Delay Tolerant Users: A solution to end-to-end network energy efficiency

Samantha Gamboa

THÈSE / Télécom Bretagne
sous le sceau de l'Université européenne de Bretagne

pour obtenir le grade de Docteur de Télécom Bretagne
En accréditation conjointe avec l'Ecole Doctorale Matisse
Mention : Informatique

présentée par
Samantha Gamboa
préparée dans le département Réseaux, sécurité et multimédia
Laboratoire Irisa

Delay Tolerant Users : A solution to end-to-end network energy efficiency

Thèse soutenue le 26 juin 2015
Devant le jury composé de :

Jean-François Hélard
Professeur, INSA/IETR – Rennes / président

Michela Meo
Maître de Conférences, Politecnico di Torino - Italie / rapporteur

David Grace
Professeur, University of York – United Kingdom / rapporteur

Steven Martin
Professeur, LRI/Université de Paris-Sud / examinateur

Alexander Pelov
Maître de conférences, Télécom Bretagne /examinateur

Nicolas Montavont
Maître de conférences (HDR), Télécom Bretagne / directeur de thèse

Xavier Lagrange
Professeur, Télécom Bretagne / invité

N° d'ordre : 2015telb0359

Sous le sceau de l'Université européenne de Bretagne

Télécom Bretagne

En accréditation conjointe avec l'Ecole Doctorale Matisse

Ecole Doctorale – MATISSE

Delay Tolerant Users – A solution to end-to-end energy efficiency

---

Thèse de Doctorat

Mention : Informatique

Présentée par **Samantha Gamboa**

Département : Réseaux, Sécurité et Multimédia (RSM)

Laboratoire : IRISA

Directeur de thèse : Nicolas Montavont

Soutenue le 26 juin 2015

Jury :

Mme. Michela Meo, Maître de Conférences, Politecnico di Torino (Rapporteur)
M. David Grace, Professeur, University of York (Rapporteur)
M. Jean-François Hélard, Professeur, IETR (Examinateur)
M. Steven Martin, Professeur, IRI (Examinateur)
M. Nicolas Montavont, Maître de Conférences, HDR, TELECOM Bretagne (Directeur de thèse)
M. Alexander Pelov, Maitre de Conférences, TELECOM Bretagne (Encadrant)
M. Xavier Lagrange, Professeur, TELECOM Bretagne (Invité)

To my parents, Maria and Julio, for all their love and unconditional support.

Acknowledgements

First of all, I would like to express my most sincere gratitude to my supervisors Dr. Alexander Pelov and Dr. Nicolas Montavont, for believing in me despite the adversities and for their guidance throughout these years of work. Their great knowledge, valuable advices, helpful exchanges and continuous support had contributed to my academic development as well as to my personal growth.

I would like to thank all the members of the jury, for having accepted being part of this thesis committee. Specifically to Dr. Michela Meo, Prof. David Grace, Prof. Jean-François Hélard, Prof. Steven Martin and Prof. Xavier Lagrange.

I would like to thank Prof. Xavier Lagrange and Dr. Patrick Maille for their kindly help and collaboration all along the thesis.

I would like to thank to Hussein Al Hal Hassan and Dr. Loutfi Nuaymi who helped me to enlarge my vision of the research topic.

I would like to address special thanks to the other professors and colleagues who currently work at Telecom Bretagne campus Rennes, or have worked there during the past three years, with whom I shared pleasant moments and unforgettable experiences.

I would like to thank the Institute Mines Telecom with the program Futur et Ruptures, and the Cominlabs excellence laboratory for the financial support of this thesis.

I would like to thank also the members of my family and friends who support me in the distance.

Finally, but most importantly, I want to thank Tanguy for his support and encouragement words in difficult moments. His company and love during this period gave me the needed strength to achieve this goal.

Abstract

Cellular networks have been traditionally designed to keep the network infrastructure always operational. This is done in order to ensure ubiquitous service availability and enough capacity to serve the peak of usage of the customers. Recently, the concern about the energy efficiency of this paradigm has increased, and a different approach has concentrated the research efforts of industry and academy. In this new paradigm, the infrastructure is dynamically adapted to the temporal and spatial traffic variations, reducing the energy wastage. The majority of these studies make the adaptation of the infrastructure unnoticeable to the users. However, with the appropriate interactivity and incentives, some users may be willing to offer their cooperation to the network.

In this thesis we consider the user cooperation in the design and control of energy efficiency techniques. We consider a specific type of cooperation in which the users are able to offset the start of their services for a bounded and known-in-advance delay. Based on proactive interaction with the users, the network may ask them to delay the start of their services if an energy efficiency technique is applied in the area where they are located (e.g. a base station is switched off). Thus, a portion of the traffic is shifted and the network can optimize the resource utilization in order to consume less energy.

First, we present an overview and classification of the literature covering the main domains of the thesis, namely energy efficiency in cellular networks and user demand shaping. We describe as well the most recent cellular network architecture - LTE. Then, we propose two different strategies to control the network resources depending on the cooperation of the users and their delay tolerance, and we evaluate the impact of such cooperation schemes in different energy efficiency techniques. Afterwards, we propose a theoretical framework for the analytical evaluation of the proposed strategies. We obtained the theoretical bounds of the attainable energy savings when employing different energy efficiency techniques, and we investigated the trade-off between the waiting time bounds proposed to the users and the energy gains. We observed that increased delay tolerance leads to more energy gains, and that the gains have an upper bound determined by the system serving capacity. We also noted that delaying opportunistically the user services depending on the system conditions is more beneficial than systematically delaying all of them. Finally, we evaluated the strategies under more realistic conditions using system level simulations. We corroborated the theoretical trends and we observed that the attainable gains are limited by the duration of the network reconfiguration process.

Résumé

Les réseaux cellulaires fonctionnent traditionnellement en maintenant l'infrastructure du réseau toujours opérationnelle, de manière à satisfaire non-seulement la disponibilité du service, mais aussi les capacités réseaux nécessaires pour gérer les pics de charge utilisateur. Ce paradigme est limité en terme d'efficacité énergétique et une approche différente est actuellement prise dans le cadre des recherches industrielles et académiques. Dans ce nouveau paradigme, l'infrastructure est dynamiquement adaptée en fonction des variations temporelles et spatiales du trafic, réduisant de ce fait les pertes d'énergie.

Dans cette thèse, nous prenons en compte la coopération de l'utilisateur dans la mise en œuvre et le contrôle des techniques d'efficacité énergétique. Nous considérons un type spécifique de coopération où l'utilisateur est capable de décaler l'utilisation de son service pendant un temps fixé et borné, et connu à l'avance. En utilisant une interaction proactive avec les utilisateurs, le réseau peut alors leur demander de décaler l'utilisation de leur service si une technique d'efficacité énergétique est appliquée dans leur zone de localisation (ex : une station de base étant désactivée). Ainsi, une portion du trafic n'est pas générée et le réseau peut optimiser l'utilisation des ressources, rester une plus longue période en utilisant un ensemble de ressources limité, et donc entrainer une moindre consommation d'énergie.

Nous présentons d'abord une vue d'ensemble et une classification de le littérature des domaines abordées dans le cadre de cette thèse : l'efficacité énergétique dans les réseaux cellulaires et l'adaptation du trafic. Ensuite, nous proposons deux stratégies différentes pour contrôler les ressources du réseau, fonction de la coopération des utilisateurs et de leur tolérance aux délais, et nous évaluons l'impact d'un tel schéma de coopération pour différentes techniques d'efficacité énergétique. Après ça, nous proposons un modèle théorique afin d'évaluer analytiquement les stratégies proposées. Nous obtenons alors les limites théoriques d'économie d'énergie atteignables par utilisation de différentes stratégie d'efficacité énergétique and nous évaluons le compromis entre les limites du temps d'attente proposé aux utilisateurs, et les économies d'énergie atteignables. Nous avons observés qu'une augmentation dans la tolérance au délai entraine un meilleur gain énergétique, et que ce gain a une limite maximale déterminée par la capacité du système. Nous avons aussi noté que retarder de manière opportuniste le service de l'utilisateur en fonction des conditions du système est plus bénéfique que de les retarder tous systématiquement. Finalement, nous avons évalué par simulation les stratégies sous des conditions plus réalistes. La simulation confirme les tendances observées avec le modèle théorique et nous avons noté que les gains atteignables sont limités par les temps d'adaptation du système lors des phases de reconfiguration.

List of Publications

Journals

- S. Gamboa, A. Pelov, P. Maille, X. Lagrange, and N. Montavont, Reducing the energy footprint of cellular networks thanks to delay-tolerant users, accepted in *IEEE Systems Journal*

International Conferences

- 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,* December 2013.
- S. Gamboa, A. Pelov, P. Maille, and N. Montavont, Exploiting user delay-tolerance to save energy in cellular network: an analytical approach, in *IEEE International Symposium on Personal Indoor and Mobile Radio Communications PIMRC*, September 2014.
- S. Gamboa, A. Pelov and N. Montavont, Changing paradigms for green cellular networks: the case of delay-tolerant users, accepted to *International Symposium on Modelling and Optimization in Mobile, AdHoc and Wireless Networks WiOpt*, May 2015.
- H. Al Haj Hassan, S. Gamboa, L. Nuaymi, A. Pelov and N. Montavont, The smart grid and future mobile networks: integrating renewable energy sources and delay tolerant users, accepted in *IEEE Vehicular Technology Conference VTC-Fall*, September 2015.

Contents

1 Introduction

1.1 Context and motivations

Traditionally cellular networks have been dimensioned for providing seamless coverage and delivering maximal performance, with little regard to the energy consumption. Recently, the focus has been moved to seek ways to increase the energy efficiency by better adapting the resources to the users behaviour. However, the users energy concerns are focused on their battery operated equipments, without paying attention to the impact their services have on the network energy consumption and the associated environmental consequences.

In some other fields this paradigm is already shifting, making the users aware of the impact their services have, and interacting with them in order to use the resources more efficiently and to achieve mutual benefits. For example, this is one of the pillars of the development of the Smart Grid. The strategy, called *demand response*, stands for managing the customers electricity demand in response to supply conditions, e.g. reduce the consumption in critical periods in order to balance the power generation and consumption, or to adapt the consumption to the electricity market prices. To do so, a communication line is established between suppliers and consumers, allowing them to interact and to take decisions to optimize their services. This approach does not substantially change the total energy demand since a large fraction of the energy saved during the load reduction period is consumed at a more opportune period, shifting the demand in time [YQST13].

A closer example in cellular networks, following the same paradigm, is called *demand shaping*. The purpose is to influence the user or application behaviour in order that the traffic is generated (or not) in certain periods and/or conditions [HSJW+12]. This technique is mostly studied to anticipate and alleviate the periods of congestion, i.e. when the network resources are not enough to satisfy a given QoS target level, for the ongoing services or the estimated incoming requests. Furthermore, this strategy is useful as well to balance the load between network equipments and to optimize the resource utilization of current deployed resources [DRSW06].

Energy efficiency in cellular networks has become a central point in recent research efforts as it became evident that the operation of the Information and Communications Technology (ICT) sector is responsible for the generation of a non-

negligible part of the global Green House Gas (GHG) emissions and global electricity consumption, contributing in 2007 with 1.3% and 3.9% of the total, respectively. Moreover, cellular networks represented at the time 15% of the ICT GHG contribution and 8.5% of the ICT electricity consumption [MML+10]. More recent studies show that the relative share of the ICT sector in the total worldwide electricity consumption has increased to 4.6% in 2012 [VLL+14]. This effect can be pronounced further with the deployment of new technologies, such as Long Term Evolution (LTE) and high-capacity dense deployments to cope with the increasing traffic demand [Cis14].

The biggest contributor to this energy footprint is the access part of the cellular network, composed mainly of base stations, which represents 57% of the total cellular network energy consumption [HHA+11]. The base stations were traditionally designed to be always operational, while consuming significant amounts of energy even when not carrying any traffic. In current standards the control information is transmitted periodically to maintain the service availability. Moreover, the access network is dimensioned to have enough capacity to support high traffic peaks which rarely occurs, as the traffic varies in time and space following the daily patterns of the customers and required services. Thus, the combination of these traditional approaches leads to resource underutilization and considerable energy wastage.

To operate the access network in a more energy-efficient way, a management paradigm consists in adapting the radio resources to the temporal and spatial traffic variations. Recently, a large number of studies, including efforts from industry and academia, have been proposed toward this idea of resource adaptation. The first step is designing base station components to be more efficient when using them, or to be deactivated when they are not needed. For example, the development and integration of adaptive transceivers and smart antennas in the base station. The adaptive transceivers allow to adapt the base station power requirements to the signal load and deactivate some components when there is no information to transmit [GFW+11]. The smart antennas allow to concentrate the radio resources in a effective area and automatically adjust the required parameters to follow the spatial variations of the traffic, increasing the energy efficiency [CPB+13].

A further step, is the development of high level energy-efficient techniques and protocols to exploit this hardware flexibility as much as possible at the different network levels: from the base station radio resource management, organizing the way the data should be transmitted in order to minimize the input power needed for transmission; to the reconfiguration of the entire access network, deactivating as much network resources as possible, e.g. switching off some antennas, sectors, entire bases stations or groups of them, and executing the required compensation techniques to avoid compromising the service availability [MSES12].

1.2 Goals and objectives

The majority of existing works on energy efficiency in cellular networks follow the classical approach of optimizing the access network under the condition of having minimal to no effect on the users. For example, a radio resource management technique uses radio fast deactivation only in periods where no data is transmitted, still requiring to transmit control and reference signals, which limits its benefits [FMM+11]. Furthermore, the network reconfiguration strategies base the decision of activating and deactivating the resources on load thresholds, which are often chosen based on strict quality constraints, in order to make the reconfigurations unnoticeable to the active users, and to prevent the dissatisfaction of expected incoming arrivals [GO13]. This results in reconfigurations happening only in very low load levels, which reduces the periods of low energy consumption and the effectiveness of the strategies.

Recent studies show that users can be more tolerant and cooperative if they are aware of the network status. The acceptance can be reinforced if they are informed about the possible benefits their cooperation can bring to them (e.g. discounts, later improved service quality) or the deterrents their non-efficient usage can cause (e.g. surcharges, poor service quality) [SBM+12b, HSJW+12] . This can ultimately lead to adapting the spatial and temporal usage patterns to more efficient ones for both, the network and the customers. Thus, approaches combining high interaction between the user equipment and the operator management entities can result in more efficient network optimizations for minimizing the energy consumption.

The objective of this thesis is to evaluate the potential energy gains a network operator can achieve with the active cooperation of its customers. In particular, we consider the case in which the users are willing to offset the start of their services for a given bounded delay. Thus, the aim of this thesis is to study if such tolerance can impact positively the execution of the energy efficiency techniques, improving their performance towards minimizing the energy consumption of the cellular network. Furthermore, we investigate the trade-off between the waiting time bounds proposed to the users and the attainable energy reductions.

However, we will not discuss the exact mechanisms of how the information should be presented to the users, neither the incentives and motivations influencing their participation. The implementation of the interactive mechanisms is also out of the scope of this work. Our analysis is focused on investigating the potential gains if the requests can actually be shifted, assuming complete knowledge of the users willingness, and controlling the energy saving strategies accordingly.

1.3 Novelty and contributions

This thesis contributes on the enhancement of energy efficiency in cellular networks by reducing the energy consumption of the access network. The novelty of our proposal lies in considering the willingness of the users to cooperate in the control loop of the energy saving strategies, with the intention of extending the periods of low energy consumption.

In particular, this thesis provides the following contributions:

- We present the basis of user awareness and cooperation for cellular network energy efficiency. Specifically, we introduce the concept of Delay Tolerant User (DTU), employed to designate the cooperative users willing to offset the start of their services for a given bounded delay. Such cooperation is verified to be helpful for optimizing the network resource utilization and reducing the energy consumption.
- We propose two strategies controlling the energy saving techniques and considering the delay tolerance of the users. The strategies react to load variations deactivating some access network resources when possible, and rely on the delay tolerance of the users to extend the periods of low energy consumption.
- We quantify the potential energy reductions attainable with the usage of the proposed strategies. We apply the strategies to different energy saving techniques, such as capacity adaptation and standalone or coordinated cell switching, and we estimate the power and energy consumption of the network if different levels of delay are proposed to the cooperating users.

1.4 Methodology and document structure

First, we overview the existing literature in the context of this thesis. The main concepts and the relevant scientific publications are presented in Chapter 2. We present the most recently deployed cellular network technology: Long Term Evolution (LTE) and the power consumption models of the different base stations forming an LTE access network. Afterwards, we describe the different approaches on energy efficiency in cellular networks and provide a general classification for them. Finally, we review the different studies considering the user awareness and cooperation for optimization purposes in cellular networks.

Second, we propose two strategies considering the delay tolerance of users to reduce the energy consumption of cellular networks, which are described, modelled and evaluated analytically in Chapter 3. We model a portion of the access network when using an energy efficiency strategy as a system which can be in different states depending on the active network resources. We characterize these states as a function of the number of simultaneous users the system can serve. The traffic and

system dynamic are modelled using Markov Chains (MC). We calculate the steady state probability distribution of each MC which represents the long term behavior of the modelled system. We deduce the required set of MCs representing the dynamic of the system when considering the DTU traffic, and applying the proposed strategies. Using the steady state probability distributions we quantify the average power and energy consumption of the system. We do this based on the power consumption models proposed in the literature, which we adapt to the considered energy efficient strategies. Then we infer the theoretical bounds of the energy savings, comparing the DTU-aware strategies energy performance to the consumption of the traditional strategies.

Third, we implement and evaluate the proposed strategies in a system level simulator, as described in Chapter 4. We developed our simulation platform based on the LTE module of the Network Simulator 3 (ns-3), which allowed us to consider a more realistic scenario for the LTE access network functioning. We implement a coordinated cell switching algorithm controlled by the proposed DTU-aware strategies, where we account for the impact the reconfigurations have on the users and on the performance of the strategies. We simulate the behavior of the network under several scenarios and compared the performance to the theoretical results.

Finally, we present in Chapter 5 the final conclusions of this thesis. We summarize the thesis outcome and identify the prospective work directions.

1

2 LTE and existing energy efficiency approaches

2.1 Introduction

In this chapter we present an overview of the most recent cellular network architecture - LTE, and the efforts of the research community to improve its energy efficiency at different network management levels. We present studies that address efficient network resource utilization, making the users aware of the impact of their services, and how such techniques can be used for improving the energy efficiency. Figure 2.1 presents a detailed map of the concepts that will be explained in this chapter, and we also position our proposal regarding the different categorized domains.

In Section 2.2, we describe the architecture of LTE. We highlight the need for structural models for the access network elements and the importance on the estimation of their power consumption. We present one of the recent and widely used Base Station (BS) component model as well as the corresponding individual and integrated power consumption figures.

Recently, a large number of studies have been proposed toward the idea of radio resource adaptation to operate the access network in a more energy efficient way. In Section 2.3 we present and explain our literature classification in this subject, based on the time scale at which the different approaches affect the access network. We describe in detail the proposed categories and we classify and characterize some of the most representative strategies we found in the literature. In Section 2.4 we analyse the operation and design of the Network Reconfiguration Strategy (NRS), which is the category of energy efficiency strategies we mainly focus in this thesis.

The majority of existing works towards energy efficiency in cellular networks follows the classical approach of optimizing the access network under the condition of having minimal to no effect on the customers. Other studies show that the customer behavior can be influenced toward a more efficient usage of the network resources. Section. 2.5 overviews these strategies mainly developed to avoid congestion, which is indirectly related to energy efficiency. We also present several studies in this field directly designed for this purpose. In Section. 2.6 we summarize the important points discussed in this chapter and we relate them to the proposal of this thesis.

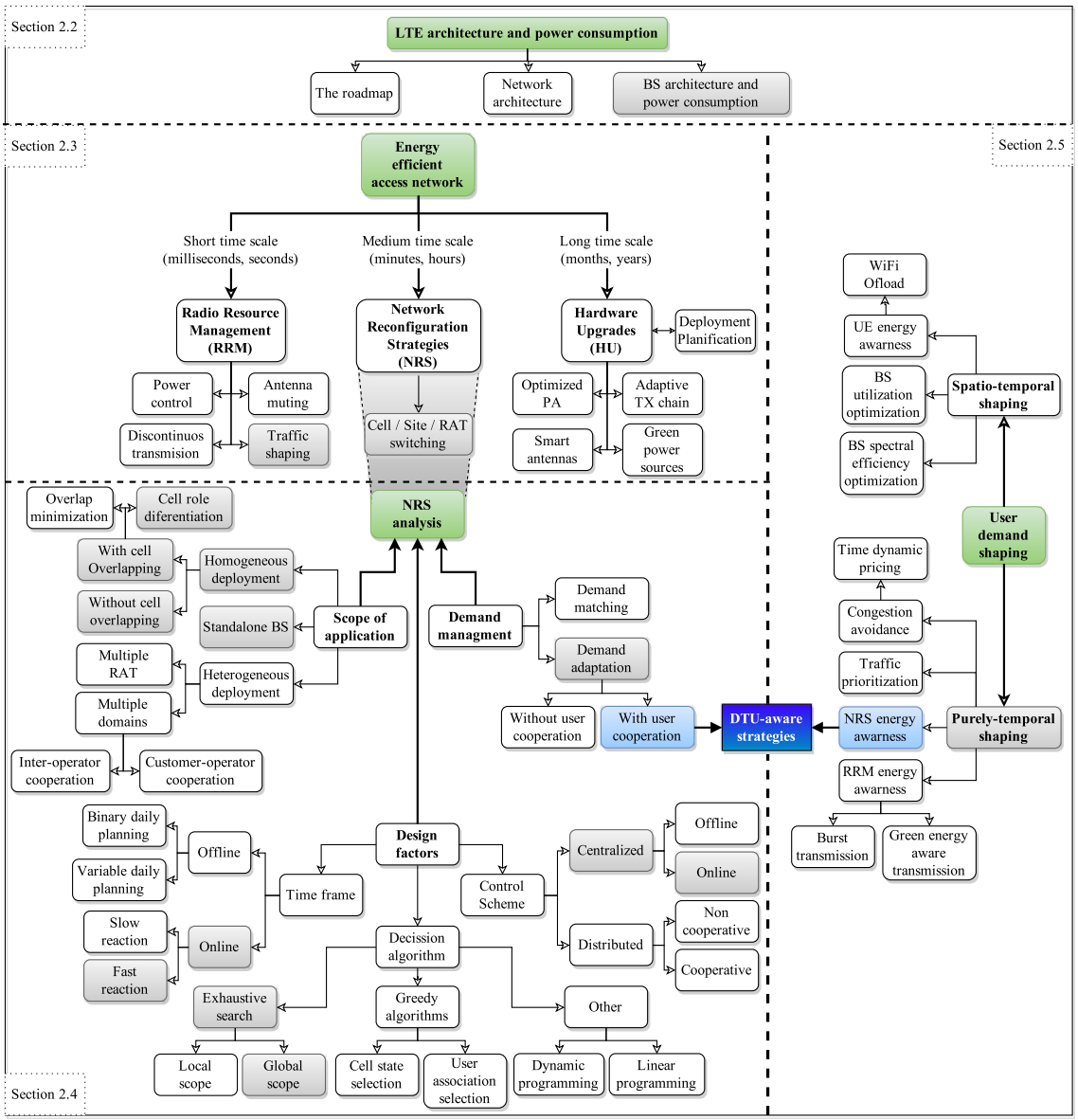


Figure 2.1: Summary of this chapter. Green boxes define the domains, which we further classify into categories. Blue boxes identify our proposal in this thesis. Gray boxes are studied in detail and used for the definition and evaluation of our proposal. The arrows indicate specialization (light arrows) or parametrization (bold arrows). Each domain is described in a dedicated section, with a detailed discussion of its classification.

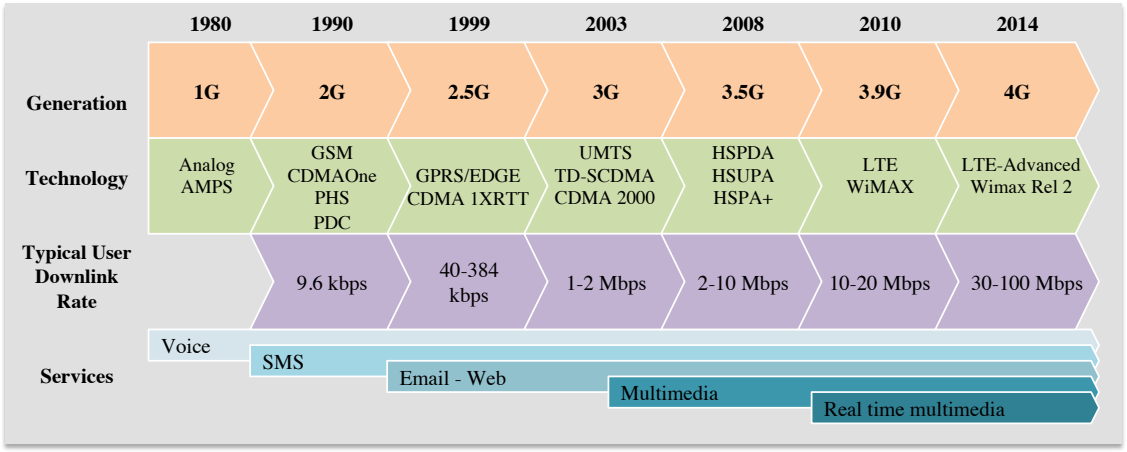


Figure 2.2: Evolution of cellular systems. Approximated release/deployment dates and typical user downlink rates.

2.2 LTE ARCHITECTURE AND POWER CONSUMPTION

The roadmap to LTE:

During the 80s, analog wireless telephone networks were introduced worldwide, representing the first generation of what is now called *cellular networks*. All along these three decades of evolution, variety of specifications, standards and products have been developed [DLLX02]. Each new generation of cellular standards is characterized by more capacity, higher data rates and broader groups of services (Fig. 2.2). The voice call service, for which the cellular networks were initially conceived, now coexists with multiple data services, such as text messaging, mobile internet access and multimedia Internet-based services. The adoption of mobile data services was initially slow, but at the end of the 2000s, the data use started to increase dramatically. The evolution of the technologies allowing higher data rates (e.g. 3G, 3.5G) and the emergence of user-friendly and application-driven end devices boost the data traffic. This trend rapidly congested the networks, leading to the clear requirement of increasing the cellular networks capacity for the generations to come.

Thus, the specification of the fourth cellular network generation focuses in providing a peak data rate of at least 600 Mbps on the downlink and 270 Mbps on the uplink, in the wider bandwidth. These requirements were published in 2008 by the International Telecommunication Union (ITU) in order to drive the development of the new technologies [Int08]. However, some technologies under development at the time did not fulfil these requirements, but still represent a significant improvement from the 3G. Thus, the so called 3.9 generation was established with two main technologies: Worldwide Interoperability for Microwave Access (WiMAX) (IEEE 802.16) and Long Term Evolution (LTE). Latter, in 2010, the ITU allowed the use of the term 4G to describe LTE, WiMAX and any other technology with substan-

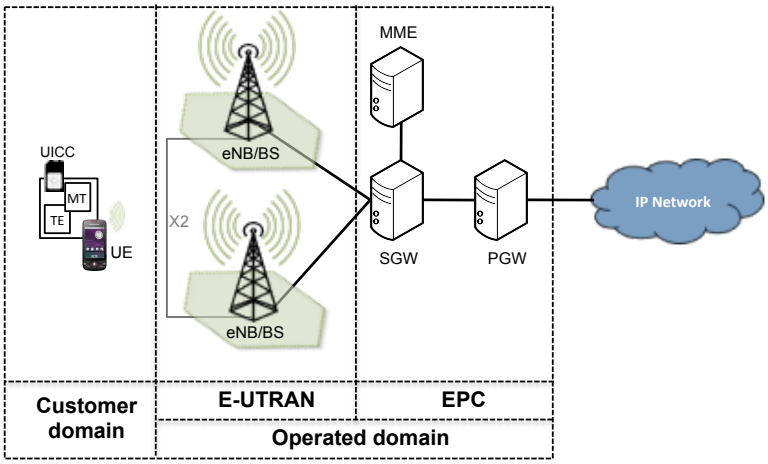


Figure 2.3: LTE architecture.

tially better performance than the 3G systems [Int10]. Further versions of these technologies, i.e. IEEE 802.16m and LTE-Advanced, are considered the real 4G of cellular system, as they satisfy the ITU requirements.

Within these two technologies, LTE has by far the greater support amongst network operators and equipment manufacturers and is likely to be the world's dominant mobile communication technology for some years to come [Cox12]. This is mostly due to the fact that WiMAX lacks backwards compatibility with previous cellular network generations, while LTE is fully compatible with most of the widely deployed 2G and 3G standards.

LTE network architecture:

The simplified reference LTE network architecture we will consider throughout this document, is depicted in Fig. 2.3. Two different domains are identified: the customer domain, composed by the User Equipment (UE), and the operated domain composed by the core network and the radio access network. The core network is called the Evolved Packet Core (EPC), while the access network is denoted as Evolved Universal Terrestrial Radio Access Network (E-UTRAN). The EPC communicates with packet data networks in the outside of the LTE network, such as the Internet or private corporate networks.

The UE is composed of three elements: the Mobile Termination (MT) which handles all the communication functions, the Terminal Equipment (TE) which generates and consumes the data streams, and the Universal Integrated Circuit Card (UICC) which is a circuit card identifying the subscriber. It runs an application called Universal Subscriber Identity Module (USIM) which keeps information about the user's phone number, home network identity and security keys.

The EPC is a packet only core network and offer connectivity and inter-system mobility with legacy access networks. The main logical nodes of the EPC are: the

Packet Data Network Gateway (PGW), the Serving Gateway (SGW) and the Mobility Management Entity (MME). The PGW provides connectivity from the UE to external networks. It is in charge of the UE IP address allocation and perform per-user-based packet filtering. The SGW routes and forwards UE data packets, manages and stores UE contexts, support the local UE mobility procedures, and provides connectivity between LTE and other technologies (2G/3G). The MME is in charge of all the control plane functions related to subscriber and session management. This includes the UE tracking and paging procedures, the establishment of the connection and security procedures between the network and UE (e.g. authentication, identification, access control, etc.) and the establishment, maintenance and release of UE service sessions. The MME also controls the other elements of the network, by means of signalling messages that are internal to the EPC.

The E-UTRAN is only composed of several types of one logical element: the Evolved Node B (eNB), which is the equivalent of the Base Station (BS) of previous technologies. However, LTE removed the Radio Network Controller (RNC) present in previous cellular generations, hence distributing the control of the access network between the eNBs. Throughout this document we will refer to the term eNB and BS indistinctly.

The E-UTRAN host several functions [3GP09] [3GP10c]:

- Radio resource management (RRM), which covers all functions related to radio bearer control, radio admission control, connection mobility control and dynamic allocation of radio resources (scheduling) to UEs in both uplink and downlink.
- Subscriber and equipment trace and positioning
- Connection setup and release
- Measurement and reporting configuration for mobility and scheduling.
- IP header compression and encryption of user data stream
- Routing of user plane data towards SGW and the other way around
- Scheduling and transmission of paging messages (originated from the MME)
- Self Configuring and Self-Organising Networks (SON) functionality, which covers coverage and capacity optimization, inter-cell interference coordination, mobility robustness optimization, load balancing and energy savings procedures

BS reference architecture and power consumption:

As stated by Han et al. [HHA+11], the access network and its principal (and unique in LTE) component, the BS, is the most power consuming part of cellular networks. However, quantifying the individual power contribution of each BS is not a trivial task for several reasons: first, in order to protect their industrial designs, equipment manufacturers are not likely to reveal the architecture of their products, to the point of forbidding the operators to explore their devices components, which prevents the fully understanding of their power contribution. Second, manufacturers

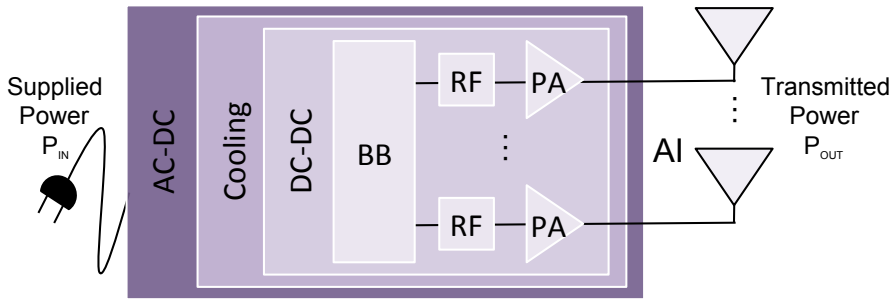


Figure 2.4: LTE BS architecture [AGD+11].

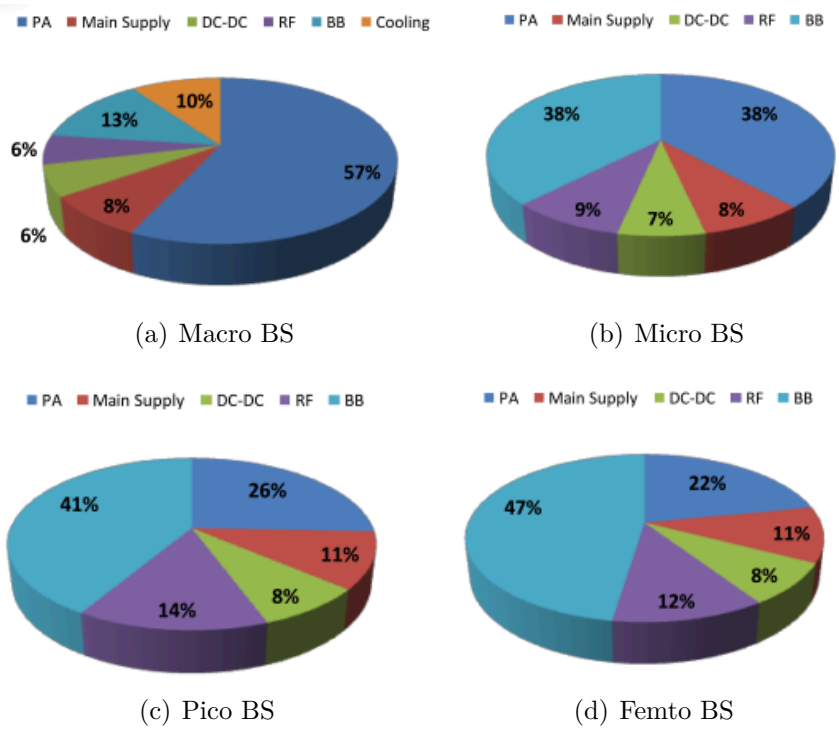
have different designs, such that a power model may be valid only for a single brand or type of BS. Third, architectures and technologies continually evolve, limiting the long-term applicability of power models, e.g., a particular model may only be valid for a certain generation of BSs [HAH11].

Considering these issues, the research community, including efforts from industry, academy and manufacturers, agreed in a general common architecture for a LTE BS without revealing competitive technical differences. This is one of the contributions of the project EARTH to the research community. Moreover, actual measurements of the power consumption allowed to the project members to quantify and model the power consumption of the different types of LTE BSs. These results and derived models are presented by Holtkamp [HAH11], and were also published as part of the EARTH project deliverables [EAR12c], and in abbreviated form by Auer et al. [AGD+11]. Since then, these models have been widely adopted as reference models when evaluating power and energy reduction strategies in cellular networks, and they are used in the following chapters of this manuscript as well.

The general architecture of an LTE BS is showed in Fig. 2.4. This reference architecture aims to define a high-level block diagram with the main radio hardware components that can be generalized to all BS types. A BS may contain one or multiple Transceiver (TRX), each of which serves one antenna. A TRX chain has the following components:

- Antenna Interface (AI)
- Power amplifier (PA)
- Radio Frequency (RF) small-signal transceiver module
- Baseband (BB) engine: receiver (uplink) and a transmitter (downlink) section
- DC-DC power supply
- Active Cooling system
- AC-DC unit (Main supply) for connection to the electrical power grid

A BS site has one or more sets of antennas, through which it communicates with the UEs in one or more *sectors*. A sector can contain several antennas, e.g., when using MIMO [4GA12]. In cellular networks, the word *cell* can be used in two



(a) Macro BS (b) Micro BS

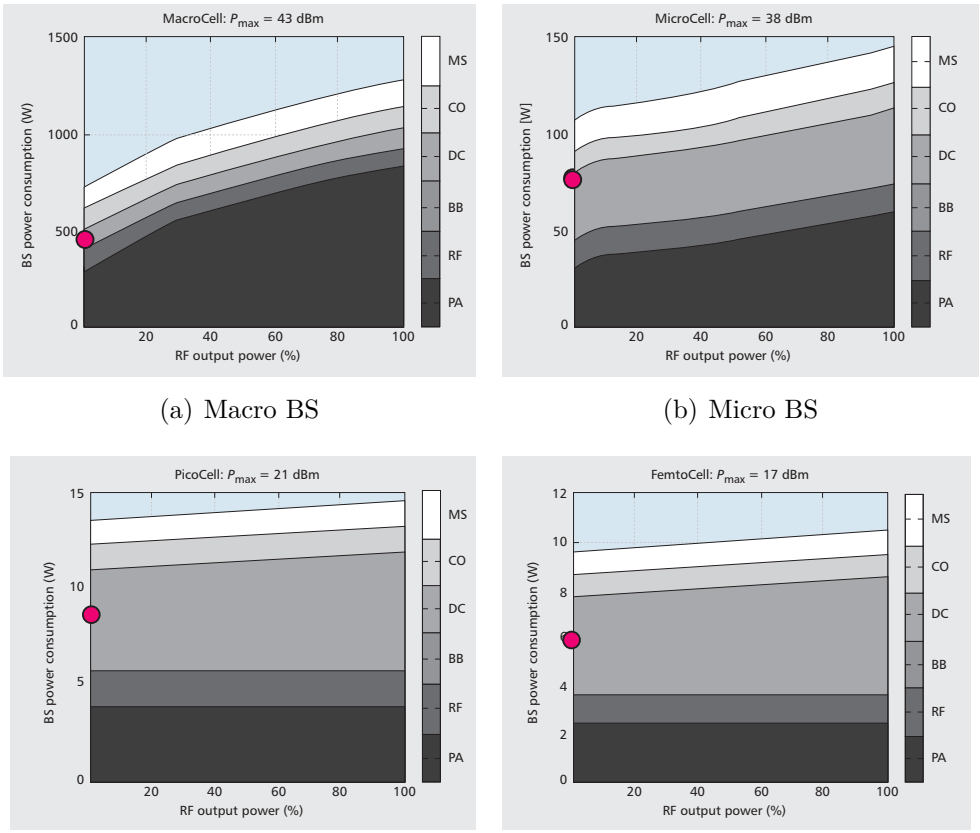
(c) Pico BS (d) Femto BS

Figure 2.5: BS power consumption breakdown for the different types of LTE BSs. Source: [EAR12b].

different ways: to refer to a group of sectors controlled by the same BS, which is the convention used in USA; or as a synonym of sector, as used in Europe [Cox12]. We will use the latter definition and we will refer to a sector and a cell indistinctly.

Different types of BS are considered in the literature, depending – among other features – on the radiated power at the antenna P_{out}, which determines the BS coverage range [3GP10a]. BS covering a wide area (e.g., >500m) are referred as Macro BS. Medium range coverage (e.g., 250m) is provided by Micro BS, while local area coverage (e.g., 100m) is provided by Pico BSs. Short range and indoor coverage is provided by Femto BSs.

The different types of BS imply different component selection within the general architecture of Figure 2.4. Energy efficiency also varies among BS types, for example small BS (pico and femto) may use more efficient and dedicated components, while large BSs (macro and micro) may require more reconfigurability, for example using more programmable and less energy efficient integrated circuits (e.g. FPGA) [DDG+12]. A further evolution of the wide range BS is the Remote Radio Head (RRH) BS. In this type of BS the PA is located close to the AI and is connected to the BB by means of an optical link, allowing a physically distributed antenna architecture and avoiding the power losses in the RF feeder cables. Usually, the cooling system is avoided in this type of BS as the PA is cooled by natural air circulation.



(a) Macro BS (b) Micro BS

(c) Pico BS (d) Femto BS

Figure 2.6: LTE BS power consumption depending on the signalling load. Source: [AGD+11].

A further evolution of the RRH are the Active Antenna Systems where the RF is also collocated directly next to the radiating antenna elements [4GA12].

The BS power consumption breakdown for the different LTE BS types are depicted in Fig. 2.5. These values are calculated when the BS is working at full load. However, a part of the power consumption varies depending on the signal load. In the literature this part is often called the dynamic power consumption [HBB11] [ARF10]. In LTE systems, the downlink transmission scheme uses orthogonal frequency-division multiplexing (OFDM). Thus, the radiated power at the antenna P_{out} depends on the BS physical resource allocation in the downlink and the corresponding generated signals. The evaluation presented by Auer et al. [AGD+11] and depicted in Fig. 2.6, shows that mainly the PA scales with the BS signal load and the corresponding output power. Moreover, this largely depends on the BS type, mainly impacting the Macro and Micro BS, because the PA represents a considerable part of their total power consumption. The work presented by Auer et al. [AGD+11] provides a linear power model that relates P_{out} with the total power needed by the

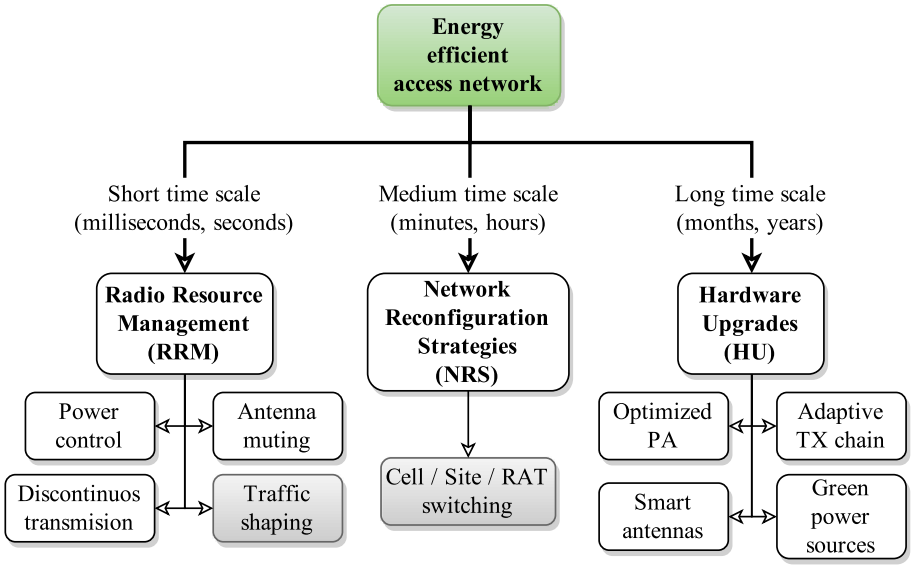


Figure 2.7: Energy efficiency strategies classification.

BS to operate (P_{in}). This work was extended by Desset et al. [DDG+12] providing power models for the different BS components and types to further capture the evolution of the E-UTRAN and the combination of different BS architectures.

2.3 Energy efficient access networks

The existing approaches on energy efficiency rely on the adaptation of the operated part of the cellular network to the traffic load variations, in order to avoid unnecessary energy consumption. For example, changing the BS transmission settings according to the traffic levels, or even changing the access network layout, i.e. determining which BSs should remain active and which can be deactivated.

The following surveys, cited in chronological order of publication, show the evolution and the increased interest of the research community in adaptive cellular access networks: Correia et al. [CZB+10], Hasan et al. [HBB11], Suarez et al. [SNB12], Budzisz et al. [BGR+14], Rao et al. [RF14], and De Domenico et al. [DCC14]. Deliverables of research projects, such as EARTH, are also a condensed literature source in this field [EAR12c] [EAR12d]. We developed a new classification of the energy efficient approaches, using as criteria the time scale in which the different strategies operate. This criteria indirectly determines the part of the network the strategies affect.

Figure 2.7 summarize the main categories of our classification. We found that the improvements concerning the hardware itself affects the cellular network in a *long time scale*, i.e. months, years, as it implies the integration of new hardware

to the network. Some of the studies in this category denoted as Hardware Upgrades (HU) are described in Section 2.3.1. Fast adaptation mechanisms affect the network in a *short time scale*, in the order of milliseconds to seconds, reacting to instantaneous load variation. This category includes the Radio resource management (RRM) strategies, which decide for each instant which of the hardware resources the BS should use. Some of the RRM strategies designed for power consumption reduction are described in Section 2.3.2. Finally, approaches operating in a *medium time scale*, affect the access network in the order of tens of seconds, minutes to hours. They are denoted as *Network Reconfiguration Strategy (NRS)* and adapt the access network to temporal and spatial load variations. We overview this category in Section 2.3.3, while a more detailed review about the main characteristics of the NRS as well as their design factors and representative literature is presented in Section 2.4.

2.3.1 Hardware Upgrades (HU)

Traditionally, the BSs were designed to be always operational, i.e. all the components are active all the time. In order to satisfy the increasing capacity need, more and more BS will be deployed in the coming years [And13], which increase the energy consumption of the whole access network. The first step to face this situation and control the energy consumption to sustainable levels, is designing BS components to be more efficient when using them or to be deactivated when they are not needed. These HU are integrated to the BSs and remain unchanged for long periods after the BS deployment, in some cases throughout the lifespan of the BS. A further step, explained in Section 2.3.2 and 2.3.3, is the development of higher level management techniques and protocols to exploit these features as much as possible.

The most consuming element in macro and micro BSs is the PA. The energy efficiency of the PAs can be optimized for full signal load and maximum output power, thanks to the use of signal conditioning techniques: Digital Pre-Distortion (DPD) increases the PA linearity and Crest Factor Reduction (CFR) reduces the Peak-to-Average Rower Ratio (PAPR) [Xu10]. However, when the signal load decreases different levels of output power are possible, and the PA not always requires the high levels of input power needed in the full load regime.

Gonzalez et al. [GFW+11] propose an adaptive transceiver chain for macro BSs. The architecture of the PA, RF and the DC-DC power supply is designed to allow *signal load adaptivity* as well as *fast component deactivation*. This architecture requires an additional component: the Digital Signal Processing and Control (DSPC). It analyses the signals coming from the BB unit and controls the other elements of the transceiver chain to adapt their operation to the instantaneous signal load. The PA is adjusted to the signal load using adaptive clipping, which generates different levels of PAPR, allowing the reduction of the required power supply voltage to reach an efficient operating point for the carried signal load [FBZ+10]. The PA and the

RF are enabled to fast activate and deactivate some of the transmit blocks during empty symbol periods. The DC-DC power supply is optimized to provide the set of voltages required by the adaptive PA, and it is able to switch between them rapidly and efficiently. Experiments for the validation of the adaptive transceiver chain are presented by EARTH [EAR12e], showing significant power consumption reductions for the different load regimens.

Another hardware upgrade for LTE systems is the use of adaptive or "smart" antenna systems. One of the characteristics of these systems is the use of multiple antennas per sector, also known as Multiple-Input and Multiple-Output (MIMO). When all antennas are active, several transmission modes are possible, which allow serving UEs under different radio conditions and take advantage of the multipath rich environment. The BS scheduler has the capability to optimally select the MIMO scheme that suits the UE channel conditions. The presence of several antennas allows to improve the performance in poor signal conditions, in order to exploit *spatial diversity*: the same signal stream is transmitted from each antenna but with different coding/frequency applied, increasing the robustness of the received signal due to redundancy. In good signal conditions the *spatial multiplexing* creates a number of independent transmission channels between the transmitter and receiver, which enables two or more different signal streams to be transmitted simultaneously, which increases the throughput and the transmission energy efficiency.

