UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MATIAS ARTUR KLAFKE SCHIMUNECK

Adaptive Monte Carlo Algorithm to Global Radio Resources Optimization in H-CRAN

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Prof. Dr. Juergen Rochol
Coadvisor: Prof. Dr. Cristiano Bonato Both

Porto Alegre
August 2017

CIP – CATALOGING-IN-PUBLICATION

Schimuneck, Matias Artur Klafke

Adaptive Monte Carlo Algorithm to Global Radio Resources Optimization in H-CRAN / Matias Artur Klafke Schimuneck. – Porto Alegre: PPGC da UFRGS, 2017.

94 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2017. Advisor: Juergen Rochol; Coadvisor: Cristiano Bonato Both.

1. Radio resource allocation. 2. Cellular network. 3. Adaptative monte carlo. 4. 5G. 5. Energy efficiency. I. Rochol, Juergen. II. Both, Cristiano Bonato. III. Título.

ACKNOWLEDGMENTS

Only in Brazilian Portuguese. See the page 4.

AGRADECIMENTOS

Agradeço primeiramente, e acima de tudo, a minha família. Meu pai, meu grande motivador. Minha mãe, minha eterna protetora. Minha avó, pelas infinitas orações e velas queimadas. Minha irmã, melhor amiga e confidente. Minha namorada, a qual tem sido uma fonte de paciência, companheirismo e amor constante. Vocês são minha maior motivação para sempre seguir em frente. Muito obrigado por tudo que vocês fazem por mim.

Meus sinceros agradecimentos ao meu orientador Juergen Rochole por todos os valorosos ensinamentos que me foram dados ao longo deste curso. Agradeço também ao meu coorientador Cristiano Bonato Both pelas oportunidades e confiança a mim creditadas durante o desenvolvimento deste trabalho. Estendo os meus agradecimentos a todos os professores da Universidade Federal do Rio Grande do Sul, os quais me proporcionaram a indescritível oportunidade de me tornar um profissional e ser humano melhor.

Agradeço ao Grupo de Redes de Computadores da UFRGS e a todos os amigos parêntes que compartilharam comigo suas angústias e alegrias ao longo destes anos. Um obrigado muito especial ao pessoal dos Lab-210 e Lab-212. Foi nesses dois laboratórios onde gastei grande parte do meu tempo participando de discussões, conversas e debates que me mantiveram em contato, tanto com a pesquisa, quanto com os valores humanos que desejo levar para minha vida.

Olhando para o passado, sou muito grato a todas as pessoas que desempenharam alguma influência em minhas decisões. Sem eles, o presente não seria possível. Durante toda a minha trajetória de vida, tenho tido o privilégio de conviver com pessoas incríveis, muitas das quais levo como amigos para a vida toda. Agradeço, especialmente, aos membros do Bloco Tipo Àtoa e a Galera do Esquenta por todos os momentos de risadas e cumplicidade. Mesmo longe, vocês estiveram comigo em todos momentos dessa jornada.

Muito Obrigado!!

"Any sufficiently advanced technology is indistinguishable from magic."

— Arthur C. Clarke

"The best way to predict the future is to invent it."

— Alan Kay.

ABSTRACT

Up until 2020 it is expected that cellular networks must raise the coverage area in 10-fold, support a 100-fold more user equipments, and increase the data rate capacity by a 1000-fold in comparison with current cellular networks. The dense deployment of small cells is considered a promising solution to reach such aggressive improvements, once it moves the antennas closer to the users, achieving higher data rates due to the signal quality at short distances. However, operating a massive number of antennas can significantly increase the energy consumption of the network infrastructure. Furthermore, the large insertion of new radios brings greater spectral interference between the cells. In this scenery, the optimal management of radio resources turn an exaction due to the impact on the quality of service provided to the users. For example, low transmission powers can leave users without connection, while high transmission powers can contribute to inter radios interference. Furthermore, the interference can be raised on the unplanned reuse of the radio resources, resulting in low data transmission per radio resource, as the under-reuse of radio resources limits the overall data transmission capacity. A solution to control the transmission power, assign the spectral radio resources, and ensure the service to the users is essential. In this thesis, we propose an Adaptive Monte Carlo algorithm to perform global energy efficient resource allocation for Heterogeneous Cloud Radio Access Network (H-CRAN) architectures, which are forecast as future fifth-generation (5G) networks. We argue that our global proposal offers an efficient solution to the resource allocation for both high and low density scenarios. Our contributions are threefold: (*i*) the proposal of a global approach to the radio resource assignment problem in H-CRAN architecture, whose stochastic character ensures an overall solution space sampling; (*ii*) a critical comparison between our global solution and a local model; (*iii*) the demonstration that, for high density scenarios, Energy Efficiency is not a well suited metric for efficient allocation, considering data rate capacity, fairness, and served users. Moreover, we compare our proposal against three state-of-the-art resource allocation algorithms for 5G networks.

