	α	$D(s)$	Traditional	DTU-aware	Traditional	DTU-aware	Traditional	DTU-aware
Dynamic Sectorization: One ↔ Three sectors	0.3	50	65%	**67%**	27%	**40%**	15%	**19%**
Dynamic Sectorization: Two ↔ Three sectors	0.75	30	33%	33%	33%	33%	26%	**31%**
Capacity Adaptation OP2	0.75	30	6%	6%	6%	6%	4%	**6%**
Capacity Adaptation OP3	0.6	30	9%	9%	9%	9%	5%	**6%**
Capacity Adaptation OP4	0.4	40	13%	13%	9%	**12%**	4%	**5%**

TABLE II: Average daily energy reductions regarding to the system in Always On operation

and the daily energy savings of the traditional strategies and the further gains when including the DTU-aware mechanism. The results indicate that even with modest delays (up to $50s$), the performance of the traditional NR strategies can be improved. Up to thirteen percentage points of increased daily energy reductions are achieved when using the DTU-aware mechanism. The results also show that higher user delay tolerance does not always lead to increased gains. Future extensions of this work will include the use of system level simulations to evaluate the impact of the strategy considering the user radio link conditions and mobility.

References

[1] S. Gamboa, A. Pelov, P. Maille, X. Lagrange, and N. Montavont, "Energy efficient cellular networks in the presence of delay tolerant users," in *IEEE Global Telecommunications Conference GLOBECOM*, 2013.

[2] EARTH, "Final Report on Green Network Technologies ," Energy aware radio and network technologies (EARTH), Tech. Rep., 2012.

[3] D. Ferling, T. Bitzer, T. Bohn, D. Wiegner, and A. Pascht, "Power efficient transceivers to enable energy-efficient mobile radio systems," *Bell Lab. Tech. J.*, vol. 15, no. 2, pp. 59–76, Sep. 2010.

[4] EARTH, "Final Report on Green Radio Technologies," Energy aware radio and network technologies (EARTH), Tech. Rep., 2012.

[5] P. Skillermark and P. Frenger, "Enhancing energy efficiency in lte with antenna muting," in *Vehicular Technology Conference (VTC Spring), 2012 IEEE 75th*, May 2012, pp. 1–5.

[6] F. Cardoso, S. Petersson, M. Boldi, S. Mizuta, G. Dietl, R. Torrea-Duran, C. Desset, J. Leinonen, and L. Correia, "Energy efficient transmission techniques for lte," *Communications Magazine, IEEE*, vol. 51, no. 10, pp. 182–190, October 2013.

[7] 3GPP, "Evolved universal terrestrial radio access (e-utra); carrier aggregation; base station radio transmission and reception," 3rd Generation Partnership Project (3GPP), Tech. Rep. 36.808 V10.1.0, Jun. 2013.

[8] H. Chen, Y. Jiang, J. Xu, and H. Hu, "Energy-efficient coordinated scheduling mechanism for cellular communication systems with multiple component carriers," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 5, pp. 959–968, May 2013.

[9] L. Hevizi and I. Godor, "Power savings in mobile networks by dynamic base station sectorization," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2011 IEEE 22nd International Symposium on*, Sept 2011, pp. 2415–2417.

[10] G. Micallef, P. Mogensen, and H.-O. Scheck, "Cell size breathing and possibilities to introduce cell sleep mode," in *Wireless Conference (EW), 2010 European*, april 2010, pp. 111 –115.

[11] M. Ajmone Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal Energy Savings in Cellular Access Networks," in *2009 IEEE International Conference on Communications Workshops*. IEEE, Jun. 2009.

[12] S. Videv, J. Thompson, H. Haas, and P. Grant, "Resource allocation for energy efficient cellular systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, p. 181, 2012.

[13] M. Proebster, M. Kaschub, T. Werthmann, and S. Valentin, "Context-aware resource allocation for cellular wireless networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, pp. 1–19, 2012.

[14] M.-R. Ra, J. Paek, A. B. Sharma, R. Govindan, M. H. Krieger, and M. J. Neely, "Energy-delay tradeoffs in smartphone applications," in *Proceedings of the 8th MobiSys*. New York, NY, USA: ACM, 2010.

[15] EARTH, "Energy efficiency analysis of the reference systems, areas of improvements and target breakdown ," Energy aware radio and network technologies (EARTH), Tech. Rep., Jan. 2012.

[16] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Communications*, Oct. 2011.

[17] A. Pattavina and A. Parini, "Modelling voice call inter-arrival and holding time distributions in mobile networks," *Proc. of 19th International Teletraffic Congress*, pp. 729–738, 2005.

[18] *How much bandwidth does Skype need?* [Online]. Available: https://support.skype.com/en/faq/FA1417/how-much-bandwidth-does-skype-need

[19] P. l. Frenger, P. Moberg, J. Malmodin, Y. Jading, and I. Godor, "Reducing Energy Consumption in LTE with Cell DTX," in *2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*. IEEE, May 2011.