Another characteristic of the smart antennas is the *Beamforming*, i.e. to adapt the radiation patterns of the transmitted radio signals. When beamforming is used in medium time scale (minutes to hours) the purpose is to change the shape of the cell and its coverage area by means of modifying the beam tilt in elevation, beam pointing and beam width in azimuth. The purpose is to dynamically adapt the link budget of the BS in order to achieve higher level of spectral efficiency in a given area, during a given period. Beamforming algorithms that are active on a small time scale (ms–scheduling) allow concentrating the antenna radiated power on a per-user basis, using adaptive weights in the antenna configuration parameters which are updated thanks to the UE feedback, reducing interference and overall power consumption requirements [CPB+13].

While new BSs are designed to be more power efficient, another trend in the access network research is the use of alternative energy sources to power them. Coupling cellular BSs to Renewable Energy (RE) sources can reduce the power grid electricity consumption, support the cellular network in case of power grid failure and allow cellular coverage in areas with no/limited power grid connectivity (i.e. islands, deserts, etc.). The main constraint in the use of RE for powering BSs is the electricity generation intermittence. For example, the solar panels performance is determined by the sunlight intensity, cloud coverage, smog, air density, etc; the wind energy production can be affected by the air temperature, mechanical obstructions, altitude, etc [HNP13]. To ensure the reliability of the BS using RE, backup methods are usually employed. Batteries are used to store the excess in the energy production,

and to provide the required energy in periods of production deficit. Another solution is to employ backup diesel generators for the off-grid BS operation in periods when RE is not available. In places with power grid connection, the BS can operate in one or both modes, on-grid and off-grid, depending on the conditions and the operator policies. The effective deployment of RE BSs strongly depends on local information. Some studies in this field focus in finding the appropriated dimensioning of the RE hardware infrastructure (photovoltaic panels, wind turbines, etc.) depending on the location conditions. For example, Paudel et al. [PSN+11] studied the deployment of RE powered BSs in Nepal, and Moury et al. present a study made for Bangladesh [MNH12]. Piro et al. [PMF+13] present a general analysis about the benefits a cellular operator can achieve integrating RE infrastructure in their networks. The study shows promising long term cost and CO_2 emission reductions, encouraging operators towards this practice. Han et al. [HA14] present an overview of research issues related to optimal usage of RE powered BSs. For example, offloading the users to RE powered BSs, prioritizing the RE powered BSs in cooperative transmissions schemes, or adapting the BSs link budget and the access network layout depending on the RE availability.

2.3.2 Radio Resource Management (RRM)

The purpose of the RRM algorithms and protocols is to efficiently utilize the limited spectrum and the BS hardware resources, using adaptive techniques while satisfying the users Quality of Service (QoS) needs [PK09]. The RRM strategies determine when and which of the physical resources (transceiver chains, antennas, subcarriers, transmission power, etc.) the BS should use to send the traffic over the wireless link.

One energy efficient RRM strategy is the Power Control (PC). This technique adapts the transmission power radiated at the antenna depending on the channel conditions. Besides reducing the power consumption, this technique is also beneficial to link adaptation and interference reduction. Kivanc et al. [Kiv03] propose a sub-carrier allocation strategy which reduces the overall power consumption by means of a power control strategy.

As presented in Section 2.2, the major power consumption of an operational BS is independent of the transmission power. Thus, recent energy efficient RRM techniques take into consideration the new flexible hardware features explained in Section 2.3.1. The downlink Discontinuous Transmission (DTX) consists in an interruption of signal transmission, allowing the momentary component deactivation of the transceiver chain or a part of it. If the interruption is short enough, it can be unnoticeable to a receiver. Frenger et al. [FMM+11] were the first to apply this concept to a LTE OFDM subframe. During the periods where there is no data to transmit DTX is applied. However, the control signals are transmitted frequently

within each subframe, with service availability and synchronization purposes. Hence, the authors went a step further by proposing the scheduling of Multicast and Broadcast Single Frequency Network (MBSFN) sub frames in periods of low load. The MBSFN have reduced control signal overhead which increase the overall attainable power reductions if DTX is used. Holtkamp et al. [HAH11] propose a downlink resource allocation scheme that combines PC and DTX, such that downlink power consumption is minimized while still upholding the required user QoS .

The adaptation of the number of active antennas is another RRM strategy considered for power and energy reductions. Kim et al [KC09] consider to switch the transmission mode between MIMO and SIMO antenna schemes in the UE to reduce their energy consumption when lower data rates can be used. A more general approach for the BSs is presented by Skillermark et al. [SF12], where the concept of *antenna muting* is proposed: one or several antennas of a MIMO system can be deactivated, reducing the power consumption of the BS but impacting the achievable capacity due to the reduction of the set of transmission modes the BS can use.

Gupta et al. [GS12] propose the addition of a Traffic Shaping Module (TSM) to the RRM schedulers to reduce the BS power consumption. The purpose of the TSM is to buffer data to further transmit it in bursts, thus providing longer DTX periods between burst. On the contrary, the *capacity adaptation* technique spreads the data in time, in order to reduce the instantaneous needed bandwidth and adapting the operating point of the PA to the signal load to reduce the power consumption. Results presented by EARTH [EAR12d] show that the performance of this two traffic shaping schemes depends on the traffic conditions and the considered deployment.

2.3.3 Network Reconfiguration Strategies (NRSs)

The purpose of a Network Reconfiguration Strategy (NRS) is to adjust the access network configuration to temporal and spatial load variations, in order to reduce the overall energy consumption. The difference between the RRM techniques and the NRSs is the time scale at which the reconfiguration is done and the network scope it impacts. The RRM techniques react to almost instantaneous load variations, in the order of milliseconds (e.g. LTE OFDM subframe), and they affect only the cell applying the technique, without requiring reactions of the neighboring cells. NRSs react on a longer time scale, e.g. several seconds or minutes to hours, and may require coordination and cooperation of a group of cells to preserve coverage and service availability.

The NRSs are denoted with different terms in the literature. For example, cell/BS switching (on/off) [MCCM12, OSK13], cell/BS sleep mode [ESC11, GWOF13], cell breathing [MMS10], cell zooming [NWGY10], cell wilting and blossoming [CFC+11], dynamic sectorization [HG11], among others. However, all these approaches share the same principle for energy efficiency in the access network, which is to deactivate

(partially or completely) unneeded radio resources (e.g. cells, complete BSs, group of BSs, complete access networks) when possible, in order to adapt the active resources to the traffic demand in a specific geographic area. The traffic demand can vary in two dimensions: temporal and/or spatial. Human activity is the principal factor affecting the temporal variation of the traffic load. Naturally, during the late night periods the load is very low, while during the day the usage increases considerably, with some noticeable periods of very high load. The latter are often called the *busy hour* and cellular operators select, position and configure their BSs to have enough capacity to be able to afford the traffic in these scarce periods. Mobility and customer routines are the principal reasons for the load spatial variation. BSs located in business areas exhibit different load patterns than the BSs located in residential areas. Moreover, social places such a shopping centers or restaurants, and social events such as concerts or sport encounters, create temporary customer concentrations where the load served by the cellular networks is very high.

The design of a NRS has three challenges: minimize the access network energy consumption, guarantee the service availability over the area, and minimize the perceived degradation of the user experience. The general procedure of the execution of a NRS is as follows. Firstly, the required information about the afforded traffic load is retrieved by means of measurement, historical data or estimations. Then, this information is processed to determine if there is the need or the opportunity of executing a reconfiguration in the access network. If so, the different available reconfiguration options are evaluated and the appropriated configuration is selected depending on the target performance level.

A network reconfiguration is needed when the current network state cannot satisfy acceptable levels of performance. If the load increases, some cells may need to be activated, as there is the risk of not satisfying the pre-defined and acceptable service quality for the active users in the network and the expected new arrivals. If the load decreases, some cells could be deactivated, as the load can be handled by some other cells which remain actives. This cell activation/deactivation is usually done conservatively in the literature, triggering the reconfiguration process on load thresholds that are still manageable by the network, in order to prevent an unaffordable trend.

Once the reconfiguration need is identified, the NRS algorithms should determine which of the different reconfiguration options is the most appropriated to perform given the pre-defined performance target and the system constraints. Depending on the scope that the algorithm controls (i.e. the number of cells), the hardware degrees of freedom (i.e. the number of configurable parameters), and the user and load spatial distribution, a large number of reconfiguration possibilities may be available. The impact of the different reconfigurations is evaluated estimating the values for different performance metrics if the reconfiguration was applied, either considering the current state of the network, or estimating the future states given the load trend. This is usually formulated as an optimization problem where the

objective is to find the set of reconfiguration parameters that minimizes the network energy consumption, while satisfying all the performance metric constraints.

In this thesis we mainly focus on this type of techniques. Thus, in the next section we present in more detail the different NRSs we found in the literature as well as the main hypothesis and design factors for their formulation and evaluation.

2.4 Network Reconfiguration Strategies analysis

A Network Reconfiguration Strategy (NRS) adapts the access network layout and capacity depending on the traffic conditions. To do so, the set of active radio resources changes over the time. The NRSs determine which resources should be activated or deactivated and when a given configuration should be applied. An important criteria to determine when designing or studying a NRS is the scope of application. This will determine which resources the NRS can control, e.g. some NRSs control the state of the transmitters and cells within a single BS, while some other NRSs control multiple BSs of the deployment or even all BSs of the considered access network.

When multiple BSs are considered, the architecture and deployment arrangement determine the degrees of freedom the NRS can have, e.g. when the BSs are densely deployed allowing cell overlapping, the number of candidate cells for deactivation is high, as well as the complexity of the reconfiguration decision process; while in a coverage constrained scenario, the set of candidates is small but techniques for preserving the service availability should be employed. Section 2.4.1 describe more in detail these issues as well as how they are addressed in the reviewed literature.

A reconfiguration decision is taken when the current access network configuration cannot satisfy the required performance, either in terms of energy consumption, or in terms of service quality. A system with all resources activated but no customer activity is not optimal in terms of energy, while a system with just a few resources activated but high demanding customer activity is not optimal in terms of service quality. Thus, a trade-off between these performance metrics should be found by the NRSs. In Section 2.4.2 we describe how these metrics are defined in the literature.

Section 2.4.3 present the different criteria we found in the literature which determine the NRS operation. For example, which entities are capable of taking a reconfiguration decision, how the traffic variations are considered, how frequent the system is evaluated to determine a reconfiguration need and how the different reconfiguration possibilities are evaluated to choose the most suitable given the system state.

We present in Annex A a complementary detailed table with a summary of the reviewed literature, pointing for each work the different criteria the authors consider within the classification provided in the following sections.

2.4.1 Scope of application

The type of BSs in the network, as well as their architectural and logical configuration, define different points to consider when designing and evaluating NRSs. In this section, we first present the studies concerning NRSs acting on a local BS scope, i.e. controlling only the parameters of a single BS site. Afterwards, we present the strategies acting on a global access network scope. We divide the types of deployment depending on the number of access network layers the NRS should control: if the deployment is *homogeneous*, the NRS affects only one access network layer, while in *heterogeneous* deployments the NRS can control multiple access network layers. Each of these categories is further divided depending on the architectural characteristics of the different deployments and the approaches found in the literature to apply NRSs on them. In Table 2.1 we present a summary of the criteria we use to characterize the scope and deployment types considered in the literature, as well as the references to the representative reviewed works for each of them.

2.4.1.1 Standalone BS

Some studies focus on the design of NRSs for individual BSs in order to adapt its resources to the local traffic condition within the BS scope, without interaction with the rest of the network, i.e. no actions are requested or performed by neighboring BSs, and no considerations are made about the impact of the reconfiguration in the network global scope. The BS resources controlled by the NRS vary depending on the technology and the available hardware. For example, Saker et al. [SES09, SE10, SEC10] and Elayoubi et al. [ESC11] present a general analytical model of a NRS strategy intended for adapting the number of active GSM transmitters and 3G active carriers depending on the traffic afforded by the entire BS. Each of these resources is seen as a portion of capacity, and the authors focus on providing activation/deactivation policies to maintain the probability of user dissatisfaction low, e.g. triggering the activation of another resource in loads still manageable by the current active ones, in order to conservatively prevent lack of capacity when serving the time varying traffic.

Tomaselli et al. [TSPB13] go a step further and test their proposed NRS in a controlled test bed environment, considering an operational BS. The authors set up such BS with three overlapped cells, which are deactivated or activated depending on the number of active UEs in the system. The authors show that the reactivity of LTE BS hardware is adequate to perform fast activation/deactivation of the cells and that the access network management equipment can be used to implement automatic cell switching schemes, as it provides traffic and cell status monitoring, as well as the required interfaces for implementing the switching procedures. However, the impact in the QoS of the served UEs was not investigated by the authors.

Table 2.1: Summary of the reviewed NRSs literature, classified by the NRS scope of application, as well as other related relevant criteria.

Criteria				Representative references
Scope of application	Standalone BS			[SES09] [SE10] [SEC10] [ESC11] [TSPB13] [HG11] [HHA+13] [HAB+13]
	Homogeneous deployment	Without cell overlapping		[CCMM08] [CCMM09] [ACCM09] [SKB10] [STKB11] [RRAF13] [GO12] [GWOF13] [HSL13] [TGA13]
		With cell overlapping	Considering overlap minimization	[ZGY+09] [OK10] [OKLN11] [NWGY10] [Niu11] [MMS10] [CZZN10] [GZN12] [GWOF13] [CLHS14]
			Considering cell role differentiation	[CCMM09] [SMES12] [MSES12] [GWOF13][HMJ11] [MCCM11] [CFC+11] [STKB11] [CFGU12] [MCCM12] [CJXH13] [OSK13] [SNB14]
	Heterogeneous deployment	Considering multiple RATs		[SES09]
		Considering multiple domains	Inter-operator cooperation	[HMJ12] [MM13] [OKLN11]
			Customer-operator cooperation	[SNB14]

Some other research work consider not only the BS capacity issues of deactivating/activating a resource, but also the spatial implications of it. For example, Hevizi et al. [HG11] consider a BS covering a given area with a set of different cells/sectors, each of them configured with directional antennas, taking care of a portion of the BS covered area in an almost not overlapped fashion. The authors propose a NRS in which depending on the traffic load served by the entire BS, some of these sectors may be deactivated. To maintain the BS coverage the remaining active sectors should reconfigure, mainly changing their beam form and other antenna parameters in order to enlarge or shrink the spatial coverage of the radio resources.

Cellular operators offer almost universal service availability in urban and suburban areas, and the NRSs should be able to continue ensuring it when operating in such scenarios. However, there are some other critical scenarios where the priority is not the universal service availability but the energy consumption. Heimerl et al. [HHA+13, HAB+13] design and implement a high range BS with minimal energy

consumption using virtual coverage: the BS is activated only on demand to satisfy users request. The system was deployed in rural Papua, Indonesia, contributing to bring service availability in zones with difficult access to electricity, i.e. an off-grid village powered only by solar panels and diesel generators.

2.4.1.2 Homogeneous deployments

Homogeneous deployments are composed of several BSs of the same technology, belonging to a unique layer, controlled exclusively by one operator. They are deployed over the service area to provide universal coverage and support seamless mobility. Depending on the arrangement of the BSs and the radio configurations, the cells of an homogeneous access network may overlap.

In the case of a *Non-Overlapping* configuration, cells are set up according to careful radio resource planning in order to not overlap their coverage zones. In this kind of deployments the critical factor to consider by the NRSs is the service availability. The deactivation of a resource in this coverage-limited scenario may compromise the service availability and/or the minimal service quality if the appropriated complementary actions are not performed, e.g. the execution of a coverage preservation technique (Section 2.4.2.1). Chiaraviglio et al. [CCMM08, CCMM09] evaluates the feasibility of a deactivation scheme estimating the propagation limits of the remaining active cells depending on their radio characteristics (e.g. transmission power). If these limits are over a minimal required threshold, the deactivation pattern is feasible as the coverage is maintained. This estimation is widely used in the literature to check if the coverage constraints are met when cells may be deactivated in a Non-Overlapping deployment configuration, e.g. [SKB10, STKB11, RRAF13]. In addition to the propagation limits, a minimal signal quality within the cells boundaries is established by some authors for considering the coverage constraint satisfied in a Non-overlapping deployment. For example, Guo et al. [GO12, GWOF13] and Han et al. [HSL13] consider that a deactivation scheme is acceptable if the SINR is above a certain minimum threshold, which is required for a predefined minimal data throughput in the cell boundaries of the cells that will remain active.

In the *Overlapping* case, the coverage area of different cells can intersect, normally using different parts of the wireless spectrum and/or smart coordination in order to avoid interference and support seamless mobility between cells. In this scenarios, the coverage can be maintained in different ways, which makes the deactivation/activation schemes more complicated to select. We identify two kind of approaches in the literature for doing this selection when considering overlapping scenarios.

The first kind of approaches aims at minimizing the overlapping condition considering all cells as candidates for the deactivation. For example, Oh et al. [OK10, OKLN11] investigate the optimal BS density and parameters to transform an overlapped scenario in a non-overlapped one. The authors also present an estimation

of the overlapping reduction attainable in a real deployment – a portion near the city center of Manchester, United Kingdom. The authors sequentially deactivate the BS with the minimum distance to the nearest active neighbor, and show that using the same fixed cell size for all BSs, a considerable number of BSs can be deactivated (between 10% and 70% depending on the radio configuration), while still maintaining the coverage constraints dictated by propagation limits. However, in dense urban scenarios, such as the city center of Manchester, the access network is dimensioned accounting the traffic spatial distribution in the busy hour, which results in different overlapping BSs types with a large variety of cell sizes. Thus, the uniform deployment proposed by the authors may cause traffic imbalance, neglecting the traffic spatial variations. Another strategy for minimizing the overlapping is presented by Zhou et al. [ZGY+09] and Niu et al. [NWGY10, Niu11]. The authors propose to dynamically adapt the cell sizes in order to distribute the traffic between the overlapped cells, concentrating the traffic in cells with high load, which makes possible to deactivate the low loaded ones. Such strategy account the traffic spatial distribution but may require coordination in order to avoid congestion and user dissatisfaction in the cells which remain active.

The second kind of approaches divide the cells into two sets: critical cells, which will remain active to maintain coverage, and flexible cells, which can be deactivated when required [BGR+14]. This can be done in a fixed way, i.e. the NRS should only chose when and which flexible BSs should be activated/deactivated; or in a variable way where the set of critical and flexible cells may change depending on the network conditions. These approaches are sometimes taken when considering small BSs, e.g. Micro, Pico or Femto BSs deployed under the coverage of a Macro BSs. Moreover, it is usually assumed that the Macro BS is the critical one guaranteeing coverage. For example, Saker et al. [SMES12] study the impact of the deployment of Pico cells under the Macro coverage to increase the capacity of the network. For reducing the energy consumption, the authors consider that only the Pico cells are candidate to be deactivated in periods of low load. However, as pointed by Micallef et al. [MSES12] and Guo et al. [GWOF13], in zones when the density of small cells is high or when they are conscientiously deployed, the required coverage can be provided by them, and may be convenient to deactivate the Macro BSs when appropriated for minimizing the energy consumption. Samdanis et al. [STKB11] propose to dynamically select which BSs are critical and which are flexible depending on the spatial load distribution. The authors select the high loaded BSs as critical BSs, as they have less probability to be deactivated. This approach is also used when considering distributed NRSs where the cells decide themselves their status. For example, in the NRS proposed by Oh et al. [OSK13] the neighboring cells transactionally agree in their roles, either a cell is allowed to be flexible and deactivate, or it is allowed to remain active to take care of the traffic and coverage.

Capone et al. [CFGU12] go a step further and propose an overlapped deployment architecture for future networks, in which high range BSs are deployed only with

signalling and control purposes and hence they cannot be deactivated. In this architecture a dense small cells layer is deployed with service purposes, and each cell is activated or deactivated depending on the user activity detected by the high range BS. If the implementation barriers are overcome (e.g. synchronization, backhaul architecture constrains and bottlenecks), this approach can be highly efficient, as the cell reactivity dimension is reduced, i.e. the small cell coverage may also represent small number of users to serve and to react to, which makes the energy consumption more proportional to the carried traffic. Moreover, only few high energy consuming BSs will be needed in the network, reducing the fixed part of the access network energy consumption.

2.4.1.3 Heterogeneous deployments

The NRSs considering heterogeneous deployments may control or make use of different layers of access nodes, which may belong to different technologies or administrative domains. The principal factor the NRSs needs to consider in this kind of deployment is the traffic distribution between the different layers, which is often conditioned by the capacity difference between layers, the user attachment capability and/or the administrative rights.

Some studies focus on allowing the interoperability between different Radio Access Technology (RAT) (e.g. 2G/3G/4G). In this case all layers are under the administrative control of one operator, and the BSs are often collocated, which facilitates the NRS control. For example, Saker et al. [SES09] studied the load distribution between 2G and 3G technologies to minimize energy consumption while preserving the QoS. The authors point that legacy technologies are less energy efficient, consuming more energy while providing less capacity. Thus, the intuitive and more efficient approach in such heterogeneous deployment is to deactivate the 2G BSs as much as possible. This highlights the importance of having compliance between the cellular generations to be able to successfully apply such NRS without compromising the service availability for the users using a given technology.

When considering different administrative domains two cases are considered in the literature: the studies considering that the NRS can control BSs from different operated access networks, i.e. inter-operator cooperation; and the studies that consider the NRS can make use of some customer BSs (e.g. Home Evolved Node B (HeNB) or WiFi access points) to offload the operator traffic, i.e. customer-operator cooperation. Issues about migration of the energy consumption and quality of service of the customers should be considered in the design of NRS treating with this type of network inter cooperation.

In the case of inter-operator cooperation, we found two different approaches for applying NRSs in the reviewed literature. On one hand, Oh et al. [OKLN11] and Hossain et al. [HMJ12] apply the NRS in this heterogeneous scenario as if it was

a big overlapped homogeneous scenario taking in consideration all the BSs of the different operators. Thus, BSs belonging to different operators can be active at the same time. On the other hand, Marsan et al. [MM13] considers to progressively deactivate and activate the complete set of BSs belonging to a given network, so that BSs belonging to a given operator are either all operative or all inoperative. The first approach is more complex, as it intends to be more dynamic, which may need strong interactivity between the different access networks. The second approach only requires the operators to agree on predefined switching schedules. However, as it is pointed by Marsan et al. [MM13], a large number of open issues affecting the NRSs need to be additionally considered for this kind of heterogeneous deployments, e.g. control complexity, UE roaming possibility, consumer and commercial protection policies, etc.

In the case of customer-operator cooperation, the operator has the objective of relocating the traffic to customer's BSs when needed in order to deactivate some of his cells. For example, Suarez et al. [SNB14] consider a heterogeneous deployment composed of operator Macro BSs and customer Femto BSs or HeNBs. The authors rely in the traffic distribution between the two types of cells to create Macro cell deactivation opportunities,. To do so, they consider changing the cell size of the operated BSs and establishing user association policies for the HeNBs. The QoS of the HeNB owners is preserved as the authors consider they can limit the amount of resources available for public use, i.e. the amount of resources available for operator traffic offload. Thus, the number of Macro cells that the NRS can deactivate depends not only of the density of deployed HeNB, but it is also proportional to the cooperation willingness of their owners.

2.4.2 Objectives and constraints

NRSs primary objective is to adapt the access network configuration depending on the traffic load. Usually, the load is measured or estimated on a per-cell basis, aggregating the individual cell measurements a posteriori if more high-level metrics are required, e.g. site load, cell cluster load, geographic area load, or the load afforded by the entire deployment. The load is calculated either by observation or estimation of one or multiple parameters. For example, the number of active users in a cell, the aggregated traffic generated by the active users - measured as throughput or as number of request. The downlink resource block occupation is also a load metric used in the literature.

The ultimate objective of the NRSs is to increase the energy efficiency of the system in consideration. Such energy efficiency is characterized in the literature using different metrics. For example:

- Energy or power reduction compared to a baseline, which is usually the *Always On* paradigm in which all the cells of the modelled system are permanently

active.
- Relation between network performance and energy consumption, e.g., bits/Joule, etc.
- Relation between network coverage and power consumption, e.g., km^2/W, subscriber/W, etc.

The first metric is widely used and is also presented in terms of number of deactivated cells. This is often the case when considering equivalent power consumption for the modelled cells, i.e. only one type of BS is considered. The last two metrics are often used when considering the overall performance given a non uniform deployment, i.e. with different cell sizes and/or BS types, which can result in different performances depending on the approach taken for cell deactivation and coverage preservation, e.g. different realizations of a NRS with different criteria for the selection of the critical and the flexible cells.

The NRSs should respect two general constraints when deciding to change the access network configuration: the service availability should be ensured over the service area of the operator, and the service quality should be over the predefined acceptable standards. In the following sections we present how these parameters are characterized in the literature and the techniques employed to satisfy them when applying NRSs.

2.4.2.1 Service availability and coverage

In the reviewed literature, the coverage of a cell can be defined using two different but complementary criteria: the propagation limits and the signal quality. When considering only the propagation limits, the coverage of a cell is the geographical area where the BS can establish effective communication with the UEs. This is, UEs placed in this area receive pilot and control signals from the cell and are able to connect to it. When additionally considering the signal quality in the definition, the coverage area of a cell is reduced to the geographical area where the UEs can effectively establish communication achieving a minimal level of quality of service.

When applying a NRS, the resulting access network configuration should ensure that all geographical points of the operator service area are covered by at least one cell. This is often simplified to be equivalent to ensure that all considered UEs in the system are covered by at least one cell. Nevertheless, some algorithms define a flexibility margin, considering that a given configuration is acceptable if a given percentage of UEs are covered, e.g. 95% or 98% are often selected as acceptable values.

In some critical conditions, the coverage metric is expressed as purely *Service availability*. For example, Heimerl et al. [HHA+13, HAB+13] elaborate mechanisms to offer coverage on demand in zones with limited energy resources. The BS will be available only when a request arrives, without transmitting any signal otherwise.

However, in urban scenarios, deactivating a cell may create an unacceptable coverage hole in the network where the service is unavailable for the users. In order to avoid this, the cells that remain active may reconfigure to compensate the coverage in the required area. Activating a cell may create interference with the BSs already covering the zone, which is solved by changing the compensating BSs coverage (de-compensate). Both cases require modifications in the hardware parameters to achieve the target coverage configuration. The flexibility in the coverage changes depends on the BS hardware and the propagation conditions of the service area environment. Two main parameters are considered in the literature to achieve this coverage change:

- Transmission power: Controlling the intensity of the transmitted signals, increasing or decreasing the propagation limits.
- Antenna configuration: Controlling the radio beam geometry to direct the signals to specific spatial areas.

In the case of homogeneous overlapping networks the coverage may be guaranteed without compensation needs. An example is the configuration presented by Saker et al. [SMES12] and Micallef et al. [MSES12] in which Pico BSs are deployed under the coverage of a Macro BS, which is always active. The Pico cells are deactivated without considering coverage issues, reacting only to capacity requirements. However, it may also be considered that BSs deployed for capacity duties can also benefit from the coverage compensation techniques to capture the spatial variations of the load. For example, Cardoso et al. [CPB+13] propose to change the coverage parameters of the cells in order to direct the radio resources to the areas in need of increased capacity in a given moment. This approach presents a further level of adaptation to load spatial variations.

Some authors consider as well that coverage compensation actions can be avoided, using instead strategies to improve the cell-edge UE performance, such as Coordinated Multi-Point transmission/reception (CoMP) and relays. For example, Cao et al. [CZZN10], Han et al. [HSL13] and Guo et al. [GO13] propose to use CoMP schemes to provide the minimal required coverage and capacity to the UEs located in the area of an absent cell. Such UEs may detect different neighbor active cells but having poor signal conditions. The principle behind the proposal is that the signals transmitted from the different neighboring active cell can be combined. Thus, the UEs perceived performance is increased to satisfactory levels.

2.4.2.2 Service quality

The Quality of Service (QoS) metrics represent the perceived performance by the UEs when using a given service. The QoS can be either measured or estimated by one or several of the following parameters:

- Throughput: rate of successfully transmitted information, usually measured

in bits per second (bit/s or bps).

- Delay: this parameter usually refers to the end-to-end delay, i.e., the time that the user needs to transfer a packet to a destination, crossing the network. But it can also refer to the radio interface delay, which consider only the time the packet takes for crossing the radio protocol stack.
- Packet delay variation or Jitter: variation in the reception time between consecutive packets.
- Packet loss: proportion of packet which fails to reach their destination
- Bandwidth: portion of the spectrum used to transmit information (Hz). Sometimes used to denote the maximal throughput the channel can support (bit/s or bps)
- Spectral efficiency: throughput divided by the used bandwidth to transmit the information (bit/s/Hz)

An important figure in the presentation of the metrics and the formulation of the system constraints when applying NRSs is the *Outage*, which represents the ratio of users that do not reach the minimum acceptable level of the QoS metric to the ones than are satisfied. This metric is sometimes called *Blocking probability* depending on the system model and the considered technology. The execution of NRS should keep the outage considerably low, i.e. below a predefined threshold, in order to avoid general user dissatisfaction.

When considering mobility and/or network reconfigurations, two important metrics are evaluated in the NRS studies: the *Dropping probability* and the *Handover failure rate*. The dropping probability expresses a measure of the possibility that the ongoing user sessions quality degrades considerably, until reaching unacceptable levels. For example, Samdanis et al. [STKB11] uses the dropping probability to determine if it is acceptable to perform a given reconfiguration considering that the traffic is distributed to neighboring cells that may not have enough capacity to serve the ongoing services. When considering system level approaches, the handover failure rate express the ratio of UEs which do not finish the handover procedure to the users that complete it successfully. For example, Marsan et al. [MCCM11] and Conte et al. [CFC+11] take this into account for the design of progressive cell switching, which ensures the smooth transfer of users from/to the neighboring cells when a cell is activated or deactivated.

2.4.3 Algorithm design factors

Once the system model is established and the performance metrics have been selected and modelled, the system evaluation and decision processes should be defined. In particular three main aspects should be considered. First, how often and with which system load information accuracy the system is evaluated to identify the need of a reconfiguration. These two factors are often related. On one hand

Table 2.2: Summary of the reviewed NRSs literature, classified by the time frame in which the strategies are applied, as well as other related relevant criteria.

Criteria			Representative references
Time frame	Offline	Binary daily planning	[CCMM08] [CCMM09] [MMS10] [ACCM09] [MM13]
		Variable daily planning	[CZZN10] [OK10] [MCCM12] [SMES12] [GWOF13] [HSL13] [OKLN11] [RRAF13]
	Online	Slow reaction	[SE10] [SEC10] [ESC11] [GO12] [CJXH13] [GO13] [DUGK14] [SKB10] [STKB11] [MCCM11][CFC+11]
		Fast reaction	[ZGY+09] [NWGY10] [Niu11] [SES09] [GZN12] [OSK13] [SMES12] [HHA+13] [HAB+13] [TSPB13] [CFGU12] presumed: [HG11] [HMJ11] [HMJ12] [OSK13] [CLHS14]

we found techniques that base the decision purely on load information coming from traffic statistics. These NRSs may execute the decision process only a few times, calculating the optimal configurations and reconfiguration instants for a given period, e.g. a day or a week. On the other hand, the NRSs reacting to the actual load variations in the system tends to decide the configuration and reconfigure more frequently. Second, the control scope of the NRS should be established, defining which entities will take the decisions and based on which information. This will define the scope of the impact caused by a given reconfiguration decisions, as well as the number of variables to consider in the decision process. And last but not least, the decision process should be defined, determining how to take the best decision depending on the system state and considering all the configuration possibilities. In the following sections we describe the different approaches we found in the literature for these three design factors.

2.4.3.1 Time frame and traffic estimation

A NRS algorithm may be differentiated depending on how the traffic variations are considered, how frequent the system is evaluated and for how long a given configuration will be applied. Table 2.2 summarizes the main categories we identified for characterizing the different approaches. The references to the representative literature are presented as well. *Offline* strategies determine when and for how long to apply a given access network configuration based in traffic statistics, while *Online* strategies determine which configuration to apply given the actual traffic conditions and possibly the near future trends. These categories are further described in the following presenting some relevant examples as well.

Offline:

In this type of algorithm the reconfigurations are performed based on a predefined schedule. This schedule is the algorithm output, taking as inputs previously measured and processed information such as traffic statistics. Offline algorithms have low complexity and low processing overhead. However, they present the risk of over- or under-estimating the load during unexpected events, e.g. a short period of sudden traffic increase. Due to spatial and temporal load variations, different switching schedules may be required at different locations in the network, e.g. in business areas, the periods of low load are longer during the weekend than during the week [MM13].

This kind of algorithms is particularly suitable for BS hardware with low dynamism. For example, most of the 2G and 3G BSs deployed nowadays were not designed for dynamic switching, and turning them off comes at the cost of long wake-up times. Moreover, too frequent switching can considerably shorten their lifespan. The early works on NRSs considered these constraints in the strategy design, and based the switching decisions on historical data. For example, Chiaraviglio et al. [CCMM09], Micallef et al. [MMS10] and Ajmone-Marsan et al. [ACCM09, MM13] propose to reconfigure the access network once per day, which produces two system states: minimal energy consumption in the period of the day where very low load is expected, e.g. the late night and early morning, and full consumption the rest of the day.

More adaptive schedules are presented by Cao et al. [CZZN10], Oh et al. [OK10, OKLN11], Marsan et al. [MCCM12], Han et al. [HSL13], Guo et al. [GWOF13], Rengarajan et al. [RRAF13] and Saker et al. [SMES12] . The authors propose to define several periods during the day, and depending on the expected peak load level over each period, a given configuration is associated and scheduled for that period.

Online:

Online algorithms estimate the load and other required inputs based on measurements of the current system state. The resource switching decisions closely follow the variations of the traffic load, making the adaptation of the access network more energy efficient and allowing the reactivity to unexpected load fluctuations. However, the constant measurement and reconfiguration make online algorithm more complex and costly in terms of processing. Moreover, they often require information and coordination from several access network elements in order to estimate the impact of the reconfiguration, and the required actions to perform (e.g. handover users, compensate coverage, etc.).

The important criteria to consider by the online NRS is to determine if a reconfiguration is needed or worth given the current system state, and if the chosen configuration will satisfy near future system state conditions. Thus, two periods are identified in the execution of online strategies, which also determine the degree of reactivity and adaptation to load fluctuations: the decision period and the steady state period. Four processes are identified in the decision period: measurement,

coordination (if any), decision and reconfiguration.

An online algorithm is considered having *Slow reaction* when the decision process is time consuming. The studies considering this time frame require that the reconfiguration decision is carefully taken. It should be also verified if the resulting configuration will be maintained for acceptable and long enough steady state periods. For doing this, the measurement periods have to be large enough to collect a statistically significant amount of information to support the decision process. For example, in the strategy proposed by Guo et al. [GO12, GO13], each BS continuously monitors its load level, and the NRS take the reconfiguration decisions each 15 minutes based in average load values of the previous 15 minutes period. Similarly, Chen et al. [CJXH13] monitor the load for 30 minutes in order to determine the appropriated component carrier scheme to apply. However, no study about the suitability of the strategy for the next period is made by these authors. Contrary, Dawoud et al. [DUGK14] developed a prediction scheme that collect the traffic information during the measurement period and process it to predict the future traffic trend and select the appropriated configuration accordingly.

Concerning the reconfiguration process, Saker et al. [SE10, SEC10] and Elayoubi et al. [ESC11] select the length of the reconfiguration and steady state periods based on estimations in order to avoid too frequent switching for the constrained hardware the authors consider. Samdanis et al. [SKB10, STKB11] discuss about performing progressive reconfiguration process which are time consuming but no estimation of the time frame in which the algorithms will be executed is provided. However, some authors focus on the design such progressive reconfigurations. Marsan et al. [MCCM11] studied the progressive cell deactivation process, in which the transmission power of the cell of interest is reduced progressively by a given amount in different time steps. The authors estimates the number of UEs that will perform handover depending on their position, and each time step is selected appropriately by estimating the time needed by the UEs to successfully perform the required handover procedure to attach to a neighboring cell. The process is shown to be time consuming for the small cells considered by the authors, taking up to 1 minute depending on the layout of the network and considering uniformly distributed users. The work is complemented by Conte et al. [CFC+11] designing the complementary activation process as well.

Fast reaction algorithms have very short decision periods. The measurement process is done instantaneously, i.e. the algorithm takes a sample of the current state of the system, and the decision process is triggered using this input to determine the most appropriate network configuration satisfying the performance targets. For example, Tomaselli et al. [TSPB13] evaluate in a test bed a fast measurement scheme making the hardware reactive to the instantaneous variation in the number of active users in the BS.

The factor that determines the adaptivity of fast reaction algorithms is the steady

state period. The NRS is more more adaptive to load variations when the steady state periods are short, but frequent switching can be detrimental to the service quality, as the constant variation of the radio conditions impacts the signal quality perceived by the UEs. Two approaches were identified in the literature to avoid the constant switching: predefine the length of the steady state periods, or apply the algorithms in periods of low load variability. In the first case, Niu et al. [NWGY10, Niu11] propose fixed steady state periods, while Gong et al [GZN12] propose a tunable parameter that determines the length of the steady state period and the reactivity of the strategy. However, too long steady state periods also come with the risk of underestimating the load. In order to tackle this trade-off, protection margins are selected in the literature, for which the optimal network configuration is obtained using as inputs modified values of the measured system load or the configurations performance. For example, Zhou et al. [ZGY+09] uses a protection margin in which the configuration is selected for load corresponding to the current load scaled by a predefined factor. Instead, Oh et al. [OSK13] modify the decision condition itself, adding an hysteresis margin to the reconfiguration threshold, which reduces the level of load for which a given configuration is acceptable. In the second case, Micallef et al. [MSES12] suggest to apply their NRS only in periods of low load variability, i.e. during the 12-hour period representing low traffic, in order that the reconfigurations are rarely triggered once a first low consumption configuration was established.

Some other studies design their NRSs to support constant switching. For example, Heimerl et al. [HHA+13, HAB+13] implemented a fast reaction system in their BS for critical energy conditions. The BS activate once a user request is detected, and deactivate when no requests are ongoing. In the futuristic access network architecture proposed by Capone et al. [CFGU12], the small cells providing services to the users are activated and deactivated when commanded by the controller cell. In order to serve the users without delay, these small cells should be highly reactive when activating as well as when deactivating, in order to make the energy consumption closely proportional the network usage.

In some of the reviewed literature, neither the time frame in which the decision mechanisms are executed, nor the length of the steady states are discussed. But in some cases, a fast reaction time frame can be presumed given the NRS scope and/or the formulation of the reconfiguration process. For example, Hevizi et al. [HG11] base the reconfiguration decisions on the instantaneous number of users with queuing information in the BS, which implies an almost instantaneous decision period. In the distributed NRSs proposed by Hossain et al. [HMJ11, HMJ12] and Oh et al. [OSK13], each flexible BS decides its status based on the instantaneous load of itself and its neighbors.

Table 2.3: Summary of the reviewed NRSs literature, classified by the type of control scheme used by the strategies, as well as other related relevant criteria.

Criteria				Representative references
Control scheme	Centralized	Offline		[CCMM08] [CCMM09] [MMS10] [CZZN10] [OKLN11] [MCCM12] [MM13] [RRAF13] [HSL13] [GWOF13] [SMES12]
		Online	Local scope	[CJXH13] [CFGU12]
			Global scope	[GO12] [GO13] [ZGY+09] [NWGY10] [Niu11] [SKB10] [STKB11] [GZN12] [MSES12] [CLHS14] [DUGK14]
	Distributed	Non cooperative	Empty cell	[ZGY+09] [NWGY10] [Niu11]
			Standalone scope	[SES09] [SE10] [SEC10] [ESC11] [TSPB13] [HG11] [HHA+13] [HAB+13]
		Cooperative		[GO12] [GO13] [OK10] [OSK13] [HMJ11] [HMJ12]

2.4.3.2 Control Scheme

When considering the different NRS in the literature, the scope of control as well as the scope of the information used for the decision process may be different, which defines different control schemes. We classified the literature depending on the control scheme as shown in Table 2.3. *Centralized* NRSs represent algorithms having complete information about the entire system state and selecting the configuration of any cell accordingly. In *Distributed* NRSs each cell in the considered access network has the needed information and decision mechanisms to choose its own configuration in a given moment.

Centralized:

In this control scheme, only one entity – a controller, is in charge of collecting the metrics information, processing it and executing the reconfiguration of any cell participating in the NRS. This means that the controller have full information of the system state, and it can choose the appropriated configuration as well as the time when it need to be executed. However, the processing overhead of this kind of NRS scales with the number of involved entities, i.e. with the number of cells, BSs or access networks controlled by the NRS, and the number of metrics needed to be considered for the decision process.

Centralized control is often related to offline algorithms [CCMM08, CCMM09, MMS10, CZZN10, OKLN11, MCCM12, MM13, RRAF13, HSL13, GWOF13, SMES12]. This is because the overhead of the load measurement is avoided. Moreover, offline decision algorithms will be executed limited number of times, e.g. once per day or once per week, in order to obtain the required reconfiguration schedule, without

constraints associated to the execution time of the optimization problems.

Centralized online algorithms aim to generate the optimal solution for the network configuration, as the controller has a global view of the network. However, depending on the scope and the granularity of the information to consider, the complexity of such algorithms varies. When the scope is local, e.g. only a few BSs, centralized online algorithms are highly efficient. For example, Chen et al. [CJXH13] divide the network in clusters with centralized control, i.e. one of the BSs is in charge of collecting the information about the cluster state and deciding their configurations. This reduces the control scope to a few BSs, which considerably reduce the complexity for the decision process. Similar approach is considered by Capone et al. [CFGU12] for their futuristic network architecture. The authors consider that the high range BS is in charge of controlling the small cells under its coverage, and to decide when they need to be activated or deactivated depending on the position of the active users.

When the control scope is extended to a complete access network, the complexity of the algorithms augments. Several approaches referring to this case were found in the literature, depending on the granularity of the information considered for the decision process. Guo et al. [GO12, GO13] use a global vision of the network system state, considering the aggregated afforded load and selecting the set of active BSs from predefined patterns. This optimization is not highly complex as the decision space is reduced when considering activation patterns. On the contrary, when full information about each cell load and coverage is used, the algorithms are highly complex. In this case, the authors define heuristics to reduce the decision space and rapidly find a (sub)optimal network configuration, and/or trigger the decision process periodically in time frames adjusted to the algorithm complexity.

For example, Micallef et al. [MSES12] and Samdanis et al. [SKB10, STKB11] iteratively select the state of each cell in the network depending only in their load condition without considering how the load is distributed. Even though this can be costly, each cell is evaluated only one time during the decision process and the authors apply the NRS only in periods of very low load variability, so that the decision process will be triggered only a few times during the time the NRS is active. Dawoud et al. [DUGK14] considers a similar heuristic for the deactivation of the cells but using as an input traffic prediction for the next period of 10 minutes, so that the variability of the load and the load distribution is accounted. Chang et al. [CLHS14] propose an algorithm for finding the global scope network configuration in polynomial time, depending on the number of involved BSs. Furthermore, the authors propose to trigger the decision process periodically in intervals of one hour. Zhou et al. [ZGY+09], Niu et al. [NWGY10, Niu11] consider full knowledge about the afforded load in a per UE basis. The authors focus in determining the optimal cell size and UE redistribution for emptying the larger amount of cells and be able to deactivate them. This implies a high complexity leading to hard to solve optimization problems, especially if considering wide areas with large number of BSs.

In this case, the authors propose distributed versions of the algorithms to reduce the complexity.

Distributed:

In the distributed schemes, each cell or BS participating in the NRS can take its own activation/deactivation decision, either because it only consider its own local conditions, or because it has exchanged useful information with its neighbors and the reconfiguration is acceptable. Thus, two categories were defined depending on the considered factors for the reconfiguration decision: non-cooperative and cooperative.

In the *non-cooperative* case, no interaction with the neighboring cells is performed. On one hand, some studies consider each cell deactivates itself when empty. However, the strategies using this approach rely on the UEs cell selection to concentrate the load only in a few cells, creating the required deactivation opportunities for the other cells. For example, the distributed versions of the algorithms of Zhou et al. [ZGY+09] and Niu et al. [Niu11] propose a modified UE association policy prioritizing the attachment of the UEs to the moderate and high loaded cells. To do so, the cells broadcast to the UEs their load condition, which is used by the UEs as indicator to select the appropriated cell to connect. This strategy is conducive for overlapped scenarios were the UEs have high probability to have more than one cell that can provide acceptable QoS. Even thought this is a distributed approach for the cell deactivation, the mechanisms to decide the activation of a cell are not discussed by the authors. Such mechanism may require either a centralized controller which will remove the distributed attribution, or neighboring cell actions which will transform the distributed approach in cooperative.

On the other hand, we have the case when the NRS scope is limited to a standalone BS. According to our categorization, the NRSs controlling a standalone BSs are centralized, as the NRS controller has full information about the system state, i.e. about the complete BS, and it can control all BS resources. However, the ultimate goal of this type of NRSs is to be applied at several BS in the network in order to maximize the attainable energy reductions. Thus, the approaches presented in Section 2.4.1.1 can be considered as the main contributions to distributed non-cooperative control [SES09, SE10, SEC10, ESC11, TSPB13, HG11, HHA+13, HAB+13] .

In the *cooperative* case, each cell exchanges information with some other cells and takes the reconfiguration decisions that have the most positive impact for itself and its pairs. Some strategies divide the access network into fixed groups, and the cells within each group negotiate their configuration depending on the load and coverage of each one, and the impact that its activation or deactivation may have in the group. Finally, the cells transactionally agree on a given satisfactory configuration for the group. For example Guo et al. [GO13] present a strategy in which an non-overlapping access network is divided into a group of three BSs according to predefined coverage preservation schemes. Each BS belonging to a group estimates

the impact of taking a given action depending on the load conditions. This impact is quantified and each BS will execute the most profitable action, e.g. the value of taking a given action is over a given decision threshold. The cell can be influenced to take a given action manipulating the threshold, either to ensure a given fixed level of performance, or to improve the performance using learning mechanisms. The authors also present a version of their distributed algorithm in which the groups are not fixed. Each BS exchange information with all its neighbors BSs and try to pair with the most convenient neighbor, i.e. the one which complementary action represents the most benefit, e.g. one BS with high positive impact taking the decision to deactivate itself with a BS with high positive impact taking the decision of extend its coverage.

Similar approach is taken by Oh et al. [OK10, OSK13] but considering an overlapping scenario where deactivating a cell impacts only the capacity of the neighboring cells to afford the combined load. A given cell will request to deactivate when the amount of transferred traffic to the neighbors BS is reasonable, i.e. bellow the acceptable threshold of each of them. The cell will deactivate only if it receives positive response of all its neighbors. An active BS with sleeping neighbors will send an activation request to one of them when the traffic in the shared coverage zone increases over a given threshold. Hossain et al. [HMJ11, HMJ12] propose to extend the periods of neighbors deactivation allowing the traffic distribution between active cells by means of coverage adjustment. An overloaded cell with sleeping neighbors, proceed first to try to distribute the traffic between the active neighbors. If the overloaded cell determines that an active neighbor can afford the resulting load, i.e. the total traffic is bellow a given threshold, it sends a coverage extension request. If none of the active neighbors can satisfy the constraint, the overloaded cell proceed to request the activation of the sleeping neighbor.

To avoid conflicts, i.e. non compatible neighbor actions being executed at the same time, some mechanism should be defined. For example, Oh et al. [OSK13] propose a transactional mechanism in which once a partnership is defined, or an action is confirmed, the neighboring BSs are blocked, i.e. they are not allowed to perform or even request any action until the cell executing the blocking action confirm them that the process finished.

2.4.3.3 Decision algorithm

The way NRSs choose the appropriated reconfiguration among all possibilities varies considerably, as it depends among other factors on the decision problem formulation, i.e. how the impact of a given configuration is quantified, and the scope of the decision, i.e. which elements can/should be configured. Numerical methods can be used to generalize the application of the NRSs and provide as output wide-scope networks dimensioning given the energy efficiency target, e.g. cell density.