Keywords: Radio resource allocation. Cellular network. Adaptative monte carlo. 5G. Energy efficiency.

Algoritmo de Monte Carlo Adaptativo para Otimização dos Recursos de Rádio em H-CRAN

RESUMO

Até 2020 espera-se que as redes celulares aumentam em dez vezes a área de cobertura, suporte cem vezes mais equipamentos de usuários e eleve a capacidade da taxa de dados em mil vezes, comparada as redes celulares atuais. A densa implantação de pequenas células é considerada uma solução promissora para alcançar essas melhorias, uma vez que aproximar as antenas dos usuários proporciona maiores taxas de dados, devido à qualidade do sinal em curtas distâncias. No entanto, operar um grande número de antenas pode aumentar significativamente o consumo de energia da infraestrutura de rede. Além disso, a grande inserção de novos rádios pode ocasionar maior interferência espectral entre as células. Nesse cenário, a gestão dos recursos de rádio é essencial devido ao impacto na qualidade do serviço prestado aos usuários. Por exemplo, baixas potências de transmissão podem deixar usuários sem conexão, enquanto altas potências elevam a possibilidade de ocorrência de interferência. Além disso, a reutilização não planejada dos recursos de rádio causa a ocorrência de interferência, resultando em baixa capacidade de transmissão, enquanto a subutilização de recursos limita a capacidade total de transmissão de dados. Uma solução para controlar a potência de transmissão, atribuir os recursos de rádio e garantir o serviço aos usuários é essencial. Nesta dissertação, é proposto um algoritmo adaptativo de Monte Carlo para realizar alocação global de recursos de forma eficiente em termos de energia, para arquiteturas *Heterogeneous Cloud Radio Access Network* (H-CRAN), projetadas como futuras redes de quinta geração (5G). Uma solução eficiente para a alocação de recursos em cenários de alta e baixa densidade é proposta. Nossas contribuições são triplas: (*i*) proposta de uma abordagem global para o problema de atribuição de recursos de rádio na arquitetura H-CRAN, cujo caráter estocástico garante uma amostragem geral de espaço de solução; (*ii*) uma comparação crítica entre nossa solução global e um modelo local; (*iii*) a demonstração de que, para cenários de alta densidade, a Eficiência Energética não é uma medida adequada para alocação eficiente, considerando a capacidade de transmissão, justiça e total de usuários atendidos. Além disso, a proposta é comparada em relação a três algoritmos de alocação de recursos de última geração para redes 5G.

Palavras-chave: Alocação de Recursos de Rádio; Redes Celulares; Monte Carlo Adaptaivo; 5G; Eficiência energética.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

1G	*First-Generation*
2G	*Second-Generation*
3G	*Third-Generation*
3GPP	*3rd Generation Partnership Project*
4G	*Fourth-Generation*
5G	*Fifth-Generation*
AMC	*Adaptive Monte Carlo*
BBU	*Base Band Unit*
BPS	*Bits Per Second*
BS	*Base Station*
CAPEX	*CAPital Expenditure*
CDMA	*Code Division Multiple Access*
CoMP	*Coordinated Multi-Point*
C-RAN	*Cloud Radio Access Network*
EE	*Energy Efficiency*
FDD	*Frequency Division Duplex*
FDM	*Frequency-Division Multiplexing*
FDMA	*Frequency Division Multiple Access*
GHz	*Gigahertz*
HARQ	*Hybrid Automatic Repeat Request*
H-CRAN	*Heterogeneous C-RAN*
HD	*High Density*

HetNet	*Heterogeneous Network*
HPN	*High Power Nodes*
Hz	*Hertz*
LD	*Low Density*
LO	*Locally Optimal*
LPN	*Low Power Nodes*
LTE	*Long-Term Evolution*
MCS	*Monte Carlo Steps*
MD	*Moderate Density*
MHz	*Megahertz*
MIMO	*Multiple Input Multiple Output*
MINLP	*Mixed-Integer Non-Linear Programming*
ms	*Millisecond*
NP-hard	*Non-Deterministic Polynomial-time hard*
OFDMA	*Orthogonal Frequency-Division Multiple Access*
OPEX	*Operating Expenditure*
QAM	*Quadrature Amplitude Modulation*
QPSK	*Quaternary Phase Shift Keying*
RB	*Resource Blocks*
RE	*Resource Element*
RF	*Radio Frequencies*
RRC	*Radio Resource Control*
RRH	*Remote Radio Heads*
SB	*Scheduling Block*
SE	*Spectral Efficiency*
SINR	*Signal-to-Interference-Plus-Noise Ratio*
SNR	*Signal-to-Noise Ratio*
TDMA	*Time Division Multiple Access*
UE	*User's Equipment*