Table 2.4: Summary of the reviewed NRSs literature, classified by the type of decision algorithm they use, as well as other related relevant criteria.

Criteria				Representative references
Decision algorithm	Numerical derivation			[TGA13]
	Exhaustive search	Local Scope		[HHA+13] [HAB+13] [SES09] [SE10] [SEC10] [ESC11] [TSPB13] [HG11] [GO12] [GO13] [OK10] [OSK13] [SMES12]
		Global Scope	Predefined patterns	[CCMM08] [CCMM09] [ACCM09] [HSL13] [HMJ11][HMJ12] [MCCM12]
			Variable patterns	[GWOF13]
	Greedy algrithms	Cell state selection		[MMS10] [DUGK14] [MSES12] [SKB10] [STKB11] [CJXH13] [OKLN11]
		User association selection		[ZGY+09] [NWGY10] [Niu11]
	Other formulations	Dynamic programming		[CLHS14] [RRAF13] [GZN12]
		Linear programming		[CZZN10]

When the needed output of the decision process is the set of active/deactivated cells, the optimization problems are defined considering the specific characteristics of the system, e.g. location of the cells, radio configurations, location of the users, user association policies, etc. There are different ways to solve such optimization problems depending on the NRS control scope and configuration options. Table 2.4 summarizes the different techniques we found in the literature and the associated representative references. The use of exhaustive search algorithms may be suitable for fixed and reduced decision spaces, while the usage of greedy algorithm heuristics can systematically reduce the searching space until finding a (sub)optimal decision. These different optimization techniques as well as some other approaches depending on the problem formulation are presented in the following.

Numerical derivation:

Some authors obtain numerically the required wide-scope access network configuration given the traffic conditions. For example, Tsilimantos et al. [TGA13] determine the the proportion of active cells given some performance target constraints. The authors use stochastic geometry for their analysis to derive the optimal density of active BSs, which provide a general view of any NRS applied in any network with a given activation policy and BS and traffic density. However, no switching dynamic is studied, i.e. which cells should be activated or deactivated for a given configuration.

Exhaustive search:

Using exhaustive search, all possible system configurations are systematically tested and the optimal solution for the given system state and performance target

are obtained. The applicability of a NRS using exhaustive search depends on the number of possible configurations the system can adopt. In the reviewed literature this depends on the scope of the NRS.

When considering a *local scope*, the activation/deactivation decision is taken locally by the BS, either because the NRS considers a standalone system, i.e. only the BS of interest; or because the BS is participating in a distributed NRS. In both cases, the degrees of freedom for the reconfigurations are limited to the operation states of the BS. In the standalone case it can vary from binary activation/deactivation, e.g. the complete BS [HHA+13, HAB+13] or a given BS RAT [SES09], to multiple system operational states depending on the number of active resources, e.g. the number of carriers or transmitters [SE10, SEC10, ESC11] or the number of cells [TSPB13, HG11]. In the distributed case, the cells are able to chose between more options than the deactivation/activation decision, which lightly increases the space state of the exhaustive search, e.g. the cells can decide to extend coverage [GO13] or to distribute the traffic to the neighbors [OK10, OSK13].

When considering a *global scope*, which includes the entire considered access network, the authors reduce the degree of freedom of the system in order to efficiently apply an exhaustive search algorithm. This is done defining a finite and restricted number of configurations using switching patterns. Thus, the number of reconfiguration possibilities is considerably reduced and the decision process can efficiently select the optimal option using the exhaustive search technique. We found two different approaches depending on how this patterns are established: predefined or variable.

When considering predefined patterns two strategies can be found found in the literature. On one hand, some algorithms use wide scope regular switching patterns, i.e. the activated and deactivated cells are chosen for the entire access network according to regular patterns guaranteeing a given global performance. Chiaraviglio et al. [CCMM09] and Ajmone-Marsan et al.[ACCM09] consider that the network can switch between two operational states, while Han et al. [HSL13] and Marsan et al. [MCCM12] consider a set of multiple regular patterns, which attain different network performance levels, and the NRS can choose between them depending on the estimation of the traffic load afforded by the entire network. Notice that these solutions are executed in a global scope and capture and react only to temporal load variations affecting the entire considered network. On the other hand, Hossain et al. [HMJ11, HMJ12] define groups of cooperating cells in the access network, and each group can be configured depending on switching patters, i.e. the same cells are activated/deactivated depending on the group configuration. Using this approach, different realization of the NRS are possible in the access networks, i.e. in each group of cells, which can ultimately account for the spatial traffic variations in the network.

Guo et al. [GWOF13] consider variable switching patterns to be selected us-

ing exhaustive search. The authors define an uniform deactivation policy, which will be evaluated in all the cells, and any cell satisfying the given condition will be deactivated. This generates different network patterns depending on the spatial distribution of the traffic in a given moment. Thus, the NRS will decide using exhaustive search which of the predefined policies may satisfy the performance target. However, this strategy has a mayor drawback, as the performance of a given policy will vary depending on the state of the system when the policy was applied, i.e. the spatial traffic distribution. Thus, further information should support the establishment of the deactivation policies, e.g. the impact of the deactivation of a cell in the neighboring cells.

Greedy algorithm heuristic:

The reconfiguration decision space becomes very large when a given configuration of the access networks requires the decision of each cell state, e.g. activated or deactivated, as well as its radio configuration parameters, e.g. selecting the transmission power or the antenna beam form. In this case, heuristics can be used to find rapidly a suboptimal configuration achieving good performance given the system state. In the reviewed literature, some authors use greedy algorithms following the heuristic of choosing the best option for each step of analysis, regardless to the other steps outcome. In NRSs this is translated in systematically evaluating if the application of a given policy satisfies a given local performance target and choose the policy generating more benefits.

In the case of cell state selection, all cells are analyzed sequentially, and each analysis determines either the state of the cell of interest, or the state of the cells associated to it, e.g. neighboring cells. When the next cell is analyzed, the set of possible configurations is reduced, as it considers the system after applying the activation/deactivation decisions taken in the analysis of the previous cells. The algorithm stops when a decision was taken for all cells of the deployment, which may produce an optimal set of active/deactivated cells or a suboptimal with acceptable performance. The order in which cell analysis is performed depends on the approach taken by the NRS. For example, Micallef et al. [MMS10, MSES12] propose to deactivate low loaded cells, i.e. the afforded traffic is under a given threshold, no matter how the load is distributed. Thus, the authors choose randomly the least loaded cell to analyze and mark it for deactivation. Afterwards, the system state is recalculated without that cell moving on to the next one. The algorithm stops either when no more cell are considered low loaded, or when the deactivation of the remaining low loaded ones violates the performance constraints. A similar approach is considered by Dawoud et al. [DUGK14] but using predicted traffic conditions for the analysis of each cell.

Samdanis et al. [SKB10, STKB11] propose to concentrate the traffic in the high loaded cells. Thus, their greedy algorithms first sort the cells by load level and start the analysis from the most loaded one. The analysis of each cell consists

in examining the set of neighboring cells, starting from the less loaded one, and verifying if the analysed cell can satisfy the selected neighbor coverage and traffic. If it is the case the neighboring cell is marked for deactivation. After examining all neighboring cells, the newly most loaded cell is analysed. The algorithm stops when no more cells can be deactivated without violating the coverage and outage conditions. Similarly, Chen et al. [CJXH13] concentrate the traffic in the most loaded component carriers of a BS cluster, but only considering the local condition of each component carrier.

Oh et al. [OKLN11] intend to maintain the access network coverage with an uniform cell configuration, reducing the cell density. Thus, the authors order the cells using inter-cell distance as criteria, and analyse the cell with minimum distance to its nearest neighbor. The analysis consists in examining if its deactivation would not violate the coverage constraint.

Zhou et al. [ZGY+09] and Niu et al. [NWGY10, Niu11] applies a greedy algorithm to determine the cell user association in an overlapping scenario which provides the most empty cells in a deployment. Each active UE is analysed by looking the possible cells to associate to, and the algorithm decides the association to the most loaded cell satisfying the QoS requirements. The algorithm ends when all UEs are associated to a cell, and the empty cells are deactivated.

Other formulations:

Some authors define the cell configuration decision as a *dynamic programming* problem in which a complex problem is solved breaking it down into a collection of simpler overlapped sub-problems, and the sub-solutions are recursively combined to find the optimal solution. Chang et al. [CLHS14] intends to determine the cell configuration satisfying coverage an minimizing the energy consumption. To do so, the authors define the decision process as a mixed integer optimization problem, which cannot be efficiently solved. Thus, the authors break the problem into two sub-problems that are solved in tandem. The first sub-problem consists in finding the optimal cell parameters that minimize the area power consumption, while the second sub-problem consists in minimizing the overlapping between active cells. Rengarajan et al. [RRAF13] aim to find the optimal cell configuration and user association allowing to deactivate the larger number of cells in the access network. The decision process is defined as a non-convex problem as well, so the following sub-problem division is considered: determine the optimal BS transmit powers given a user association policy, and find a complementary user association policy that enables the reduction of the network power consumption. The authors iteratively search the policy evolution that leads to the optimal decision. Gong et al. [GZN12] take a temporal dynamic programming problem approach in which the goal is to find the set of successive configuration decisions leading to have the optimal performance over a given period. This is done looking the optimal policy at the end of the period and finding recursively the set of actions that can lead to that policy. However,

given the authors formulation, the problem cannot be solved efficiently. Thus, some heuristics are considered by the authors to influence the selection of the action in each sub-period. Cao et al. [CZZN10] defines the configuration decision problem as a *linear programming* problem in which the authors intend to find the combination of active fixed cells and relays to satisfy a given performance target given by linear objective functions and constrains.

2.4.4 Demand management

In the previous sections we have described the operation of the different NRSs depending on deployment and traffic conditions, and we have shown how the different strategies characterize these conditions to take the network reconfiguration decisions. The majority of the strategies consider than an active user represents a given amount of generated traffic, with a service quality, which has to be satisfied instantaneously and at all cost. Thus, the algorithms are designed to match the activation and deactivation of resources to the immediate traffic demand.

The opposite technique is to explicitly adapt the traffic demand to match the resource availability. On one hand, the network can employ user-unaware strategies, in which the radio resources are activated/deactivated according to the energy requirements, but regardless to the user activity. Even when this approach may cause user dissatisfaction due to degraded service quality, it is a solution to decrease the energy consumption of the networks using demand adaptation.

On the other hand, in this thesis we propose another category of demand adaptation techniques for NRSs. In our approach the network interacts with the users, profiting of their cooperation to shape the traffic generation, and thus perform the resource adaptation in a more flexible fashion. In the next section we present how this traffic shaping has been address in the literature to better utilize the network resources, and how some authors address energy efficiency issues using these techniques.

2.5 User demand shaping

Strictly speaking, the term *demand shaping* refers to strategies that aim to influence the demand to match planned supply. In wireless networks these techniques are mostly studied to anticipate and alleviate the periods of congestion, i.e., when the radio resources are not enough to satisfy the QoS of the served users and incoming requests; or to balance the load and the utilization of resources within the network. The demand shaping strategies aim to influence the user or application behavior in order that the traffic is generated, or mostly not generated, in certain periods and/or network conditions. Several types of demand shaping in wireless environments has been identified in the literature. We classify them in two types, according to the

principal factors driving the shaping: *purely-temporal shaping*, and *spatio-temporal shaping*. Some representative studies for each category are described in the following sections.

In Section 2.5.3 we present some of the studies performed to measure and quantify the willingness of the users to participate actively and consciously in cellular services using demand shaping. In Section 2.5.4 we describe some of the few studies we found in the literature, which were conducted in this field with energy efficiency purposes. Finally, we discuss how avoiding congestion thanks to demand shaping techniques has an indirect impact in the long term energy efficiency of the networks.

2.5.1 Purely-temporal shaping

Purely temporal shaping techniques influence the user/application behavior in order to reduce the traffic in congestion periods. The idea is to shift the traffic from peak or busy hours to off-peak periods. Operators are interested in this technique as they dimension their networks according to busy hour demand. If the traffic in busy hours increases, the operators need to deploy new resources to satisfy the user demand with acceptable QoS. If this trend continues uncontrollably, the operator will arrive either at unsustainable and no-profitable deployments, or at the limit of the technology capacity.

A way of performing temporal shaping is adapting the price of the services depending on the time and network status. This strategy is called Time Dynamic Pricing (TDP) [SJWHC12]. Schoenen et al. [SY13] propose to persuade the user (using pricing surcharge) to not use the service in congestion periods. The authors model using Markov chains and Pietri Nets the dynamic of the traffic demand, and adapt the user arrival rate when the system is in congestion depending on a control function that accounts the surcharge in the price if the service is used. This control function is derived from previous empirical work (presented in Section 2.5.3) supporting the intuitive postulate: if the price increases considerably the users are more likely to not use the service. However, the impact of the session deferral is not studied.

Ha et al. [HSJW+12] focused on elastic data services and present the architecture, implementation and field tests of the strategy involving all entities of a cellular network. The UE front-end allows the user to choose to use the service right away for a given price (dynamically calculated by the operator, higher in congestion periods) or schedule it for later use with a discount in the price. In this way, the deferral of the traffic is considered, as users who scheduled the service are prioritized at the moment of the service starting. We will refer to this strategy as *request shift*. The authors state that in order to establish time-dependent prices, operators should survey users and monitor their traffic patterns, with and without TDP activated. This data is then used to estimate the willingness of the users to shift their traffic in time in

exchange for a monetary discount, and if so, how much. Then, the time-dependent prices for the next day are calculated, which establishes a closed loop for the price control, with the objective of reducing the utilization in congestion periods. Similar approach is presented by Gabale et al. [GDK+13] for cellular downlink traffic. In addition to the request shift strategy presented before, the authors study another type of temporal shaping, called *delivery shift*. This strategy consist on processing the user request right away, and providing a longer-time-frame estimate of when the content will be delivered. This is, shifting the completion of the service instead of the start of it. Obviously, this strategy is not suitable for real time applications such a voice or video calls, but it can be appropriated for data synchronization, bulk transfers or video content delivery.

Another way of implicit temporal traffic shaping is the QoS aware RRM scheduling, which exploits the UE information on a per flow basis. Proebster et al. [PKWV12] design a scheduler which estimates the traffic flow delay requirement depending on the application and the system information reported by the UE, e.g., if the application is in the background or the foreground of the screen. The scheduler then uses this information to prioritize the flows of the applications according to their window state and their delay budget. As a result, the transmission of the uncritical information is shifted in time.

2.5.2 Spatio-temporal shaping

Spatio-temporal shaping techniques aim to profit of the spatial diversity of the wireless environment and BS deployment to achieve user and network benefits. To do so, some temporal shaping is required as well, mostly delaying the generation and/or transmission of the traffic until reaching a given area.

Dawson et al. [DRSW06] propose mechanisms to maximize the BS utilization, shifting the traffic to underutilized, and otherwise unnecessary operational, cells. The authors propose dynamic pricing based on policies that account the geographical position of the cells, their utilization and operational cost. When the UE is identified to be under the coverage of a underutilized or low-cost cell, the user is notified of the possible discounts that can be currently applied if the service is performed from the current location. Moreover, the authors propose the entities and automated mechanisms to support such strategy. There is an implicit temporal shaping in this strategy, as it also encourages the users to defer their non-critical services if they are under the coverage of a high-cost or congested cell. Making the user aware of the spatial characteristics of the network, stimulates his mobility to underutilized cells, creating some behavior patterns. For example, after several times of using the service, most users will learn that discounts are usually available in a specific time and place, i.e. specific cells in the frequented areas.

Schoenen et al. [SYW11] propose to explicitly motivate users to change of lo-

cation to achieve better signal quality than the one available in their current position. The benefit is mutual, as users will obtain higher data rates or improve some other QoS metric, while the operator increases the spectral efficiency in his network [BSY12]. This last is achieved reducing the number of radio resources dedicated to serve the users, as for example in a OFDM scheme, the operator has to provide more resources to cell edge users compared to cell center users in order to provide fair capacity share. This study implies some temporal shifting too, as the movement to another location probably will take some time. However, the expected benefits with the application of these strategies are dominated by spatial characteristics more than temporal ones. For example, moving to improve spectral efficiency can benefit in both, low load periods and congestion periods.

Celis et al. [CDML14] proposed a less explicit mechanism to achieve similar objectives. The authors developed and implemented in an operational 3G network, a cellular application that aims to efficiently utilize the network resources. When a user wants to use bulk data services (e.g. download a video or transfer a file), a request is transferred through the network, and the network itself decides when it is the right moment to perform the transfer. The network perform such decision based on: network profiles (e.g., historical cell load data, current network status, cells load, etc.), UE instantaneous information (e.g., current location and signal strength) and user-specific profiles. These last contain user mobility information as well as the characteristic of her/his subscription. The objective is to exploit the user and network information to deferral the user requests until the network conditions are suitable for it. In this study, this change of conditions comes from mobility, e.g. the user moved to a zone with better signal strength.

Another spatio-temporal demand shaping strategy is considered in the context of *WiFi offload*. The wide coverage ranges of cellular networks contribute to its ubiquitous nature. But the proliferation and deployment of WiFi BSs make available both access networks in most of the locations frequented by the users. Moreover, most of the current cellular UEs have also a WiFi interface. Thus, opportunistically using the WiFi BSs when available is an efficient way to exploit the spatial diversity of wireless networks and most of all to reduce the traffic in the cellular network. This is already the case with some current UE applications, where the user can select to send synchronization or backup traffic (e.g. Dropbox, Google+, DSphoto, etc.) only over WiFi. A different approach is presented by Ra et al. [RKN10]. The authors developed and implemented an online algorithm in the UEs that chooses at every instant whether to use any of the available interfaces (i.e. WiFi or cellular) to transfer data and, if so - which of them; or to delay the transmission in anticipation to a more efficient connection becoming available in the future, without increasing delay indefinitely. The final objective is to deal with the energy-delay trade-off, as is stated that the cellular interface consumes more battery for transmission than the WiFi one. The algorithm is compared with a static delay algorithm in which, an initial delay is establish for the start of the service. If no WiFi BS is found within

this delay, the transfer starts using the cellular network. The length of the initial delay depends on the application type and its delay tolerance, and on the knowledge about the WiFi availability. Similar concept is evaluated by Lee et al. [LLY+13]. However, in their work the use of the WiFi interface is always preferred and each data transfer is associated with a completion deadline. Whenever the UE gets under the coverage of a WiFi BS the data transfer is resumed. If the transfer does not finish within its deadline, the cellular network completes the transfer.

2.5.3 User willingness

Surveys and field trials have been performed by the research community to test the readiness of the users to the demand shaping paradigm.

Schoenen et al. performed several surveys to test the willingness of the users to use a cellular system under demand shaping control. The first survey was conducted in the summer of 2011 among 60 university students [SBM+12a], while the second one was conducted in autumn of 2011 among around 100 students [SBM+12b]. No seasonal correlation is expressed by the authors. The first survey tests the following points for the voice call, video streaming applications:

- Which is the maximum distance (d) in meters, the user is willing to walk in return for a discount of (m) percent ($d \in [0, 50]$ and $m \in [20, 80]$).

and for web browsing and bulk downloading applications:

- Which is the maximum distance (d) in meters, the user is willing to walk in return for a speed up of aX in the connection ($d \in [0, 50]$ and $a \in [2, 4]$).

In the second survey the authors introduced the notions of surcharge, i.e. paying more if the user still uses the service given the conditions. The survey questioned about the following points for the voice call, video streaming and data applications:

- Which is the maximum surcharge, expressed in price multiplier s, the user is willing to pay to use the service right away no matter the network conditions ($s \in [1, 5]$).
- Which is the maximum distance (d) in meters, the user is willing to walk in return for a discount of (m) percent ($d \in [0, 100]$ and $m \in [20, 80]$).
- Which is the maximum distance (d) in meters, the user is willing to walk for avoiding to pay a surcharge, expressed in percentage (n) of the service price ($d \in [0, 100]$ and $n \in [20, 100]$).
- Which is the maximum distance (d) in meters, the user is willing to walk in return for a speed up of aX ($d \in [0, 100]$ and $a \in [2, 4]$).
- Which is the maximum amount of time (t) in minutes, the user is willing to wait before using the service in return for a discount of (m) percent ($t \in [0, 60]$ and $m \in [20, 80]$)
- Which is the maximum amount of time (t) in minutes, the user is willing to

wait before using the service for avoiding to pay a surcharge, expressed in percentage (n) of the service price ($t \in [0, 60]$ and $n \in [20, 100]$).

The results confirm the general intuitive trend: the acceptance drops with effort (distance d or waiting time t) and stronger incentives or deterrents are followed with more cooperation. Voice call users are more willing to move than data and video users. The opposite is observed when considering temporal shifting: data and video users are more willing to wait to start their services than voice users. Moreover, voice users are more willing to tolerate shorter delays than the video and data users. However, a significant portion of the polled users (between 40% and 80%, depending on the control variables, i.e. incentives or deterrents) are able to tolerate the minimum delay proposed of 10-15 minutes in the start of their voice services. Finally, the authors fit the data with exponential expressions relating the trade-off between the control variables, i.e. distance, waiting time, discounts and surcharges, and the type of service.

In the first survey the authors included a last question to measure the concern the users have about the wireless carbon footprint of their services, which might motivate them to move or wait even without explicit financial incentive or deterrent, just by the spirit of being "greener". The results show that more than 50% of the polled users were above the indicator expressing moderated concern. This is an encouraging point to study use-aware demand shaping algorithms for network energy efficiency purposes. However, this direction is not further investigated by the authors of the study.

Ha et al. [HSJW+12] implemented a TDP infrastructure in an operative cellular network to perform field trials. The authors recruited 50 participants and evaluated their 3G data usage under Time Independent Pricing (TIP) service fares e.g. flat rate, during 3 months. Afterwards, they evaluated the data consumption using a static time dependent pricing (Static-TDP) during 3 weeks. Three different tariffs were offered to the users during the day: full fare, 10% of discount and 40% of discount. The last part of the experiment tested the dynamic time dependent pricing (Dynamic-TDP) in which the users data price is calculated depending on the usage and network status, which varies among the day. For both TDP strategies the price is presented to the users in real time using the application developed by the authors, warning them by means of a color code the price of using their data applications in each instant. The results shows that the users participating in the trial shifted their data usage from high to low price periods. Correlated to this, users decreased the peak-to-average hourly traffic ratio, spreading their traffic over the day. Moreover, the total daily usage increased, probably because users consume more data when is cheaper. The authors state that this last result is an excellent incentive for operators searching to maximize their profits. Extensions of this work are presented by Joe-Wong et al. [JwHSC15], increasing the experiment duration to more than a year and simultaneously controlling the two groups (to avoid the learning-effect). Similar results are obtained in that last study.

2.5.4 Applications in green networking

Although the demand shaping strategies in wireless networks were developed mainly to relieve congestion and better utilize the network resources, they are also useful for other purposes. Among them, some studies consider the energy efficiency of the network components, either as a side effect, or as a design characteristic.

For operators, the uncontrolled traffic growing in busy hours and the congestion, is an economical concerning. However, this have an indirect consequence in the energy consumption of the networks. Operators dimension their networks according to busy hour peak traffic. If the network is constantly in congestion in these periods, operators need to increase the capacity, mostly adding more BSs, increasing operators expenses and the network energy consumption. Alleviating the congestion in peak hours using demand shaping techniques can ultimately contribute to the energy efficiency of the network, delaying the deployment of more energy consuming BSs.

The first studies directly focused in energy issues that worth mentioning are those performed for the energy efficiency of UEs. Most of the UEs are powered by batteries, thus, their functioning time is limited by the battery charge level, which obviously decreases depending on the usage of the UE. Spatio-temporal demand shaping strategies intend to extend the functioning periods of UEs, reducing their power consumption. In particular the delay-aware WiFi offload techniques developed by Lee et. al [LLY+13] and Ra et al. [RKN10], envisage the preservation of the UEs battery charge, minimizing the time the UE transmit using the 3G interface. In fact, while the power consumption on the two kinds of radios can be comparable, the achievable data rates on these interfaces differ significantly. For the same amount of data, the 3G interface would take longer to transmit, consuming more power in total. Thus, shaping the traffic in order to use opportunistically the WiFi interface can preserve the UE battery charge.

Other studies focus on the purely-temporal demand shaping of the downlink traffic in the BSs. Gupta et al. [GS12] propose a RRM scheduler with packet shaping in order to reduce the BS energy consumption. The authors propose to buffer the data before transmit it. The data is transmitted once the transmission condition is fulfilled (e.g., having enough data, buffered data arriving to deadline, etc.). Thus, the data is shaped in bursts. The power consumption reduction comes from the fact, that if the data is transmitted in bursts, the inter-burst periods are longer, i.e. the periods when there is no data to transmit. This allows to take more advantage of the adaptability of the hardware, specifically the fast deactivation, producing longer cell DTX periods, and further reducing the power consumption of the BS.

Gabale et al. [GS14] propose an energy aware RRM scheduler with purely-time traffic shaping and delivery shift. The authors consider a cellular system in which

the BSs are powered by both renewable energy and non renewable energy sources. The objective of the energy aware scheduler is to maximize the data transmitted in periods when renewable energy is available, while satisfying the service completion deadline offered to the users. Thus, reducing the power consumption from polluting sources.

The proposal of this thesis goes a step further and aims to contribute to the energy efficiency in the entire access network scope. We propose a new purely-temporal demand shaping category, in which the users are aware about the energy efficiency techniques applied in the network. Depending on the network status, the users are persuaded to delay the start of their services, which may allow the network to remain in a low energy consumption state.

2.6 Summary and discussion

In this chapter we introduced the fourth generation of cellular network technology with highest penetration on the market, namely LTE. We presented the architecture and power consumption model of its access network components: the Base Station (BS)s. We presented some of the strategies that are devised in the literature to reduce the energy consumption of the BSs and more generally, of the whole access network. We classified these strategies in three categories depending on the time scale they affect the access network. *Hardware upgrades* are incorporated to the networks in a *long time scale* (months, years) but represent technological innovation that substantially contribute to the energy efficiency of the network. Moreover, in combination with some other management techniques, the reduction of the power consumption can be prolonged. These management techniques constitute the other categories in our classification. RRM strategies reacts in a *short time scale* (milliseconds, seconds) adapting the radio resources of the individual BSs to the load carried by them. NRS reacts in a *medium time scale* adapting the active/inactive BSs to the temporal and spatial load variation within the whole access network or a part of it. NRS are naturally more complex as often involves multiple entities to control. However they provide the highest energy reductions, as complete cell or BSs can be deactivated. For this reason we delved in the wide NRS literature identifying the design characteristics and methods used in the elaboration and evaluation of NRS algorithms. Finally, we presented some of the user demand shaping concepts used in cellular networks to avoid congestion, highlighting important findings that encourage their usage for energy efficient networking.

In the next chapters we present our contribution towards that direction, combining energy efficiency strategies with demand shaping techniques and user network-awareness, in order to further reduce the energy consumption of the access network. We propose a proactive user-network interaction in which the network require the users cooperation when applying an energy efficiency strategy in a given area. The

users cooperate delaying the start of their services for a given bounded time, which offset the traffic generation, and allow the network to remain for longer periods using limited resources and consuming less energy. In the next chapter we describe two strategies following the proposed paradigm, and we present their analytical evaluation.

2

3 Exploiting user delay-tolerance to save energy in cellular networks: An analytical approach

3.1 Introduction

In this chapter we present our proposed strategies for the energy efficiency of cellular networks. These strategies combine user demand shaping techniques with energy efficient dynamic management of the access network. In particular, we shape the user traffic generation in order to create opportunities to apply energy efficiency techniques in the access network and thus reduce the energy consumption.

In Section 3.2 we motivate our choice concerning the demand shaping technique: the request shifting, in which the start of users services is offset for a given (bounded) time. In Section 3.3 we present the general hypothesis on the system model of a cellular access network using radio resource adaptation, and the traffic and capacity assumptions we made for the evaluations.

The two strategies that we developed are studied in Section 3.4 and Section 3.5. The main difference between the strategies is how the users are considered depending on the access network state, which directly impacts the way the resources are used. The first strategy considers the choice of the users to participate or not in the cooperative scheme. The management decisions are taken based on the conditions of the users willing to delay their services, and giving priority to the utilization of dynamic resources, i.e. resources that can be deactivated/activated, while the service of the impatient users is ensured using a set of resources which are available even when the dynamic resources are not active. We denote these last resources as static resources. The second strategy makes no distinction between users, and the delay of the service requests is made in an opportunistic fashion, i.e. only when it is required by the system load conditions. Thus, this strategy serves users giving priority to the utilization of the currently available resources, relying on the shifting of user traffic to delay the activation of the dynamic resources. For each strategy, we present the mathematical model of the traffic dynamic and we analyze the strategy functioning in scenarios where different energy efficiency techniques are applied, showing the trade-off between the strategy parameters. In Section 3.6 we compare the performance of the strategies in common scenario and we quantify the daily energy consumption reductions. Finally, in Section 3.7 we summarize and discuss the principal findings of the work presented in this chapter.

3.2 Motivation

When considering dynamic adaptation of the access network such as RRM techniques or NRSs, the network capacity in a given zone changes. For elastic and best-effort services, the *completion shifting* technique can be beneficial, as the traffic can be split and partially shifted to the moment when more capacity is available (e.g., due to the activation of more resources), and/or prioritized (e.g. in low capacity periods) to met the completion deadlines. This is not the case when considering real time applications such as voice and video calls. Real time applications establish sessions and the network should maintain a constant bit rate during the whole duration of the service, which establishes strict constraints in the application of energy efficiency techniques. On one hand, the capacity change due to the deactivation of network resources can considerably affect the service quality, degrading the QoS of ongoing sessions and eventually causing call dropping. On the other hand, the activation of resources is traditionally performed keeping unnoticeable the network changes to the users, triggering the resource activation process in traffic levels which are still affordable. Thus, the duration of the periods of low energy consumption is bounded, which limits the energy savings resulting from the application of RRM and NRS.

The *request shifting* seems to be a good solution to address these constraints. If the users are able to wait a predefined and known-in-advance time to perform their real time calls, the network could delay the activation of resources, remaining for longer periods in low capacity, and in low energy consumption states. This idea is also supported by the studies of Schoenen et al. [SBM+12a] [SBM+12b], Ha et al. [HSJW+12] and Joe-Wong et al. [JwHSC15] already presented in Section 2.5.3. These studies showed that users can change their usage patterns to better match network requirements. Moreover, the users are willing to delay the start of their real-time services for a given time. In addition, the impact of their services in the environment was identified as a recognizable concern.

It is important to note that, contrary to these studies, we will not address the details about the incentives needed to influence the cooperation of the users and make them willing to delay their network access; neither how it should be implemented or presented to them. Our intention is to estimate the potential gains, in terms of energy consumption reduction, if real-time requests can actually be shifted. Thus, our approach establishes the relationships and trade-offs between the time the request can be shifted or delayed and the possible energy gains if the access network can remain in low capacity and low energy consumption states.

In this chapter we approach this idea from a theoretical point of view. We use well known user traffic distributions and we present the developed strategies to control the access network resources based on the user estimated behavior. Finally we obtain the theoretical bounds of the benefits that can be achieved combining traffic shifting and dynamic resource management for energy efficiency.

3.3 General Assumptions

In this thesis, we are studying a cellular network able to use radio resource adaptation in a given set of its BSs. We focus on the subsystem composed by these BSs and the UEs in the area covered by them. We use a centralized control scheme, so that the decisions are taken locally for the group of concerned base stations. In this section we give a general description of the system model which is common to the two proposed strategies. We provide specific details on the energy efficiency strategies used for the numerical evaluations later in this chapter.

The system is modelled to be in one of two operational states, depending on the active available resources. The system is in *all-On* state when all components of the access network are operational and available. In this state the system has an available maximum capacity denoted by $C_{\max}$. The system is in *min-On* state when only a given subset of radio resources provides coverage and satisfies a minimum load level. Such resources are denoted as *static resources*. In this state the system has a capacity C_{sta} which varies depending on the strategy. The set of radio resources that can be deactivated are denoted as *dynamic resources* and provide the system with C_{dyn} resources such that:

$$C_{\max} = C_{\text{sta}} + C_{\text{dyn}} \tag{3.1}$$

The system switches between states depending on the level of load, i.e. the switching is triggered when the level of load reach a certain predefined threshold. The conditions for the selection of the different thresholds depend on the employed management strategy. In this chapter we assume that the state switching is done instantaneously, i.e., the reconfiguration periods are not considered in the evaluations. Coverage and system availability issues are assumed assured by the radio planing and the design of the specific energy efficiency strategy we will use for each evaluation scenario.

We evaluate the system under homogeneous traffic, considering only one type of service offered to the users. Thus, the system load varies proportionally to the number of ongoing users session. We model the users' dynamic using ergodic and homogeneous continuous time *Markov Chains (MC)*. The parameters used to describe these processes are the following: the session interarrival time is exponentially distributed with parameter λ. The session service time is exponentially distributed with parameter μ. The system capacity is fixed and depends on the set of resources the MC considers. Finally, the method of service is FIFO. We assume that during the entire service time, a session consumes a fixed average number of resources which is the same for all users, i.e. independent of their position. Thus, we can express the offered load and the system capacity in terms of simultaneous active sessions.

The users participating in the demand shaping and willing to offset the start of their services are called Delay Tolerant User (DTU). The system proposes to the

DTU a maximal initial delay denoted by D. Users not willing to wait under any circumstance are called Non-Delay Tolerant User (N-DTU).

3.4 Strategy One: Persistent DTU

3.4.1 Strategy description

This strategy assumes that the user willingness to cooperate with the network is predefined, and the network knows which users are DTU and which ones are N-DTU. When a given user requests a service, the decision to serve her/him depends on the system state and the type of user. When the system is in *min-On* state, the DTU arrivals are put on hold, i.e. the service request is accepted by the network, but the start of the service is delayed. N-DTU arrivals are served right away if possible, i.e. if the load is under the capacity limit in this system state (C_{sta}), and blocked otherwise. When the system is in *all-On* state, all arrivals (DTU and N-DTU) are served immediately if there is enough available capacity left, and blocked otherwise. The switching between system states is done depending on thresholds in the number of users in the system. The system turns to *all-On* state when the number of waiting DTUs reaches the predefined threshold $U_{\max}$. Then, the $U_{\max}$ DTUs on hold and all new arrivals are served by the recently activated dynamic resources, while users already in service continue to be served by the static resources. To ensure the service after the initial delay, the system should not accept more DTU than the dynamic capacity (C_{dyn}), which gives an upper bound for the value of $U_{\max}$. The system turns to *min-On* state whenever the number of active users falls below another predefined threshold $U_{\min}$. The $U_{\min}-1$ users with ongoing services are taken over by the static resources.

An example of this strategy is illustrated in Fig. 3.1. The two values of $U_{\min}$ illustrate the functioning of the system in varying traffic conditions (i.e. different arrival rates of N-DTU, limiting the number of users that can be absorbed when switching to *min-On* state). From t_0 to t_1 the DTU arrivals are put on hold. At t_1 the system switches to *all-On* state as $U_{\max}$ is reached. At t_2, the system returns to *min-On* state, as the $U_{\min}-1$ users can be absorbed by the static resources.

The selection of the strategy thresholds depends on the traffic conditions, the maximum delay defined by the network (D) and the desirable system quality of service. Thus, $U_{\max}$ and $U_{\min}$ should be selected in order to ensure the service of the DTUs before D. Likewise, the system should be capable of continuing serving the ongoing sessions when it turns to *min-On* state, in order to keep low the call dropping levels. These points are treated in the following section.

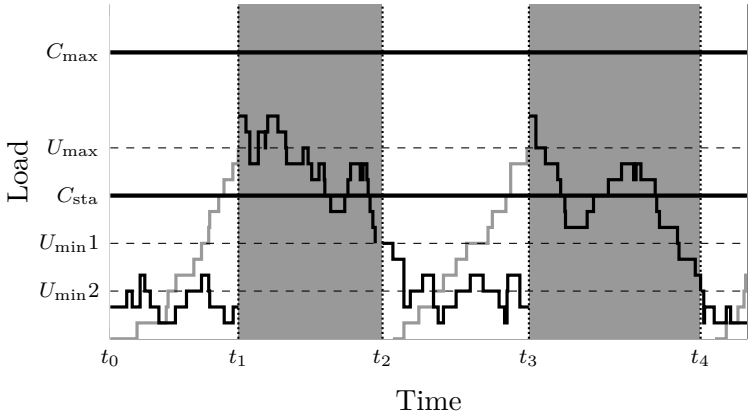


Figure 3.1: **Load dynamic example of Strategy One**. White periods: system in *min-On* state – Serving N-DTUs and Delaying DTUs. Dark gray periods: system in *all-On* state – No delay.

3.4.2 Mathematical model

We modelled the strategy presented in the previous section with two MCs. The general representations of the MCs state space are depicted in Figure 3.2 and are explained and solved in the first part of this section. The following parts describe the different criteria used for the selection of the strategy thresholds, namely the waiting time and quality of service constraints.

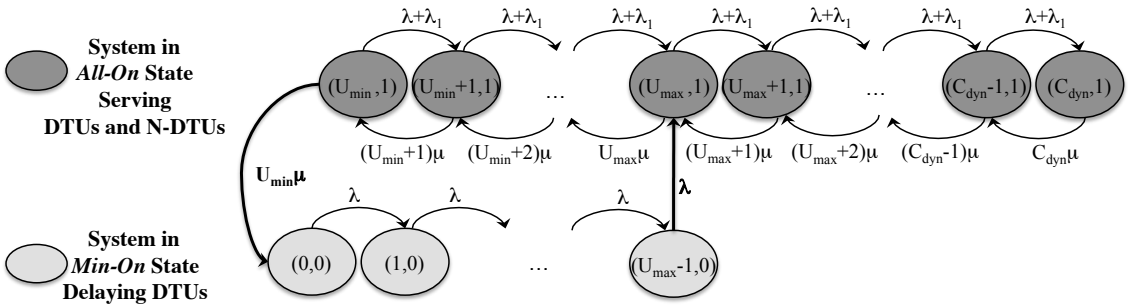
3.4.2.1 Markov chains

The state space of the MCs is $S = \{(i, j)\}$ where i represents the number of users in the system while j indicates the dependency of the system state. If the system is in *all-On* state, $j = 1$, otherwise the system is in *min-On* state and $j = 0$. If the MC state is independent of system state, $j = -1$. The possible values of i are constrained by the strategy thresholds, as well as the resource capacity.

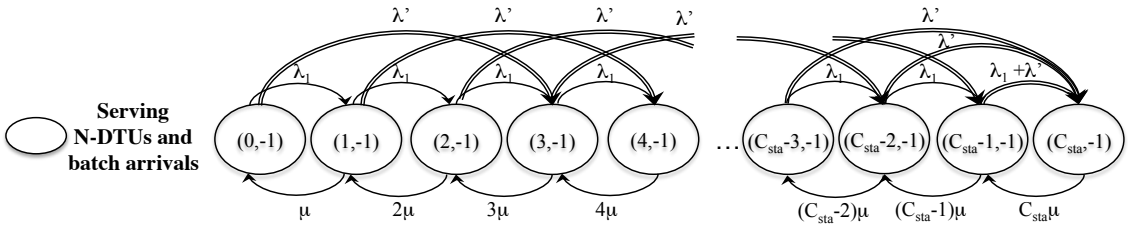
The two type of system resources are independent and they are represented separately. Thus, the strategy is modelled using the two MCs depicted in Fig. 3.2. The DTU and N-DTU interarrival rate is represented by λ and λ_1 respectively. In this strategy, the service rate is 0 for the DTU when waiting and $i\mu$ otherwise, for each state of the two MCs.

Dynamic resources:

The MC depicted in Figure 3.2(a) represents the traffic behaviour of the users associated to the dynamic resources, i.e. when the system is in *all-On* state serving all type of users, and when it is in *min-On* state and DTUs are waiting. Furthermore,



(a) Dynamic resources



(b) Static resources

Figure 3.2: Markov Chain of the user dynamic using Strategy One, Batch size $b = 3$.

the system state transitions are represented in this MC as well, as they depend on the dynamic resource utilization:

- The switch *all-On* $\rightarrow$ *min-On*: the state is $(U_{\min}, 1)$ and a departure occurs
- The switch *min-On* $\rightarrow$ *all-On*: the number of waiting users reaches $U_{\max}$

The balance equations of the MC modelling the dynamic resources are:

$$\begin{aligned}
(\lambda+\lambda_1+i\mu)p_{(i,1)} &= (i+1)\mu p_{(i+1,1)} && \text{if } i = U_{\min}\\
(\lambda+\lambda_1+i\mu)p_{(i,1)} &= (i+1)\mu p_{(i+1,1)} + (\lambda+\lambda_1)p_{(i-1,1)} && \text{if } U_{\min} < i < U_{\max}\\
(\lambda+\lambda_1+i\mu)p_{(i,1)} &= (i+1)\mu p_{(i+1,1)} + (\lambda+\lambda_1)p_{(i-1,1)} + \lambda p_{(i-1,0)} && \text{if } i = U_{\max}\\
(\lambda+\lambda_1+i\mu)p_{(i,1)} &= (i+1)\mu p_{(i+1,1)} + (\lambda+\lambda_1)p_{(i-1,1)} && \text{if } U_{\max} < i < C_{\text{dyn}}\\
i\mu p_{(i,1)} &= (\lambda+\lambda_1)p_{(i-1,1)} && \text{if } i = C_{\text{dyn}}\\
\lambda p_{(i,0)} &= U_{\min}\mu p_{(U_{\min},1)} && \text{if } i = 0\\
\lambda p_{(i,0)} &= \lambda p_{(i-1,0)} && \text{if } 0 < i < U_{\max}
\end{aligned} \tag{3.2}$$

From Equations (3.2) we can deduce the transition matrix Q_{dyn}. The MC is irreducible and consists of positive recurrent states. The unique steady-state probability vector $\pi_{\text{dyn}} = \{p_{(i,j)}\}$ is given by:

$$\pi_{\text{dyn}} Q_{\text{dyn}} = 0 \tag{3.3}$$

$$\sum_{i=U_{\min}}^{C_{\text{dyn}}} p_{(i,1)} + \sum_{i=0}^{U_{\max}-1} p_{(i,0)} = 1 \tag{3.4}$$

Static resources:

The MC depicted in Figure 3.2(b) represents the static resources which are always available. These resources serve N-DTUs which arrive individually with an interarrival rate λ_1. Additionally, when the system switches to *min-On* state, all ongoing sessions served by the dynamic resources are absorbed by the static resources. This can be seen as a batch arrival to the static resources. The size of this batch is given by:

$$b = U_{\min} - 1 \tag{3.5}$$

The batch arrival rate represents how often the system switches to *min-On* state. This arrival rate is not necessarily exponentially distributed, as it depends on the behavior of the dynamic resources. However, in order to be consistent with the intended Markovian analysis, we approximate it by the following equation:

$$\lambda' = p_{(U_{\min},1)} U_{\min} \mu \tag{3.6}$$

where $p_{(U_{\min},1)}$ is the probability that the dynamic resources are serving $U_{\min}$ sessions and $U_{\min}\mu$ is the probability that a departure occurs, triggering the switching to *min-On* state and transferring the b sessions to the static resources.

The balance equations of the MC modelling the static resources is:

$$\begin{aligned}
(\lambda' + \lambda_1) p_{(i,-1)} &= i\mu p_{(i+1,-1)} && \text{if } i = 0 \\
(\lambda' + \lambda_1 + i\mu) p_{(i,-1)} &= (i+1)\mu p_{(i+1,-1)} + \lambda_1 p_{(i-1,-1)} && \text{if } 0 < i < b \\
(\lambda' + \lambda_1 + i\mu) p_{(i,-1)} &= (i+1)\mu p_{(i+1,-1)} + \lambda_1 p_{(i-1,-1)} + \lambda' p_{(i-b,-1)} && \text{if } b \leq i < C_{\text{sta}} \\
i\mu p_{(i,-1)} &= \lambda' \sum_{n=C_{\text{sta}}-b}^{C_{\text{sta}}-1} p_{(n,-1)} + \lambda_1 p_{(i-1,-1)} && \text{if } i = C_{\text{sta}}
\end{aligned} \tag{3.7}$$

From Equation (3.7) we can deduce the transition matrix Q_{sta}. The MC is irreducible and consists of positive recurrent states. The unique steady-state probability vector $\pi_{\text{sta}} = \{p_{(i,-1)}\}$ is given by:

$$\pi_{\text{sta}} Q_{\text{sta}} = 0 \tag{3.8}$$

$$\sum_{i=0}^{C_{\text{sta}}} p_{(i,-1)} = 1 \tag{3.9}$$

3.4.2.2 Waiting time

As stated in Section 3.3 we define a *maximal initial delay* (D). For each user arriving in the system, the probability of waiting more than D should be controlled.

In this strategy, DTUs will wait for their services if they arrive when the system is in *min-On* state, and they will be served when the system turns to *all-On* state. To do so, the number of waiting users should reach the threshold $U_{\max}$, as explained in Section 3.4.1.

Then, the waiting time W_i of the user i arriving to the system, when there are already $i-1$ users waiting, is the sum of the interarrival times of the remaining $U_{\max}-i$ users that should arrive to turn the system to *all-On* state. We model the interarrival time as an exponentially distributed variable. The sum of this kind of random variables is another random variable which follows an *Erlang Distribution* [Wik]. Thus, the distribution of W_i with shape k and rate λ is given by:

$$f_{W_i}(W_i;k,\lambda) = \frac{\lambda^k W_i^{k-1} e^{-\lambda W_i}}{(k-1)!} \qquad \text{for } W_i, \lambda \geq 0 \tag{3.10}$$

where

$$k = U_{\max} - i \tag{3.11}$$

The Complementary Cumulative Distribution Function (CCDF) of W_i is given by:

$$\bar{F}_{W_i}(W_i;k,\lambda) = \sum_{n=0}^{k-1} \frac{1}{n!} e^{-\lambda W_i} (\lambda W_i)^n \tag{3.12}$$

The waiting time is conditioned on the states during which the user enters the system. The user i has a probability $p_{(i-1,0)}$ of entering the system while it is in state $(i-1,0)$, i.e. $i-1$ users are already waiting. And the probability to wait more than D is obtained from the CCDF of the waiting time (3.12). Thus, the probability that the user i waits more than D after entering to the system is given by:

$$P(W_i > D) = p_{(i-1,0)} \bar{F}_{W_i}(D;k,\lambda) \tag{3.13}$$

Generalizing for all the users that can experience some delay, the probability that they wait more than D for starting their services is given by:

$$\gamma = \sum_{i=1}^{U_{\max}-1} P(W_i > D) \tag{3.14}$$

Our objective is to keep γ under acceptable levels. Thus, the selection of the thresholds have to satisfy the following constraint:

$$\gamma \leq \gamma_{\max} \tag{3.15}$$

where $\gamma_{\max}$ is the target probability. For example, if $\gamma_{\max} = 0.05$, at least 95% of the users entering in the system should wait less than D. In 5% of the cases the user may encounter longer delays.

3.4.2.3 Quality of service

While the DTUs are willing to offset the start of their services, once their services started, the system should ensure their continuity. Thus, the system dimensioning not only has to control the waiting time, but also keep an acceptable level of Quality of Service (QoS). The dissatisfaction metric we use is a linear combination of the user *dropping* and *blocking* probabilities, p_{drop} and p_{block} respectively [Lag00]:

$$\delta = \beta p_{\text{drop}} + (1-\beta) p_{\text{block}} \tag{3.16}$$

A new arrival will be blocked with probability p_{block} when the system is not able to serve it, i.e. the system is already using all the available resources. This happens when the dynamic resources are in the state $(C_{\text{dyn}}, 1)$ with probability $p_{(C_{\text{dyn}},1)}$ or when the static resources are in the state $(C_{\text{sta}}, -1)$ with probability $p_{(C_{\text{sta}},-1)}$. An user with an active call can be dropped with probability p_{drop} if the system switches to *min-On* state and there are not enough resources to absorb him. Given the dynamic of the strategy and generalizing for all users in the batch, this probability is given in Equation (3.17) where $p_{(u,-1)}$ is the probability of having u sessions using the static resources and the factor at the right represents the dropped proportion of the batch in the state $(u, -1)$.

$$p_{\text{drop}} = \sum_{u=C_{\text{sta}}-b+1}^{C_{\text{sta}}} p_{(u,-1)} \frac{u+b-C_{\text{sta}}}{b} \tag{3.17}$$

In order to keep an acceptable QoS, these two probabilities should be controlled. As a trade-off parameter, we use $\beta = 0.9$, e.g. user dropping is heavily penalized. Please note that a higher value of the dissatisfaction metric leads to worst performance of the system, e.g. a system which can guarantee $\delta = 0$ is perfect, whereas a system with $\delta = 1$ is non-working. Thus, we define the maximum acceptable value for the metric as $\delta_{\max}$, and the dimensioning of the strategy should satisfy the constraint:

$$\delta \leq \delta_{\max} \tag{3.18}$$

3.4.3 Numerical evaluation

In this section we present the numerical evaluation of the presented DTU-aware strategy adapted to a NRS. We consider three different deployment scenarios allowing to perform cell switching. They are described in the first parts of this section. After we show how to adapt the model presented in the preceding sections to these characteristics and how to estimate the average power consumption of the system when using the DTU-aware strategy. Finally, we provide the optimal results in terms of power consumption and we show how the power reductions depend on the strategy parameters.

Table 3.1: System parameters for the evaluation of Strategy One

Parameter	Macro	Micro
BS power consumption [AGD+11]		
N_{TRX}	6	2
P_{max} [W]	20	6.3
P0 [W]	130	56
Δ_P	4.7	2.6
P_{sleep} [W]	75	39
Transmission Bandwidth	10 MHz	
Antenna configuration	2x2 MIMO	
Downlink [3GP10b]		
N_{RB}	50	
$N_{\text{sc}}^{\text{RB}}$	12	
$EPRE$ [dBm]	15	10
Session characterization		
$N_{\text{RB}}^{\text{AS}}$	10	
W	20	
BS session capacity (C)	100	
Average service time [s] (μ^{-1}) [PP05]	81.083	

3

3.4.3.1 Deployment scenarios

We consider a homogeneous access network in which the BSs are deployed allowing coverage overlapping. Three configurations are identified:

- **Full redundancy:** Two BSs are covering the same area using all radio resources, e.g. two BS in a heavy traffic hotspot using a frequency reuse scheme or carrier aggregation, or two BS from different operators on the same site. We consider the case where both BSs are Macro BSs.
- **Partial redundancy:** The coverage area of a BS is overlapped by the neighboring BSs, e.g. in an urban dense deployment. We consider the case of a Macro BS partially overlapped by its three Macro BSs neighbors.
- **Heterogeneity:** BSs of different types overlap their coverage, e.g. small BSs in hotspots to boost the capacity. In particular, we consider the case of six Micro BSs under the coverage of a Macro BS. We also consider the illustrative scenario where six Macro BSs are under the coverage of a long-range BS that we denoted as Mega BS.

3.4.3.2 Radio resource adaptation

We consider that a subset of BSs in the deployment can be switched to *Sleep Mode (SM)*. A BS enters in SM when all of its cells are turned to a low power consumption state. In particular, we assume that some of the BSs are using the component deactivation hardware upgrade presented in Section 2.3.1 for this purpose. We consider that no control signals are transmitted during the period when the BS is in SM, so the power consumption remains in low state during the entire period. A BS capable of entering in SM is denoted as Sleep-Capable Base Station (SC-BS). Given

the characteristics of the considered deployments, we assume that the coverage in the area is ensured by the remaining active BSs which overlap the area of a SC-BS in SM. Furthermore, the BSs which remain active provide the required radio resources for guaranteeing a minimum capacity in the area. This dynamic system matches to the system model presented in Section 3.3:

- The system is in *min-On* state when the SC-BSs are in SM. The C_{sta} static resources are provided by the BSs that remain active.
- The system is in *all-On* state when the SC-BSs are operational and providing C_{dyn}. The available capacity $C_{\max}$ is the sum of the resources provided by all BSs covering the area.

3.4.3.3 Capacity estimation

We consider that the maximal number of concurrent sessions depends on the downlink physical resource allocation of the cell. In LTE systems the downlink transmission scheme uses Orthogonal Frequency-Division Multiplexing (OFDM). The smallest physical resource that can be allocated to a session is a *Resource Block (RB)*. This corresponds to a given number of subcarriers in the frequency domain and one subframe (i.e. 1 ms) in time domain. We denote as N_{RB} the number of resource blocks of the entire downlink bandwidth, W is the number of considered subframes in the periodic allocation, and $N_{\text{RB}}^{\text{AS}}$ is the number of resource blocks per active session. We model $N_{\text{RB}}^{\text{AS}}$ as a constant number to represent an average behaviour of the users in the cell. In reality this number depends on the UE channel condition and the adequate modulation and coding scheme used for reliable communication. Thus, the session capacity of the considered BS downlink resource grid is given by Equation (3.19). The capacities (C_{sta} and $C_{\max}$) of the different scenarios varies depending on the number of active BSs in each system state.

$$C = \frac{N_{\text{RB}} W}{N_{\text{RB}}^{\text{AS}}} \tag{3.19}$$

3.4.3.4 Power consumption

The BS power consumption model we use was developed in the context of the EARTH Project and introduced by Auer et al. [AGD+11]. This model relates P_{out} (the output power radiated at the antenna) and P_{in} (the total power needed by the BS to operate) for each type of LTE BS. The power model is well approximated by:

$$P_{\text{in}} = \begin{cases} N_{TRX}(P_0 + \Delta_P P_{\text{out}}) & 0 < P_{\text{out}} < P_{\max} \\ N_{TRX} P_{sleep} & P_{\text{out}} = 0 \end{cases} \tag{3.20}$$

where N_{TRX} is the number of transceiver chains, 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 BS and P_{sleep} represents the power consumption of the BS in SM. However, in our model the condition of sleep mode power consumption is not conditioned by P_{out}, but by the explicit system state as it will be explained in the following. The different parameter values we consider for the evaluation are given in Table 3.1. We model the power consumption of the Mega BS as a Macro BS due to a lack of appropriate model.

In order to have a better approximation of P_{in} we calculate P_{out} as a function of the number of active sessions in the BS. This is possible because P_{out} depends on the BS physical resource allocation. The BS determines the downlink transmit power per data RB (P_{RB}), which is constant on average in each modulated symbol and depends of the energy per resource element ($EPRE$) fixed in the BS.

The calculation of the average output power is dependent only on the number of allocated RBs and does not depend on their time/frequency distribution. Thus, P_{out} is given by Equation (3.21), where N_{RB} is the number of resource blocks of the entire downlink bandwidth and $\alpha(i, C) \in [0; 1]$ is a factor representing the fraction of assigned resources depending on the number of concurrent sessions.

$$P_{\text{out}}(i, C) = \alpha(i, C) N_{\text{RB}} P_{\text{RB}} \tag{3.21}$$

For the considered model, α is given by (3.22) where i is the number of active sessions in the BS and C the session capacity of the considered BS downlink resource grid.

$$\alpha(i, C) = \frac{i}{C} \tag{3.22}$$

Combining Equations (3.21) and (3.20) provides P_{in} as a function of the number of sessions in service. We denote the BS power consumption when serving i users as $P_{\text{in}}(i)$. When the BS is in sleep mode, the power consumption P_{in} is independent of the number of sessions on hold and is constant. In this case we denote it as $P_{\text{in}}(\text{sleep})$.

In order to estimate the average power consumption of the BS, we use the steady state probabilities of the corresponding MC which models its traffic behaviour. When considering a SC-BS representing the dynamic resources of the system, the average power consumption is given by:

$$\overline{P_{\text{in}}^{\text{dyn}}} = \sum_{i=U_{\min}}^{C_{\text{dyn}}} p_{(i,1)} P_{\text{in}}(i) + \sum_{i=0}^{U_{\max}-1} p_{(i,0)} P_{\text{in}}(\text{sleep}) \tag{3.23}$$

Considering the always active BSs, representing the static resources of the system, the average power consumption is given by:

$$\overline{P_{\text{in}}^{\text{sta}}} = \sum_{i=0}^{C_{\text{sta}}} p_{(i,-1)} P_{\text{in}}(i) \tag{3.24}$$

3.4.3.5 Scenario considerations

For the considered scenarios, the aggregated offered load of the system is given by:

$$A = \frac{\lambda_{\text{sys}}}{\mu} \tag{3.25}$$

The parameter μ is fixed and its value is given in Table 3.1, thus the different offered load levels we consider are result of the variation of the interarrival time distribution. The proportion of DTU in the system is denoted with the parameter η. Thus, the respective system arrival rates for DTU and N-DTU users are:

$$\lambda_{\text{DTU}} = \eta\lambda_{\text{sys}} \tag{3.26}$$

$$\lambda_{\text{N-DTU}} = (1-\eta)\lambda_{\text{sys}} \tag{3.27}$$

The different scenarios are composed of several SC-BSs and Always Active BSs as shown in Table 3.2. We denote as n and m the number of SC-BSs and Always Active BSs in the scenario respectively. In the case of m BSs taking care of the load of the SC-BS when entering in SM, we consider that given the regular patterns of the considered scenarios, the batch of users is evenly distributed between the m Always Active BSs. Thus, the batch size is given by:

$$b = \left\lceil \frac{U_{\min} - 1}{m} \right\rceil \tag{3.28}$$

In the scenarios where n SC-BSs can enter in SM, the Always Active BSs receive batch arrivals more frequently. Thus, the batch arrival is given by:

$$\lambda' = np_{(U_{\min},1)}U_{\min}\mu \tag{3.29}$$

We consider that the load varies homogeneously in the scenario. However, the N-DTU arrivals are served by any type of BSs, while the DTU arrivals are only considered by the SC-BSs. Thus, the parameters for the model are:

$$\lambda = \frac{\lambda_{\text{DTU}}}{n} \tag{3.30}$$

$$\lambda_1 = \frac{\lambda_{\text{N-DTU}}}{n+m} \tag{3.31}$$

3.4.3.6 Optimization

For a given scenario, offered load (A), maximal tolerable delay (D) and proportion of DTU (η), we are only interested in the combination of strategy parameters that minimize the total scenario average power consumption, while satisfying the waiting

Table 3.2: Scenarios for the evaluation of Strategy One.

	Full Redundancy	**Partial Redundancy**	**Heterogeneity**			
Scenario	1	2	3	4	5	6
SC-BSs type	Macro BS	Macro BS	Macro BS	Micro BS	Mega BS	Macro BS
n	1	1	1	6	1	6
C_{dyn}	100	100	100	600	100	600
AA-BSs type	Macro BS	Macro BS	Micro BS	Macro BS	Macro BS	Mega BS
m	1	3	6	1	6	1
C_{sta}	100	300	600	100	600	100
C_{max}	200	400	700	700	700	700
Maximum system load for strategy suitability						
$\eta = 1.0$	$0.5C_{\text{max}}$	$0.25C_{\text{max}}$	$0.14C_{\text{max}}$	$0.9C_{\text{max}}$	$0.14C_{\text{max}}$	$0.9C_{\text{max}}$
$\eta = 0.5$	$0.66C_{\text{max}}$	$0.4C_{\text{max}}$	$0.24C_{\text{max}}$	$0.95C_{\text{max}}$	$0.24C_{\text{max}}$	$0.95C_{\text{max}}$
$\eta = 0.2$	$0.83C_{\text{max}}$	$0.63C_{\text{max}}$	$0.45C_{\text{max}}$	$5.82C_{\text{max}}$	$0.45C_{\text{max}}$	$5.82C_{\text{max}}$

time and the dissatisfaction constraints. Thus, we perform an exhaustive search of the parameters generating the MCs that solve the following optimization problem:

$$\begin{aligned} \underset{U_{\text{min}},U_{\text{max}}}{\text{minimize}} \quad & n\overline{P_{\text{in}}^{\text{dyn}}} + m\overline{P_{\text{in}}^{\text{sta}}} \\ \text{subject to} \quad & \gamma \leq \gamma_{\text{max}} \\ & \delta \leq \delta_{\text{max}} \\ & U_{\text{max}} \leq C_{\text{dyn}} \\ & U_{\text{min}} \leq U_{\text{max}} \end{aligned} \tag{3.32}$$

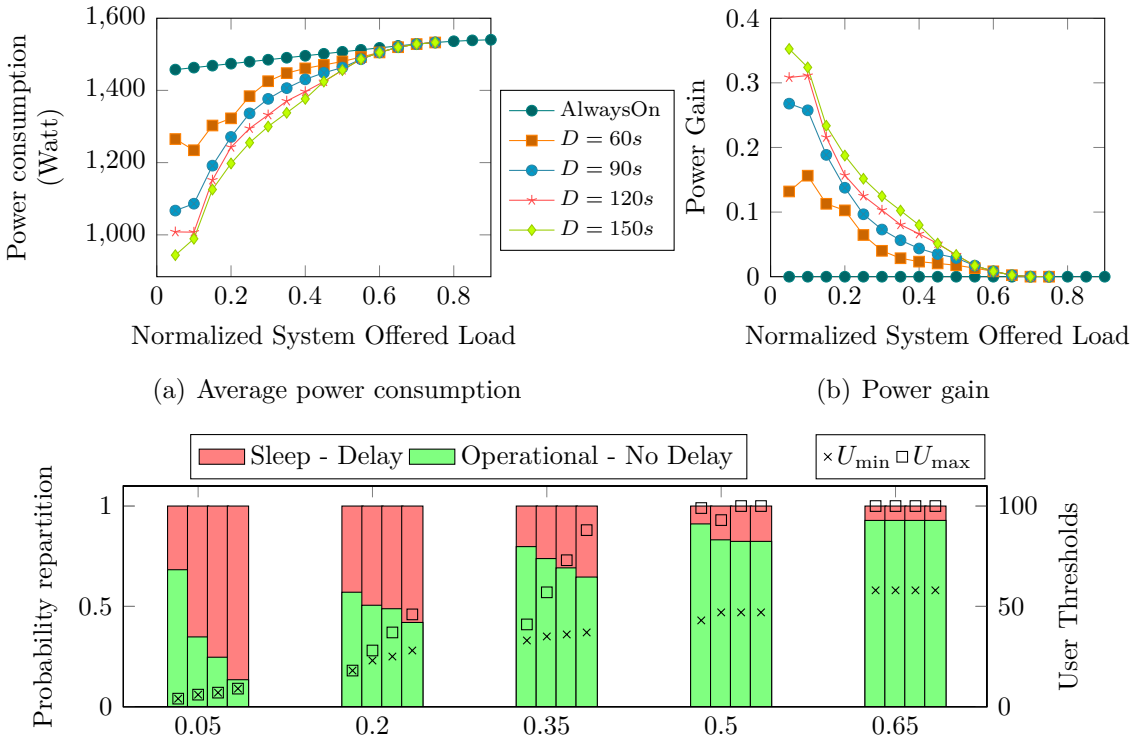
Thus, for each combination of U_{min} and U_{max} the corresponding MC is generated and solved. The performance parameters, i.e. the probability the users wait more than D (γ) and the dissatisfaction metric (δ), are calculated based in the resulting steady state probabilities. Finally, the constraints are evaluated and for each set of strategy thresholds U_{min} and U_{max} satisfying the constraints, we choose the optimal as the one with the minimal scenario average power consumption.

3.4.3.7 Results

In this section we present the optimal results obtained through numerical evaluation for the scenarios summed up in Table 3.2. We present the impact of the strategy parameters on its performance using Scenario 4 as reference. Afterwards, we discuss the capacity difference between scenarios and how it affects the numerical evaluation and the suitability of the DTU-aware strategy. Detailed result graphics for the other suitable scenarios are given in Annex B.

Impact off the strategy parameters

Figure 3.3 presents the results for the Scenario 4, considering four different D proposed to the users, and full user cooperation ($\eta = 1$). The strategy is compared to the scenario in which all the BSs remain operational affording the same load repartition than the proposed strategy. This last strategy is labelled as Always On.



(a) Average power consumption

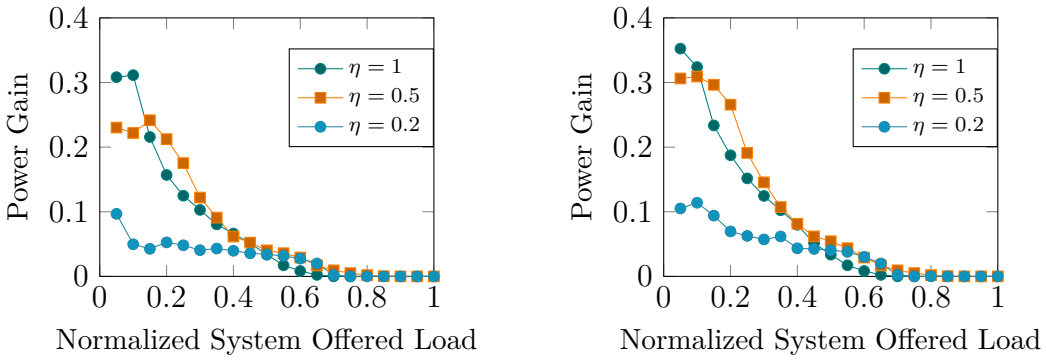
(b) Power gain

(c) Dynamic resources operation probability and strategy thresholds depending on A and D

Figure 3.3: Results for the Scenario 4, $\eta = 1$,$\gamma_{\max} = 0.05$, $\delta_{\max} = 0.05$.

Figure 3.3(a) presents the average power consumption and Figure 3.3(b) presents the power gain respect to the baseline scenario, i.e. the proportion of average power that is saved when using the proposed strategy. Finally, Figure 3.3(c) presents the optimal strategy thresholds for achieving these results and the repartition of the operation probabilities for the dynamic resources. We observe power gains up to 35% when using the proposed strategy. Notice that the power gain is due to the fact that the SC-BSs switch to SM. However, the power consumption in SM is still relatively high. In the case of Macro BSs it represents 58% of the power consumed by an operational BS without carrying traffic, and 70% in the case of Micro BSs. These values are expected to decrease with the constant technology innovations. Thus, greater gains can be envisaged when more load proportional hardware will be used.

In general we observe that increasing D until a given upper bound increases the probability that the dynamic resources are deactivated, as shown in Figure 3.3(c). Note however that the maximum D providing additional benefits is upper-bounded for a given A. The optimization problem we use identifies the combination of thresholds that minimizes the average power consumption while satisfying the delay con-



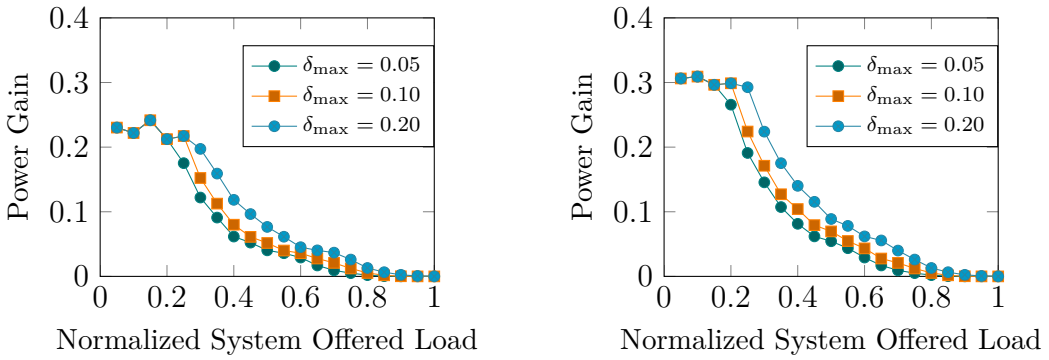
(a) $D = 120s$ (b) $D = 150s$

Figure 3.4: Power gain variation for different levels of η. Scenario 4, fixed $\delta_{\max} = 0.05$ and $\gamma_{\max} = 0.05$.

straint. In some cases, for a given A, there are no combination of thresholds that can provide better performance for a larger D. Thus, users wiling to wait more time do not provide further gains for the system given the performance constraints. For example, in Fig. 3.3(c), when the offered load is at 50% of the system capacity, the optimal thresholds (and the performance of the strategy) are the same for $D = 120s$ and $D = 150s$. Thus, it makes no sense to ask users to wait longer than $120s$ for this level of load. In very low loads, less arrivals occur and the system needs to activate the dynamic resources more often to satisfy the delay constraint. For example, this is the case of $D = 60s$ in Figure 3.3(c) where the performance of the strategy is worse for a normalized offered load of 0.05 than for 0.2.

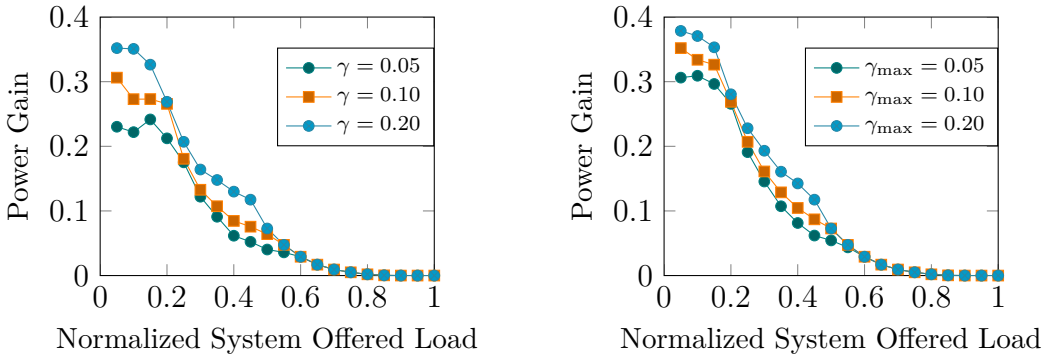
Figure 3.4 shows the variation of the power gain if the proportion DTU η changes in the system. The maximal gain is observed in the lowest loads and decreases with η. However, while the system offered load increases, the gains increase when η decreases. This is because the activation of the dynamic resources is governed by the DTU traffic. If the DTU arrivals decrease, in low loads the system should activate the dynamic resources to satisfy the delay constraint as explained before. When the load increases, the strategy becomes efficient and the dynamic resources are operating with less load if η decreases, which explains the overall better performance when $\eta = 0.5$ in Figure 3.4. However, when η is small the gain decreases as the dynamic resources are in low load regimen, and the resources are activated often to satisfy the delay constraint of the DTUs.

Figure 3.5 and Figure 3.6 show that when the dissatisfaction and delay constrains are relaxed, the strategy provides more benefits to the system in terms of power consumption reduction. We also confirm that in low loads, the performance of the strategy is bounded by the delay constraint. Contrary to the dissatisfaction metric, more power savings are achieved for these levels of load when the delay constraint is relaxed. The opposite is observed in increased level of loads, where the strategy



(a) $D = 120s$ (b) $D = 150s$

Figure 3.5: Power gain variation for different levels of δ_{max}. Scenario 4, fixed $\eta = 0.5$ and $\gamma_{max} = 0.05$.



(a) $D = 120s$ (b) $D = 150s$

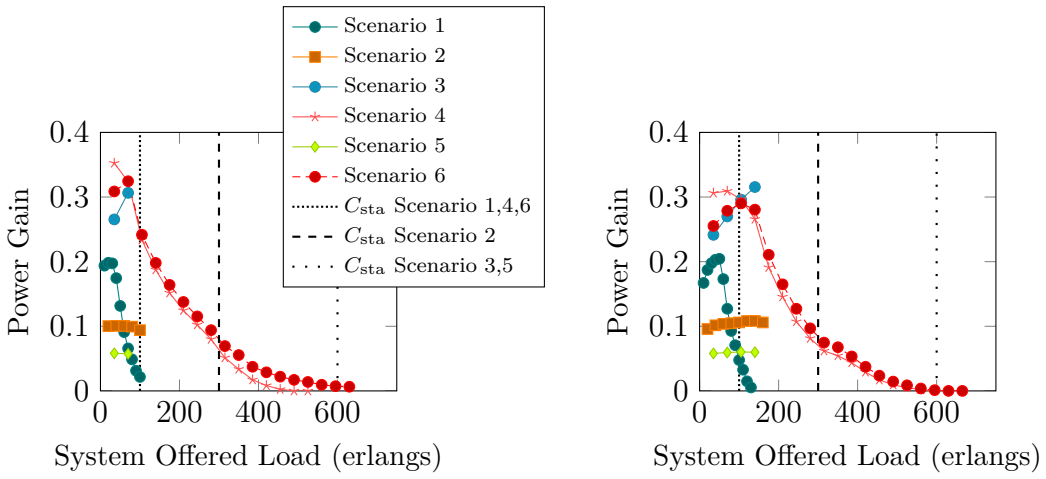
Figure 3.6: Power gain variation for different levels of γ_{max}. Scenario 4, fixed $\eta = 0.5$ and $\delta_{max} = 0.05$.

performs similarly for different delay constraints, but obviously perform better for higher levels of tolerated dissatisfaction.

Resource capacity impact

The modelled DTU-aware strategy controls the dynamic resources of the system maximizing their utilization while taking benefits of their adaptability to reduce the overall power consumption. Thus, the strategy is active and stable when the offered load experienced by the dynamic resources is under their capacity. The maximum system offered load for which the strategy is active depends of η as well, as the it is distributed between both type of resources. These values are given in Table 3.2.

In order that the application of the strategy provides benefits to the system, it should be active and representing gains in system offered loads beyond the capacity of the static resources. Otherwise, the system can afford the load using only the



(a) $\eta = 1$ (b) $\eta = 0.5$

Figure 3.7: Power gain variation for the different scenarios. Two different η. Fixed $D = 150s$, $\delta_{\max} = 0.05$ and $\gamma_{\max} = 0.05$.

static resources, and the prioritization of the utilization of the dynamic resources by the proposed strategy is not worth. As observed in Table 3.2 and in Figure 3.7 Scenarios 2, 3 and 5 do not meet this condition. Thus, the strategy is not suitable for them. We still evaluate their performance for the offered loads in which the strategy is active, as observed in Figure 3.7. The power gain is inferior to the contribution in power consumption of the dynamic resources, e.g. for the Scenario 2, the average power reduction when using the DTU-strategy is 10%, while the consumption of the dynamic resources represents 25% of the total consumption when active. In this case is better to completely turn off the dynamic resources, and only activate them in case of congestion of the static resources. The DTU-aware strategy proposed in the next section follows this approach.

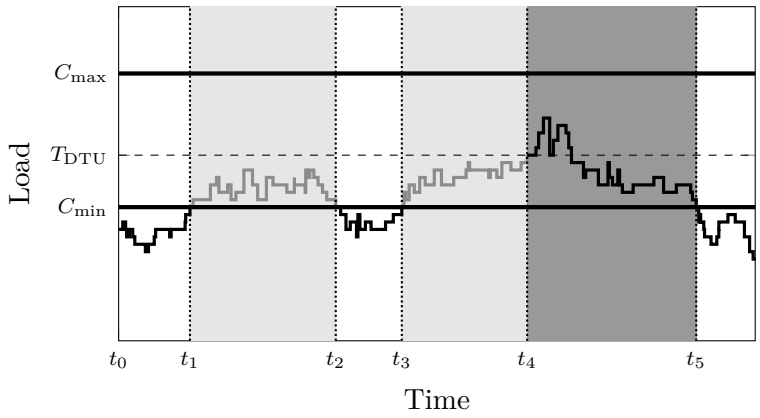


Figure 3.8: **Load dynamic example of the Strategy Two**. White periods: system in *min-On* state – No delay. Light gray periods: system in *min-On* state – Delaying users. Dark gray periods: system in *all-On* state – No delay.

3.5 Strategy Two: Opportunistic DTU

3.5.1 Strategy description

This strategy makes no distinction between the user types (DTU or N-DTU). The delay in the start of the services is then conditioned on the number of users in the system and not by a persistent user choice. When the system is in *min-On* state and the number of active users is below C_{sta}, all arrivals are served right away. Users will wait only in periods of congestion of the static resources. This is, users will be accepted in the system even if (instantaneously) there are not enough available resources to serve them. Thus, the start of their service will be shifted until the needed resources become available. This condition can be satisfied either due to the liberation of some static resources, or by the activation of the dynamic ones. We define a threshold in the number of users $T_{\text{DTU}} > C_{\text{sta}}$. During the periods where the number of users is in-between C_{sta} and T_{DTU}, i.e. when there are some users waiting, the system will stay in *min-On* state. The actual switching to *all-On* state depends on the number of waiting users and it will occur when the threshold T_{DTU} is reached. To ensure the service after the initial delay, the system should not accept more users than the capacity. Thus, T_{DTU} is bounded by $C_{\max}$. The system is switched to *min-On* state when no more extra capacity is needed, i.e. when the number of users falls below C_{sta}.

An example of this strategy is depicted in Figure 3.8. In the period from t_1 to t_2 the system is in *min-On* state and above its capacity, thus some users are on hold. At the end of this period, the system load descends under C_{sta} which means that the service of the waiting users started without the need of turning the system to

all-On state. On the contrary, at t_4 the system has to switch to *all-On* state (T_{DTU} is reached) to serve the waiting users and the new arrivals.

Users waiting in the system will be served either because some static resources become available, i.e. some user departures occur, or because the system increases its capacity to $C_{\max}$ (T_{DTU} was reached and the C_{dyn} resources were added). Thus, the selection of T_{DTU} is critical to ensure the service before the maximal delay D. In the following section, we present the MC modelling this strategy and the analysis of the selection of T_{DTU} providing an user waiting time less than D.

3.5.2 Mathematical model

We model this strategy using the MC which state space is depicted in Figure 3.9. This MC is explained and solved in the first part of this section. The following parts describe the different criteria used for the selection of the strategy threshold in order to bound the waiting of the users and ensuring the quality of service for them.

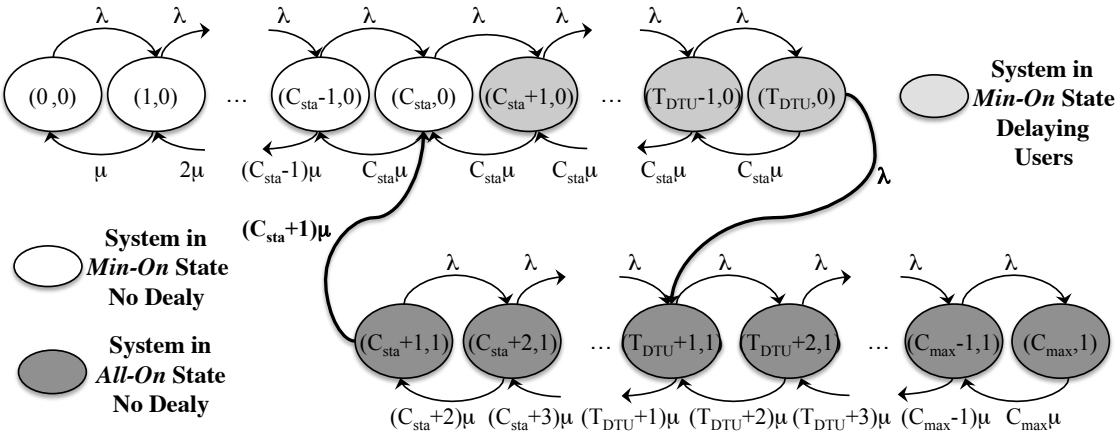


Figure 3.9: Markov Chain of the user dynamic using Strategy Two.

3.5.2.1 Markov chain

This strategy is modelled using the MC depicted in Fig. 3.9. Only one arrival rate λ is represented as there is no predefined distinction between N-DTUs and DTUs. In this strategy, users will wait only in congestion periods when the system is in *min-On* state. The model represents this in the states $(i, 0)$ when $C_{\text{sta}} \leq i \leq T_{\text{DTU}}$. In these states, the service rate is limited to $C_{\text{sta}}\mu$ as the system is only capable of serving up to C_{sta} simultaneous users when in *min-On* state. Thus, some arrivals are delayed. The service rate is $i\mu$ otherwise. The system state transitions are represented with the thicker arrow in Fig. 3.9 and they are:

- The switch *all-On* $\rightarrow$ *min-On*: the state is $(C_{\text{sta}}+1, 1)$ and a departure occurs
- The switch *min-On* $\rightarrow$ *all-On*: the number of users in the system surpasses T_{DTU}

The balance equations of the MC modelling the system user dynamic this strategy are:

$$\begin{aligned}
\lambda p_{(i,0)} &= i\mu p_{(i+1,0)} && \text{if } i = 0 \\
(\lambda + i\mu)p_{(i,0)} &= (i+1)\mu p_{(i+1,0)} + \lambda p_{(i-1,0)} && \text{if } 0 < i < C_{\text{sta}} \\
(\lambda + i\mu)p_{(i,0)} &= \lambda p_{(i-1,0)} + i\mu p_{(i+1,0)} + (i+1)\mu p_{(i+1,1)} && \text{if } i = C_{\text{sta}} \\
(\lambda + C_{\text{sta}}\mu)p_{(i,0)} &= (i+1)\mu p_{(i+1,0)} + \lambda p_{(i-1,0)} && \text{if } C_{\text{sta}} < i < T_{\text{DTU}} \\
(\lambda + C_{\text{sta}}\mu)p_{(i,0)} &= \lambda p_{(i-1,0)} && \text{if } i = T_{\text{DTU}} \\
(\lambda + i\mu)p_{(i,1)} &= (i+1)\mu p_{(i+1,0)} && \text{if } i = C_{\text{sta}} + 1 \\
(\lambda + i\mu)p_{(i,1)} &= (i+1)\mu p_{(i+1,1)} + \lambda p_{(i-1,1)} && \text{if } C_{\text{sta}} + 1 < i < T_{\text{DTU}} + 1 \\
(\lambda + i\mu)p_{(i,1)} &= (i+1)\mu p_{(i+1,1)} + \lambda p_{(i-1,1)} + \lambda p_{(T_{\text{DTU}},0)} && \text{if } i = T_{\text{DTU}} + 1 \\
(\lambda + i\mu)p_{(i,1)} &= (i+1)\mu p_{(i+1,1)} + \lambda p_{(i-1,1)} && \text{if } T_{\text{DTU}} + 1 < i < C_{\max} \\
i\mu p_{(i,1)} &= \lambda p_{(i-1,1)} && \text{if } i = C_{\max}
\end{aligned} \tag{3.33}$$

From Equation (3.33) we can deduce the transition matrix Q_{two}. The MC is irreducible and consists of positive recurrent states. The unique steady-state probability vector $\pi_{\text{two}} = \{p_{(i,j)}\}$ is given by:

$$\pi_{\text{two}} Q_{\text{two}} = 0 \tag{3.34}$$

$$\sum_{i=0}^{T_{\text{DTU}}} p_{(i,0)} + \sum_{i=C_{\text{sta}}+1}^{C_{\max}} p_{(i,1)} = 1 \tag{3.35}$$

3.5.2.2 Waiting time

In this strategy a waiting user has two possible ways of being served. We consider the user i arriving when there are already m users waiting in the system, i.e. there are $n = C_{\text{sta}} + m$ users already in the system. The service of the user i will take place either when m departures occur, i.e. the user is in front of the waiting queue; or when the system arrives to the state $(T_{\text{DTU}} + 1, 1)$ and the switching to *all-On* state is performed. Thus, there are different sequences of events (arrivals and departures) which can lead the user i to be served, and the service will take place after the sequence of event that occurs first, i.e. the sequence that takes the minimum amount of time to happen. Moreover, in each state, the next event will take place after a time t_e which is the minimum between the time a new arrival occurs and the time an ongoing service finishes. As the interarrival time and the service time are exponentially distributed, the minimum between these two random variables is

another random variable which is exponentially distributed as well, and it is given by:

$$f_{t_e}(t_e;\phi) = \phi e^{-\phi t_e} \qquad t_e \geq 0 \tag{3.36}$$

where

$$\phi = C_{\text{sta}}\mu + \lambda \tag{3.37}$$

Each event is an arrival or a departure with probability $p_A = \lambda/\phi$ or $p_D = \mu/\phi$ respectively. We consider a set of events S_a^d containing n_a arrivals and n_d departures, such as there are l_a^d sequences of these events that can lead the user i to be served. Notice that $l_a^d = 0$ if there are no possible combinations of the set of events leading the user to be served. Thus, the occurrence probability of S_a^d is given by:

$$p_{S_a^d} = l_a^d({p_A}^{n_a}{p_D}^{n_d}) \tag{3.38}$$

The time after which a sequence of S_a^d occurs ($T_{S_a^d}$) is the sum of the time each event takes (t_e). As t_e is exponentially distributed, the sum of this kind of random variable is a random variable which follows an *Erlang Distribution* [Wik] and is given by:

$$f_{T_{S_a^d}}(T_{S_a^d};k,\phi) = \frac{\phi^k {T_{S_a^d}}^{k-1} e^{-\phi T_{S_a^d}}}{(k-1)!} \qquad \text{for } T_{S_a^d}, \phi \geq 0 \tag{3.39}$$

where

$$k = n_a + n_d \tag{3.40}$$

The Complementary Cumulative Distribution Function (CCDF) of $T_{S_a^d}$ is given by:

$$\bar{F}_{T_{S_a^d}}(T_{S_a^d};k,\phi) = \sum_{n=0}^{k-1} \frac{1}{n!} e^{-\phi T_{S_a^d}} (\phi T_{S_a^d})^n \tag{3.41}$$

The waiting times are conditioned on the states during which the user enters the system. The user i has a probability $p_{(i-1,0)}$ of entering the system while in state $(i-1,0)$. The probability she/he waits more than D if she/he is served by the sequence of events S_a^d is obtained from the CCDF of $T_{S_a^d}$ (3.41). Thus, the probability that the sequence of events S_a^d leads the user i to wait more than D after entering the system is given by:

$$P(T_{S_a^d} > D) = p_{(i-1,0)} p_{S_a^d} \bar{F}_{T_{S_a^d}}(D;k,\phi) \tag{3.42}$$

Generalizing for all possible sequences of events and denoting with W_i the waiting time of the user i, the probability that the user i waits more than D after entering the system is given by:

$$P(W_i > D) = \sum_{a=0}^{h} \sum_{d=0}^{m} P(T_{S_a^d} > D) \tag{3.43}$$

where h represents the maximum number of arrivals that a sequence leading the user to be served can contain, and it is calculated using Equation (3.44). Note that if m departures occur, the user is served as he/she is in the front of the waiting queue. However, the dynamic for being served by the dynamic resources when switching to *all-On* state is not that simple, as it is dependent on the number of users in the system. A sequence containing departures, needs to be compensated by more arrivals, in order that the number of users in the system reach T_{DTU} and lead the user to be served by the system in *all-On* state. Thus, the calculation of h accounts for the worst case sequence.

$$h = T_{\mathrm{DTU}} - C_{\mathrm{sta}} + m \tag{3.44}$$

Generalizing to all users that can experience some delay, the probability that they wait more than D for starting their services is given by:

$$\gamma = \sum_{i=C_{\mathrm{sta}}+1}^{T_{\mathrm{DTU}}} P(W_i > D) \tag{3.45}$$

Our objective is to keep γ under acceptable levels. Thus, the selection of T_{DTU} has to satisfy the following constraint:

$$\gamma \leq \gamma_{\max} \tag{3.46}$$

where $\gamma_{\max}$ is the target probability. For example, if $\gamma_{\max} = 0.05$, at least 95% of the users entering the system will wait less than D. In 5% of the cases the user may encounter longer delays.

3.5.2.3 Quality of service

This strategy switches to *min-On* state when the C_{sta} resources are enough to serve the users in the system. Thus, no drop of ongoing communication is expected. In the same way, new arrivals are delayed and not blocked when the system is over the capacity in *min-On* state. The system will refuse to serve a new arrival only when the capacity of the system is reached. Thus, the dissatisfaction metric used in this strategy is the blocking probability when the system is in *all-On* state, and is given by:

$$\delta = p_{\mathrm{block}} = p_{(C_{\max},1)} \tag{3.47}$$

In order to keep an acceptable QoS, this probability should be controlled. Thus, we define the maximum acceptable value for the metric as $\delta_{\max}$, and the dimensioning of the strategy should satisfy the constraint:

$$\delta \leq \delta_{\max} \tag{3.48}$$

3.5.3 Numerical evaluation

In this section we present the numerical evaluation of the proposed DTU-aware strategy adapted to two different energy efficiency strategies which take action in a standalone BS scope, optimizing the operation of the BS with advanced and adaptive hardware. Afterwards, we show how to adapt the model presented in the preceding sections to these characteristics and how to quantify the average power consumption when using the DTU-aware strategy. Finally, we provide the optimal results in terms of power consumption for different levels of offered load.

3.5.3.1 Deployment

The reference scenario is a flat deployment of 3-sectorised Macro BSs with 10MHz of bandwidth and 2x2 MIMO antennas, each of them transmitting up to 20W of output power. Each BS is using adaptive transceiver chains and smart antennas (see Section 2.3.1). We consider a BS site density corresponding to dense urban scenario with 500 meters of inter site distance. We analyse a network region composed of five BS sites.

3.5.3.2 Radio resource adaptation

We consider two scenarios depending on the radio resource adaptation technique that is used:

- **Dynamic sectorization:** it corresponds to a NRS in which the number of active sectors/cells of the BS varies depending on the traffic conditions [HG11].

Table 3.3: System Parameters for the evaluation of Strategy Two.

Deployment [EAR12b]	
Deployment Type	Urban
Inter Site Distance $[m]$	500
Site Area $[Km^2]$	0.2165
Number of BS sites	5
BS type	Macro 3-sector Reconfigurable
Transmission Bandwidth [MHz]	10
Antenna configuration	2x2 MIMO
Total Resource Blocks	50
BS power consumption [AGD$^+$11]	
N_{TRX}	6, 4, 2
P_{max} [W]	20
P0 [W]	130
Δ_P	4.7
Traffic characterization	
System capacity $[Mbps/Km^2]$ [EAR12c]	115
Session target throughput $[kbps]$	500
System session capacity (C_{max})	235
Average service time [s] (μ^{-1}) [PP05]	81

Some cells of the BS are deactivated, reducing the overall power consumption. The remaining active cells are modified using the beamforming capabilities of the smart antennas to compensate coverage. This BS dynamic behaviour fits to the system model presented in Section 3.3:
  - The system is in *min-On* state when some sectors of the BS are not active. The C_{sta} resources are provided by the sectors that remain active.
  - The system is in *all-On* state when all the BS sectors are operational. The available capacity $C_{\max}$ is the sum of the resources provided by all the BS sectors.
- **Capacity adaptation:** it corresponds to a RRM strategy in which the number of usable RB are limited, allowing the adaptation of the PA operation point, reducing the power consumption of the cells [EAR12d]. In order to fit to the system model presented in Section 3.3, we consider the case in which the transceiver chain switches only between two operating points:
  - The system is in *min-On* state when the number of RB is limited. The C_{sta} resources are provided by the usable RBs.
  - The system is in *all-On* state when the cell can use all the bandwidth providing a total capacity of $C_{\max}$.

3.5.3.3 Capacity estimation

The infrastructure of LTE allows the operators to provide data based services with real-time quality of service constraints, such as video calls. The system capacity ($C_{\max}$) we consider is obtained from a similar scenario in the literature and is given in Table 3.3. This capacity is calculated in order to provide a minimum quality of service for high definition video transmission. However, we consider a minimum of 500 kbps as the target for an acceptable video session quality [Mic]. When a video session 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 UE. Considering static uniformly distributed users we account for the average behaviour assuming that each video call uses in average approximately the same number of RB. Therefore, we consider that the capacity and the offered load are linear functions of the number of simultaneous sessions. The system session capacity is given in Table 3.3. These values are obtained considering homogeneous traffic demand. Thus, the capacity and the load are uniformly distributed between BSs. The values of C_{sta} depend on the applied radio resource adaptation strategy and the considered parameters for each of them. The scenarios we evaluated are summarized in Table 3.5 and the description of the parameters are given in the following section.

3.5.3.4 Power consumption

The BS power consumption model we use for the evaluation of the strategies is based in the model of Auer et al. [AGD+11], which we already used in Section 3.4.3.4. However, we adapt it to the power efficient strategies we are considering. The parameters values for the baseline model are given in Table 3.3, while the strategies specific parameters and equations are given in this section.

Dynamic sectorization

When using the dynamic sectorization some transceiver chains are completely deactivated and their power consumption is near zero Watt. Thus, the BS power consumption depends on the number of active transceiver chains N_{TRX}. Furthermore, C is the maximum number of sessions the BS can serve with the N_{TRX} active resources. Thus, the BS power consumption is given by:

$$P_{\mathrm{in}}^{N_{\mathrm{TRX}}}(i,C) = N_{\mathrm{TRX}}(P_0 + \Delta_P P_{\mathrm{out}}(i,C)) \tag{3.49}$$

We remind the assumptions made in the Section 3.4.3.4, where we have considered that the calculation of the average output power is dependent only on the number of allocated RBs and does not depend on their time/frequency distribution. Thus, the output power is given by Equation(3.50), where N_{RB} is the number of resource blocks of the entire downlink bandwidth and $\alpha(i,C) \in [0;1]$ is a factor representing the fraction of assigned resources depending on the number of concurrent sessions, which is calculated using Equation (3.51).

$$P_{\mathrm{out}}(i,C) = \alpha(i,C) N_{\mathrm{RB}} P_{\mathrm{RB}} \tag{3.50}$$

$$\alpha(i,C) = \frac{i}{C} \tag{3.51}$$

We denote the number of active transceiver chains when the system is in *min-On* state or in *all-On* state as N_{min} and N_{all} respectively. Thus, the average power consumption is given by:

$$\overline{P_{\mathrm{in}}} = \sum_{i=0}^{C_{\mathrm{sta}}} p_{(i,0)} P_{\mathrm{in}}^{N_{\mathrm{min}}}(i, C_{\mathrm{sta}}) + \sum_{i=C_{\mathrm{sta}}+1}^{T_{\mathrm{DTU}}} p_{(i,0)} P_{\mathrm{in}}^{N_{\mathrm{min}}}(C_{\mathrm{sta}}, C_{\mathrm{sta}}) + \sum_{i=C_{\mathrm{sta}}+1}^{C_{\mathrm{max}}} p_{(i,1)} P_{\mathrm{in}}^{N_{\mathrm{all}}}(i, C_{\mathrm{max}}) \tag{3.52}$$

Capacity adaptation

When using the capacity adaptation strategy, the power consumption of the BS depends on the number of usable cell RBs, as the PA is optimized to operate below a maximal signal load level. Depending on the used operating point (n), the maximal signal load is limited to ϕ_n and its power consumption is reduced by a factor denoted

Table 3.4: Adaptive transceiver chain operating points (n) and their associated maximal signal load (ϕ) and power reduction factor (θ). Source: [EAR12d].

n	1	2	3	4	5	6
ϕ_n	1	0.79	0.63	0.5	0.39	0.31
θ_n	0	0.06	0.09	0.13	0.18	0.23

as θ_n. The different values for these parameters are shown in Table 3.4 Thus, the BS power consumption is given by:

$$P_{\text{in}}^{\theta,\phi}(i) = (1-\theta)N_{\text{TRX}}(P_0 + \Delta_P P_{\text{out}}(i,\phi)) \tag{3.53}$$

In this case the output power also depends on the fraction of assigned RBs which is proportional to the number of concurrent sessions, represented by the factor $\alpha(i,\phi) \in [0;1]$ given in Equation (3.55).

$$P_{\text{out}}(i,\phi) = \alpha(i,\phi)N_{\text{RB}}P_{\text{RB}} \tag{3.54}$$

$$\alpha(i,\phi) = \frac{i}{\phi C_{\max}} \tag{3.55}$$

When in *all-On* state the system will use the operating point $n = 1$, while when in *min-On* state different cases are evaluated as shown in Table 3.5. In general, the average power consumption is given by:

$$\overline{P_{\text{in}}} = \sum_{i=0}^{C_{\text{sta}}} p_{(i,0)} P_{\text{in}}^{\theta_n,\phi_n}(i) + \sum_{i=C_{\text{sta}}+1}^{T_{\text{DTU}}} p_{(i,0)} P_{\text{in}}^{\theta_n,\phi_n}(C_{\text{sta}}) + \sum_{i=C_{\text{sta}}+1}^{C_{\max}} p_{(i,1)} P_{\text{in}}^{\theta_1,\phi_1}(i) \tag{3.56}$$

3.5.3.5 Optimization

For a given scenario, offered load (A) and maximal tolerable delay (D), we are only interested in the combination of parameters that minimize the total scenario average power consumption, while satisfying the waiting time and the dissatisfaction constraints. Thus, we perform an exhaustive search of the parameters that solves the following optimization problem:

$$\begin{aligned} \underset{T_{\text{DTU}}}{\text{minimize}} \quad & \overline{P_{\text{in}}} \\ \text{subject to} \quad & \gamma \leq \gamma_{\max} \\ & \delta \leq \delta_{\max} \\ & T_{\text{DTU}} \leq C_{\max} \end{aligned} \tag{3.57}$$

Thus, for all the possible values of T_{DTU}, the corresponding MC is generated and solved. The performance parameters, i.e. the probability the users wait more than D (γ) and the dissatisfaction metric (δ), are calculated based in the resulting steady state probabilities. Finally, the constraints are evaluated and for each T_{DTU} satisfying them, we choose the optimal as the one with the minimal scenario average power consumption.