CONTENTS

1 INTRODUCTION

Projections on cellular network service requirements foretell a 10-fold expansion in the coverage area, a 100-fold increase in User's Equipment (UEs), as well as a 1000-fold rise in data rate transmission capacity by 2020 in comparison with current cellular networks specifications (ERICSSON, 2013)(YANG et al., 2015). Current network architectures are unable to meet the preconized requirements. Luckily, the dense deployment of small cells is reckoned as a promising solution to reach such stringent improvements since it moves the antennas closer to UEs, thus achieving higher data rates by taking profit of better signal quality at short distances (PENG et al., 2015). However, operating a massive number of antennas can significantly increase the energy consumption of the network (SHI; ZHANG; LETAIEF, 2014). Moreover, insertion of a large number of new radios means a buildup on spectral interference between the cells, potentially challenging the gain in spectral capacity (TANG et al., 2014).

Fortunately, Energy Efficiency (EE), *i.e.* the relation between spectral capacity and energy consumption can be managed through a radio resource allocation algorithm that controls transmission power and distribution of spectral Resource Blocks (RBs), *e.g.*, in a Generalized Frequency Division Multiple Access Allocation Scheme (XIONG et al., 2011). Nonetheless, granting service to UEs through the optimal management of radio resources is a challenging task. For instance, while low transmission powers may preclude UE's connection, high transmission powers may increase interference, thus hindering Spectral Efficiency (SE), *i.e.* the ratio between transmission capacity and channel bandwidth (PENG et al., 2014). Similarly, unplanned reuse of RBs may also intensify interference, resulting in low data transmission per RB, while under-reuse of RBs will limit the overall data transmission capacity. These competing effects are at the core of the high complexity of the management of radio resources, and motivate the design of a solution that encompasses control of transmission power, assignment of the spectral RBs, while ensuring service to the UEs.

The management of radio resources has been investigated in the context of Orthogonal Frequency-Division Multiple Access (OFDMA) based systems for traditional and scattered networks (XIONG et al., 2011)(WANG; ZHANG; ZHANG, 2014). Recently, this problem has attracted renewed attention within the community in connection with the trends on cloud-based centralized cellular architectures, *i.e.* Cloud Radio Access Network (C-RAN) and Heterogeneous C-RAN (H-CRAN). These architectures adopt Remote Radio Heads (RRH) that digitize and forward the signal samples to be processed at a centralized Base Band Unit (BBU) pool (PENG et al., 2015). RRHs are simple antennas with a lower cost of deployment and operation when compared to traditional base stations. In practice, several RRHs acting as Low Power Nodes (LPNs) will be deployed in the coverage area of High Power Nodes (HPNs) to increase the network bit rate in hot spot areas, *i.e.* holding a large number of UEs and high traffic demand (SHI; ZHANG; LETAIEF, 2014). Fortunately, the coupling of RRHs to a centralized BBU enables the use of cloud computing concepts to centralize decisions and achieve global optimization of resources (MAROTTA et al., 2015).

Despite the advantages offered by RRH malleability, studies indicate that the densification of RRHs can significantly increase the energy consumption of the network (SHI; ZHANG; LETAIEF, 2014). Besides linearly scaling with the number of RRHs, energy consumption also depends on energy demands of a high-performance optical network connecting RRHs to remote BBUs (PENG et al., 2015). Reducing energy consumption is possible by dynamically decreasing the RRHs transmission power through global cloud control solutions (SHI; ZHANG; LETAIEF, 2014). Nevertheless, energy consumption alone cannot be taken as a metric for network efficiency. Indeed, limiting the transmission power reduces Signal-to-Noise Ratio (SNR), possibly restraining the data transmission capacity, thus demanding more spectral RBs to meet a target UE demand (SHI; ZHANG; LETAIEF, 2014)(DAHROUJ et al., 2015). In this case, the network EE is pointed as an important metric to evaluate the system performance, balancing the energy consumption and ensuring service to the UEs.

Currently, recent research found in the literature present partial solutions to the radio resource allocation problem for C-RAN and H-CRAN architectures (PENG et al., 2015)(SHI; ZHANG; LETAIEF, 2014). Shi, Zhang, and Letaief (SHI; ZHANG; LETAIEF, 2014) presented a global optimization model for C-RAN aiming to reduce the energy consumption of the network through a joint selection of active cell and minimization of the transmit power. Peng et al. (PENG et al., 2015) investigated the H-CRAN architecture problem aiming at a single cell (*i.e.* local) EE optimization through iterative RBs distribution with minimal power assignment. However, the proposed solutions have not yet fully explored the new control possibilities crucially on global network knowledge allowed for those architectures.