Table 3.5: Scenarios for the evaluation of Strategy Two. $C_{\max} = 235$.

Scenario	**1**	**2**	**3**	**4**
Power efficient strategy	Dyn Sect	Dyn Sect	Cap Adap	Cap Adap
***min-On* state parameters**	$N_{\min} = 1$	$N_{\min} = 2$	$\theta_2 = 0.06, \phi_2 = 0.79$	$\theta_4 = 0.13, \phi_4 = 0.5$
***all-On* state parameters**	$N_{\text{all}} = 3$	$N_{\text{all}} = 3$	$\theta_1 = 0, \phi_1 = 1$	$\theta_1 = 0, \phi_1 = 1$
C_{sta}	70	176	141	94

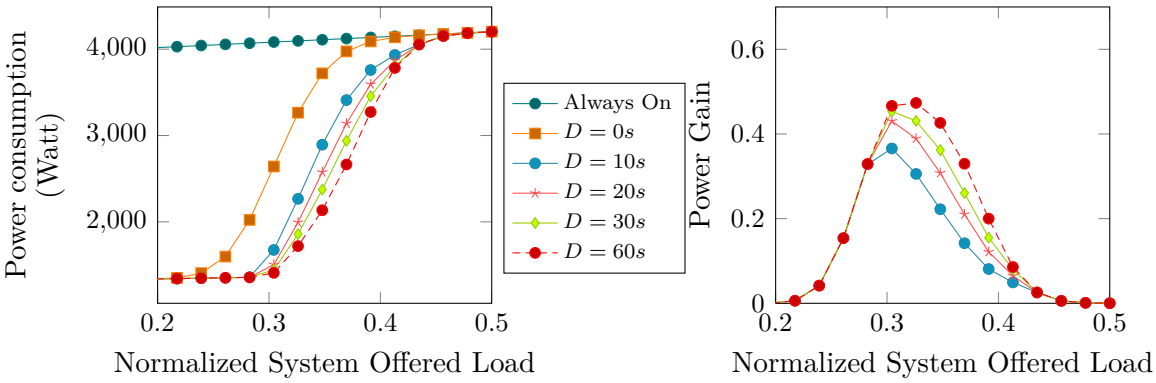
3.5.3.6 Results

In this section we present the optimal results obtained through numerical evaluation for the scenarios summed up in Table 3.5. We use Scenario 1 as reference to present the strategy performance and major findings. The same trends are observed for the other scenarios, scaling the power gain depending on the energy efficiency strategy applied in each scenario. Results for the other scenarios are given in Annex C.

Figure 3.10 presents the optimal results in terms of average power consumption for the Scenario 1, considering four different D proposed to the users. The strategy is compared to two baselines. The first baseline is the scenario in which the system persistently works using all the resources, and it is labelled as *Always On*. The second baseline is the scenario in which the dynamic resources are activated when the number of active sessions reaches C_{sta}, so that no user is delayed. The later baseline is labelled as $D = 0s$. Figure 3.10(a) presents the scenario average power consumption and Figure 3.10(b) presents the power gain regarding to the scenario without delay. Finally, Figure 3.10(c) presents the optimal strategy thresholds and the resulting system operation probability.

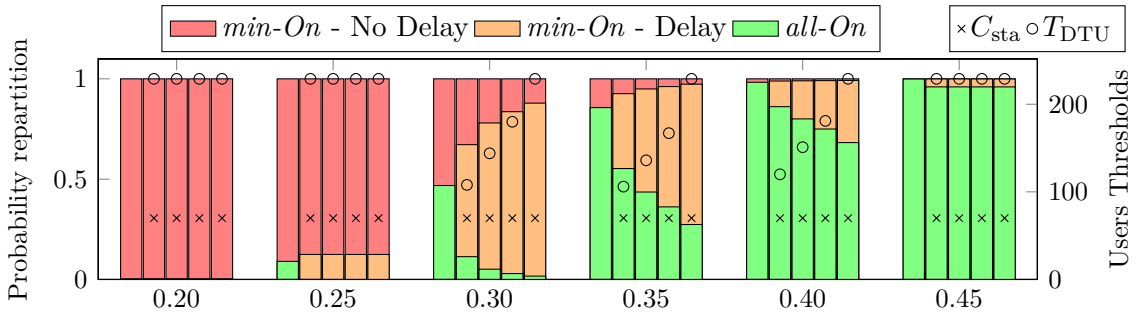
The gains regarding the Always On scenario strategy for loads below to C_{sta} are proportional to the power consumption of the dynamic resources when active. This is not surprising as they are not activated as they are not needed. Thus, even the strategy without any delay is efficient for these levels of load representing a power gain of up to 65% regarding the Always On baseline for the Scenario 1.

The benefits of delaying the start of the user services become evident in loads close to C_{sta} and above, e.g., when the system offered load represents 20% of the capacity, i.e. 0.25 in Figure 3.10. We observe up to 47% of power gain under increased delay tolerance for this scenario, compared to the strategy without delay. This is because the system has larger probability to remain in *min-On* state as the users can wait, even when affording higher offered load levels, e.g. larger than C_{sta}. This is shown in Figure 3.10(c) where the probability of operation is dominated by the *min-On* and delaying system state for normalized loads above to 0.25 and inferior to 0.35. For loads superior to 45% of the system capacity the gains are negligible, as the static resources are not enough to serve all arrivals, even if they can wait, so the dynamic resources are active all the time. Notice that when using the baseline $D = 0s$, the probability of activation of the dynamic resources starts to increase for normalized



(a) Average power consumption

(b) Power gain regarding the $D = 0s$ baseline



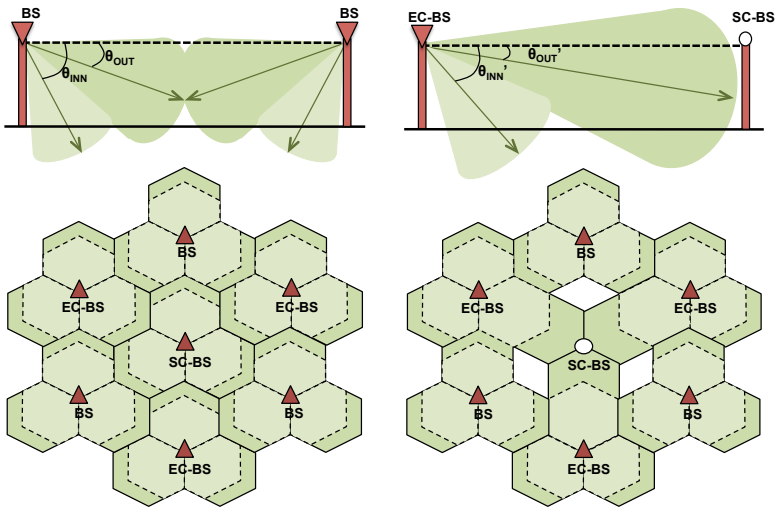
Normalized System Offered Load, each bar corresponds to (left to right): $D = 0s, 10s, 20s, 30s, 60s$

(c) System operation probability repartition and strategy thresholds depending on A and D

Figure 3.10: Results for the Scenario 1,$\gamma_{\max} = 0.05$, $\delta_{\max} = 0.05$.

load levels above to 0.21. In the case of the DTU-aware strategy with $D = 60s$, this is the case for loads above 0.30. Thus, 9 percentage points of further system load can be served without activating at all the dynamic resources, thanks to the proposed DTU-aware strategy.

For normalized load levels of 0.25 and 0.45 the optimal thresholds are the same. However, the probability distribution is different given the level of offered load. In the first case, all waiting users are served by the static resources, so that T_{DTU} is never reached (the probability that the system is in *all-On* state is 0). In the second case, the system is most of the time in *all-On* state to serve the increased offered load. What is more important to notice, is that T_{DTU} is the same for all values of D. This is because in the formulation of the optimization problem we search for the thresholds that minimize the power consumption under the constraint of a maximal delay. For these levels of offered load, there are no other values for the threshold that can ensure performance better that the one attained at $D = 10s$. Thus, users willing to wait more time do not represent further gains for the system.



(a) *all-On* state (b) *min-On* state

Figure 3.11: System configurations: (a) all the BSs are operational and using the same configuration (b) SC-BS is in SM and the EC-BSs extend their coverage modifying the antenna tilt.

3.6 COMPARATIVE EVALUATION

In this section we compare the performance of the two strategies presented in this chapter in the same scenario. In Section 3.6.1 we describe the deployment and energy efficiency strategy selected for the evaluation, as well as the traffic characterization and capacity estimations of the scenario. We evaluate the proposed strategies for different traffic levels following a daily pattern characteristic of an European country. Finally, we present the results in Section 3.6.2 showing the power and energy gains of the strategies and their efficiency depending on the different periods of the day.

3.6.1 Scenario

3.6.1.1 Deployment

The reference deployment is a homogeneous non-overlapping hexagonal deployment of sectorised BSs proposed by Guo et al. [GO13]. Each BS site is composed of six adaptive antennas, each one representing a sector. Two different kinds of sectors can be distinguished, as depicted in Figure 3.11(a): the *inner* sectors with antenna tilt θ_{INN} and the *outer* sectors with antenna tilt θ_{OUT}. Further details are summarized in Table 3.6.

Table 3.6: System parameters for the numerical comparision between strategies.

Deployment [GO13]	
Deployment type	Hexagonal Macro
Inter Site Distance $[m]$	500
Site Area [Km^2]	0.2165
Number of BS sites	7
BS type	6-sector Reconfigurable
Sectors operation frequency [MHz]	2600 inner, 800 outer
Bandwidth [MHz]	10
Antenna configuration	1x2 MIMO
Antenna tilt	normal $\theta_{\text{OUT}} = 3°$ $\theta_{\text{INN}} = 12°$ expanded $\theta_{\text{OUT}}\prime = 0°$ $\theta_{\text{INN}}\prime = 9°$
Baseline BS power consumption [AGD$^+$11]	
N_{TRX}	6
P_{max} [W]	20
P0 [W]	130
Δ_P	4.7
P_{sleep} [W]	75
Traffic characterization	
Session target throughput [Mbps]	1
Site capacity [Mbps] [GO13]	33 (*all-On* state) 10 (*min-On* state)
Site session capacity	$C_{\text{max}} = 33$ $C_{\text{sta}} = 10$
Session duration [s] [PP05] (μ^{-1})	81

3.6.1.2 Radio resource adaptation

We consider that the central BS of the deployment (Figure 3.11(a)) can be switched to *Sleep Mode (SM)*. This BS is denoted as Sleep-Capable Base Station (SC-BS) in the following. As we consider a non-overlapping scenario, the deactivation of the SC-BS will create a coverage hole in the area. Thus, the surrounding active BSs need to apply a coverage preservation technique to ensure the service availability and a minimal capacity in the affected area. The neighboring BSs in charge of compensating the absence of a sleeping BS are called Expand-Capable Base Station (EC-BS). In the considered adaptive deployment, the EC-BSs expand themselves changing the tilt of their inner and outer sector to $\theta_{\text{INN}}\prime$ and $\theta_{\text{OUT}}\prime$ respectively. Thus, the EC-BSs compensate the coverage of the SC-BS as depicted in Fig 3.11(b). Please note that with this configuration, the coverage compensation is done without the need of increasing the transmission power of the EC-BSs [GO13].

This dynamic system partially fits to the system model presented in Section 3.3: the system is in *min-On* state when the SC-BS is in SM. The C_{sta} static resources are provided by the expanded EC-BSs. However, as no redundancy or overlapping is considered, when the SC-BS is turned on the EC-BSs shrinks, removing their resources from the coverage area of the SC-BS. Some of the users with ongoing communications may need to perform handover to the SC-BS and continue being served by it, as it offers the better signal condition for them given their location. Strategy Two considers this, as no explicit access control is performed for the users, and no distinction of resources is made for their services. Thus, when the system

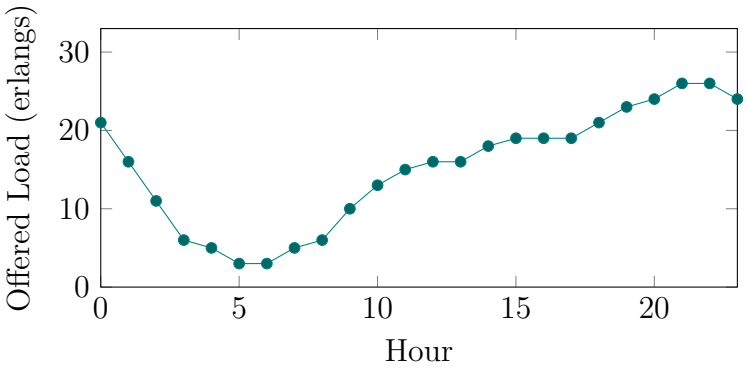


Figure 3.12: Base station site daily traffic profile for a European dense urban scenario [EAR12b].

turns to *all-On* state, waiting users and users with ongoing communication are considered in the same way. However, the way that Strategy One was modelled does not take this into consideration. Thus, additional assumptions are made in order to compare the strategies in this scenario. First, we only consider the case where $\eta = 1$. Thus, when using Strategy One, the EC-BSs will only serve the batch arrivals from the SC-BS and no other traffic will be held by them. We also assume that all these sessions will finish before the SC-BS is activated again. Finally, we consider that the system capacity when in *all-On* state corresponds to the capacity provided by the SC-BS.

3.6.1.3 Capacity estimation

The different system capacities are obtained from [GO13] and correspond to the data rate perceived in the worst-case user location for the reference system. When the system is in *all-On* state, the capacity $C_{\max}$ corresponds to the data rate perceived at the cell edge of the outer sectors of the SC-BS (Fig. 3.11(a)). When the system is in *min-On* state, the capacity C_{sta} is the data rate perceived at the cell edge of the expanded outer sectors of the EC-BSs (Fig. 3.11(b)). Thus, we focus our evaluation in the dynamic part of the deployment, i.e. the area covered by the SC-BS when active. Similar assumptions to the ones made in Section 3.5.3.3 are made about the user session characterization and the parameters are summarized in Table 3.6.

3.6.1.4 Traffic profile

The traffic profile used for the evaluation correspond to an European dense urban scenario with a population density of 3000 citizen/km^2, served uniformly by three

cellular operators. The percentage of data subscribers for this scenario is 75%. We used the typical European data traffic profile to determine the level of user activity throughout a complete day [EAR12b]. Considering uniform distribution of the users and the parameters in Table 3.6, the resulting BS site user density is 163 data subscribers/site/operator. Thus, for each BS site in our reference system, the traffic load varies according the profile depicted in Fig. 3.12. For example, the number of active users in the busy hour (21h) accounts 16% of the subscribers, i.e. 26 active users/site/operator.

3.6.1.5 Power consumption

We evaluated the average power consumption of the SC-BS when applying the two proposed DTU-aware strategies. For Strategy One, this corresponds to Equation (3.23). For Strategy Two, the power consumption is calculated using Equation (3.49), considering the power consumption when the system is in *min-On* state equal to $P_{\text{in}}(\text{sleep})$.

3.6.2 Results

The objective of the DTU-aware strategies is to maximize the probability that the SC-BS is in SM. Figure 3.13 and Figure 3.14 show the performance of the proposals when the delay tolerance varies. These figures present the optimal strategies thresholds and the resulting probability repartition of the SC-BS operation mode, considering different D proposed to the users. Figure 3.15 presents the optimal results in terms of average power consumption for a complete day. The strategy is compared to two baselines. The first baseline is the scenario in which the SC-BS is always active, labelled as *Always On*. The second baseline is the scenario in which no delay is experienced by the users and the SC-BS is activated when the number of active sessions served by EC-BSs reaches $0.8C_{\text{sta}}$, so that no user is blocked when the system is in *min-On* state.

As expected, Strategy Two has a better overall performance, allowing the SC-BS to remain in SM until offered load levels above C_{sta} (e.g., 12 erlangs, $D = 20, 30, 40s$). In the busy hours (offered load levels larger than 23 erlangs) the benefits of the delay tolerance of the users is negligible as the SC-BS should remain active all the time to absorb the traffic load. In Strategy One, the fact that the SC-BS must wake up to serve DTUs limits considerably its performance, compared to Strategy Two where DTUs can be served by EC-BSs. Note that all users in Strategy Two must be tolerant to delay, but they will effectively experiment delays only when the system is in a given state upon arrival. Thus, we see that opportunistically delaying users is better in terms of energy consumption than systematically delaying a part (or even all) of them.

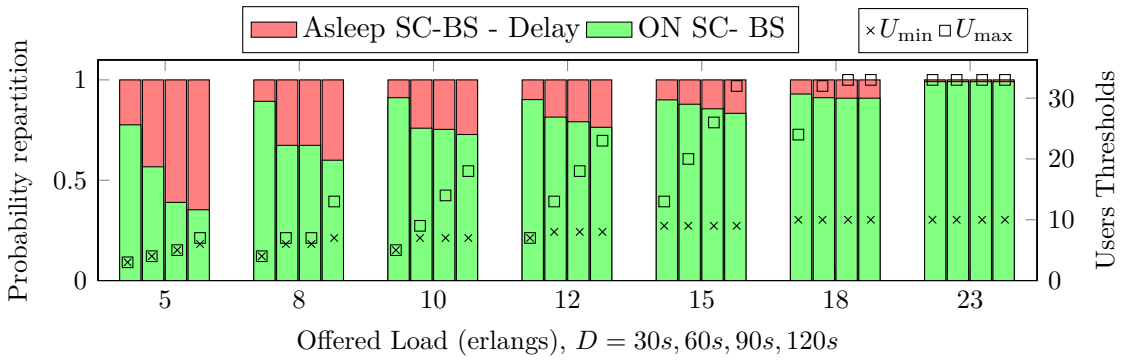


Figure 3.13: Sleep-capable base station operation probability and strategy thresholds depending on the traffic load. Strategy One in a dense urban scenario. Four fixed maximum tolerable delay (D), $\eta = 1$, $\gamma_{\max} = 0.05$, $\delta_{\max} = 0.05$.

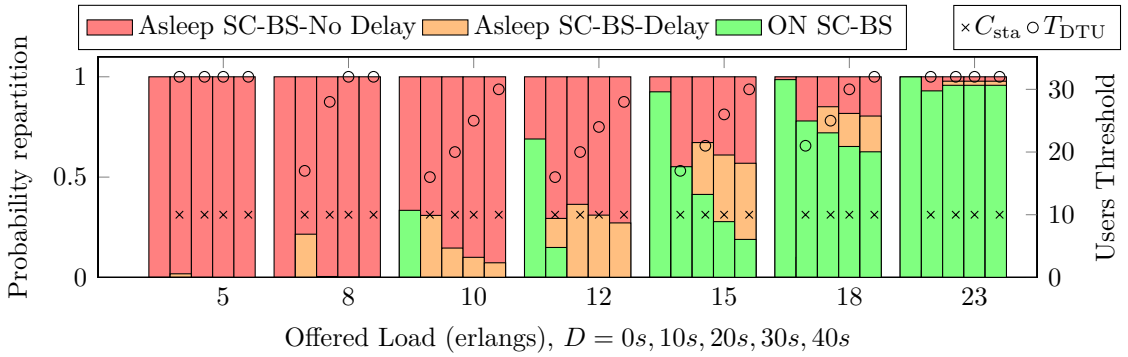
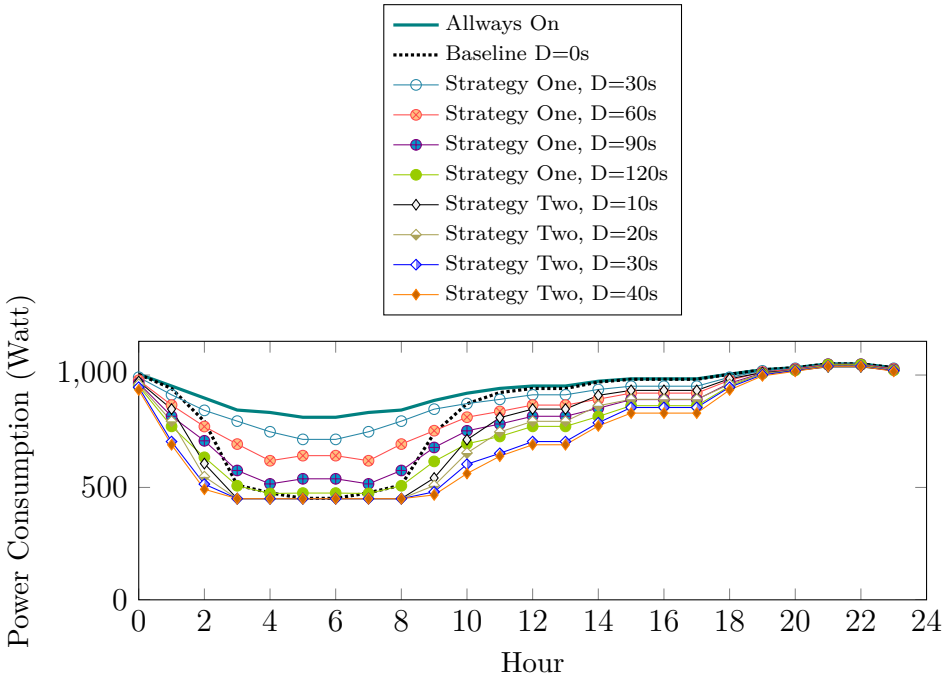
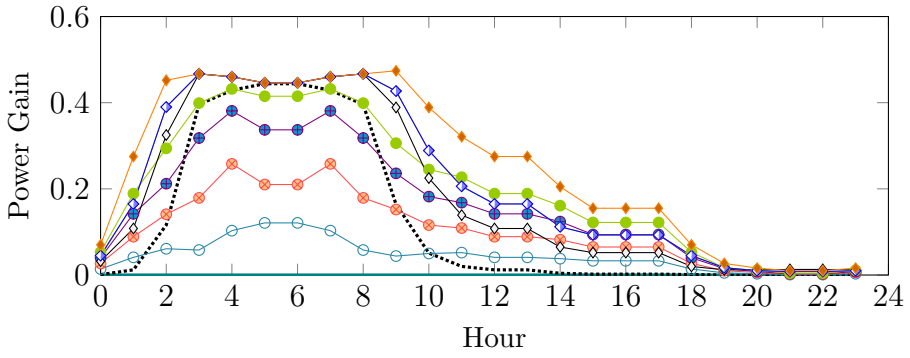


Figure 3.14: Sleep-capable base station operation probability and strategy threshold depending on the traffic load. Strategy Two in a dense urban scenario. Four fixed maximum tolerable delay (D) are considered altogether with the baseline strategy ($D = 0s$) and $\gamma_{\max} = 0.05$, $\delta_{\max} = 0.05$.

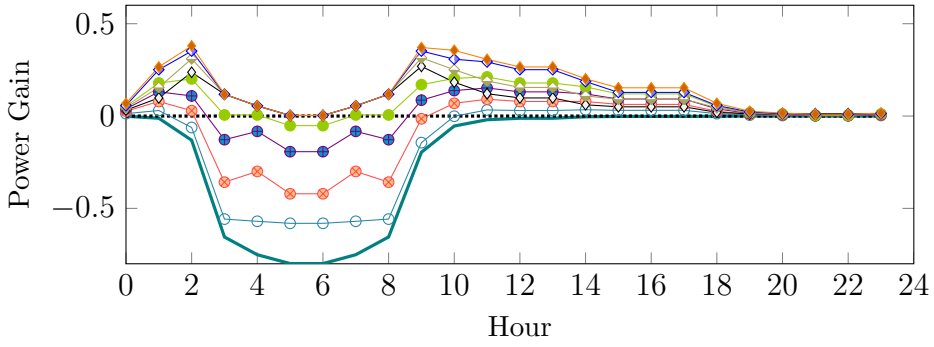
Figure 3.15 shows the results regarding the daily power consumption evaluation made for the SC-BS when the different DTU-aware strategies are applied. These results consider fast activation/deactivation and $P_{\text{sleep}} = 75W$. As expressed before in Section 3.4.3.7, it is likely in the future to have SM with lower power consumption which will further increase the gain presented in this section. Figure 3.15(b) shows up to 48% of power consumption reduction compared to the Always On strategy if the DTU-aware strategies are used. In low traffic periods, the baseline $D = 0s$ shows considerable gains, outperforming Strategy One. For the rest of the day, the two DTU-aware strategies have a better performance than the baseline $D = 0s$ as shown in Figure 3.15(c). However, Strategy One needs users to tolerate higher delays (D=90,120s) to provide a daily benefit for the operator, as exposed in Table 3.7. Finally, Strategy Two is the most advantageous as it better uses the system resources, the users are asked to wait only when needed, the waiting time is smaller and the



(a) Average power consumption



(b) Average power gain compared to the Always On strategy



(c) Average power gain compared to the Traditional SM strategy (D=0s).

Figure 3.15: Daily SC-BS power consumption evaluation for the different strategies, $\gamma_{\max} = 0.05$ and $\delta_{\max} = 0.05$.

Table 3.7: Average daily energy consumption $\overline{E}$ depending on the strategies. Energy reduction factor compared to the Always On strategy (ξ_{ON}) and compared to the Traditional Sleep Mode Strategy (ξ_{BL}).

Strategy	$\overline{E}$**[kWh]**	$\xi_{\mathbf{ON}}$	$\xi_{\mathbf{BL}}$
Always On	22.60	0.00	-0.12
Baseline D=0	20.13	0.11	0.00
One D=30s	21.65	0.04	-0.08
One D=60s	20.45	0.10	-0.02
One D=90s	19.23	0.15	0.04
One D=120s	18.34	0.19	0.09
Two D=10s	18.70	0.17	0.07
Two D=20s	18.13	0.20	0.10
Two D=30s	17.43	0.13	0.13
Two D=40s	17.17	0.21	0.15

strategy represents 15 % of more daily energy savings than the baseline $D = 0s$, with an energy reduction of 21% compared to the Always On strategy.

3

3.7 SUMMARY AND DISCUSSION

In this chapter, we presented an analytical formulation and evaluation of the potential power and energy gains that can be achieved with the combination of demand management techniques and energy efficiency strategies in cellular access networks. We first exposed the motivation of choosing request shifting as a demand management strategy, as it can increase the flexibility of the traditional energy efficiency strategies. As we analyse the problem from a theoretical point of view, we introduced the general system model and the dynamic resource management used in the formulation of our strategies. We also presented the traffic characterization we perform, using well known traffic distributions and assumptions.

Afterwards we described, analysed and evaluated the two strategies we proposed, which combine traffic shifting and energy efficiency techniques. These strategies are based on switching between system states depending on the carried traffic. Thresholds in the load are defined to trigger the reconfigurations and users are accepted in the system even when they are not served right away. We defined two type of resources: static and dynamic. The first are available when needed to ensure the service availability and a minimal access network capacity. The second can be activated/deactivated when required, with certain degree of freedom depending on the used hardware and energy efficiency strategy.

Our first proposed strategy has the objective of maximizing the utilization of the dynamic resources. To do so, users are asked to wait when the dynamic resources are not active, time in which impatient requests are served by the static resources. When the dynamic resources are active all users are served indistinctly. However, this strategy requires an explicit differentiation about the user tolerance to delays, and we observed that the benefits of this strategy varies with the proportion of

traffic that can be delayed. The decisions of activating the dynamic resources are dependent on the waiting users, as the waiting time should be bounded. Decisions to deactivate them depend on the availability of the static resources to serve the ongoing communications, in order to avoid user dissatisfaction. We observed that the higher the waiting time constraint, the higher the energy gain. However, as the access network has limited capacity, this maximal delay has an upper limit as the service needs to be ensured after the waiting time. The drawback of this strategy is the tight link between DTUs and the dynamic resources. The latter are always activated if there are DTUs in the system, no matter if there are enough static resources to serve them. This makes the strategy inefficient for low loads, i.e. loads below to the capacity of the static resources, situation in which considerable gains can be achieved by simply deactivating the dynamic resources and serving all users without delay. However, the benefits of the strategy are evident when the load is beyond the static resources capacity, as it allows to distribute and balance the load between both types of resources. Thus, the dynamic resources can be deactivated and the activation can be shifted in time due to the cooperation of the users, maximizing the low energy consumption periods.

This strategy is also suitable in the case where there are no static resources. For example, in an off-grid battery limited system, the BSs should minimize the time it is active, or conversely maximize the periods when it is deactivated. This is achieved by Strategy One which offsets the start of the services until having enough requests to justify the activation. The ideal deactivation condition should be when no user is active in the system (as there are no static resources to take care of ongoing communications). However, in this kind of critical systems, the targeted level of quality of service is more flexible and some communication drops can be tolerable or even necessary/unavoidable by the operator. Thus, the selection of the strategy parameters may depend not only on traffic conditions but also on other system characteristics (e.g., battery level).

The second strategy we developed makes no distinction between the different type of users, as the waiting condition only depends on the system state and not on a persistent user choice. Thus, all users are considered DTUs, but they will wait only in a given condition, i.e. when the static resources are exhausted. All users are served immediately if the load is below the static resource capacity and the activation of the dynamic resources is done based on load thresholds. Using traditional techniques the threshold values are below the static resources capacity, to avoid any user dissatisfaction caused by the congestion of the limited available resources. The basis of the strategy we proposed is to push the activation threshold beyond this capacity, thanks to the delay tolerance of the users. Thus, we can extend the periods when the system can remain using only the static resources, delaying the activation of the dynamic resources or even avoiding it, e.g. in the case of a short period of increased traffic. As well as in the previous strategy, the condition to activate the resources depends on the number of waiting users in the system.

We observed that the waiting times are small, as users can be served by the static resources and not solely by the activation of the dynamic resources. The drawback of this strategy is that urgent traffic is not explicitly considered. In practical systems, this type of traffic can be prioritized on a scheduling level, as the static resources are operational. However, this may have some impact on the waiting users if any. This strategy overcomes the limitations of the persistent DTU strategy and the energy gains are maximized in periods of low load. Moreover, the range of loads where the strategy represents benefits is extended compared to the traditional strategies.

In the last part of this chapter we evaluated the performance of the strategies for different levels of load representing the daily traffic variation in a typical European cellular network. We observed that in almost 80% of the day the strategies can bring some benefits compared to the Always On paradigm in the evaluated scenario. This means that for 80% of the day the considered access network is over-provisioned for the evaluated traffic profile, which highlight the importance of the dynamic radio resource adaptation to make the access network energy consumption more proportional to the carried traffic. The daily energy consumption is reduced by 11% if a traditional strategy is used. Reductions up to 19% are observed if a persistent DTU strategy is employed with a maximal delay of 2 minutes, and up to 21% if an opportunistic DTU strategy is used, with a maximal delay of 40 seconds.

In Chapter 2 we show that several strategies are proposed in the literature towards the idea of resource adaptation for energy efficiency. In this chapter we presented a generic and simple model of the traffic dynamics of the access networks when using these techniques. The traffic assumptions we made allow to easily adapt the model to different energy efficiency strategies, based in the notion of capacity variation depending on the available radio resources. We considered throughout this chapter four different scenarios: BS switching in a homogeneous overlapped deployment, BS capacity adaptation and BS dynamic sectorization in a standalone BS scope, and coordinated BS switching in a homogeneous non-overlapped deployment. Thus, we presented different use cases in which the proposed strategies can be applied when having the appropriated hardware flexibility. Each set of parameters for each energy efficiency technique, represents a different system dynamics and consequently a different energy consumption modelling and energy gain quantification regarding the traditional approaches. We showed with the evaluation of the different scenarios, that independently of the used energy efficiency technique, there are further gains to be made with the user side cooperation in the form of initial delay tolerance.

The results presented in this chapter represent the theoretical bounds for the possible gain in the evaluated scenarios. We can expect that these gains vary when considering more realistic system characteristics, e.g. location dependent radio conditions, heterogeneous traffic and distributions, mobility, etc. Moreover, the strategies represent changes in the radio environment for a given area and users in it. In the next chapter we present the implementation of the strategies in a system level simulator, allowing us to evaluate them in a more realistic radio environment.

4 System level evaluation

4.1 Introduction

In this chapter we describe the implementation and evaluation of the DTU-aware strategies in a system level simulator, which allowed us to consider a more realistic scenario for the LTE access network functioning. In Section 4.2 we elaborate in the motivations of using this approach for the sake of having more realistic approximations of the gains attainable by the DTU-aware strategies. In particular we are interested in combining the DTU-aware mechanisms with a fast-reaction cell switching algorithm in order to further increase its energy reductions.

We developed our simulation platform based on the LTE module of the Network Simulator 3 (ns-3). In Section 4.3 we present the main reasons driving this decision and we describe some relevant points of the model used for our evaluation. We present the general architecture of the simulation models implemented in the ns-3 platform, which emulates a functional and operative LTE network. The LTE compliant data protocol stack implemented in ns-3 is discussed as well. We rely on the handover algorithms designed for supporting user mobility to ensure the continuity of services when applying cell switching algorithms. Thus, we describe the corresponding handover protocol implemented in the LTE module of ns-3.

In Section 4.4 we present the scenario and algorithms we implemented for the evaluation of the DTU-aware strategies in the simulation environment. First, we describe the considered system model and coordinated cell switching algorithm. We present as well the spatial characterizations we use for the execution of the DTU-aware strategies, e.g. the definition of a DTU zone in the deployment, which limits the execution of the DTU-aware algorithms to the given area in which the coordinated cell switching is applied. Afterwards, we remind the operation of the DTU-aware strategies and we describe the corresponding algorithms implemented in the simulator. We also describe how the traffic is generated considering a pseudo voice application for the simulated UEs, and a tunable call arrival and service time generator following a given random variable distributions. We present as well the algorithms corresponding to the traffic monitoring functionality, which allow to trigger the reconfigurations depending on the applied DTU-aware strategy. Finally, we define the metrics we use for the evaluation of the performance of the strategies,

comprising the QoS perceived by the UEs, the experienced waiting time resulting from the strategies application and the average power consumed by the access network.

In Section 4.5 we present the system level evaluation examining the performance of the DTU-aware strategies. We present the simulated scenario and the scenario-specific system characterization in terms of capacity and coordinated cell switching. Finally, we present the results of the simulations showing the strategies dynamics, the effectiveness in maintaining acceptable levels of QoS and waiting times, and in reducing the average power consumption of the system in study. We show as well the impact of the cell switching reconfiguration time in the attainable power reductions. Finally in Section 4.6 we summarize and discuss the main findings of the work presented in this chapter.

4.2 Motivation

In the previous chapter we presented two DTU-aware strategies to further reduce the energy consumption of the access network when energy efficiency techniques are employed. We approached the problem from a theoretical point of view, where we made several assumptions and simplifications about the cellular access network functioning, e.g. instantaneous reconfiguration and fixed radio resources per service. Thus, we decided to move towards a more realistic scenario, using a system level simulator modelling a functional LTE access network. Two main motivations drove this decision. First, we wanted to investigate the impact the dynamic resource management has on UEs' QoS and on the general access network behavior, and how it affects the performance of the DTU-aware strategies. Second, we wanted to evaluate the behavior and effectiveness of the DTU-aware strategies in online conditions, taking into consideration the system and traffic time dependent variations. We also want to investigate the time needed to perform the network reconfigurations without affecting considerably the user's perceived QoS and waiting time. These points, among other secondary motivations, are discussed in the following paragraphs.

The radio link quality of the UEs under the coverage of a given BS depends, among other parameters, on their relative position to the serving BS and to the neighboring BSs interfering with the serving signal. LTE BSs have different data transmission modes in order to achieve reliable communication in a variety of radio link conditions, taking advantage of the OFDM multi-user diversity, which allows the allocation of resources in a per-UE basis. To do so, a closed control loop is established between the BS and the UE. The UE reports a KPI indicating the quality of the channel to the BS, which chooses the appropriate modulation and coding scheme for transmission. When considering energy efficiency strategies, the radio environment conditions become highly dynamic. Thus, the channel conditions perceived by the UEs can vary during the time they are within a cell. These changes

are well studied and considered in the design of the access network protocols, as traditionally they were the result of the mobility of the users. However, when using radio resource adaptation, the channel conditions change even if the UE is static. Using system level simulations we can investigate how to perform the radio environment reconfigurations having the least impact on users QoS with ongoing communications, as well as how the current access network protocols reacts to it, e.g. the handover protocols designed for mobility.

We decided to focus on the cell switching technique as it is the energy efficiency strategy that provides the higher gains, as shown in Chapter 3. When a cell is deactivated, users need to perform handover to the remaining active cells in order to continue their communications, or to simply have network availability in case no service is ongoing. This process implies control signalling exchange between the three involved entities: the UE, the source cell, i.e. the cell to which the UE is initially attached, and the target cell, i.e. the cell to which the UE needs to attach. If the deactivation of the source cell is done suddenly, the control information can be lost, compromising the handover protocol. Thus, this process needs to be done progressively, which is time consuming. Our approach is based on the idea of extending the periods of low energy consumption thanks to the delay tolerance of the users, i.e. when the cells are not active. However, a maximal waiting time is proposed to the users, and extra delay in the start of their services can be experienced when considering the reconfiguration periods. Moreover, the low energy consumption periods start when the deactivation process finishes. Thus, it is possible that the waiting periods and deactivation periods overlap, reducing the length of the low consumption periods if the reconfiguration takes too long.

Simulator choice:

An online evaluation of the DTU-aware strategies can provide some insights of the points discussed above, and the trade-off to be made in order to maximize the energy gain resulting of the application of the strategies. A discrete event simulator appears convenient for our purpose, as it allows to evaluate the behavior of the system in time. Several LTE simulators were developed in the academia and are available for free download and use. For example, the LTE Vienna simulator [IV] developed by the Vienna University of Technology, or the LTE-Sim platform [Tel] developed by the Telematics Lab of the Politecnico di Bari. The models conceived for these simulators are specific to the evaluation of LTE protocols. General purpose network simulators incorporate LTE modules in their implementations as well. For example, the LTE specialized module of the OPNET simulator (commercial), SimuLTE of the OMNEST/OMNeT++ simulation platform (open source) [Omn], or the LTE Module of the Network Simulator 3 (ns-3) (open source) [NS3]. It is worth to mention that the last one utilizes estimations for the physical layer model which are derived from the LTE Vienna simulator and it has the active development collaboration from the LTE-Sim team. Thus, we decided to use the LTE module of ns-3 for our evaluation, which will be presented in the following section.

4.3 NS-3 LTE SIMULATION PLATFORM

The ns-3 platform is an open-source discrete-event network simulator and is publicly available for research, development, and use. The platform is split over numerous modules containing one or more models for real-world network devices and protocols. The LTE module was initially developed by the Centre Tecnologic de Telecomunicacions de Catalunya under the name of LENA [Cen] and was further integrated to the public release of ns-3. We use the ns-allinone-3.21 version [NS3], which at the time of writing is the latest release. The main reasons for using this simulator are the following:

- It is a discrete-event simulator allowing to investigate the behavior of the network entities, traffic and protocols given the logical sequence of events in time;
- It allows the simulation of access networks with multiple BSs and several UEs connected to each BS;
- It implements core network interfaces, protocols and algorithms providing realistic data transmission between the UEs and remote hosts located in external networks;
- It allows the implementation of different applications in the end nodes, i.e. in the UEs and remote hosts; which allows the simulation of different traffic dynamics within the LTE network;
- It comprises an adequate model approximation of the radio channel conditions in a cellular environment;
- It implements radio resource scheduling with a granularity of one resource block, allowing a better estimation of the instantaneous transmitted power;
- It allows to manipulate the radio environment in simulation time. In particular, the cell transmission power can be changed allowing the simulation of cell switching techniques;
- It implements automatic inter-cell handover algorithms ensuring the continuity of the communication when the UE detects a more suitable cell to connect, e.g. during a cell switching procedure.

In the following sections we show an overview of the general characteristics of the ns-3 LTE module simulation model. In Section 4.3.1 we present the entities modelled in the simulation platform which allow to approximate the functioning of a LTE network. For our evaluations, we consider the UEs establish communication sessions with some host in a external network. In Section 4.3.2 we present how is the flow of data traffic within the simulated entities. Finally, we rely in the standard handover procedures for ensuring the continuity of the UE sessions when a cell switching is applied. Thus, the handover algorithms implemented in the simulator are presented in Section 4.3.3.

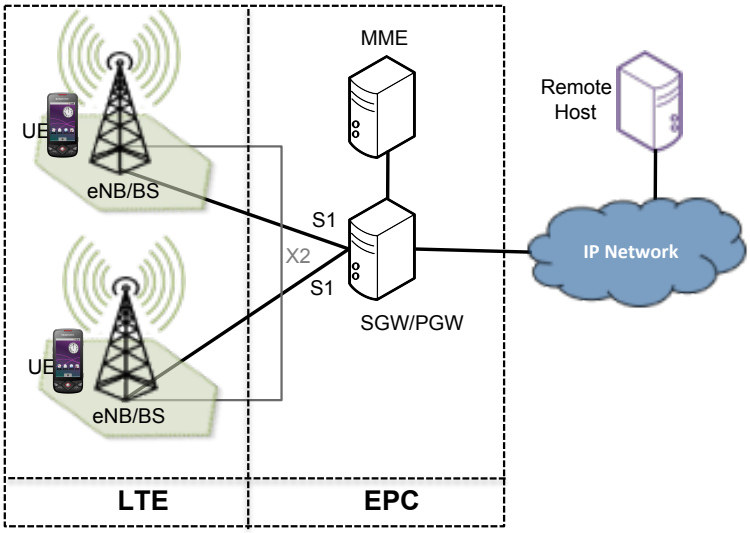


Figure 4.1: LTE-EPC simulation model.

4.3.1 General architecture

The general architecture of the model implemented in the ns-3 LTE module is depicted in Fig. 4.1 and is divided into two major parts. The LTE model comprises the radio protocol stack implemented in the BS nodes and in the UE nodes. The Evolved Packet Core (EPC) model includes the core network interfaces, protocols and entities, which are implemented in the Serving Gateway (SGW), the Packet Data Network Gateway (PGW) and Mobility Management Entity (MME) nodes, and in the BS nodes. The SGW and PGW functional entities are simplified and implemented within a single node, which is unique for each simulation, i.e. no inter-SGW mobility is supported by the simulator. The MME is a logical node inside the SGW/PGW. The BSs are interconnected with each other by means of a point-to-point link and communicate between them using the X2 interface. Each BS is connected to the SGW/PGW by means of a point-to-point link and communicate using the S1 interface. The simulator allows to simulate topologies comprising up to tens of BSs and hundreds of UEs. It is highly configurable allowing to choose the desired parameters for the protocols and algorithms compatible with the standards, for both the LTE and the EPC models. The main objective of the EPC model is the simulation of end-to-end IPv4 connectivity over the LTE model. This is, between the UEs and remote hosts situated in external IP networks. However, it also supports the UE inter-cell mobility thanks to the X2 interface which allows the execution of automatic handover algorithms.

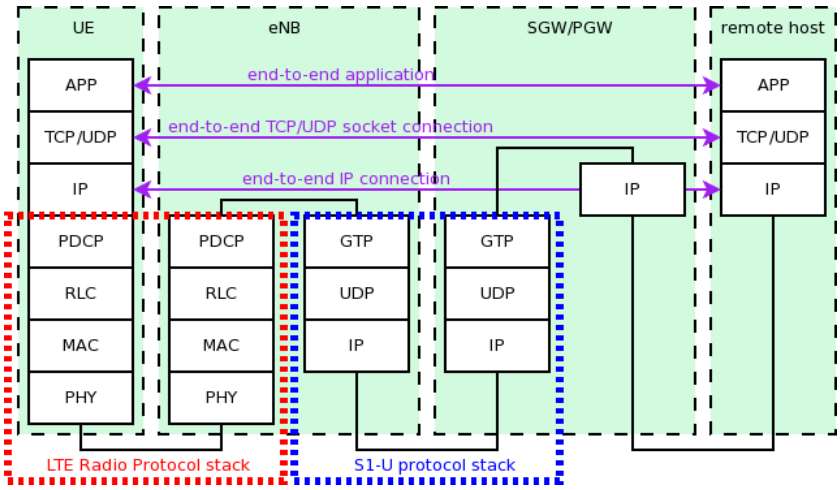


Figure 4.2: LTE-EPC data plane protocol stack [NS3].

4.3.2 Data plane protocol stack

The data plane protocol stack implemented in the simulator is depicted in Fig. 4.2 and is briefly overviewed in this section. Any ns-3 application can be installed on an end node, as long as it uses UDP or TCP over IPv4. The LTE core network is an IP network in which each BS is connected to the SGW/PGW using a point-to-point link. The S1-U protocol stack is in charge of tunnelling the IP packets between the external IP network and the corresponding BS to which the communicating UE is attached. The IP packets from/to the end-nodes are encapsulated in GPRS Tunnelling Protocol (GTP) packets containing the information of the BS or PGW of interest, which are sent over the local network using an UDP socket.

The radio protocol stack is in charge of sending/receiving the end-user IP packets over the radio interface. In first instance, the IP packets are encapsulated in Packet Data Convergence Protocol (PDCP) packets. The simulator implements a simple PDCP protocol which only transfers the data to the other layers and maintain the protocol sequence numbers. Thus, the data is transferred to the RLC layer without additional processing featured for the PDCP layer in the standards (e.g. header compression, ciphering, etc.). The Radio Link Control (RLC) layer receives the data and ask for transmission opportunities to the inferior layers. When the Medium Access Control (MAC) layer permits transmission, the RLC layer segments or concatenates the data coming from the PDCP layer forming RLC packets which are sent to the MAC layer.

The MAC layer determines how and when the data will be transmitted. It is in charge of multiplexing the data coming from the RLC layer into Transport Block (TB). The size of the TBs, i.e. the amount of data that can be aggregated in each of them, depends on the Modulation and Coding Scheme (MCS) to be used for the

transmission. The smallest unit of resource allocation in the Physical layer (PHY) is the Resource Block (RB) and the channel conditions on each of them is different depending on the UE location and interfering signals. The MCS and the TB size for each RB are mapped from the Channel Quality Indicator (CQI) reported by the UE. The CQI is itself obtained by mapping the spectral efficiency calculated by the UE using the perceived Signal-to-Interference-plus-Noise Ratio (SINR) and a modified *Shannon* formula accounting the difference between the theoretical bound and the performance of real MCS [PBM11]. The CQI is reported periodically by the UE and provides an overall estimation of the perceived radio conditions.