The above solutions represent an important step forward. Given resource competition between the cellular radios, EE landscape is highly complex. In view of this competition, we argue that the approaches in EE optimization to H-CRAN environments, the investigation of a global and heterogeneous solution is an exaction. Moreover, due to the heterogeneity of antennas and UEs relative spatial distribution, consideration of a fairness constraint in the optimization problem is necessary to avoid EE optimal solutions to favor UEs closest to antennas, being detrimental to the farthest UEs. Evidently, service is guaranteed by a UE demand constraint.

We propose a global approach to the radio resource allocation problem in a multi-cell H-CRAN model. As the main contribution of this thesis, we present an Adaptive Monte Carlo

(AMC) algorithm to the optimal radio resources allocation model. Stochastic approaches, such as AMC, provide an efficient way of sampling highly complex solution spaces, as is the case for the EE landscape in the resource allocation problem for H-CRAN. Moreover, we also demonstrate that for high density scenarios EE is not a well suited metric for efficient allocation, considering data rate capacity, fairness, and served users. We evaluate our solution in a scenario defined by the 3rd Generation Partnership Project (3GPP) for the latest definition of Long-Term Evolution (LTE-A) with small cell deployment (3GPP TR 36.872, 2013). We compare the AMC proposal against three state-of-the-art resource allocation algorithms. The results of our evaluation show that the global solution improves the efficiency of the radio access network when compared to a local model.

The remainder of this thesis is organized as follows. In Chapter 2, we provide a brief description of the central background concepts associated with wireless communication and cellular networks. In Chapter 3, we present a review of the most relevant research and related works for this thesis. In Chapter 4, we present the radio resource allocation problem formulation and our proposed AMC solution in detail. In Chapter 5, we present the evaluation of our proposed approach with the developed prototype. Finally, in Chapter 6, we conclude this thesis presenting final remarks and the perspective for future work.

2 BACKGROUND

In this chapter, a bibliographic review of wireless network and radio resources were presented. We focus on the OFDMA systems, currently implemented in 4G and considered as the base of the future generation. We also were given a review of new cloud-based network architectures projected for the 5G. The Section 2.1 presents the key concepts for wireless networks. The evolution and the current mobile radio resources are explored at the Section 2.2. The Section 2.3 review some new architectures for the next generation of cellular network.

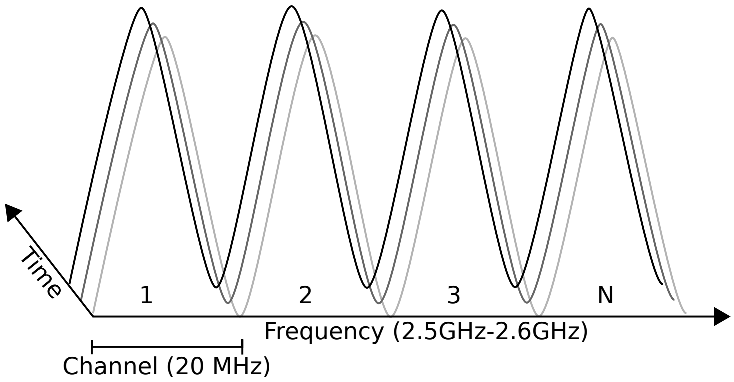
2.1 Key Concepts for Wireless Networks

In this section, we will review the main concepts of wireless transmission focusing on a cellular network and Radio Frequencies (RF) signals. First, we present a study of frequency division and frequency reuse. Afterward, we consider the essential about signal quality and channel transmission capacity.

2.1.1 Frequency Channel

A cellular network is composed of distributed and fixed cells that communicate with mobile devices through wireless links. A wireless link is created using different RF channels, which, in general, cellular networks are licensed by government agencies through predefined frequency ranges (TAROKH; SESHADRI; CALDERBANK, 1998). For each frequency range n different channels can be created, for example, at 2.5 Gigahertz (GHz) to 2.6 GHz, ten channels with 10 Megahertz (MHz) or five channel with 20 MHz can be used. The Figure 2.1 shown the concepts of frequency and the channel division example.

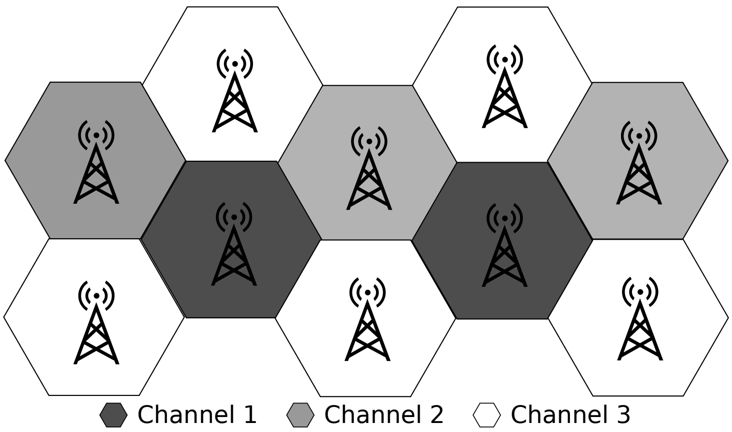
Figure 2.1: Radio Frequency and Wireless Channel.



Source: the author (2017)

Wireless channels are defined by using Frequency-Division Multiplexing (FDM) to provide multiple channels with a capacity of carrying different signals (PAHLAVAN; LEVESQUE, 1994). Mobile operators require multiple channels to deploy a cellular network due to the necessity of a channel per cell and the limited spatial frequency reuse. The reuse of the same channel for adjacent cell cause signal interference and poor quality of the cellular connections (TAROKH; SESHADRI; CALDERBANK, 1998). Fortunately, different channels can be used to avoid interference between adjacent cells (ROBERTS et al., 2006). The Figure 2.2 represents a traditional cellular deployment and channel reuse for a mobile operator.

Figure 2.2: Traditional cellular deployment and channel reuse.



Source: the author (2017)

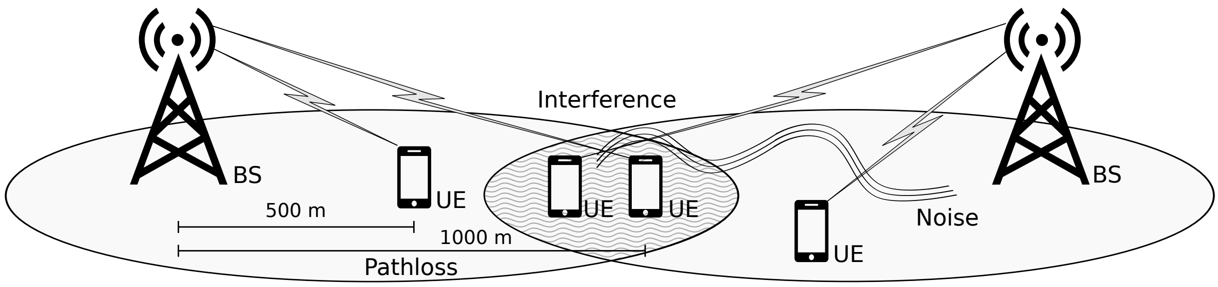
A cellular deployment is represented by hexagonal cells, and, in the example of Figure 2.2, just three channels are enough to assign no adjacent interference cells. The channel assignment and reuse have been stressfully researched for many years to control the interference, and well organize the cells (PAHLAVAN; LEVESQUE, 1994)(ROBERTS et al., 2006). However, the channel reuse can variate with the network cell deployment and adopted transmission technology (FRENCH, 1979). Furthermore, each deployed cell physically consists of a Base Band Unit (BBU) and a Base Station (BS)(LIU et al., 2016).

The BS are radio towers that transmit and receive data from the User Equipment (UE). The UEs are any mobile device that uses the cellular network to communicate, being represented primarily by smartphones, tablets, and laptops (LIU et al., 2016). Traditionally, BBUs are placed on the base of each BS to process and forward all BS-UE communication (3GPP TR 36.872, 2013). The sequence of this chapter presents the review of the most important aspects of a BS-UE communication, *i.e.* wireless signal quality, and transmission capacity.

2.1.2 Signal Quality and Transmission Capacity

The quality of the transited signal is essential for a mobile system transfer information over a wireless channel. The signal power used by the BS in the transmission determines the original signal quality (DAI; YU, 2016). As higher as the transmission power is, the higher is the quality and coverage area of a wireless signal. However, as higher as the transmission power is, the higher is the energy cost of a cellular connection (SHI; ZHANG; LETAIEF, 2014). Moreover, the wireless signal is sensitive to the degradation of the environmental signal caused by path loss, noise, and interference (3GPP TR 36.872, 2013). The Figure 2.3 exemplify cellular environment considering the referred signal degrades.

Figure 2.3: Environmental signal degradation caused by path loss, noise, and interference.



Source: the author (2017)

The path loss is caused by the attenuation of the signal during its propagation in the space between the transmitting BS and the UE receiver (LAFORTUNE; LECOURS, 1990). The space attenuation causes the loss of power in the signal, *i.e.* the power of the signal received by the UE is always less than the power initially transmitted. The signal losses are related to different factors, such as refraction, diffraction, reflection, and absorption of the environment (LAFORTUNE; LECOURS, 1990). However, the main factors of the signal loss are distance and transmission frequency (3GPP TR 36.814, 2010). Thus, simple attenuation models for path loss are defined considering the used channel frequency and the distance to be covered by the signal.