4.3.3 Handover procedure

The simulator implements handover procedures compliant with the LTE standard. The UEs report periodically information about the cells they can detect given their position. The KPIs measured by the UEs are the Reference Signal Received Quality (RSRQ) and the Reference Signal Received Power (RSRP) for each detected cell. The RSRP is the linear average received power of the signals carrying cell-specific reference signals over the entire bandwidth. Thus, it measures the strength of the cell-specific signal. The RSRQ is the ratio between the RSRP and the total received power including all cells interference and noise. Thus, this is a measure of cell-specific signal quality. The handover procedure in LTE is BS driven, thus, the involved BSs decide when to trigger the handover and direct its execution. There are two event-based handover triggers implemented in the simulator.

A2-A4-RSRQ handover algorithm:

This algorithm defines an acceptable signal strength threshold. The UE will stay in the current serving cell if the RSRQ is above this threshold. Thus, two events should occur to trigger the handover:

- Event A2: The RSRQ of the serving cell becomes worse than the threshold, i.e. the UE is experiencing poor signal quality from the serving cell.
- Event A4: The RSRQ of a neighbor cell becomes better than a threshold, i.e. a suitable neighboring cell is detected.

Strongest cell handover algorithm:

This algorithm triggers the handover as soon as a better cell is detected. Regarding the specification, the event that should occur to trigger the handover is:

- Event A3: the RSRP of a neighbor cell becomes better than the RSRP of the serving cell.

To avoid frequent unnecessary handover, the algorithm defines two parameters: the *hysteresis*, which represents how much the neighboring RSRP should be better than the serving cell RSRP in order to trigger the handover; and the *time-to-trigger*, which is the amount of time that must elapse after the hysteresis threshold is reached in

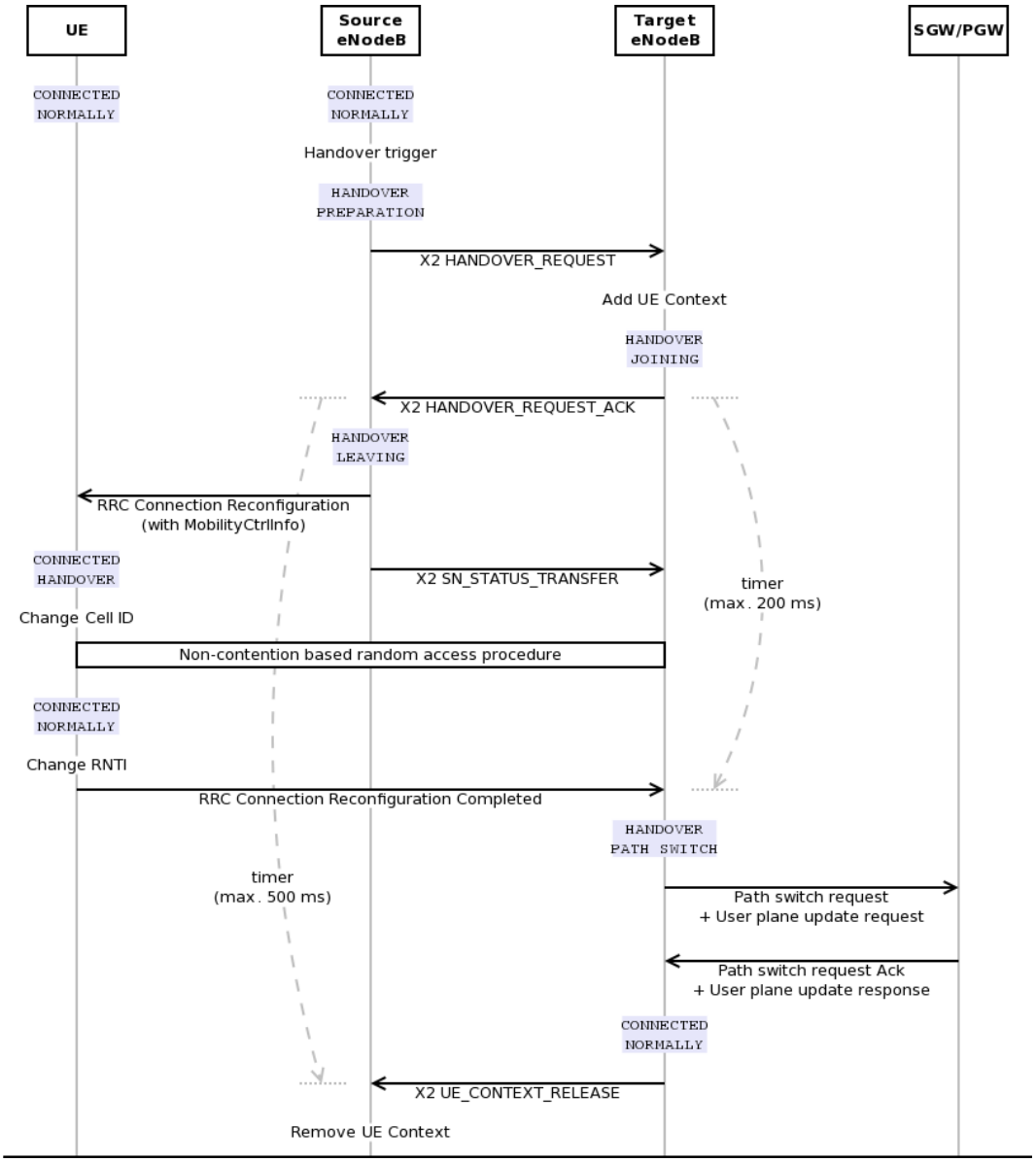
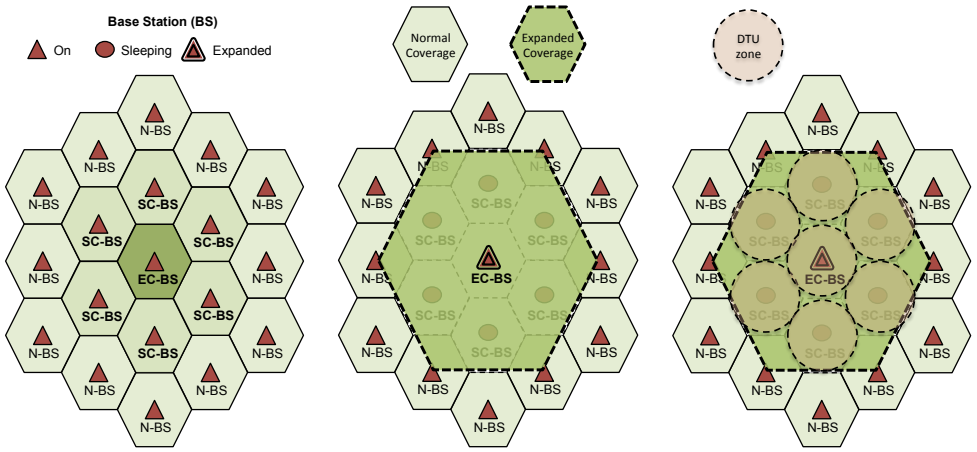


Figure 4.3: Sequence diagram of the handover procedure implemented in NS-3. Source: [NS3].



(a) *all-On* state (b) *min-On* state (c) *min-On* state, localized DTU zone

Figure 4.4: System configurations and identification of the DTU zone.

order to finally trigger the handover if the triggering condition was maintained. After the handover is triggered, the procedure presented in the Figure 4.3 is executed. If some of the timers expire, the handover is considered failed.

4.4 DTU-AWARE STRATEGIES IN NS-3

In this section we present the setup of the simulation environment that allow us to evaluate the performance of the DTU-aware strategies using a system level approach. In Section 4.4.1 we describe the access network deployment configuration we consider for the simulations, as well as the BSs role and behavior in the coordinated cell switching algorithm we implement. Section 4.4.2 present the details about the systematic design of the DTU-aware strategies in combination with the coordinated cell switching. Section 4.4.3 provides some complementary information about the traffic generated by the UEs in the simulations, as well as how the BSs monitor this traffic in order to take the switching decisions. Finally, Section 4.4.4 describe the metrics we use to evaluate the performance of the DTU-aware strategies in the simulation environment.

4.4.1 System model

We consider a group of neighboring BSs of a flat hexagonal access network, depicted in Figure 4.4(a). Following the model imposed by the simulator and depicted in Figure 4.1, all the BSs are connected to an unique SGW/PGW and interconnected between them using point-to-point links. Each BS is able to dynamically set its downlink transmission power. We consider a cell switching algorithm and, as in

the previous chapter, three types of BSs are identified. A BS designated to enter in sleep mode is denoted as *Sleep-Capable Base Station (SC-BS)*. A BS designated to compensate the coverage of the SC-BSs when sleeping, is called *Expand-Capable Base Station (EC-BS)*. Finally, a BS which does not participate in the dynamic algorithm is denoted as *Normal Base Station* (N-BS). The execution of the cell switching algorithm is controlled by one of the BSs, which has complete knowledge of the load of the participating BSs. The two system states are differentiated as well. The system is in *all-On* state when all the BSs are operational. In this state the system has an available maximum capacity denoted by $C_{\max}$. The system is in the *min-On* state when the SC-BSs are sleeping. In this state the system has a capacity $C_{\min} < C_{\max}$, provided by the coverage compensation made by the EC-BS (Figure 4.4(b)). In a multi-cell environment, each cell is identified by a parameter called *Cell ID*. In this chapter we focus the evaluation on omnidirectional micro BSs containing only one cell. Thus, we will refer to BSs and cells indistinctly. In the group of BSs of interest, the set of SC-BSs, EC-BSs and N-BSs are identified and Θ_{SC}, Θ_{EC} and Θ_N, respectively.

DTU Zone: The area covered by the BSs in Θ_{SC} and Θ_{EC} when the system is in *all-On* state is defined as the *DTU zone*. When the system is in *min-On* state, i.e. the SC-BSs are in sleep mode, users located in this area could be asked to collaborate with the network, delaying the start of their services. For simplicity, we approximate the coverage area of the corresponding BSs using a circular model, with center in the BS position and with a radius $r = \frac{\text{ISD}}{2}$, were ISD is the Inter Site Distance in the hexagonal deployment. An example is depicted in Figure 4.4(c). The system is designed to support a maximal delay (D) for users in the DTU zone. In case some delay of the users service is required, the network informs them that the waiting time will not be longer than D, and the cell switching algorithm is configured to satisfy this constraint. The exact mechanisms and protocols that can be used to perform this interaction between the users and the network are not discussed in this thesis. Thus, we adopted an over-the-top approach in which the state of the network and the belonging to the DTU zone are verified each time a UE intends to start a service. Afterwards, the control decision to start or delay the service is delegated to the network.

Coordinated cell switching:

We consider a coordinated BS switching algorithm in which all involved BSs start the reconfiguration at the same time. Moreover we consider a 6/7 configuration as presented by Ajmone-Marsan et al.[ACCM09], in which the central EC-BS covers the area of the 6 neighboring SC-BSs when they are switched into sleep mode, as depicted in Figure 4.4(b). In this BS configuration we assume that the EC-BS is the controller of the DTU-aware cell switching algorithm, as it has complete knowledge of the DTU zone when in *min-On* state, and it is in direct neighbor relation with all SC-BSs.

The transitions between system states (all-On → min-On and min-On → all-On) are performed progressively. When the cell-switching controller decides to switch to *min-On* state, each SC-BS decreases its transmission power by X dBm every t_x seconds. Simultaneously, the EC-BS increase their transmission power by Y dBm every t_y seconds until reaching the maximal transmission power. Once each SC-BS has no more users associated, it turns into sleep mode.

In each transmission power change, a batch of users may need to handover to the EC-BS. Too long power steps (i.e. X and Y) or too short time steps (i.e. t_x and t_y) could produce handover failures. The handover protocol could be compromised, in the first case due to bad signal condition of the source BS (i.e. too late handover) and in the second case due to signalling overhead of a big number of user performing handover. Conversely, too long time steps or too short power steps will produce unnecessary long reconfiguration periods. Thus, a trade off between power steps and time step is needed to minimize the reconfiguration periods and to let the users successfully perform handover if needed.

When the objective is to switch to *all-On* state, the opposite procedure is performed. The SC-BSs start the procedure increasing the power to the initial level that we assume equal to the level in which no more users were associated to the BS in the previous reconfiguration process. Afterwards, each SC-BS increases the transmission power by Z dBm every t_z seconds while the EC-BS shrinks concurrently, decreasing the transmission power by W dBm each t_w seconds.

We consider that the time between the moment the SC-BS takes the decision of turning into sleep mode and the actual moment when it is deactivated, is negligible. The same assumption applies for the opposite procedure, i.e. activating the SC-BS. In general, this time may depend on the type of BS and its hardware architecture. In the case that that the activation/deactivation time is non-negligible, it should be taken into consideration in the design of the DTU-aware strategies, e.g. proposing to the users larger delays than the used for the threshold derivation.

4.4.2 DTU-aware strategies implementation

In this section, we remind the operation of the strategies in Section 4.4.2.1 and we present the implemented algorithm for each of them. Finally, in Section 4.4.2.2 we discuss the use of the theoretical optimal strategy thresholds in the system level evaluations, and how the results may differ from the theoretical bounds.

4.4.2.1 DTU-aware strategies algorithms

The cell switching strategies adapt the resources depending on the load level (L). Considering general cell switching strategies, when the system is in *all-On* state and

Table 4.1: Threshold equivalence between the generic terms used for the evaluation using the simulator and the specific thresholds of each strategy defined for the theoretical model.

Simulations	**Theoretical**	
General	**Strategy One**	**Strategy Two**
L_1	$U_{\min}$	$C_{\min}$
L_{DTU}	$U_{\max}$	T_{DTU}

the load decreases, closer to a given threshold L_1, the reconfiguration to *min-On* state is triggered. L_1 should be chosen appropriately so that the current system load can be absorbed by the EC-BS along with the new (estimated) arrivals. When the system is in *min-On* state and the load increases, surpassing the switching on threshold L_2, the reconfiguration to *all-On* state is triggered. The choice of L_2 is usually done assuming that a new arrival will be blocked if the resources in *min-On* state are exhausted. Thus, $L_2 < C_{\min}$ in order to trigger the reconfiguration before a blocking situation could arrive [HMJ11] [GO13].

When a DTU-aware strategy is employed, we consider that a part of the users are willing to cooperate with the network and they are able to delay the start of their services. Following the strategies proposed in the previous chapter, when the system is in *min-On* state, some users will be served instantaneously and some others will be put on hold, so none of them will be blocked, as long the system capacity ($C_{\max}$) is not reached. Thus, when using a DTU-aware strategy, we introduce a different wake up reconfiguration threshold denoted as L_{DTU}. When the number of active users in the system (including waiting users) surpasses this threshold, the reconfiguration is triggered to *all-On* state. As previously demonstrated, it is critical for L_{DTU} to be appropriately selected so that the waiting time of the users is bounded and inferior to D.

We consider the two DTU-aware cell switching strategies proposed in the previous chapter. The first strategy always delays DTU services if they are in the DTU zone when the system is in *min-On* state. The second strategy only delays DTU requests when there are not enough resources to serve them (i.e. $L > C_{\min}$), serving them otherwise. The threshold equivalence with the theoretical model is given in Table 4.1 for both strategies.

Each strategy adapts the threshold configuration depending on the system conditions and D. Afterwards, the dynamic cell switching algorithm is put into action, tracking the load variations (user session arrival or departure) and reacting accordingly. The controller EC-BS identifies three kind of events, namely, *Load notification*, *Arrival*, *Departure*, and may react differently to them depending on the used strategy.

In both strategies, when the system is in *all-On* state, the only triggering event is the *Load Notification*. The controller EC-BS collects the load information from

the SC-BSs, calculates the aggregated load (L) and performs the switching to *min-On* state when it is appropriated ($L < L_1$). When the system is in *min-On* state the strategies react differently. The EC-BS controller has complete and real time information about the load in the DTU zone as it is under its coverage. Thus, the EC-BS reacts to each *Arrival* or *Departure* event. Strategy One follows Algorithm 1. In this algorithm, all DTU arrivals located in the DTU zone are delayed if the system is in *min-On* state. Once the number of waiting DTUs (n_{DTU}) reaches the switching threshold (L_{DTU}), the system switches to *all-On* state and starts serving all waiting users among new arrivals.

Strategy Two follows Algorithm 2. In this strategy, DTU arrivals are delayed only if the DTU zone is congested. This happens when the number of active users in the DTU zone (N) surpass the capacity of the EC-BS (C_{min}). Departure events allow to serve waiting users in a FIFO manner. When N surpasses the switching threshold the system reconfiguration is triggered. The system state changes when the reconfiguration is finished. Thus, arrivals during reconfiguration periods are treated by the algorithm as if the system were in the previous state.

Algorithm 1: Algorithm of the Strategy One implemented for the system level evaluation

```
Data: L_1, L_DTU, State, Event, UE
if State is all-On and Event is Load Notification then
    Calculate number of active users in the DTU zone (L);
    if L < L_1 then
        Perform reconfiguration to min-On state ;
if State is min-On and Event is Arrival then
    if UE is in the DTU zone and UE cooperates then
        n_DTU += 1 ;
        if n_DTU < L_DTU then
            Push the UE in the waiting queue;
        else
            Push the UE in the waiting queue;
            Perform reconfiguration to all-On state ;
            Serve all UEs in the waiting queue;
            n_DTU = 0
    else
        Start UE service ;
```

4.4.2.2 Threshold selection

The thresholds for the DTU-aware strategies are selected depending on the estimated offered load and the maximal tolerated delay proposed in the DTU zone for

Algorithm 2: Algorithm of the Strategy Two implemented for the system level evaluation.

Data: L_1, L_{DTU}, State, Event, UE
if *State is all-On and Event is Load Notification* **then**
  Calculate number of active users the DTU zone (L);
  if $L < L_1$ **then**
    Perform reconfiguration to *min-On* state ;
if *State is min-On* **then**
  Calculate number of active users in the DTU zone (N);
  if *Event is Arrival* **then**
    if *UE is in the DTU zone* **then**
      $N+ = 1$;
      if $N > C_{min}$ *and UE cooperates* **then**
        $n_{\text{DTU}}+ = 1$;
        if $N \leq L_{DTU}$ **then**
          Push the UE in the waiting queue;
        else
          Push the UE in the waiting queue;
          Perform reconfiguration to *all-On* state ;
          Serve all UEs in the waiting queue;
          $n_{\text{DTU}} = 0$
      else
        Start UE service ;
    else
      Start UE service ;
  if *Event is Departure* **then**
    $N- = 1$;
    if *UE is in the DTU zone and* $n_{DTU} > 0$ **then**
      Start service UE in the front of the queue ;
      $n_{\text{DTU}}- = 1$;

the system in study. As we consider the same traffic distributions, we evaluate the system using the models developed in Chapter 3 to obtain the optimal thresholds which minimize the average power consumption, while satisfying the delay and system QoS constraints. However, considerable differences are expected regarding the theoretical estimations.

On one hand, the Markov Chain models presented in Chapter 3 are evaluated in steady state condition and the results constitute the average behavior of the system given the modelled conditions. This is, the solution of the balance equations represents the long term behavior of the system, which corresponds to running infinitely long simulations of it. When using a system level simulator we are interested in finite simulation lengths and in estimating the average behavior of the system (i.e. the metrics of interest) from limited amount of simulation replications. The threshold estimations made using the theoretical model may not represent the optimal for each simulated scenario as the average behavior in the short-term evaluation can differ considerably from the long term average behavior. However, limited period of execution of the strategies emulates better the application for which they are developed, as the traffic conditions of cellular networks are highly variable along time, which makes impossible to maintain the same thresholds for long periods.

On the other hand, the introduction of the reconfigurations in the simulation model change the dynamic modelled in Chapter 3. However, as the traffic model is the same, the thresholds obtained from the theoretical evaluations constitute a good basis for the strategy evaluations. Furthermore, this approach allow us to estimate the error margin between the theoretical bounds and the more realistic simulated scenario.

4.4.3 Traffic management

In this section we describe how the traffic is generated by the simulated UEs and how this traffic is monitored by the network in order to take the cell switching decisions.

4.4.3.1 Traffic generation

Before the simulation starts, the inter-arrival time and the duration of the desired service are generated for the simulated UEs, obtaining realizations of their respective distributions. The offered load is a parameter of the simulation which determines the mean values required for obtaining the realizations of the distributions. For this evaluation we used the exponential distribution for both cases in agreement with the evaluation presented in Chapter 3. But the implementation can be easily enhanced to consider another distributions. Some random UEs are selected to start their services at the beginning of the simulation to emulate a warm-up period for

the traffic following a given offered load. When a UE finishes its service, a new arrival time and service duration is generated for it. This allows to generate traffic with a given intensity no matter the length of the simulated time. It is important to notice that the arrival time is the time in which the UE attempts to use the service. The actual starting time of the service may differ from this value, depending on the DTU-aware algorithm and the system state.

The service required by the UEs is a voice call. No implementation of voice services over LTE is given in ns-3. Thus, we implemented a pseudo voice service as follows. For each call we configure two applications sending packets over a UDP socket. One of them is located on the UE node sending packets to a Remote Host (RH) on the Internet. The other application is located in the RH sending packets to the UE of interest. This setup emulates a two-way communication between the UE and the RH. Each application sends packets of a specified size periodically during the service time. We consider the end nodes are using the Codec G.711 [Cis] to encode the voice packets, which generates a payload of 160 Bytes each 20 ms. Considering the protocol overhead associated to the voice service, the UDP packet size of the pseudo voice call is 179 Bytes and is sent every 20 ms during the service time. A Packet Sink application is configured in each of the end nodes to receive and consume the packets, thus avoiding overflow.

4.4.3.2 Traffic monitoring

The DTU-aware cell switching decisions are taken based on a load measurement function implemented over the top of the module. In practice, the cell switching controller, i.e. the EC-BS, could receive this information via the X2 interface, using the Resource Status Update procedure [3GP10c]. However, this procedure is not fully implemented in the used version of ns-3. Thus, we used a simplified traffic monitoring function in which the EC-BS is aware of each arrival or departure in the BS cluster, creating each time a *Load Notification* event. Thus, when the system is in *all-On* state, the EC-BS has perfect information about the aggregated load of the SC-BSs to decide if a switch to *min-On* state is suitable. Perfect information about the position of the UEs is assumed as well. Thus, the EC-BS can detect if the UE is in the DTU-zone testing a simple geometric condition for all the positions of the BSs in Θ_{SC} and Θ_{EC}:

$$(\mathrm{UE}_x - \mathrm{BS}_x)^2 + (\mathrm{UE}_y - \mathrm{BS}_y)^2 < (\frac{\mathrm{ISD}}{2})^2 \tag{4.1}$$

were UE_x and UE_y are the coordinates of the position of the UE in the deployment and BS_x and BS_y correspond to those of the corresponding BS.

4.4.4 Evaluation metrics

In order to evaluate the operation of the DTU-aware strategies we defined some performance metrics:

Call QoS: This metric is calculated on a per-call basis and it aims to identify if a call provided a minimal level of QoS to consider the user satisfied. When using system level simulations in the literature, the target QoS for a voice call corresponds to a given maximal radio interface latency. A voice user is in outage (not satisfied) if the radio interface latency of the call is greater than 50 ms [Ngm08] [EAR12a]. Thus, we calculate the average PDCP packets latency for the call duration. If it is greater than 50 ms the call is considered in outage.

System QoS: This metric is calculated over the entire simulation time and provides a measure of the proportion of blocked and dropped calls due to the the DTU-aware algorithms and the resulting dynamic system conditions. A call is considered blocked if it is in outage and during the call at least one of the following condition was satisfied:

- No reconfiguration is in progress, the system is in *all-On* state and the number of ongoing calls is greater than $C_{\max}$.
- No reconfiguration is in progress, the system is in *min-On* state and the number of ongoing calls is greater than $C_{\min}$.

A call is considered dropped if it is in outage and a reconfiguration was in progress during the service time. Finally, we calculate the proportions of blocked p_{drop} and dropped p_{block} calls during the simulation, and the system dissatisfaction metric is given by:

$$\delta = \beta p_{\text{drop}} + (1 - \beta) p_{\text{block}} \tag{4.2}$$

As in Chapter 3 we use a linear combination of the two types of dissatisfaction conditions choosing a trade-off parameter. For the evaluation we use $\beta = 0.9$, i.e. call dropping is heavily penalized.

Waiting time: This metric is calculated on a per-call basis and represents the difference between the arrival time, i.e. the time in which the UE intends to start a call, and the actual starting time of the call.

Power consumption: The average power consumption is calculated over the entire simulation time for each BS. We calculate the instantaneous output power as:

$$P_{\text{out}} = \frac{n_{RB}^{sched}}{N_{\text{RB}}} P_{\max} \tag{4.3}$$

where $P_{\max}$ is the corresponding transmission power of the BS at the time the subframe was scheduled, n_{RB}^{sched} is the number of allocated RBs in the subframe and N_{RB} is the total number of RBs available for transmission given the bandwidth. Finally, we calculate the BS power consumption using a modified version of the

EARTH model [AGD+11] considering that the BS enters in sleep mode only when it finishes the reconfiguration process, i.e. when $P_{\max} = 0$.

$$P_{\text{in}} = \begin{cases} N_{TRX}(P_0 + \Delta_P P_{\text{out}}) & 0 \leq P_{\text{out}} \leq P_{\max} \\ N_{TRX} P_{sleep} & P_{\max} = 0 \end{cases} \tag{4.4}$$

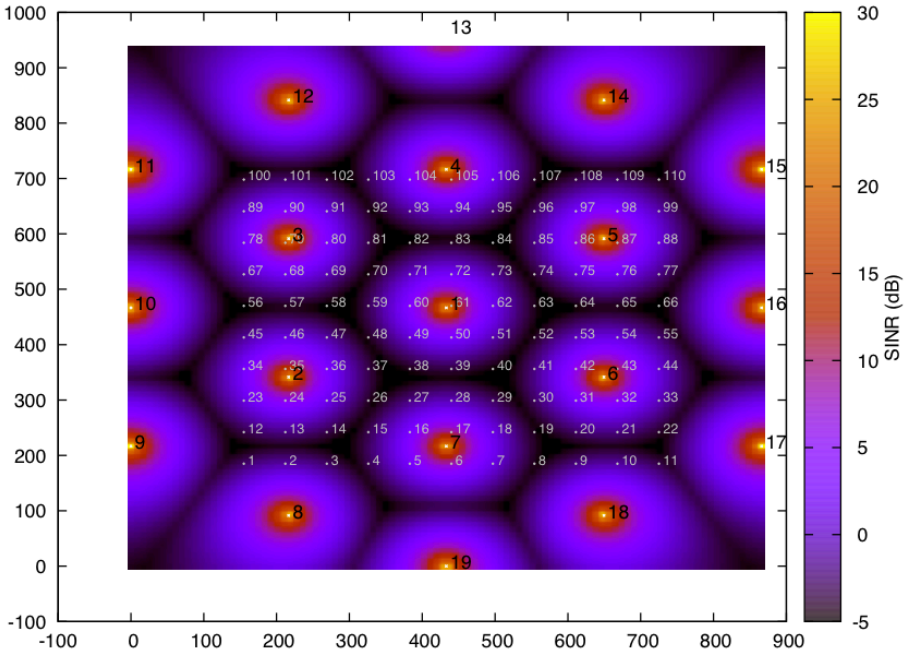
4.5 Evaluation

We simulated the 19 micro BSs of the system model described in Section 4.4.1 and depicted in Fig. 4.4. We consider static UEs, sparsely and uniformly distributed on a grid fashion with 50 meters of separation between them. We focus the evaluation in the dynamic part of the access network, i.e. the SC-BSs and the EC-BS. Thus, we only consider the UEs positioned in the area covered by them. The parameters of the simulation are summarized in Table 4.2. Some preliminary evaluations

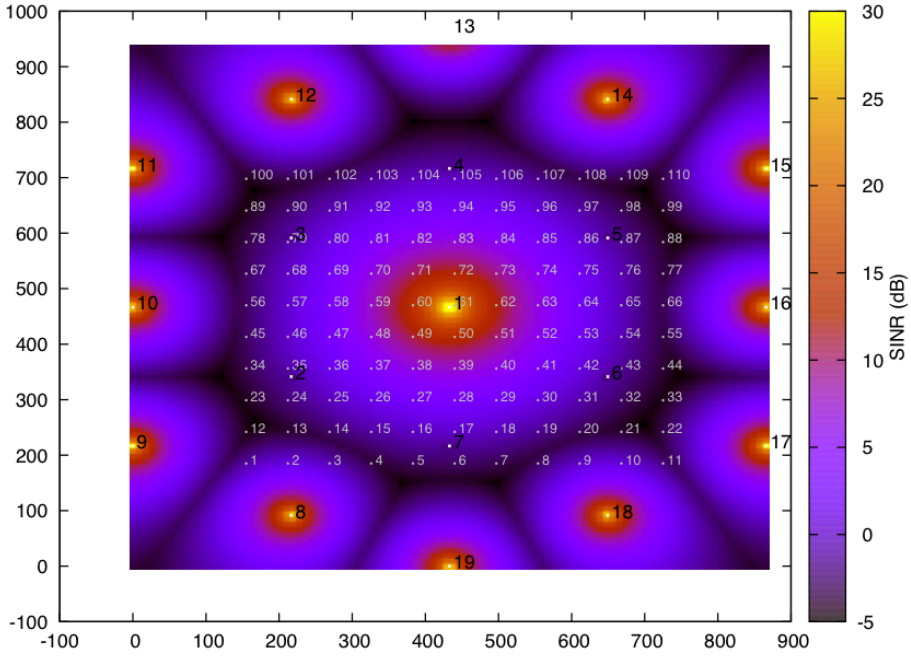
4

Table 4.2: System parameters for the system level evaluation of the strategies

Scenario (NS-3 LTE module [NS3])	
Deployment type	Hexagonal Micro
Inter Site Distance $[m]$	250
Number of BS sites	19
Antenna model	Isotropic
Path Loss model	Friis
Bandwidth [MHz]	5
Transmission Power [dBm]	38 (Expanded) 32 (Normal)
Frequency reuse	One
Scheduler	Round Robin
Handover Algorithm	A3 RSRP hysteresis 1dB time to trigger 256ms
User distribution	Uniform grid
User density	330 users/Km^2
Mobility	Constant position
Algorithm parameters	
DTU zone session capacity	$C_{\max} = 30$ $C_{\text{sta}} = 10$
γ	0.05
δ	0.05
β	0.9
Base station power consumption [AGD+11]	
N_{TRX}	2
$P_{\max}$ [W]	6.3 (Expanded) 1.6 (Normal)
P0 [W]	56
Δ_P	2.6
P_{sleep} [W]	0



(a) *all-On* state



(b) *min-On* state

Figure 4.5: Radio environment maps of the simulated scenario for the corresponding system states. The position of the BSs and UEs are indicated with white points and the corresponding identifiers are depicted at their right. Black font for the BS Cell ID and white font for the UEs IMSI.

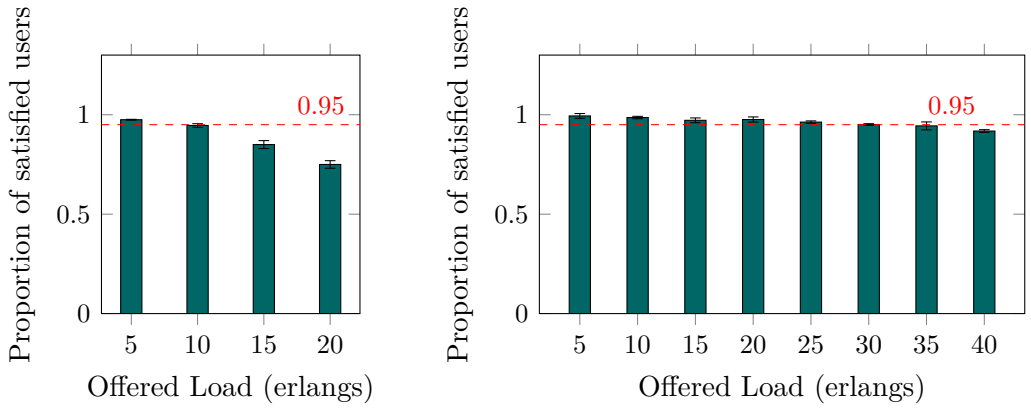
were performed in the simulator in order to obtain the needed parameters for the estimation of the DTU-aware strategy thresholds and the tuning of the cell switching algorithms. These evaluations are described in Section 4.5.1. Afterwards, the algorithms were configured and the DTU-aware strategies were put into action and evaluated in different scenarios. The results are presented in Section 4.5.2.

4.5.1 Preliminary evaluations

We performed two preliminary evaluations. First, we estimated the capacity of the system for the different states, i.e. when the system is in *min-On* state and the EC-BS is covering all the DTU zone, and when it is in *all-On* state and all the SC-BSs are operational. These parameters were needed for obtaining the strategies optimal thresholds for the considered deployment and DTU zone. Second, we derive the power reduction/increase profile of the coordinated cell switching, in order to ensure that the UEs can successfully complete the handover procedure when activating or deactivating the SC-BSs. Details about these evaluations are presented in the following.

4.5.1.1 Capacity estimation

We performed a preliminary evaluation of the scenario to determine the system capacity in terms of simultaneous ongoing calls depending on the system state. Radio environment maps of the system on both states are obtained from the simulator and are presented in Figure 4.5, showing as well the position of the simulated UEs labelled with their respective International Mobile Subscriber Identity (IMSI). We evaluated the system with static radio configuration in each state and affording different offered load levels. Afterwards, we estimated the percentage of satisfied calls. The simulated time varies between offered load scenarios, as the simulations were set to finish when each of the UE performed at least two calls. Each simulation was repeated 30 times using different seeds and the mean along with the 95 percent confidence interval is plotted in Figure 4.6 for the different offered load levels and system states. It is important to notice that when the system is in all-On state, some of the UEs are located at the cell edge, experiencing a very bad signal quality due to the high level of interference and the regularity of the UEs distribution pattern. The IMSIs of the affected UEs are: 8, 19 38, 40, 56, 57, 58, 64, 65, 66, 70 and 110; corresponding to 11% of the UEs in the evaluation. We are not interested in considering these UEs in our study as they will not be satisfied under any simulation condition, even if they were alone in the system. Thus, these UEs are excluded of all evaluations presented in this chapter. When using system level simulations, a reasonable measure of the voice system capacity is defined as the number of users in the system when more than 95% of the users are satisfied [Ngm08]. Thus, we choose



(a) *min-On state* (b) *all-On state*

Figure 4.6: Capacity evaluation: percentage of satisfied calls depending on the offered load for the different system states

the values for the parameters $C_{\max} = 30$ and $C_{\min} = 10$, given the results presented in Figure 4.6.

4.5.1.2 Coordinated cell switching power profile

The coordinated cell switching algorithm described in Section 4.4.1 was implemented with tunable time and power steps. The mobility of UEs between cells due to the cell switching is handled using the automatic handover algorithm implemented in the ns-3 module and described in Section 4.3.3, i.e. we do not manually trigger any handover procedure in the execution of the cell switching algorithm. However, the simulator does not implement a handover recovery procedure, stopping the simulation when a handover procedure fails (one of the timers depicted in Figure 4.3 expires), as it is considered as a fatal execution error. Therefore, we adopted a conservative approach for the cell switching algorithm, in which we looked for the power profile that does not cause *any* handover failure for the simulated scenario.

We simulated the system affording an offered load of 10 erlangs during 30 minutes, and we performed several switching during the simulated time. Each simulation was repeated 30 times using different seeds. We tried the power decrease/increase proposed by Marsan et al. [MCCM11] and Conte et al. [CFC+11]. The proposed profile consist in doubling or halving the transmission power $P_{\max}$ (in Watts) in each power step when switching a BS On or Off, respectively. However, contrary to their algorithm, we do not trigger any handover procedure before performing the $P_{\max}$ change. Thus, we experienced several handover failures due to random access procedure failure, which indicates that there are too many UEs performing handover at the same time to a given BS.

Table 4.3: Power profile of the coordinated cell switching reconfiguration process for the system level evaluation

	all-On→min-On		***min-On→all-On***	
	$P_{\max}$ (dBm)	t (s)	$P_{\max}$ (dBm)	t (s)
SC-BSs	-1	0.5	+1	0.5
EC-BSs	+0.5	1	-3	1

Finally, we choose the power profile presented in Table 4.3. For this profile we did not observe any handover failure and the time steps produces the shortest reconfiguration periods. For the SC-BSs we change $P_{\max}$ 1 dBm each step, which generates a reasonable batch of UEs performing handover at the same time. Smaller power steps are chosen for the EC-BS during the cell expansion. In this way the interference impact remains low, the UEs have a considerable hysteresis margin to trigger the handover, and the batch of UEs performing handover remains small. Contrary, for the shrink procedure, the batch of UEs executing the handover procedure is distributed between the SC-BSs, which diminishes the risks of handover failures due to the random access procedure. The time steps should be long enough to ensure that all handover procedures complete with success. The values were selected considering two aspects: First, the UE PHY in LTE reports the KPIs required for handover triggering (RSRP and RSRQ) to the upper layers each $200ms$, and the reported value is an average of all the measurements performed during the period [3GP12]. Second, the handover algorithm is set with a time to trigger of $256ms$. Thus, the chosen time steps allow to reflect the change in the radio conditions in the report, which may trigger adequately the handover procedure and let enough time to complete it.

4.5.2 Strategies evaluation

In this section we focus the evaluation on the execution of the DTU-aware cell switching algorithms. Thus, we assume that for each simulation, the algorithm thresholds were selected depending on accurate load estimations. We assume complete user cooperation. The simulated time for each scenario is 30 minutes, which is a reasonable period in which the network traffic can be stable around a given offered load level. Each simulation was repeated 50 times using different and independent seeds and the mean along with the 95 percent confidence interval is plotted for every parameter. We simulate the system when affording three representative offered loads given the capacity estimation presented in Section 4.5.1.1. The first offered load is equal to 7 erlangs, which is below the capacity in *min-On* state ($C_{\min}$). The second offered load is equal to $C_{\min}$, namely 10 erlangs. The last load is equal to 14 erlangs, which is superior to $C_{\min}$, but still allows the system to benefit from the DTU-aware strategies. The DTU-aware strategies are compared to two baselines strategies, simulated using the same offered load and usage patterns, which are generated before the simulation starts. The first baseline is the system operating using

the Always On strategy in which no cell switching algorithm is activated and the system is in *all-On* state during the entire simulation. The second baseline corresponds to a modest algorithm in which the cell switching is configured in order to serve the users without any delay in the start of their services, switching when the load corresponds to $0.8C_{\min}$, in order to avoid any blocking condition. This strategy is denoted as *Baseline D=0s* in the rest of this section.

4.5.2.1 User dynamics

In Figure 4.7 we present some examples of the offered load and system dynamics during the simulated time. Each plot corresponds to a simulation affording a given offered load and using a given resource management strategy, i.e. the Always On strategy, the Baseline D=0s, and the two proposed DTU-aware algorithms with different proposed maximum delay (D). All plots correspond to simulations using the same random seed. Thus, we highlight the different algorithm reactions to the same traffic distribution. From the Always On plots we can observe that the instantaneous load is highly variable, fluctuating around the expected offered load level for the entire simulated time. For an offered load of 7 erlangs, the Baseline strategy presents long periods of low power consumption followed by very short periods in all-On state. When the load increases this tend is reversed, turning into small (and inefficient) sleeping periods. When using the Strategy One, the frequency of entering in sleep mode is reduced when the load increases. However, the length of the sleeping periods is relatively uniform for the same maximal tolerated delay in the DTU zone. The simple dynamic of a waiting queue without service until turning on the SC-BSs, makes Strategy One more predictable and suitable for less dynamic hardware. The dynamic of the strategy Two is more complex, as waiting users can be served by the system in *min-On* state, causing diversity in the lengths of the sleeping periods. Strategy Two is the most efficient in low loads. For example, when the system is experiencing an offered load of 7 erlangs, the users can be served without the need of turning On the SC-BSs throughout the simulation.

4.5.2.2 Quality of service

The DTU-aware strategies aim to reduce the energy consumption of the access network without causing excessive user dissatisfaction. Two situations can degrade the perceived QoS given the dynamic of our scenario: the system is in congestion and there are not enough resources to serve a user (call blocking), or the communication is interrupted due to the cell switching transition (call dropping). The trade-off parameter between the two types of dissatisfaction conditions was set to $\beta = 0.9$, heavily penalizing the call dropping. For both strategies the user dissatisfaction metric is maintained low in the simulated scenarios, as presented in Figure 4.8.

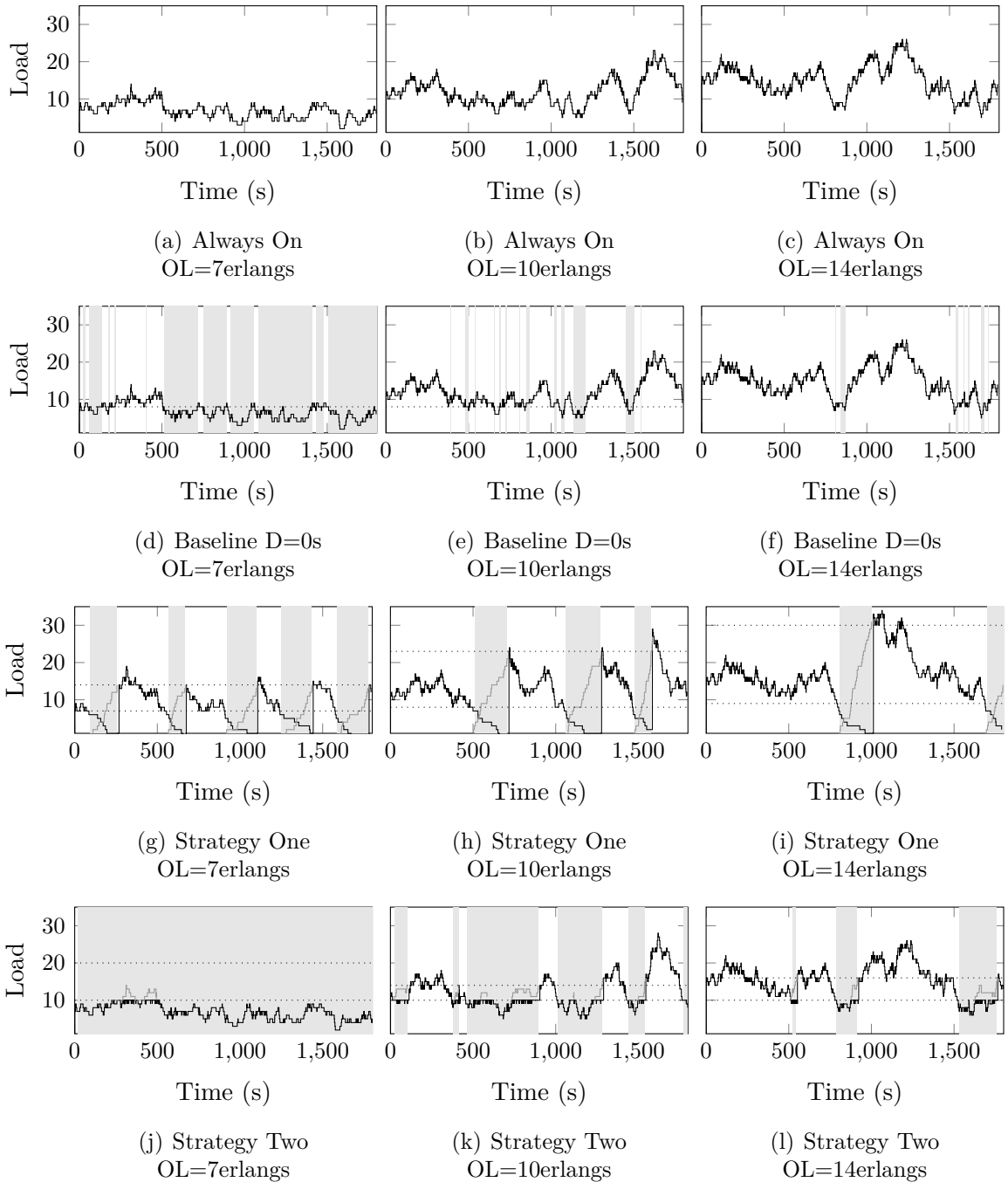
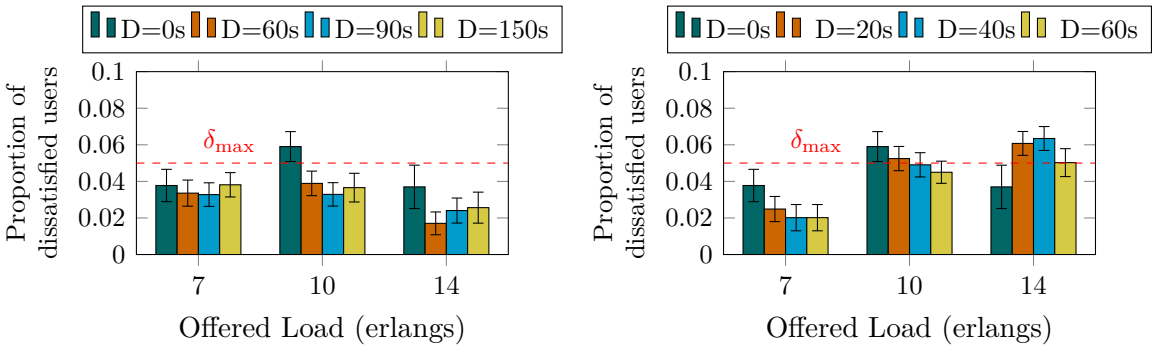


Figure 4.7: DTU zone dynamic examples. Three different offered load scenarios using the same random seed, each one simulated applying the different strategies: Always On, Baseline D=0s, Strategy One D=150s and Strategy Two D=60s. Gray lines: users waiting. Black lines: users with ongoing communications. Dotted horizontal lines: strategy thresholds. White periods: the system is in *all-On* state. Light gray periods: the system is in *min-On* state (the SC-BSs are sleeping)



(a) Strategy One

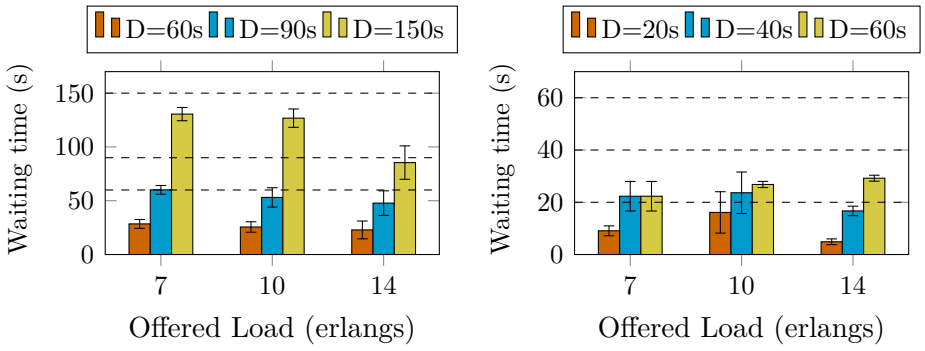
(b) Strategy Two

Figure 4.8: Proportion of dissatisfied users in the system when applying the DTU-aware strategies as well as the Baseline D=0s. Mean values and 95% confidence interval.