The signal captured by a UE is also corrupted by thermal noises and interference (TRIPATHI; REED, 2014). The thermal noise is an environment signal corruption caused by the sunlight, and artificial waves, such as sounds and electric discharges generated by electronic equipment. In general, noise is harmful to data transmission, but on a quiet and well-controlled scale (3GPP TR 36.872, 2013). Still, it is imperative to consider and recoup the noise corruption of the original signal when transmitting data on a wireless channel.

The sum of two (or more) signals transmitted by different sources at the same time, on the same frequency, and in the same geographical area causes the interference (TRIPATHI; REED, 2014). The interference corruption is dependent on the signal power and the distance of the interferer (3GPP TR 36.872, 2013). Thus, if two transmitters are located in the same geographical area and transmitting at high powers, both signals will be degraded. However, if the transmitters are separated enough and transmitting at low powers, the interference can be avoided and the same frequency used by both transmitters.

The transmission power less the path loss and the environment noise plus the interference characterize the wireless signal quality (3GPP TR 36.872, 2013). The Signal Noise Ratio (SNR) and the Signal-to-Interference-Plus-Noise Ratio (SINR) are signal quality indicators that consider the signal attenuation of the original transited signal power (BALACHANDRAN; KADABA; NANDA, 1999). The SNR and the SINR are used to determine the quality of the signal received by an UE *i.e.* the quality of the signal after applied all the environment losses. The SNR and the SINR are defined below:

$$\mathrm{SNR} = \frac{S}{N}, \qquad (2.1) \qquad\qquad \mathrm{SINR} = \frac{S}{I+N}. \qquad (2.2)$$

The SNR is defined as the ratio of the received signal S and the present environment noise N. The SINR is given by the received signal power S divided by the sum of the interference signals I and the noise N. For those two quality indicators, the path loss is previously applied at the transmitted signal $\mathrm{S} = P/L$, where P denotes the original transmission power and L the system path loss caused by the signal propagation.

The SNR and SINR are also used to estimate the transmission capacity of a wireless system. The maximum transmission capacity in bits is determined using the quality of the wireless signal and the bandwidth of the channel (ROCHOL, 2012). The Shannon theorem was the preview to establish the maximum transmission capacity of a communication link as (BENNETT et al., 2002):

$$C = B \log_2 \left(1 + \frac{S}{N}\right) \qquad (2.3)$$

The Shannon theorem determines the maximum transmission capacity of a channel through its bandwidth B and a certain SNR, or S/N (BENNETT et al., 2002). The expression provides the maximum capacity C of a channel without taking into account the transmission techniques and encoding schemes used for signal modulation (ROCHOL, 2012). It means that the transmission capacity of the Shannon theorem, defined in Equation 2.3, is the upper limit of the channel capacity C considering only physical aspects of a communication link.

2.2 Cellular Radio Resources

In this section, we will review the wireless networks focusing on the radio resources. First, we present the cellular evolution and main differences between generations. Afterward, we present the structural basis of the radio resources allocation problem for the current and next generation.

2.2.1 The Evolution of the Cellular Network

The First-Generation (1G) of cellular network utilize analogical signals that transmits voice without any handling. The frequency channels are created using Frequency Division Multiple Access (FDMA) with bands of 15 MHz partitioned into 600 channels of 30 kHz each one (ADACHI, 2001). The distribution of the radio resource in the 1G is limited to a UE-BS association and channel assignment. The UE defines the UE-BS association oriented to the closest BS, and the BS appoint a uplink channel and another downlink channel to each UE for each phone call (ROBERTS et al., 2006).

The Second-Generation (2G) was the first digital cellular systems, with enabled transmission of compressed voice and multiple UEs per channel. To behave more than one UE per channel, 2G adopted the use of Time Division Multiple Access (TDMA) (ADACHI, 2001). TDMA enables simultaneous multiple UEs transmissions at different times through the same 200 kHz channel. The 2G was also the first to introduce modulations schemes, such as the Quaternary Phase Shift Keying (QPSK)(ROBERTS et al., 2006). Meantime, due to the ossified radio resources and low UE requirements, 2G does not demand any sophisticated optimization in the distribution of the radio resources.

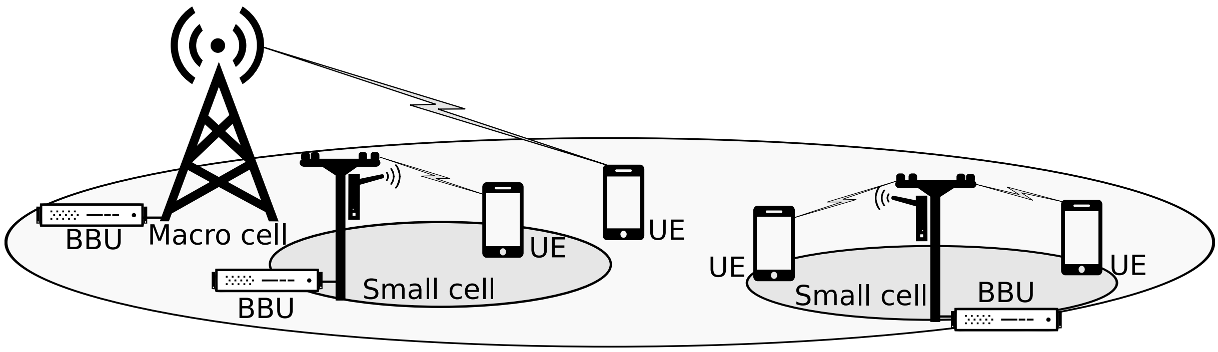
The Third-Generation (3G) was the first to require optimization solutions to manage the radio resources. The 3G establish the use of Code Division Multiple Access (CDMA), where multiple UEs can transmit at same time/frequency using different codes(ADACHI, 2001). 3G also introduced an adaptive modulation scheme that provided higher order modulations such as Quadrature Amplitude Modulation (QAM), *e.g.* 64-QAM, and, consequently each UE can assume different transmission capacities depending on its signal quality (ROBERTS et al., 2006). Thus, UEs with good signal quality can receive more Bits Per Second (bps) per Hertz (Hz) using high modulations, as far as UEs maintain simple services with low order modulation and weak transmissions powers requirements (QIAN et al., 2010).

The 3G also introduced a series of challenges for the control and management of radio resources (SAMPATH; KUMAR; HOLTZMAN, 1995). The transmission power and the modulation order turned a problem to be managed, due to the necessity to guarantee service for farm UEs, when UEs close to the BS has a overpower signal to explore all the channel capacity (QIAN et al., 2010). The 3G also introduced the necessity of an interference management, caused by nearby UEs using the same frequency channel to communicating different BSs (ROBERTS et al., 2006). Furthermore, 3G was the first to introduce the Hybrid Automatic Repeat Request (HARQ) protocol for all data transmissions (EKSTROM et al., 2006).

To overcome the presented challenges, 3G introduces for each UE a Radio Resource Control (RRC) to control the radio resource usage of the cellular system (LUO et al., 2003). Usually, an RRC for each UE is responsible for the resource allocation and release. The RRC starts selecting the closest BS with available frequency resources to communicate the UE. The RRC of each UE tries severals resources allocation as much as possible until the BS restricts it by unacceptably long delays, to behave the additional UE processing or frequency channels scarcity (QIAN et al., 2010). The RRC also execute resource release every time critical inactivity timers or UE energy high consumption are detected (LUO et al., 2003). Thus, the design of the RRC provides a dynamic resource allocation, controlled by each UE with statically configured parameters.

The current Fourth-Generation (4G) introduced Orthogonal Frequency-Division Multiple Access (ODFMA) (ROBERTS et al., 2006). The OFDMA provides a simultaneous UEs transmission by a distinct channel division into several independent sub channels (MORELLI; KUO; PUN, 2007). For each UE transmission, 29 different modulations combinations are possible depending on the transmission power and environment signal attenuation (3GPP TR 36.814, 2010). The 4G also adopts six different channels bandwidth which can be utilized independently for each BS(3GPP TR 36.814, 2010). Not enough, the 4G was the first generation to deploy a Heterogeneous Network (HetNet), that combine different cell types (*e.g.* Pico, Femto and Macro cells) to expand the network coverage and the spectrum reuse, as shown at Figure 2.4.

Figure 2.4: Traditional cellular networks architecture and main elements.



Source: the author (2017)

The architecture of the 4G cellular network is composed of the traditional BSs (*e.g.* macro cell and small cells), UEs and BBUs. The macro and small cells can joint be used to multiple transmit and receive data from the UE (3GPP TR 36.872, 2013). Small cells can be utilized to expand the network coverage and to provide high data rates at short distances with high signal quality. The resource allocation turns into a specific problem, which starts in new associations BS-UE possibilities, increases with the distribution of multiple subchannels of frequency, and ends up choosing the best modulation for each transmission (XIONG et al., 2011) (WANG; ZHANG; ZHANG, 2014). To well understand the magnitude of the problem and its elements, the next section details the radio resources used in the current cellular networks.