The dynamic of the Baseline and the Strategy Two, make them more likely to experience call dropping, as the cells switch more frequently, following closely the load variations. However, the average proportion of dissatisfied users is in most of the cases below the target of 5% selected for the strategy threshold calculation ($\delta_{\max}$). However, the dissatisfaction metric obtained from the simulations is different from the defined in the theoretical model, as we are explicitly considering the dropping of users due to the cell switching reconfigurations. Thus, $\delta_{\max}$ is only presented as a reference in this section.

4.5.2.3 Waiting time

The target probability for the waiting time constraint was set to $\gamma_{\max} = 0.05$, for the calculation of the strategy thresholds (Section 3.4.2.2 and Section 3.5.2.2). The 95th percentile of the call waiting time in the system during the simulated time is depicted in Figure 4.9, showing that in all the considered scenarios the constraint was respected: 95% of the users experienced a waiting time inferior to the one predefined in the DTU zone. The remaining 5% of the users experience slightly higher delays as can be seen in Figure 4.10. In the last figure we can also see that the higher the proposed delay, the higher the proportion of users that experience some delay. This is explained as higher delays aim to enlarge the sleeping periods, increasing the probability a user enters into the system in a sleeping period. However, in most of the evaluated scenarios, around 80% of the users did not experience any delay in the start of their calls. Thanks to the simple dynamic of Strategy One when in *min-On* state, higher delays can be proposed to the users, but with stronger guarantees of respecting the waiting time constraints. However, when the offered load increases (e.g. 14 erlangs), the system is unlikely to be in *min-On* state during



(a) Strategy One

(b) Strategy Two

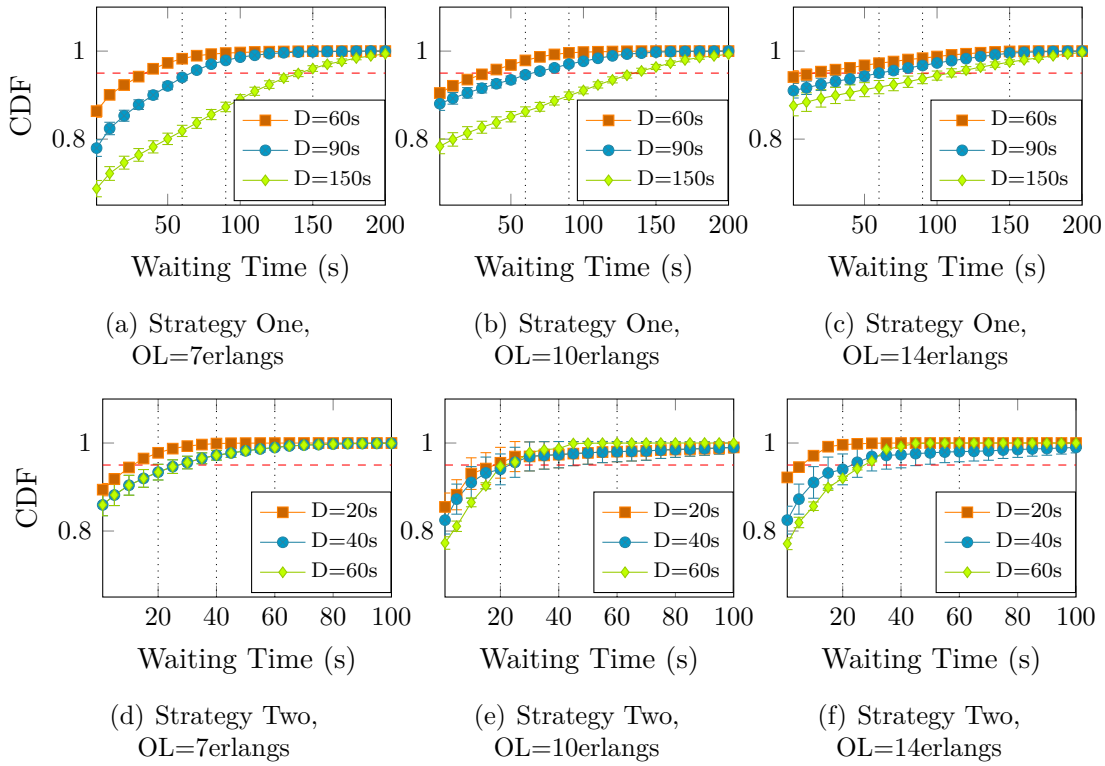
Figure 4.9: 95th percentile of the call waiting time when applying the DTU-aware strategies. Mean values and 95% confidence interval.

the simulated time, resulting in less users being delayed (Figure 4.10). Strategy Two is more complex when in *min-On* state and delay users users more actively, as they can be served under multiple conditions, i.e. departures or system state switching.

4.5.2.4 Reconfigurations impact

The reconfiguration when switching to *min-On* state takes approximately 12 seconds. This time is calculated from the moment the process is triggered until the moment the last BS switches to sleep mode. The time each BS takes for this process depends on the UE distribution, as BSs having more UEs close to them may need more time and power steps in order to satisfy the conditions for triggering the handover procedure. In our scenario this case corresponds to the BS with Cell ID 2 in Figure 4.5(a). Conversely, the reconfiguration to *all-On* state takes approximately 12.5 seconds, as the BSs need a period to adjust the transmission power to the initial value before continuing with the progressive switching On.

Figure 4.11 presents the number of switching between system states during the simulations and Figure 4.12 shows the proportion of simulated time the system is in each state, as well as the time spent performing the reconfigurations. These results confirms the intuition that the reconfigurations limit the performance of the strategies, as the system spends a non negligible part of the time reconfiguring, limiting the periods in which it can be in *min-On* state. This is more evident for the Baseline and the Strategy Two configured with small delays, for which the system switches considerably more often as shown in Figure 4.11. Strategy Two is particularly affected when the offered load level is 10 erlangs, which also corresponds to the switching off threshold of the strategy. The instantaneous load varies around this value and the strategy makes the system to switch to *all-On* state more often to satisfy the small delay constraint. Moreover, the switching off threshold is rapidly



(a) Strategy One, OL=7erlangs

(b) Strategy One, OL=10erlangs

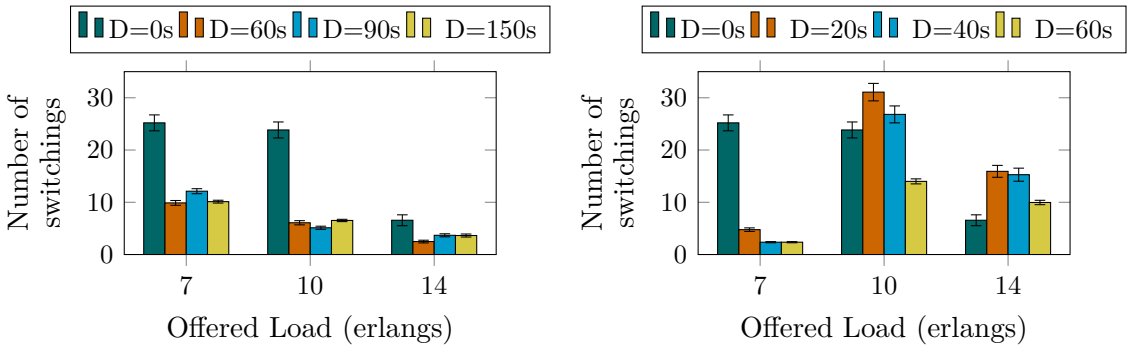
(c) Strategy One, OL=14erlangs

(d) Strategy Two, OL=7erlangs

(e) Strategy Two, OL=10erlangs

(f) Strategy Two, OL=14erlangs

Figure 4.10: Cumulative distribution function of the waiting time when applying the DTU-aware strategies. The red dashed line represents the target of 95% set for the calculation of the strategies thresholds $(1 - \gamma_{\max})$



(a) Strategy One

(b) Strategy Two

Figure 4.11: Number of transitions between system states during the simulations when applying the DTU-aware strategies as well as the Baseline D=0s. Mean values and 95% confidence interval.

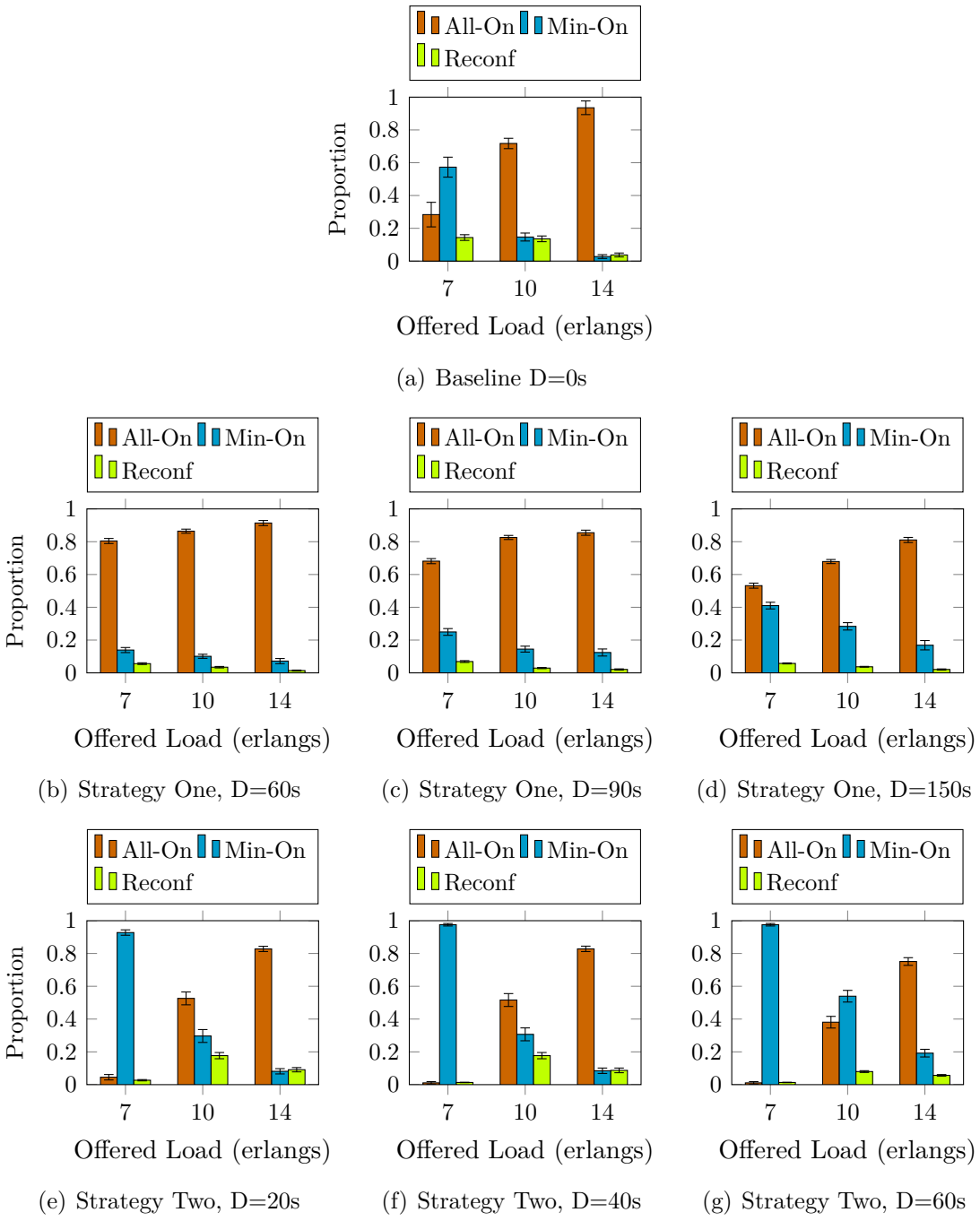
4

reached back. However, for longer delays, e.g. D=60s, the number of reconfigurations is reduced, and the system spends more time in *min-On* state.

4.5.2.5 Power consumption

Results about the average power consumption of the system are presented in Figure 4.13. The results in each figure correspond to the three resource management strategies: the Always On, the Baseline D=0s and the DTU-aware strategies configured with a given D. For each scenario, the results of the theoretical estimations are presented as well. These results were obtained solving the corresponding models as in Chapter 3, and they are represented with a white mark in each bar.

As expected for the Baseline and the DTU-aware strategies, the simulation results differ from the theoretical estimations. The differences are non negligible, representing up to 13% for the Baselines (OL=10erlangs), up to 14% for the Strategy One (D=60s, OL=7erlangs) and up to 32% for Strategy Two (D=40s, OL=10erlangs). This is mainly due to three reasons. First, the simulation model considers the reconfiguration periods that ensure that the handover procedures are completed when activating and deactivating the cells, while the theoretical evaluation assumes instantaneous reconfiguration. As showed in Figure 4.12 the system spends a non negligible part of the simulated time switching between system states, which reduces the time it can be in *min-On* state and saving power. Second, the results of the theoretical estimation represent the long term behavior of the system given the modelled conditions, as discussed in Section 4.4.2.2. The relatively short simulated time may not be representative of this long term behavior, even when the average results presented in this section are estimated over a relatively large number of independent simulation replications. Third, the calculation of the power consumption



(a) Baseline D=0s

(b) Strategy One, D=60s

(c) Strategy One, D=90s

(d) Strategy One, D=150s

(e) Strategy Two, D=20s

(f) Strategy Two, D=40s

(g) Strategy Two, D=60s

Figure 4.12: Proportion of the simulated time the system is in the different states. Mean values and 95% confidence interval.

in the theoretical estimation is based in the assumption that each call represents a fixed number of resource blocks in the downlink over a given period, while in the simulations this number is highly dependent on the BS MAC scheduler and on the radio link conditions of the UEs. As the granularity of the simulator allow us to calculate the resource allocation in a per-subframe basis, the power consumption when the BSs are operative varies in time among BSs.

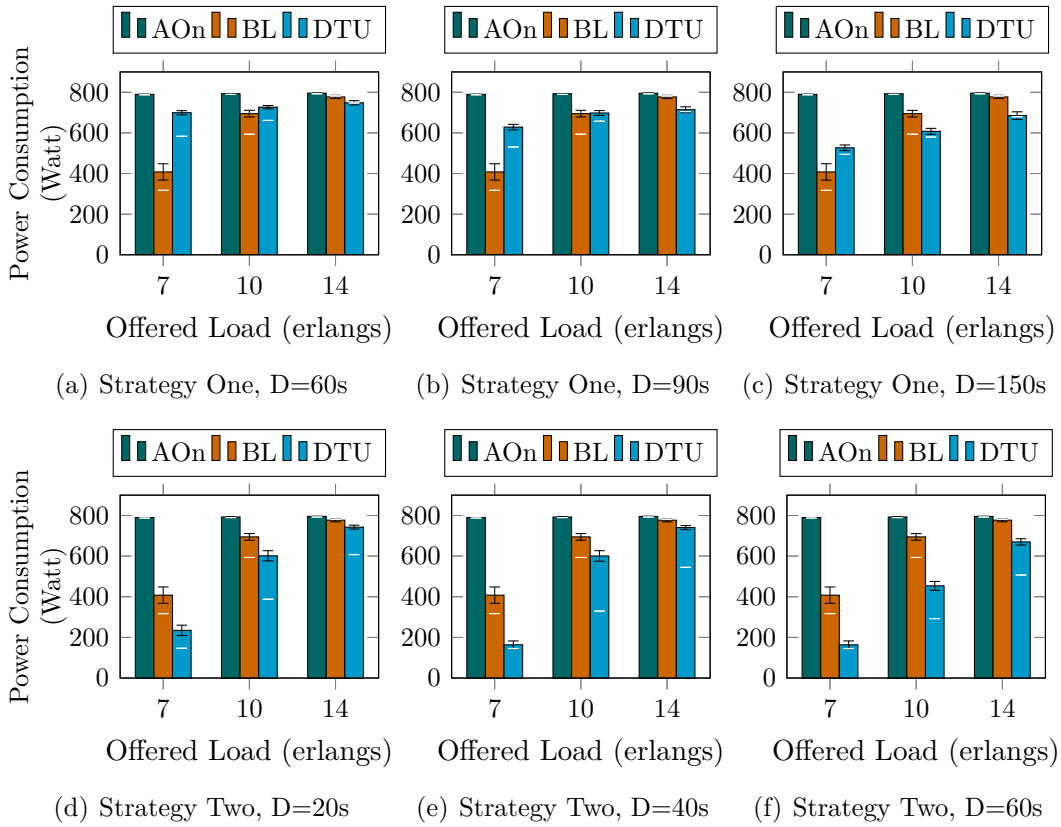
Despite the observed differences, considerable power reductions are observed compared to the Always On strategy. The results show up to 78% of power reduction for an offered load of 7 erlangs and using Strategy Two and up to 35% when using Strategy One. For 10 and 14 erlangs, reductions up to 45% and 17% are achieved respectively. Moreover, considerable reductions are observed compared the Baseline strategy, mostly with Strategy Two. It outperforms the Baseline for all considered offered load levels, exhibiting the best performance for a D=60s when affording 10 erlangs, with a difference of 30 percentage points of additional power reductions. Strategy One configured with D=150s shows a further reduction of 11% compared to the Baseline as well.

4.5.2.6 Importance of the load estimation

Our evaluation is focused on the execution of the DTU-aware cell switching algorithm, assuming a correct estimation of the offered load experienced by the system. In Table 4.4 we show the impact of the load estimation and threshold selection in satisfying the delay constraints. For example, using Strategy One and for an offered load of 10 erlangs, the thresholds are selected to propose to the users a maximum delay of 90 seconds in the DTU zone. If, instead, the offered load is 9 erlangs, the users are susceptible to wait up to 30 seconds more than the proposed by the operator, as the thresholds were selected for a different offered load. The opposite effect is observed for Strategy Two: if the load is overestimated, users will wait less, while if the load is underestimated the users can experiment 15 seconds of extra delay. Thus, the accuracy of the load estimation is an important factor to consider in the implementation of dynamic cell switching algorithms with delay constraints. Moreover, mechanisms allowing to detect excessive delay surpassing and to adapt the strategies accordingly should be considered as well.

4.6 Summary and Discussion

In this chapter we presented a simulation framework for evaluating the proposed DTU-aware strategies in scenarios approximating the real functioning of al LTE network. We presented the motivations and objectives of the system level evaluation, and described the concepts and algorithms needed for this purpose. We based our implementation on the LTE module of ns-3, which is a complete simulator with a



(a) Strategy One, D=60s (b) Strategy One, D=90s (c) Strategy One, D=150s

(d) Strategy Two, D=20s (e) Strategy Two, D=40s (f) Strategy Two, D=60s

Figure 4.13: Average power consumption of the dynamic part of the access network. Comparison of each DTU-aware strategy with the baselines scenarios (Always On and Baseline D=0s). Mean values and 95% confidence interval. The white mark in each bar represents the theoretical estimation made with the model of Chapter 3.

Table 4.4: Threshold selection for different offered loads and the impact in the maximum waiting time (D). The same thresholds define different D if the experienced offered load differ.

	Offered Load (erlangs)	**Thresholds**		D **(s)**
		L_1	$L_{\mathbf{DTU}}$	
Strategy One	9	7	14	120
	10	7	14	90
	11	7	14	70
Strategy Two	9	10	20	25
	10	10	20	40
	11	10	20	55

sufficient level of granularity, which allows us to obtain acceptable approximations of the power gain attainable with the use of the strategies.

We evaluated the DTU-aware strategies in combination with a coordinated cell switching algorithm. We simulated an access network constituted of a regular hexagonal deployment of Micro BSs with UEs uniformly distributed under their coverage, and we analyzed the system affording three different offered load levels. For each scenario, we simulated four different algorithms defining the radio resource management of the affected BSs: the Always On paradigm, the coordinated cell switching without delay and the two proposed DTU-aware strategies. For the sake of having adequate thresholds for the DTU-aware strategies, we used the same traffic distributions as in Chapter 3 and we derived using the simulator, the system session capacities for the deployment in study. Finally, we calculated the optimal thresholds using the theoretical model presented in Chapter 3. We corroborated that these thresholds were able to maintain the waiting time constraints established for the strategy during the limited periods of simulated time.

We leveraged standard handover algorithms designed for supporting user mobility in order to ensure the continuity of the services during the cell switching algorithm execution. Thus, we used a progressive reconfiguration technique performed at each system state transition in order to avoid compromising the handover protocol, while also avoiding user dissatisfaction. This process is time consuming. Moreover, due to constraints imposed by the simulator, we had to design a cell switching algorithm with zero tolerance to handover failures, which may produce even longer reconfiguration periods. This affects the power reductions attainable by the DTU-aware strategies, leading to simulation results that do not attain the theoretical bounds derived in Chapter 3. However, this is a constraint of the cell switching algorithm and not of the DTU-aware strategies. This is confirmed with the simulation results of the baseline strategy, which performance also considerably differ from the theoretical bounds due to the reconfiguration periods.

The results of the system level evaluation show the expected tends: higher delay allow higher power reductions and Strategy Two performs better than Strategy One,

given the opportunistic condition of the user collaboration and the cell switching. Even though the gains are reduced compared to the theoretical bounds, the DTU-aware strategies can represent up to 78% of reduction regarding to the Always On strategy and up to 30% of reduction regarding the baseline strategy.

4

5 Conclusions and perspectives

5.1 Thesis outcome

In this thesis we have studied the ways to improve the energy efficiency in cellular networks. In particular, we investigated methods to reduce the energy consumption of the access network considering the active cooperation of the users. The access network is composed of a large number of base stations, which represent the mayor energy consumers of the entire system. Recently, industry and academy have focused their efforts on reducing this energy consumption, driven by ecological and/or economic factors, e.g. trying to reduce the environmental impact, or attempting to decrease the operational expenses related to energy consumption.

Several strategies have been proposed towards a more energy-efficient management of the access network, where the main objective is to adapt the availability and utilization of radio resources to the temporal and spatial traffic variations. To do so, hardware upgrades and network management techniques have been proposed, attempting to deactivate the underutilized radio resources when possible to save energy. Most of the studies base their design in having minimal to no impact on users services. However, with the appropriate interactivity and incentives, the users may cooperate with the network, which may give extra flexibility to the cellular operators to optimize the resource utilization and ultimately the energy consumption of their networks

This thesis considered a specific type of user cooperation to design and control energy-efficient strategies impacting the access network. We proposed to offset the start of some users services for a given bounded delay. We called such cooperative users *Delay Tolerant User (DTU)*. Our proposal is based on a user-network interaction, in which the network may ask the users to wait to start their services if an energy-efficiency strategy could be executed in the area in which they are located.

We divided the access network resources in two categories: static and dynamic. The former are always operational as they are needed by the network to guarantee the service availability and a minimum capacity. The later provide extra capacity to the network and can be activated/deactivated depending on the policies of the employed energy efficiency technique. We proposed two strategies to control these

resources considering the delay tolerance of the users in the network, and we evaluated them to estimate the attainable energy gains the user cooperation can bring to the access network. We evaluated the strategies analytically, deriving the theoretical bounds of the attainable gains, and using system level simulations, considering more realistic conditions.

The first proposed strategy intends to maximize the utilization of dynamic resources. For this purpose, the network proposes the users to delay the start of their services when the dynamic resources are not active, accumulating service requests until having enough of them to justify the activation. The static resources are used to serve the impatient users, i.e. users without willingness to cooperate with the network, and to ensure the continuity of ongoing sessions when the dynamic resources are deactivated. The deactivation of the dynamic resources is done when the currently served traffic can be absorbed by the static resources. The activation is performed to serve the waiting users, ensuring the predefined upper bound of the delay proposed to them. We showed that this activation condition has a drawback: the dynamic resources will be systematically activated if there are cooperative users in the system. Thus, the possible gains are limited when affording low load levels, i.e. lower than the static resources capacity. In such case is better to serve all users without delay using the static resources. However, we showed that the strategy represents considerable benefits when the load increases, as it allows to distribute the load between the two types of resources and to extend the periods of low energy consumption thanks to the delay tolerance of the users.

The second proposed strategy intends to maximize the utilization of static resources, using the dynamic ones only when they are needed, i.e. when the static resources are insufficient. To do so, all users are served indistinctly if the load is inferior to the static resources capacity. When this limit is surpassed, the network will delay the start of the incoming user services, which allows the system to operate using only the static resources and consuming low levels of energy. In this strategy the waiting users have two possibilities to be served: either because some static resources become available given the completion of some services, or because the dynamic resources are activated. We showed that this strategy can provide long periods of low energy consumption, as the activation of the dynamic resources is delayed, or even avoided, in case of short periods of increased traffic.

We corroborated analytically and using system level simulations that longer delays proposed to the users provide higher energy gains. We explained how the capacity of the access network and the quality of service constraints limit the duration of the maximal delay that can be proposed to the users. Longer delays lead to higher number of users to be served when activating the dynamic resources. If this number exceeds the system capacity, the users may experience degradation in their quality of services, which is undesirable, especially when they had already waited for being served.

We considered different energy efficiency techniques in our evaluations and we showed using system level simulations that time-consuming reconfiguration periods impact the performance of the energy efficiency strategies, reducing the time the network can remain in low energy consumption. Although this limitation affects the performance of the proposed strategies, we corroborated that still further gains are attainable with the proposed user cooperation scheme.

5.2 Future work

In this section, we identify several points that may be the subject of future work for extending the contributions of this thesis.

The theoretical model presented in Chapter 3 can be enhanced to obtain a more accurate representation of the cellular access network using radio resource adaptation techniques, e.g. accounting for the impact of the reconfigurations in the system model. To do so, one may extend the state space of the proposed Markov chains, including transitional states representing the traffic and system dynamics during the reconfiguration periods. The estimation of the dissatisfaction and waiting time metrics should be adapted to consider the transitional states. Such model extension can provide a better estimation of theoretical bounds of the employment of the proposed DTU-aware strategies.

The system model presented in this thesis considers that the access network can be in one of two operational states, depending on the available resources. However, this number of states can be larger when considering more flexible techniques, and the network can activate or deactivate progressively the available resources, depending on the traffic. The model can be extended to consider these cases, adapting the Markov chain state space accordingly, e.g. adding a new set of serving and waiting states for each new capacity level resulting from the activation of a new resource. However, the complexity of the waiting time estimation in such multilevel model may be non-trivial and should be further investigated.

In Chapter 4 we presented the short-term evaluation of the proposed strategies using system level simulations. These evaluations were made assuming perfect load estimation for the selection of the switching thresholds, which remained unchanged in all predefined execution times. However, the traffic level can vary rapidly, and the thresholds should be adapted to these variations in order to satisfy the quality of service and waiting time constraints of the users in the system. For example, using advanced traffic prediction techniques to accurately estimate the load and establish the initial strategy settings. Afterwards, a closed control loop may be used to adapt and correct the thresholds and the steady state periods, depending on the measured metrics, e.g. quality of service, experienced waiting time, energy consumption. Moreover, the usage of learning mechanisms can assist this control process. The developed simulation platform allows to evaluate different traffic conditions as

well. For example, using another interarrival and service time distribution for the generation of the traffic (e.g. log-normal), and considering the user mobility in the scenarios.

Another point to consider is the definition of the scope of application of the strategies within a given access network, i.e. the definition of the DTU zone(s), the participating base stations and the differentiation of static and dynamic resources. This is highly dependent on the deployment characteristics and the radio resources adaptation technique employed to reduce the energy consumption. In this thesis we considered regular access network deployment, whereas real networks may differ from these characteristics. We also used regular patterns for the radio resource adaptation, which may not be convenient for all deployments. Thus, further studies need to be made in order to define the deployment-specific use cases of the DTU-aware strategies. These should consider the hardware flexibility and the performance targets desired by the operator, e.g. the optimal energy efficiency strategy to employ, or the optimal selection of the DTU zones to achieve a trade-off between energy reductions and number of participating base stations.

A further research direction is the adaptation of the strategies to the highly heterogeneous traffic served by the access network, i.e. different type of services with different resource utilization and quality of service constraints. Thus, the waiting policies will not only depend on the system state, but also on the type of service the customer intends to use. An interesting research direction is to investigate the trade-off between delay tolerance and service quality, e.g. offering a given limited quality of service to small-delay cooperative users, while offering upgraded services to high-delay tolerant ones. This kind of *progressive* cooperation can contribute to balance the resource utilization to further delay the activation of the dynamic resources.

Another point that needs to be considered is the management and impact of the impatient traffic. Two possible directions are envisaged: resource reservation and/or traffic prioritizing. The persistent DTU-aware strategy presented in this thesis considers the resource reservation for impatient users. However, further studies need to be made to define efficient traffic distribution and resource activation depending on both types of traffic: delay tolerant and impatient. When considering opportunistic DTU-aware strategy, the impatient traffic needs to be prioritized in periods when the system is operating with limited resources. However, such immediate resource utilization may impact the quality of service of ongoing sessions as well as the waiting time of the cooperative users. This impact should be further studied, e.g. optimizing the selection of the activation and deactivation thresholds considering estimations of the impatient traffic as well.

The strategies presented in this thesis are based on the interaction between the users and the network to achieve more efficient utilization of the network resources. The different mechanisms used to implement this interaction should be investigated

and developed. For example, one can envisage protocols and procedures in which a front-end application running in the UEs constantly communicates with a resource management controller in the network. This front-end application informs the user about the status of the network and the different options she/he has for using the network services, and the incentives or deterrents she/he can receive depending on the chosen usage. Depending on the customer usage decisions and the network state, the controller should decide if a reconfiguration is needed or if the parameters of the strategies should be adapted, e.g. adapt the thresholds or the services waiting policies.

5.3 Perspectives

In this thesis we show how the user awareness and cooperation can help to further reduce the energy consumption when employing energy efficiency techniques in the access network. To further exploit this potential, the access network should become highly dynamic. For example, dividing the control and data plane as proposed by Capone et al. [CFGU12] can be a good solution to increase the adaptivity of the access network resources. The deployment of low power, long range access technologies offers the possibility of providing control over broad areas, while the deployment of highly reactive small cells, available upon control request, provides the data services. The DTU-aware strategies protocols and procedures can be designed to operate in the control plane, to interact with the user and establish the cooperation mechanisms, as well as to control the activation and deactivation of the small cell for providing services depending on the user chosen usage.

Cellular operators have two main reasons for decrease the energy consumption of their networks. First, they aim to reduce their operational costs, as a non negligible part of them are dedicated to pay the electricity bill. Second, they intend to reduce the negative environmental impact of the operation of their networks, which is mainly produced in the generation process of the consumed electricity. The proposed strategies can contribute to these operators goals in different ways. The operators can permanently use the DTU-aware strategies in some given areas in order to reduce the overall energy consumption. For example, creating permanent DTU-zones in areas with potential underutilized BSs.

A way to reduce the ecological impact of the cellular networks operation, is to power the BSs with renewable energy sources such as solar panels or wind turbines. However, the energy generation of these sources is highly dependent on external environmental factors, which makes the availability of the energy highly variable and intermittent. The operators can rely on the proposed DTU-aware strategies, in order to adapt the network resources either to reduce the energy consumption from the electricity grid when the renewable energy is not available; or to optimize the renewable energy utilization in the case of off-grid BSs.

A further contribution to the financial goals of cellular operators reside in the context of their inclusion in the Smart Grid market. One of the pillars of the development of the Smart Grid is the dynamic management of the energy resources. When the Smart Grid detects an imbalance in the energy consumption and generation, it may ask the consumers to reduce their consumption in exchange of some financial incentives. The operators can opportunistically activate the DTU-aware strategies in some critical areas when the Smart Grid ask for these ancillary services, reducing the energy consumption and increasing their profit.

A

Summary of the reviewed literature for Network Reconfiguration Strategies

A

Table A.1: Summary of the reviewed literature for Network Reconfiguration Strategies.

Article	Contribution	Technology	Deployment	Time frame	Control	Performance constraint	Coverage Preservation	Claimed gain	Optimization	Evaluation
[ACCM09]	Estimate the energy savings of general BS switch-off schemes using analytical models. Point the trade-off between duration of the switch-off and the amount of switched cells	General	Homogeneous Non-Overlapping	Offline	Centralized	N/A	Transmission Power	Daily 25%-30%	Exhaustive search global scope	Numerical analysis
[CCMM08] [CCMM09]	Introduce the concept of dynamic radio coverage planning to reduce the power consumption of cellular networks	UMTS	Homogeneous Non-Overlapping and Overlapping	Offline	Centralized	Throughput	Transmission Power	Daily 30%-40%	Exhaustive search global scope	Numerical analysis
[SES09]	Propose load balancing techniques between radio access technologies creating opportunities to deactivate cells	GSM, UMTS	Heterogeneous Standalone BS	Online	Distributed Standalone	Channel Availability	N/A	Instantaneous 10%-50%	Exhaustive search local scope	Numerical analysis
[ZGY+09]	Present a greedy optimization algorithm executed periodically to determine the configuration of the access network	General	Homogeneous Overlapping	Online	Centralized, Distributed	Throughput	N/A	N/A	Greedy algorithm	Simulation
[SE10] [ESC11]	Propose conservative guard periods to avoid QoS degradation and unnecessary switching when applying a cell switching algorithm	GSM, UMTS, HSDPA	Standalone BS	Online	Distributed Standalone	Throughput	N/A	N/A	Exhaustive search local scope	Numerical analysis
[SEC10]	Introduce Semi-Static Sleep Mode to avoid frequent switching	GSM HSDPA	Standalone BS	Online	Distributed Standalone	Throughput	N/A	Daily 15%	Exhaustive search local scope	Numerical analysis
[MMS10]	Present a Site/Cell switching scheme in overlapping scenarios exploiting the cell-breathing effect	HSDPA	Homogeneous Overlapping	Offline	Centralized	Throughput	Antenna Configuration	Daily 25%-35%	Greedy algorithm	Simulation
[SKB10]	Propose the concept of energy partitions: subset of BSs applying a coordinated and cooperative algorithm where the load and coverage information is shared among BSs using SON schemes	LTE	Homogeneous Non-Overlapping	Online	Centralized	Throughput	Transmission Power, Antenna Configuration	N/A	Greedy algorithm	Simulation
[OK10]	Show the trade-off between network density and energy savings when applying a simple switching off algorithm	FDMA	Homogeneous Overlapping	Offline	Distributed	Throughput	None	Daily 10%-20%	Exhaustive search local scope	Numerical Analysis
[CZZN10]	Study the impact of CoMP and relaying techniques to extend coverage and further reduce the energy when a cell switching strategy is applied	LTE	Homogeneous Overlapping	Offline	Centralized	Coverage	None, Relying, CoMP	Instantaneous 3%-45%	Linear programming	Numerical Analysis
[NWGY10, Niu11]	Suggest to prioritize the association of users to cells with higher load in order to switch off the low loaded ones. Propose to reserve some bandwidth in order to react to fast variations of the load	General	Homogeneous Overlapping	Online	Centralized, Distributed	Bandwidth	None, Antenna Configuration, Relying, CoMP	20%-50%	Greedy algorithm	Simulation

Table A.1: Summary of the reviewed literature for Network Reconfiguration Strategies - Continuation I.

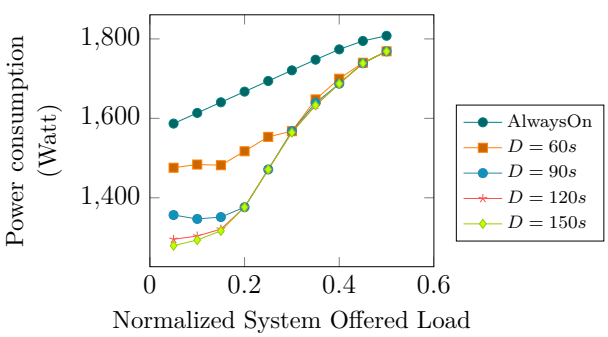
Article	Contribution	Technology	Deployment	Time frame	Control	Performance constraint	Coverage Preservation	Claimed gain	Optimization	Evaluation
[OKLN11]	Use real cellular network load and deployment data to estimate the cell overlapping degree and the potential energy reductions of NRS and inter-operator collaboration	N/A	Homogeneous Overlapping, Heterogeneous	Offline	Centralized	Coverage	Transmission Power	Instantaneous 10%-90%	Greedy algorithm	Numerical Analysis
[HMJ11]	Propose a threshold based distributed switching algorithm considering the neighbouring BSs load information and their willingness to cooperate	General	Homogeneous Overlapping	Online	Distributed	Channel availability	Transmission Power	Daily 20%-30%	Exhaustive search local scope	Simulation
[MCCM11] [CFC+11]	Design cell reconfiguration power profiles to minimize the handover failure	General	Homogeneous Overlapping	Online	N/A	Handover failure rate	N/A	N/A	N/A	Simulation
[HG11]	Propose a simple dynamic BS sectorization algorithm in which the BS sites switches from three sectorized to omnidirectional configurations to save energy	LTE	Standalone BS	Online	Distributed Standalone	Non empty buffer, Spectral Efficiency	N/A	Instantaneous 30%	Exhaustive search local scope	Simulation
[STKB11]	Complement previous work in coordinated NRS algorithms with dynamical reconfiguration and evaluation of handover failures due to them	LTE	Homogeneous Non-Overlapping and Overlapping	Online	Centralized	Throughput, Dropping probability	Transmission Power, Antenna Configuration	Instantaneous 15%-50%	Greedy algorithm	Simulation
[CFGU12]	Propose a dual layer access network architecture in which high range BSs are used for controlling while high reactive small cells provide service	N/A	Homogeneous Overlapping	Online	Centralized	Throughput	None	500%	N/A	Numerical Analysis
[HMJ12]	Suggest a traffic distribution scheme between BSs of the same and/or different radio access technologies and/or operators to create switching off opportunities	General	Heterogeneous	Online	Distributed	Channel availability	N/A	20%-40%	Exhaustive search global scope	Simulation
[GZN12]	Introduce the notion of state holding time, which is the time the system remains in a given configuration, allowing the tuning of the reactivity of the strategy	General	Homogeneous Overlapping	Online	Centralized	Throughput	N/A	N/A	Dynamic programming	Simulation
[MSES12]	Propose different level of component deactivation when reconfiguring: entire site, RAT site, sector, carrier	General LTE	Homogeneous Overlapping	Online	Centralized	Throughput	None	Daily 30%	Greedy algorithm	Simulation
[GO12, GO13]	Present an adaptive BS switching algorithm considering coverage preservation. The decision process is improved using machine learning	LTE	Homogeneous Non-Overlapping	Online	Centralized, Distributed	Cell Throughput	Antenna Configuration, CoMP	Instantaneous 50%	Exhaustive search local scope	Simulation
[SMES12]	Study the impact of the deployment of Pico BS with deactivation capabilities under the Macro coverage to increase the energy efficiency of the network	HSDPA, LTE	Homogeneous Overlapping	Offline	Centralized	Throughput	None	30%	Exhaustive search local scope	Simulation

Table A.1: Summary of the reviewed literature corresponding to Network Reconfiguration Strategies - Continuation II.

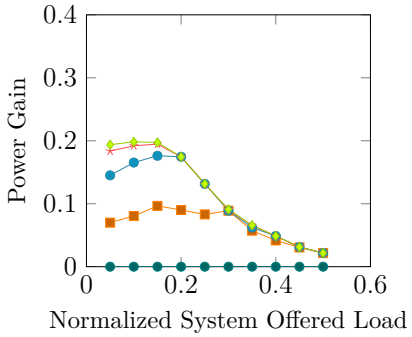
Article	Contribution	Technology	Deployment	Time frame	Control	Performance constraint	Coverage Preservation	Claimed gain	Optimization	Evaluation
[MCCM12]	Present the analytical model of multiple switching patterns to be applied progressively to adapt to different traffic conditions	General	Homogeneous Overlapping	Offline	Centralized	N/A	None	50%	Exhaustive search global scope	Numerical Analysis
[MM13]	Propose to deactivate entire access networks while the traffic is satisfied by the operators with active networks	General	N/A	Offline	Centralized	N/A	N/A	35-55%	N/A	Numerical Analysis
[RRAF13]	Study the impact of the interference in the power allocation problem when BSs are switched off	General	Homogeneous Non-Overlapping	Offline	Centralized	Throughput	Transmission Power	Instantaneous 80%	Dynamic programming	Simulation
[CJXH13]	Study coordinated scheduling mechanisms to adaptively switch off/on component carriers and/or BSs according to load variations	LTE	Homogeneous Overlapping	Online	Centralized	Cell Throughput	Transmission Power	Daily 15%	Greedy algorithm	Simulation
[HSL13]	Consider progressive BS switch-off patterns and dynamically switch among them according to the traffic load	General	Homogeneous Non-Overlapping	Offline	Centralized	Coverage, Channel availability	Antenna configuration	Instantaneous 50%	Exhaustive search global scope	Numerical Analysis
[OSK13]	Introduce the notion of network impact and defines a switching OFF/ON protocol and procedures to handle the NRS in a distributed manner	3G	Homogeneous Overlapping	Online	Distributed	Cell throughput	N/A	55%-80%	Exhaustive search local scope	Numerical analysis
[HHA+13] [HAB+13]	Design and implement a BS with low energy consumption and virtual coverage: the BS wake-up on demand to satisfy users request	GSM	Standalone BS	Online	Distributed Standalone	Service availability	Virtual coverage	Instantaneous 21% - 57%	Exhaustive search local scope	Real network
[TGA13]	Derive the theoretically optimal BS density for energy savings using stochastic geometry	General	Homogeneous Non-Overlapping	N/A	Centralized	Coverage, Throughput	Transmission Power	N/A	Numerical derivation	Numerical Analysis
[TSPB13]	Implementation of a selective switch off scheme in a real hardware controlled environment with commercial equipment (BS and UEs)	LTE	Standalone BS	Online	Distributed Standalone	N/A	N/A	Instantaneous 15%	N/A	Testbed
[GWOF13]	Propose to deactivate BSs depending on cell load thresholds and investigate the impact in the coverage considering real network deployment locations	3G	Homogeneous Overlapping	Offline	Centralized	Coverage, Cell Throughput	None	Instantaneous 13%-46%	Exhaustive search local scope	Numerical Analysis
[SNB14]	Consider to offload the operator traffic to customer deployed Femto BSs. Show the trade-off between the energy reductions and the cooperation of the customers	3G, LTE	Homogeneous Overlapping	Online	Distributed	Throughput	Transmission Power	10-70%	Exhaustive search local	Numerical Analysis
[CLHS14]	Develop a strategy to minimize the total power consumption of the network by switching cells adaptively while maintaining the network coverage	General, LTE	Homogeneous Overlapping	Online	Centralized	Coverage, Bandwidth	Transmission Power	Instantaneous 41%	Dynamic programming	Simulation
[DUGK14]	Use traffic prediction techniques to estimate the traffic trend and determine if reconfiguration is acceptable	GSM, UMTS, LTE	Heterogeneous	Online	Centralized	Cell Throughput	None	14-20%	Greedy algorithm	Simulation

B

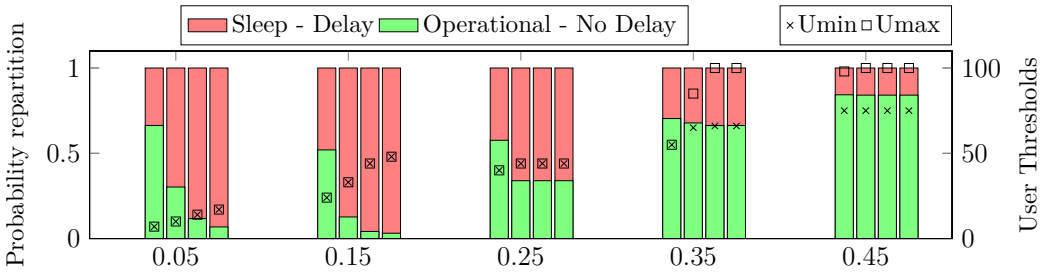
Complementary results numerical evaluation Strategy One



(a) Average power consumption



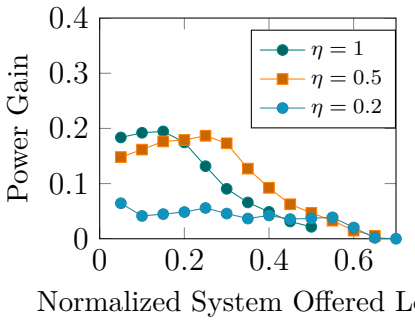
(b) Power gain



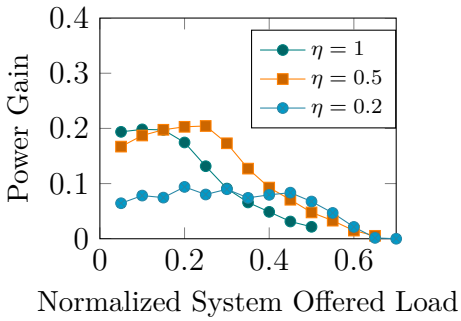
Normalized System Offered Load, each bar corresponds to (left to right): $D = 60s, 90s, 120s, 150s$

(c) Dynamic resources operation probability and strategy thresholds depending on A and D

Figure B.1: Results for the Scenario 1, $\eta = 1$,$\gamma_{\max} = 0.05$, $\delta_{\max} = 0.05$.

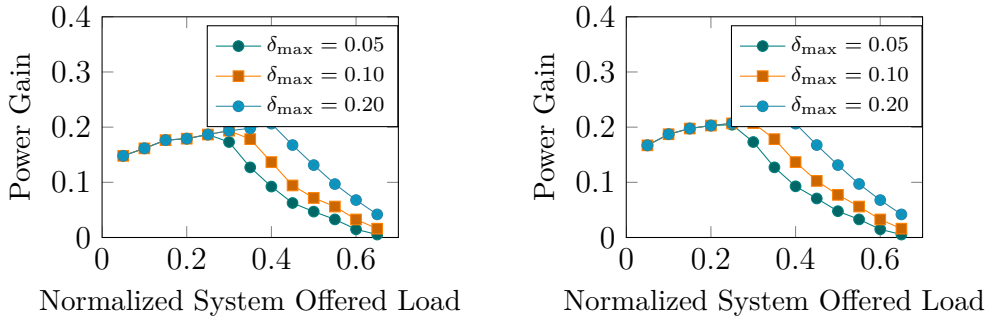


(a) $D = 120s$



(b) $D = 150s$

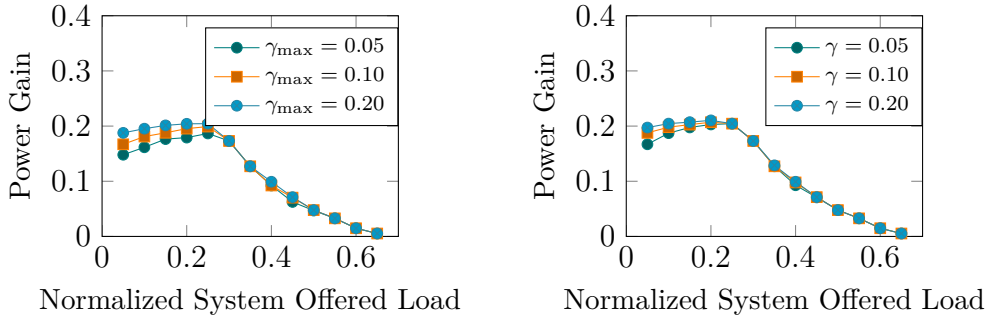
Figure B.2: Power gain variation for different levels of η. Scenario 1, fixed $\delta_{\max} = 0.05$ and $\gamma_{\max} = 0.05$.



(a) $D = 120s$ (b) $D = 150s$

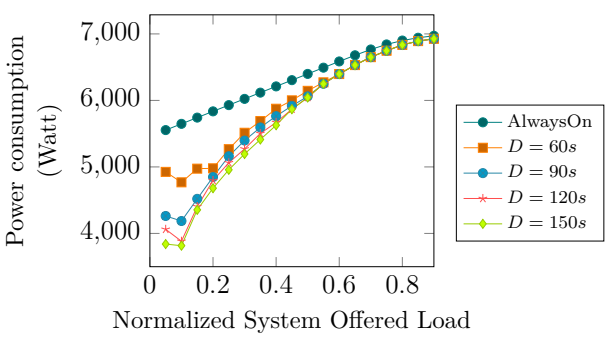
Figure B.3: Power gain variation for different levels of δ_{max}. Scenario 1, fixed $\eta = 0.5$ and $\gamma_{max} = 0.05$.

B

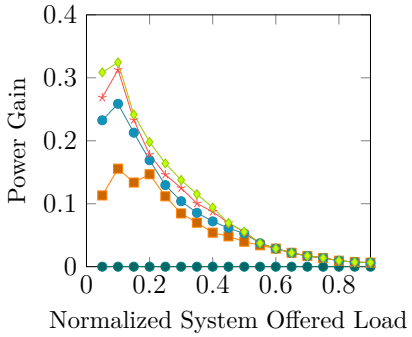


(a) $D = 120s$ (b) $D = 150s$

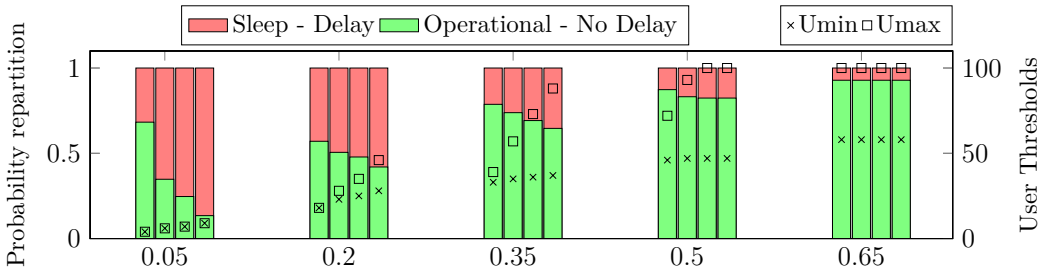
Figure B.4: Power gain variation for different levels of γ_{max}. Scenario 1, fixed $\eta = 0.5$ and $\delta_{max} = 0.05$.



(a) Average power consumption



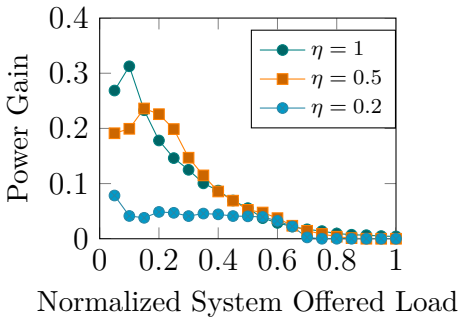
(b) Power gain



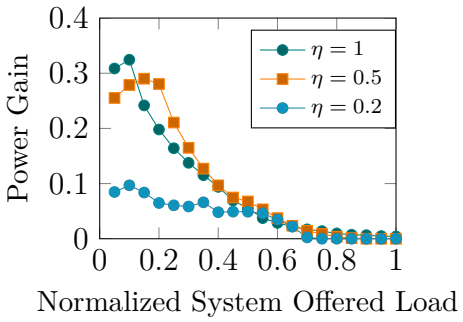
Normalized System Offered Load, each bar corresponds to (left to right): $D = 60s, 90s, 120s, 150s$

(c) Dynamic resources operation probability and strategy thresholds depending on A and D

Figure B.5: Results for the Scenario 6, $\eta = 1$,$\gamma_{\max} = 0.05$, $\delta_{\max} = 0.05$.

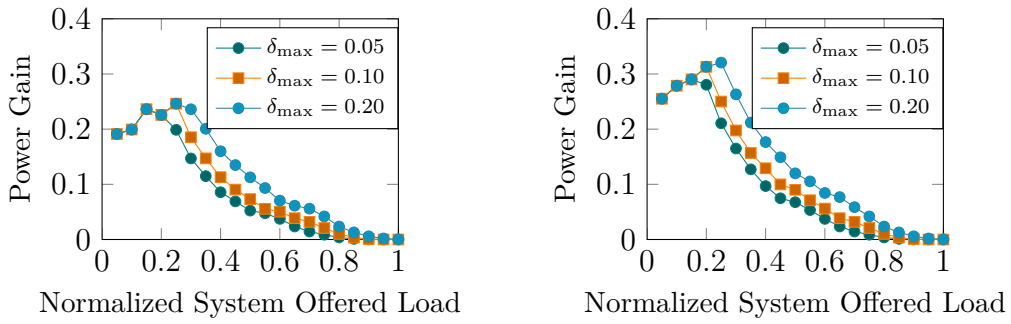


(a) $D = 120s$



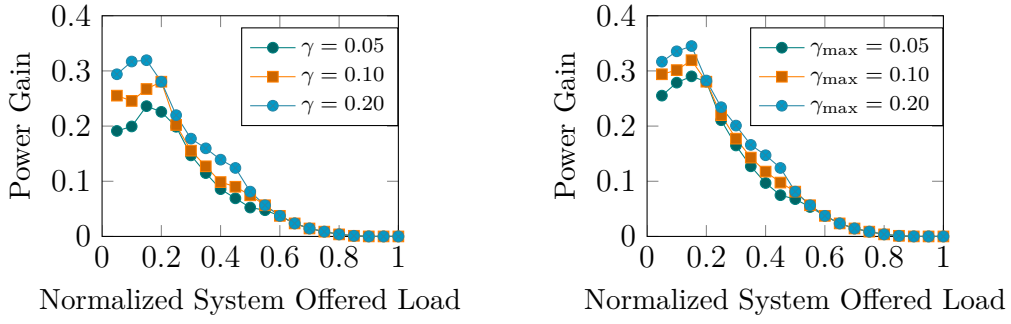
(b) $D = 150s$

Figure B.6: Power gain variation for different levels of η. Scenario 6, fixed $\delta_{\max} = 0.05$ and $\gamma_{\max} = 0.05$.



(a) $D = 120s$ (b) $D = 150s$

Figure B.7: Power gain variation for different levels of δ_{max}. Scenario 6, fixed $\eta = 0.5$ and $\gamma_{max} = 0.05$.



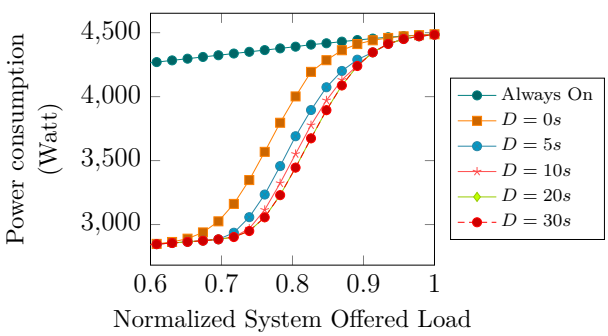
(a) $D = 120s$ (b) $D = 150s$

Figure B.8: Power gain variation for different levels of γ_{max}. Scenario 6, fixed $\eta = 0.5$ and $\delta_{max} = 0.05$.

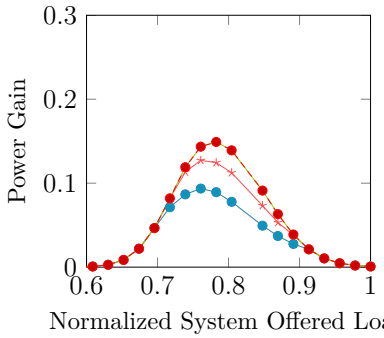
B

C

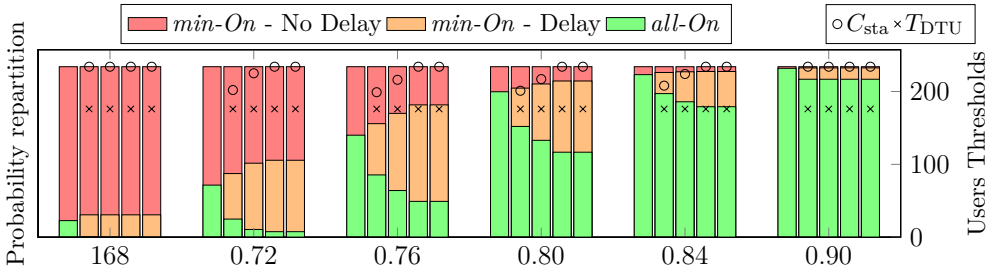
Complementary results numerical evaluation Strategy Two



(a) Average power consumption



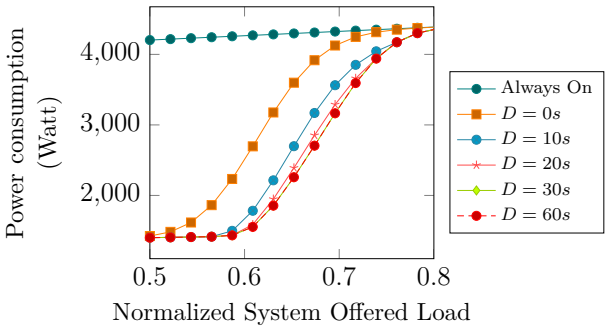
(b) Power gain regarding the $D = 0s$ baseline



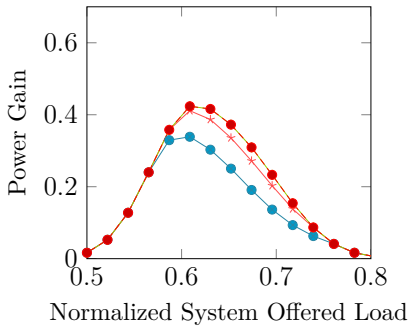
Normalized System Offered Load, each bar corresponds to (left to right): $D = 0s, 5s, 10s, 20s, 30s$

(c) System operation probability repartition and strategy thresholds depending on A and D

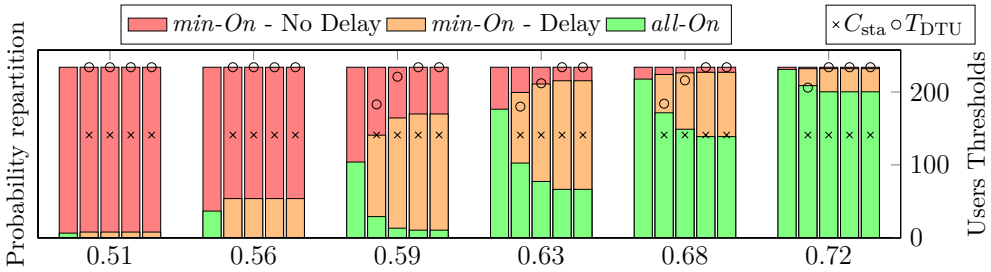
Figure C.1: Results for the Scenario 2,$\gamma_{\max} = 0.05$, $\delta_{\max} = 0.05$.



(a) Average power consumption



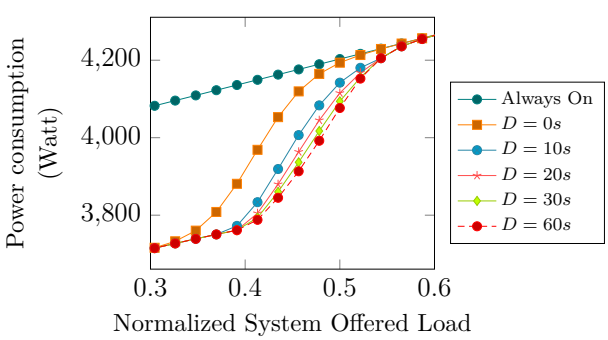
(b) Power gain regarding the $D = 0s$ baseline



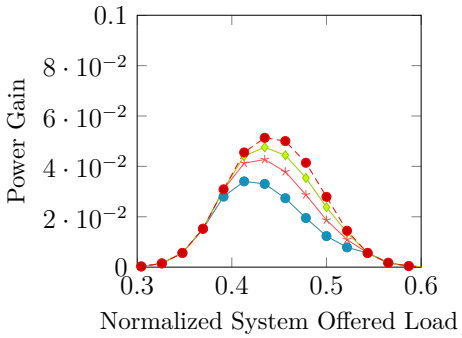
Normalized System Offered Load, each bar corresponds to (left to right): $D = 0s, 10s, 20s, 30s, 60s$

(c) System operation probability repartition and strategy thresholds depending on A and D

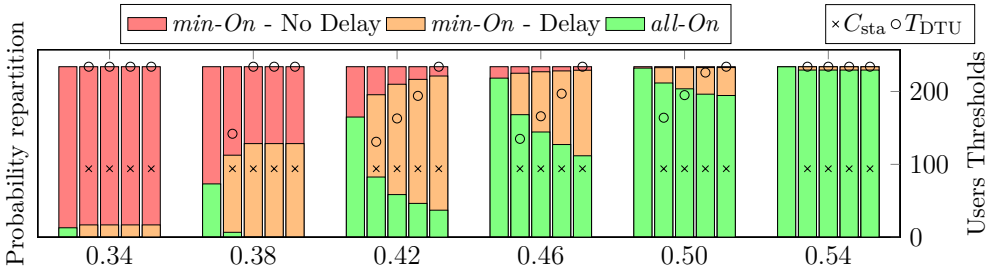
Figure C.2: Results for the Scenario 3,$\gamma_{\max} = 0.05$, $\delta_{\max} = 0.05$.



(a) Average power consumption



(b) Power gain regarding the $D = 0s$ baseline



Normalized System Offered Load, each bar corresponds to (left to right): $D = 0s, 10s, 20s, 30s, 60s$

(c) System operation probability repartition and strategy thresholds depending on A and D

Figure C.3: Results for the Scenario 4,$\gamma_{\max} = 0.05$, $\delta_{\max} = 0.05$.

List of Figures

List of Tables

Bibliography

[3GP09] 3GPP. TS 136 401 V8.6.0 - LTE; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Architecture description (3GPP TS 36.401 version 8.6.0 Release 8). Technical report, 2009. 11

[3GP10a] 3GPP. TS 125 104 V8.11.0 - Universal Mobile Telecommunications System (UMTS); Base Station (BS) radio transmission and reception (FDD) (3GPP TS 25.104 version 8.11.0 Release 8). Technical report, 2010. 13

[3GP10b] 3GPP. TS 36.213 V9.2.0 - Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 9). Technical Report Release 9, 2010. 62

[3GP10c] 3GPP. TS 36.902 V9.2.0 - Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions. 0:0–22, 2010. 11, 106

[3GP12] 3GPP. TS 36.331 V10.7.0 - LTE Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification. 0, 2012. 112

[4GA12] 4GAmericas. MIMO and Smart Antennas for 3G and 4G Wireless Systems. Technical Report October, 2012. 12, 14

[ACCM09] 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*, pages 1–5. IEEE, June 2009. 23, 31, 32, 39, 40, 100, 132

[AGD+11] Gunther Auer, Vito Giannini, Claude Desset, Istvan Godor, Per Skillermark, Magnus Olsson, Muhammad Imran, Dario Sabella, Manuel Gonzalez, Oliver Blume, and Albrecht Fehske. How much energy is needed to run a wireless network? *IEEE Wireless Communications*, 18(5):40–49, October 2011. 12, 14, 62, 63, 76, 78, 83, 108, 145

[And13] JG Andrews. Seven ways that HetNets are a cellular paradigm shift. *Communications Magazine, IEEE*, (March):136–144, 2013. 16

[ARF10] Oliver Arnold, Fred Richter, and Gerhard Fettweis. Power consumption modeling of different base station types in heterogeneous cellular networks. *Future Network and MobileSummit*, pages 1–8, 2010. 14

[BGR+14] L. Budzisz, F. Ganji, G. Rizzo, M. Ajmone Marsan, M. Meo, Y. Zhang, G. Koutitas, L. Tassiulas, S. Lambert, B. Lannoo, M. Pickavet, A Conte, I Haratcherev, and A Wolisz. Dynamic Resource Provisioning for Energy Efficiency in Wireless Access Networks: a Survey and an Outlook. *IEEE Communications Surveys & Tutorials*, PP(99):1–1, 2014. 15, 25

[BSY12] Tamer Beitelmal, Rainer Schoenen, and Halim Yanikomeroglu. On the impact of correlated shadowing on the performance of user-in-the-loop for mobility. *2012 IEEE International Conference on Communications (ICC)*, pages 7040–7044, June 2012. 46

[CCMM08] Luca Chiaraviglio, D Ciullo, M Meo, and MA Marsan. Energy-Aware UMTS Access Networks. In *International Symposium on Wireless Personal Multimedia Communications*, 2008. 23, 24, 31, 35, 39, 132

[CCMM09] L. Chiaraviglio, D. Ciullo, M. Meo, and M.A Marsan. Energy-efficient management of UMTS access networks. In *Teletraffic Congress, 2009. ITC 21 2009. 21st International*, pages 1–8, 2009. 23, 24, 31, 32, 35, 39, 40, 132

[CDML14] Eduardo Andrès Celis Muñoz, Fabrice Le Denmat, Arnaud Morin, and Xavier Lagrange. Multimedia content delivery trigger in a mobile network to reduce the peak load. *annals of telecommunications - annales des télécommunications*, December 2014. 46

[Cen] Centre Tecnologic de Telecomunicacions de Catalunya. LENA. *Online: http://networks.cttc.es/mobile-networks/software-tools/lena/. Retrieved: June 2014.* 94

[CFC+11] Alberto Conte, Afef Feki, Luca Chiaraviglio, Delia Ciullo, Michela Meo, and Marco Marsan. Cell wilting and blossoming for energy efficiency. *IEEE Wireless Communications*, 18(5):50–57, October 2011. 19, 23, 30, 31, 33, 111, 133

[CFGU12] Antonio Capone, Ilario Filippini, Bernd Gloss, and Barth Ulrich. Rethinking cellular system architecture for breaking current energy efficiency limits. In *Sustainable Internet and ICT for Sustainability (SustainIT)*, 2012. 23, 25, 31, 34, 35, 36, 129, 133

[Cis] Cisco. Voice Over IP - Per Call Bandwidth Consumption. *Online: http://www.cisco.com/c/en/us/support/docs/voice/voice-quality/7934-bwidth-consume.html. Retrieved: June 2014.* 106

[Cis14] Cisco. Cisco Visual Networking Index : Global Mobile Data Traffic Forecast Update , 2013 – 2018. pages 2013–2018, 2014. 2

BIBLIOGRAPHY

[CJXH13] Huaxia Chen, Yonglei Jiang, Jing Xu, and Honglin Hu. Energy-Efficient Coordinated Scheduling Mechanism for Cellular Communication Systems with Multiple Component Carriers. *IEEE Journal on Selected Areas in Communications*, 31(5):959–968, May 2013. 23, 31, 33, 35, 36, 39, 42, 134

[CLHS14] Chen-Yi Chang, Wanjiun Liao, Hung-Yun Hsieh, and Da-Shan Shiu. On Optimal Cell Activation for Coverage Preservation in Green Cellular Networks. *IEEE Transactions on Mobile Computing*, 13(11):2580–2591, November 2014. 23, 31, 35, 36, 39, 42, 134

[Cox12] Christopher Cox. *An introduction to LTE : LTE, LTE-advanced, SAE, and 4G mobile communications*. John Wiley & Sons, Ltd., 2012. 10, 13

[CPB+13] Filipe Cardoso, Sven Petersson, Mauro Boldi, Shinji Mizuta, Guido Dietl, Rodolfo Torrea-Duran, Claude Desset, Jouko Leinonen, and Luis Correia. Energy efficient transmission techniques for LTE. *IEEE Communications Magazine*, 51(10):182–190, October 2013. 2, 17, 29

[CZB+10] Luis Correia, Dietrich Zeller, Oliver Blume, Dieter Ferling, Ylva Jading, István Gódor, Gunther Auer, and Liesbet Der Perre. Challenges and enabling technologies for energy aware mobile radio networks. *IEEE Communications Magazine*, 48(11):66–72, November 2010. 15

[CZZN10] Dongxu Cao, Sheng Zhou, Chao Zhang, and Zhisheng Niu. Energy Saving Performance Comparison of Coordinated Multi-Point Transmission and Wireless Relaying. In *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, pages 1–5. IEEE, December 2010. 23, 29, 31, 32, 35, 39, 43, 132

[DCC14] Antonio De Domenico, Emilio Calvanese Strinati, and Antonio Capone. Enabling Green cellular networks: A survey and outlook. *Computer Communications*, 37:5–24, January 2014. 15

[DDG+12] Claude Desset, Bjorn Debaillie, Vito Giannini, Albrecht Fehske, Gunther Auer, Hauke Holtkamp, Wieslawa Wajda, Dario Sabella, Fred Richter, Manuel J. Gonzalez, Henrik Klessig, Istvan Godor, Magnus Olsson, Muhammad Ali Imran, Anton Ambrosy, and Oliver Blume. Flexible power modeling of LTE base stations. In *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 2858–2862. IEEE, April 2012. 13, 15

[DLLX02] Johan De Vriendt, Philippe Lainé, Christophe Lerouge, and Xu Xiaofeng. Mobile network evolution: a revolution on the move. *IEEE Communications Magazine*, (April):104–111, 2002. 9

[DRSW06] CJ Dawson, AHII Rick, James Wesley Seaman, and Timothy Moffett Waters. Traffic shaping of cellular service consumption through delaying of service completion according to geographical-based pricing advantages. *US Patent ...*, 2(12), 2006. 1, 45

[DUGK14] Safaa Dawoud, Abdulbaki Uzun, Sebastian Gondor, and Axel Kupper. Optimizing the Power Consumption of Mobile Networks Based on Traffic Prediction. *2014 IEEE 38th Annual Computer Software and Applications Conference*, pages 279–288, July 2014. 31, 33, 35, 36, 39, 41, 134

[EAR12a] EARTH. INFSO-ICT-247733 EARTH Deliverable D2.2: Definition and Parameterization of Reference Systems and Scenarios. Technical report, 2012. 107

[EAR12b] EARTH. INFSO-ICT-247733 EARTH Deliverable D2.3: Energy efficiency analysis of the reference systems , areas of improvements and target breakdown. Technical report, 2012. 13, 76, 84, 85, 145, 146

[EAR12c] EARTH. INFSO-ICT-247733 EARTH Deliverable D3.3: Final Report on Green Network Technologies. Technical report, 2012. 12, 15, 76

[EAR12d] EARTH. INFSO-ICT-247733 EARTH Deliverable D4.3: Final Report on Green Radio Technologies. Technical report, 2012. 15, 19, 77, 79, 149

[EAR12e] EARTH. INFSO-ICT-247733 EARTH Deliverable D5.3: Report on Validation. Technical report, 2012. 17

[ESC11] Salah-Eddine Elayoubi, Louai Saker, and Tijani Chahed. Optimal control for base station sleep mode in energy efficient radio access networks. *2011 Proceedings IEEE INFOCOM*, pages 106–110, April 2011. 19, 22, 23, 31, 33, 35, 37, 39, 40, 132

[FBZ+10] Dieter Ferling, Thomas Bohn, Dietrich Zeller, På l Frenger, István Gódor, Ylva Jading, and William Tomaselli. Energy Efficiency Approaches for Radio Nodes. *Future Network &MobileSummit*, pages 1–9, 2010. 16

[FMM+11] På l Frenger, Peter Moberg, Jens Malmodin, Ylva Jading, and Istvan Godor. Reducing Energy Consumption in LTE with Cell DTX. In *2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*, pages 1–5. IEEE, May 2011. 3, 18

[GDK+13] Vijay Gabale, UmaMaheswari Devi, Ravi Kokku, Vinay Kolar, Mukundan Madhavan, and Shivkumar Kalyanaraman. Async: De-congestion and yield management in cellular data networks. In *2013 21st IEEE*

International Conference on Network Protocols (ICNP), pages 1–10. IEEE, October 2013. 45

[GFW+11] MJ Gonzalez, Dieter Ferling, Wieslawa Wadja, Aykut ERDEM, and Philippe MAUGARS. Concepts for energy efficient LTE transceiver systems in macro base stations. *Future Network &MobileSummit*, pages 1–8, 2011. 2, 16

[GO12] Weisi Guo and Tim O'Farrell. Dynamic Cell Expansion: Traffic Aware Low Energy Cellular Network. In *2012 IEEE Vehicular Technology Conference (VTC Fall)*, pages 1–5. IEEE, September 2012. 23, 24, 31, 33, 35, 36, 39, 133

[GO13] Weisi Guo and Timothy O'Farrell. Dynamic Cell Expansion with Self-Organizing Cooperation. *IEEE Journal on Selected Areas in Communications*, 31(5):851–860, May 2013. 3, 29, 31, 33, 35, 36, 37, 39, 40, 82, 83, 84, 102, 133

[GS12] Rohit Gupta and Emilio Calvanese Strinati. Base-Station Duty-Cycling and Traffic Buffering as a Means to Achieve Green Communications. *2012 IEEE Vehicular Technology Conference (VTC Fall)*, pages 1–6, September 2012. 19, 49

[GS14] Vijay Gabale and Anand Prabhu Subramanian. GreenSlice: Enabling Renewable Energy Powered Cellular Base Stations Using Asynchronous Delivery. 2014. 49

[GWOF13] Weisi Guo, Siyi Wang, Tim O'Farrell, and Simon Fletcher. Energy Consumption of 4G Cellular Networks: A London Case Study. In *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, pages 1–5. IEEE, June 2013. 19, 23, 24, 25, 31, 32, 35, 39, 40, 134

[GZN12] Jie Gong, Sheng Zhou, and Zhisheng Niu. A Dynamic Programming Approach for Base Station Sleeping in Cellular Networks. *IEICE Transactions on Communications*, E95-B(2):551–562, 2012. 23, 31, 34, 35, 39, 42, 133

[HA14] Tao Han and Nirwan Ansari. Powering mobile networks with green energy. *IEEE Wireless Communications*, 21(1):90–96, February 2014. 18

[HAB+13] Kurtis Heimerl, K Ali, JE Blumenstock, Brian Gawalt, and EA Brewer. Expanding Rural Cellular Networks with Virtual Coverage. In *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI '13)*, pages 283–296, 2013. 23, 28, 31, 34, 35, 37, 39, 40, 134

[HAH11] Hauke Holtkamp, Gunther Auer, and Harald Haas. On Minimizing Base Station Power Consumption. In *2011 IEEE Vehicular Technology Conference (VTC Fall)*, pages 1–5. IEEE, September 2011. 12, 19

[HBB11] Ziaul Hasan, Hamidreza Boostanimehr, and Vijay K. Bhargava. Green Cellular Networks: A Survey, Some Research Issues and Challenges. *IEEE Communications Surveys & Tutorials*, 13(4):524–540, 2011. 14, 15

[HG11] Laszlo G. Hevizi and Istvan Godor. Power savings in mobile networks by dynamic base station sectorization. In *2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 2415–2417. IEEE, September 2011. 19, 23, 31, 34, 35, 37, 39, 40, 76, 133

[HHA+11] Congzheng Han, Tim Harrold, Simon Armour, Ioannis Krikidis, Stefan Videv, Peter Grant, Harald Haas, John Thompson, Ivan Ku, Cheng-Xiang Wang, Tuan Le, M. Nakhai, Jiayi Zhang, and Lajos Hanzo. Green radio: radio techniques to enable energy-efficient wireless networks. *IEEE Communications Magazine*, 49(6):46–54, June 2011. 2, 11

[HHA+13] Kurtis Heimerl, Shaddi Hasan, Kashif Ali, Tapan Parikh, and Eric Brewer. An experiment in reducing cellular base station power draw with virtual coverage. In *Proceedings of the 4th Annual Symposium on Computing for Development - ACM DEV-4 '13*, pages 1–9, New York, New York, USA, December 2013. ACM Press. 23, 28, 31, 34, 35, 37, 39, 40, 134

[HMJ11] Md. Farhad Hossain, Kumudu S. Munasinghe, and Abbas Jamalipour. An eco-inspired energy efficient access network architecture for next generation cellular systems. In *2011 IEEE Wireless Communications and Networking Conference*, pages 992–997. IEEE, March 2011. 23, 31, 34, 35, 38, 39, 40, 102, 133

[HMJ12] Md. Farhad Hossain, Kumudu S. Munasinghe, and Abbas Jamalipour. Two level cooperation for energy efficiency in multi-RAN cellular network environment. *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 2493–2497, April 2012. 23, 26, 31, 34, 35, 38, 39, 40, 133

[HNP13] Hussein Al Haj Hassan, Loutfi Nuaymi, and Alexander Pelov. Renewable energy in cellular networks: A survey. *2013 IEEE Online Conference on Green Communications (OnlineGreenComm)*, pages 1–7, October 2013. 17

BIBLIOGRAPHY

[HSJW+12] S Ha, S Sen, C Joe-Wong, Y Im, and M Chiang. Tube: time-dependent pricing for mobile data. In *ACM SIGCOMM conference on Applications, technologies, architectures, and protocols for computer communication*, pages 247–258, 2012. 1, 3, 44, 48, 54

[HSL13] Feng Han, Zoltan Safar, and K. J. Ray Liu. Energy-Efficient Base-Station Cooperative Operation with Guaranteed QoS. *IEEE Transactions on Communications*, 61(8):3505–3517, August 2013. 23, 24, 29, 31, 32, 35, 39, 40, 134

[Int08] International Telecommunication Union. ITU-R M.2134: Requirements related to technical performance for IMT-Advanced radio interface(s). Technical report, 2008. 9

[Int10] International Telecommunication Union. ITU World Radiocommunication Seminar highlights future communication technologies. *Online: http://www.itu.int/net/pressoffice/press_releases/2010/48.aspx. Retrieved: January 2015*, 2010. 10

[IV] Institute Of Telecommunication and Vienna University. Vienna LTE-A Simulator. *Online: http://www.nt.tuwien.ac.at/research/mobile-communications/vienna-lte-a-simulators/. Retrieved: June 2014.* 93

[JwHSC15] Carlee Joe-wong, Sangtae Ha, Soumya Sen, and Mung Chiang. Do Mobile Data Plans Affect Usage ? Results from a Pricing Trial with ISP Customers. In *accepted to PAM 2015*, 2015. 48, 54

[KC09] Hongseok Kim and CB Chae. A cross-layer approach to energy efficiency for adaptive MIMO systems exploiting spare capacity. *IEEE Transactions on Wireless Communications*, 8(8):4264–4275, 2009. 19

[Kiv03] D. Kivanc. Computationally efficient bandwidth allocation and power control for ofdma. *IEEE Transactions on Wireless Communications*, 2(6):1150–1158, November 2003. 18

[Lag00] Xavier Lagrange. *Les réseaux radiomobiles.* Hermes, Paris, 2000. 61

[LLY+13] Kyunghan Lee, Joohyun Lee, Yung Yi, Injong Rhee, and Song Chong. Mobile Data Offloading: How Much Can WiFi Deliver? *IEEE/ACM Transactions on Networking*, 21(2):1–14, 2013. 47, 49

[MCCM11] Marco Ajmone Marsan, Luca Chiaraviglio, Delia Ciullo, and Michela Meo. Switch-Off Transients in Cellular Access Networks with Sleep Modes. *2011 IEEE International Conference on Communications Workshops (ICC)*, pages 1–6, June 2011. 23, 30, 31, 33, 111, 133

[MCCM12] Marco Ajmone Marsan, Luca Chiaraviglio, Delia Ciullo, and Michela Meo. Multiple daily base station switch-offs in cellular networks. In *2012 Fourth International Conference on Communications and Electronics (ICCE)*, pages 245–250. IEEE, August 2012. 19, 23, 31, 32, 35, 39, 40, 134

[Mic] Microsoft. How much bandwidth does Skype need? *Online: https://support.skype.com/en/faq/fa1417/how-much-bandwidth-does-skype-need. Retrieved: June 2014.* 77

[MM13] MA Marsan and M Meo. Network sharing and its energy benefits: A study of European mobile network operators. *Global Communications Conference (GLOBECOM)*, 2016:2561–2567, 2013. 23, 27, 31, 32, 35, 134

[MML+10] Jens Malmodin, Å sa Moberg, Dag Lundén, Göran Finnveden, and Nina Lövehagen. Greenhouse Gas Emissions and Operational Electricity Use in the ICT and Entertainment & Media Sectors. *Journal of Industrial Ecology*, 14(5):770–790, October 2010. 2

[MMS10] Gilbert Micallef, Preben Mogensen, and Hans-Otto Scheck. Cell size breathing and possibilities to introduce cell sleep mode. In *2010 European Wireless Conference (EW)*, pages 111–115. IEEE, 2010. 19, 23, 31, 32, 35, 39, 41, 132

[MNH12] Sanjida Moury, M. Nazim Khandoker, and Syed Mustansir Haider. Feasibility study of solar PV arrays in grid connected cellular BTS sites. *2012 International Conference on Advances in Power Conversion and Energy Technologies (APCET)*, pages 1–5, August 2012. 18

[MSES12] Gilbert Micallef, Louai Saker, Salah E. Elayoubi, and Hans-Otto Sceck. Realistic Energy Saving Potential of Sleep Mode for Existing and Future Mobile Networks. *Journal of Communications*, 7(10):740–748, October 2012. 2, 23, 25, 29, 34, 35, 36, 39, 41, 133

[Ngm08] Ngmn Alliance. Next generation mobile networks radio access performance evaluation methodology. Technical report, 2008. 107, 110

[Niu11] Zhisheng Niu. TANGO: traffic-aware network planning and green operation. *IEEE Wireless Communications*, 18(5):25–29, October 2011. 23, 25, 31, 34, 35, 36, 37, 39, 42, 132

[NS3] NS3 Consortium. NS-3 LTE Module. *Online: http://www.nsnam.org/docs/models/html/lte.html. Retrieved: November 2014.* 93, 94, 96, 98, 108, 146

[NWGY10] Zhisheng Niu, Yiqun Wu, Jie Gong, and Zexi Yang. Cell zooming for cost-efficient green cellular networks. *IEEE Communications Magazine*, 48(11):74–79, November 2010. 19, 23, 25, 31, 34, 35, 36, 39, 42, 132

[OK10] Eunsung Oh and Bhaskar Krishnamachari. Energy Savings through Dynamic Base Station Switching in Cellular Wireless Access Networks. In *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, pages 1–5. IEEE, December 2010. 23, 24, 31, 32, 35, 38, 39, 40, 132

[OKLN11] Eunsung Oh, Bhaskar Krishnamachari, Xin Liu, and Zhisheng Niu. Toward dynamic energy-efficient operation of cellular network infrastructure. *IEEE Communications Magazine*, 49(6):56–61, June 2011. 23, 24, 26, 31, 32, 35, 39, 42, 133

[Omn] Omnet++. SimuLTE. *Online: http://omnetpp.org/. Retrieved: June 2014.* 93

[OSK13] Eunsung Oh, Kyuho Son, and Bhaskar Krishnamachari. Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks. *IEEE Transactions on Wireless Communications*, 12(5):2126–2136, May 2013. 19, 23, 25, 31, 34, 35, 38, 39, 40, 134

[PBM11] G Piro, Nicola Baldo, and Marco Miozzo. An LTE module for the ns-3 network simulator. *Proceedings of the 4th International ICST . . .*, 2011. 97

[PK09] KI Pedersen and TE Kolding. An overview of downlink radio resource management for UTRAN long-term evolution. *IEEE Communications Magazine*, (July):86–93, 2009. 18

[PKWV12] Magnus Proebster, Matthias Kaschub, Thomas Werthmann, and Stefan Valentin. Context-aware resource allocation for cellular wireless networks. *EURASIP Journal on Wireless Communications and Networking*, 2012(1):216, 2012. 45

[PMF+13] Guiseppe Piro, Marco Miozzo, Guiseppe Forte, Nicola Baldo, Luigi Grieco, Gennaro Boggia, and Paolo Dini. HetNets Powered by Renewable Energy Sources. *IEEE Internet Computing*, 2013. 18

[PP05] Achille Pattavina and Alessandra Parini. Modelling voice call interarrival and holding time distributions in mobile networks. *Proc. of 19th International Teletraffic Congress*, pages 729–738, 2005. 62, 76, 83

[PSN+11] Subodh Paudel, J. N. Shrestha, Fernando J. Neto, Jorge a. F. Ferreira, and Muna Adhikari. Optimization of hybrid PV/wind power system

for remote telecom station. *2011 International Conference on Power and Energy Systems*, pages 1–6, December 2011. 18

[RF14] Jaya B. Rao and Abraham O. Fapojuwo. A Survey of Energy Efficient Resource Management Techniques for Multicell Cellular Networks. *IEEE Communications Surveys & Tutorials*, 16(1):154–180, 2014. 15

[RKN10] Moo-ryong Ra, Martin H Krieger, and Michael J Neely. Energy-Delay Tradeoffs in Smartphone Applications. In *System*, 2010. 46, 49

[RRAF13] Balaji Rengarajan, Gianluca Rizzo, Marco Ajmone Marsan, and Barbara Furletti. QoS-aware greening of interference-limited cellular networks. In *2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, pages 1–9. IEEE, June 2013. 23, 24, 31, 32, 35, 39, 42, 134

[SBM+12a] Rainer Schoenen, Gurhan Bulu, Amir Mirtaheri, Tamer Beitelmal, and Halim Yanikomeroglu. First Survey Results of Quantified User Behavior in User-in-the-Loop Scenarios for Sustainable Wireless Networks. In *2012 IEEE Vehicular Technology Conference (VTC Fall)*, pages 1–5. IEEE, September 2012. 47, 54

[SBM+12b] Rainer Schoenen, Gurhan Bulu, Amir Mirtaheri, Tamer Beitelmal, and Halim Yanikomeroglu. Quantified user behavior in user-in-the-loop spatially and demand controlled cellular systems. In *IEEE European Wireless Conference*, pages 1–8, 2012. 3, 47, 54

[SE10] Louai Saker and Salah Eddine Elayoubi. Sleep mode implementation issues in green base stations. In *21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 1683–1688. IEEE, September 2010. 22, 23, 31, 33, 35, 37, 39, 40, 132

[SEC10] Louai Saker, Salah-Eddine Elayoubi, and Tijani Chahed. Minimizing Energy Consumption via Sleep Mode in Green Base Station. In *2010 IEEE Wireless Communication and Networking Conference*, pages 1–6. IEEE, April 2010. 22, 23, 31, 33, 35, 37, 39, 40, 132

[SES09] L. Saker, S. E. Elayoubi, and H. O. Scheck. System Selection and Sleep Mode for Energy Saving in Cooperative 2G/3G Networks. *2009 IEEE 70th Vehicular Technology Conference Fall*, pages 1–5, September 2009. 22, 23, 26, 31, 35, 37, 39, 40, 132

[SF12] Per Skillermark and På l Frenger. Enhancing Energy Efficiency in LTE with Antenna Muting. In *2012 IEEE 75th Vehicular Technology Conference (VTC Spring)*, pages 1–5. IEEE, May 2012. 19

BIBLIOGRAPHY

[SJWHC12] Soumya Sen, Carlee Joe-Wong, Sangtae Ha, and Mung Chiang. Incentivizing Time-Shifting of Data : A Survey of Time-Dependent Pricing for Internet Access. *IEEE Communications Magazine*, (November):91–99, 2012. 44

[SKB10] Konstantinos Samdanis, Dirk Kutscher, and Marcus Brunner. Self-organized energy efficient cellular networks. In *21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 1665–1670. IEEE, September 2010. 23, 24, 31, 33, 35, 36, 39, 41, 132

[SMES12] L. Saker, G. Micallef, S. E. Elayoubi, and H. O. Scheck. Impact of picocells on the capacity and energy efficiency of mobile networks. *Annals of Telecommunications - Annales Des Télécommunications*, 67(3-4):133–146, February 2012. 23, 25, 29, 31, 32, 35, 39, 133

[SNB12] Luis a Suarez, Loutfi Nuaymi, and Jean-Marie Bonnin. An overview and classification of research approaches in green wireless networks. *EURASIP Journal on Wireless Communications and Networking*, 2012(1):142, 2012. 15

[SNB14] Luis Suárez, Loutfi Nuaymi, and Jean-Marie Bonnin. Energy-efficient BS switching-off and cell topology management for macro/femto environments. *Computer Networks*, November 2014. 23, 27, 134

[STKB11] Konstantinos Samdanis, Tarik Taleb, Dirk Kutscher, and Marcus Brunner. Self Organized Network Management Functions for Energy Efficient Cellular Urban Infrastructures. *Mobile Networks and Applications*, 17(1):119–131, February 2011. 23, 24, 25, 30, 31, 33, 35, 36, 39, 41, 133

[SY13] Rainer Schoenen and Halim Yanikomeroglu. Erlang analysis of cellular networks using stochastic Petri nets and user-in-the-loop extension for demand control. In *2013 IEEE Globecom Workshops (GC Wkshps)*, pages 298–303. IEEE, December 2013. 44

[SYW11] Rainer Schoenen, Halim Yanikomeroglu, and Bernhard Walke. User in the Loop: Mobility Aware Users Substantially Boost Spectral Efficiency of Cellular OFDMA Systems. *IEEE Communications Letters*, 15(5):488–490, May 2011. 45

[Tel] Telematics Lab. LTE-Sim. *Online: http://telematics.poliba.it/index.php/en/lte-sim. Retrieved: June 2014*. 93

[TGA13] Dimitrios Tsilimantos, Jean-Marie Gorce, and Eitan Altman. Stochastic analysis of energy savings with sleep mode in OFDMA wireless networks. In *INFOCOM*, 2013. 23, 39, 134

[TSPB13] William Tomaselli, Dario Sabella, Valerio Palestini, and Valerio Bernasconi. Energy efficiency performances of selective switch OFF algorithm in LTE mobile networks. *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 3254–3258, September 2013. 22, 23, 31, 33, 35, 37, 39, 40, 134

[VLL+14] Ward Van Heddeghem, Sofie Lambert, Bart Lannoo, Didier Colle, Mario Pickavet, and Piet Demeester. Trends in worldwide ICT electricity consumption from 2007 to 2012. *Computer Communications*, 50:64–76, September 2014. 2

[Wik] Wikipedia. Erlang distribution. *Online: https://en.wikipedia.org/wiki/Erlang_distribution. Retrieved: June 2014.* 60, 74

[Xu10] Jinbiao Xu. Practical Digital Pre-Distortion Techniques for PA Linearization in 3GPP LTE. Technical report, 2010. 16

[YQST13] Ye Yan, Yi Qian, Hamid Sharif, and David Tipper. A Survey on Smart Grid Communication Infrastructures: Motivations, Requirements and Challenges. *IEEE Communications Surveys & Tutorials*, 15(1):5–20, 2013. 1

[ZGY+09] Sheng Zhou, Jie Gong, Zexi Yang, Zhisheng Niu, Peng Yang, and Development Corporation. Green Mobile Access Network with Dynamic Base Station Energy Saving. In *ACM MobiCom*, pages 10–12, 2009. 23, 25, 31, 34, 35, 36, 37, 39, 42, 132

Résumé

Les réseaux cellulaires fonctionnent traditionnellement en maintenant l'infrastructure du réseau toujours opérationnelle, de manière à satisfaire non-seulement la disponibilité du service, mais aussi les capacités réseaux nécessaires pour gérer les pics de charge utilisateur. Ce paradigme est limité en terme d'efficacité énergétique et une approche différente est actuellement prise dans le cadre des recherches industrielles et académiques. Dans ce nouveau paradigme, l'infrastructure est dynamiquement adaptée en fonction des variations temporelles et spatiales du trafic, réduisant de ce fait les pertes d'énergie.

Dans cette thèse, nous prenons en compte la coopération de l'utilisateur dans la mise en œuvre et le contrôle des techniques d'efficacité énergétique. Nous considérons un type spécifique de coopération où l'utilisateur est capable de décaler l'utilisation de son service pendant un temps fixé et borné, et connu à l'avance. En utilisant une interaction proactive avec les utilisateurs, le réseau peut alors leur demander de décaler l'utilisation de leur service si une technique d'efficacité énergétique est appliquée dans leur zone de localization (ex: une station de base étant désactivée). Ainsi, une portion du trafic n'est pas générée immédiatement et le réseau peut optimiser l'utilisation des ressources, rester une plus longue période en utilisant un ensemble de ressources limité, et donc entrainer une moindre consommation d'énergie.

Nous présentons d'abord une vue d'ensemble et une classification de le littérature des domaines abordées dans le cadre de cette thèse: l'efficacité énergétique dans les réseaux cellulaires et l'adaptation du trafic. Ensuite, nous proposons deux stratégies différentes pour contrôler les ressources du réseau, fonction de la coopération des utilisateurs et de leur tolérance aux délais, et nous évaluons l'impact d'un tel schéma de coopération pour différentes techniques d'efficacité énergétique. Après ça, nous proposons un modèle théorique afin d'évaluer analytiquement les stratégies proposées. Nous obtenons alors les limites théoriques d'économie d'énergie atteignables par utilisation de différentes stratégie d'efficacité énergétique and nous évaluons le compromis entre les limites du temps d'attente proposé aux utilisateurs, et les économies d'énergie atteignables. Nous avons observés qu'une augmentation dans la tolérance au délai entraine un meilleur gain énergétique, et que ce gain a une limite maximale déterminée par la capacité du système. Nous avons aussi noté que retarder de manière opportuniste le service de l'utilisateur en fonction des conditions du système est plus bénéfique que de les retarder tous systématiquement. Finalement, nous avons évalué par simulation les stratégies sous des conditions plus réalistes. La simulation confirme les tendances observées avec le modèle théorique et nous avons noté que les gains atteignables sont limités par les temps d'adaptation du système lors des phases de reconfiguration.

Mots-clés : Téléphonie mobile, Économies d'énergie, LTE, Modèles mathématiques, Processus de Markov, Simulation par ordinateur, Conception cross-layer, Qualité de Service (télécommunications)

Abstract

Cellular networks have been traditionally designed to keep the network infrastructure always operational. This in order to ensure ubiquitous service availability and enough capacity to serve the peak of usage of the customers.

Recently, the concern about the energy efficiency of this paradigm has increased, and a different approach has concentrated the research efforts of industry and academy. In this new paradigm, the infrastructure is dynamically adapted to the temporal and spatial traffic variations, reducing the energy wastage.

The majority of these studies make the adaptation of the infrastructure unnoticeable to the users.
However, with the appropriate interactivity and incentives, some users may be willing to offer their cooperation to the network.

In this thesis we consider the user cooperation in the design and control of energy efficiency techniques.

We consider a specific type of cooperation in which the users are able to offset the start of their services for a bounded and known-in-advance delay. Based on proactive interaction with the users, the network may ask them to delay the start of their services if an energy efficiency technique is applied in the area where they are located (e.g. a base station is switched off).

Thus, a portion of the traffic is shifted and the network can optimize the resource utilization in order to consume less energy.

First, we present an overview and classification of the literature covering the main domains of the thesis, namely energy efficiency in cellular networks and user demand shaping. We describe as well the most recent cellular network architecture - LTE. Then, we propose two different strategies to control the network resources depending on the cooperation of the users and their delay tolerance, and we evaluate the impact of such cooperation schemes in different energy efficiency techniques. Afterwards, we propose a theoretical framework for the analytical evaluation of the proposed strategies. We obtained the theoretical bounds of the attainable energy savings when employing different energy efficiency techniques, and we investigated the trade-off between the waiting time bounds proposed to the users and the energy gains. We observed that increased delay tolerance leads to more energy gains, and that the gains have an upper bound determined by the system serving capacity. We also noted that delaying opportunistically the user services depending on the system conditions is more beneficial than systematically delaying all of them. Finally, we evaluated the strategies under more realistic conditions using system level simulations. We corroborated the theoretical trends and we observed that the attainable gains are limited by the duration of the network reconfiguration process.

Keywords : cellular networks, energy efficiency, LTE, mathematical model, Markov chain, simulation, cross layer design, delay tolerance

TELECOM Bretagne

n° d'ordre : 2015telb0359
Télécom Bretagne
Technopôle Brest-Iroise - CS 83818 - 29238 Brest Cedex 3
Tél : + 33(0) 29 00 11 11 - Fax : + 33(0) 29 00 10 00