2.2.2 Cellular Radio Resources

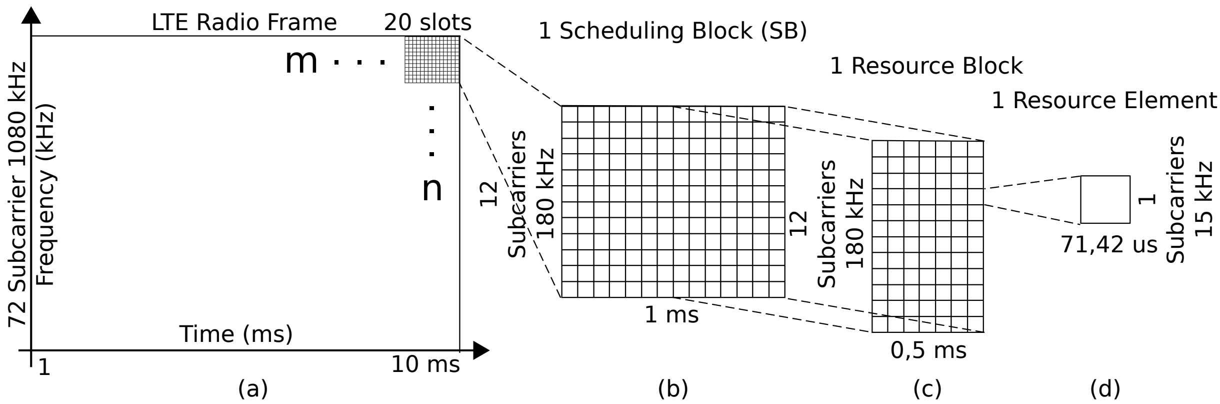
The current 4G cellular network, also called Long-Term Evolution (LTE), mainly employs Frequency Division Duplex (FDD), *i.e.* the subdivision of the spectrum occurs in the frequency domain limiting the channel bandwidth and enabling the transmission without interruption in the time domain. However, this can change depending on the standards adopted and the regulations of each country. Typically, the mobile operator in Brazil utilizes frequencies at 2.5 GHz and channels with 20 MHz. However, LTE supports different channels bandwidth, *e.g.* 1.4 MHz, 3 MHz, 5 MHz, 10 MHz, 15 MHz, and 20 MHz, from the smallest to the largest possible bandwidth in LTE. At 2.5 GHz, ‘N’ channels can be allocated considering the different channel’s bandwidth and, with FDD, all channels can be used for various mobile operators at the same time. As the transmission can co-occur, the frequency channel edges are reserved to avoid intra-channel interference. This means that for each channel bandwidth the useful part is limited to 1,08 MHz, 2,7 MHz, 4,5 MHz, 9 MHz, 13,5 MHz, and 18 MHz, respectively (NOGUEIRA; ROCHOL, 2016).

In each usable channel bandwidth, an analog signal is transmitted in its waveform format.

The waveform of the transmitted signal is dependent on the digital modulation scheme adopted to represent the information to be sent. For an LTE system, with OFDMA, the usable channel bandwidth is split into small sub-signals *i.e.* subcarriers, which are transmitted simultaneously at different frequencies. This allows an enormous amount of data to be transmitted simultaneously to the mobile receivers. However, this is only valid for downlink communication, since in LTE the uplink transmission occurs by Single-Carrier Frequency Division Multiple Access (SC-FDMA), which is not the focus of this thesis.

Considering the radio downlink LTE transmission using OFDMA, many subcarriers are transmitted simultaneously. The number of subcarriers that can be transmitted is dependent on the usable part of the channel bandwidth. For example, a channel of 1.4 MHz with a usable bandwidth of 1,08 MHz (or 1080 kHz) has a total of 72 subcarriers with 15 kHz each one. The joint transmission of the subcarriers during a period is called LTE Radio Frame. This frame can be seen as a two-dimensional matrix, where the horizontal axis represents the time domain, and the vertical axis represents the frequency domain. For a better understanding of the radio resources split in the time domain, the Figure 2.5 is divided into four parts.

Figure 2.5: Overview of the LTE radio resource in time/frequency domains.



Source: NOGUEIRA, 2016

The Figure 2.5(a) represents the previews described LTE frame with 10 millisecond (ms) per 1,080 kHz. In the Figure 2.5(b) this frame is subdivided into ten sub-frames of 1 ms per 180 kHz, *i.e.* 12 subcarriers, forming a Scheduling Block (SB). Subsequently, in the Figure 2.5(c), each SB is further subdivided into two slots of 0.5 ms per 180 kHz, also called Resource Block (RB). Finally, in the Figure 2.5(d), a Resource Element (RE) of 71.42 μs per 15 kHz, over a subcarrier, is shown.

The RE is the finest grain element in an OFDMA DL system, considering the spectral subdivision in time/frequency domains. For each RE a single value is modulated, *i.e.* the data to be transmitted are represented by a symbol, and psychically transmitted from the cell site to a mobile device. However, by the convention of LTE standardization REs from a RB can not be used for data transmission to two (or more) different receivers. This means that a RB is the smallest unit of resources that can be allocated to a mobile device. To review the set of spectral radio resources, the Table 2.1 summarize the LTE channel split for all supported LTE bandwidth.

Table 2.1: LTE Channel Bandwidth, Subcarriers and RBs

Channel Bandwidth	Usable Channel Bandwidth	Total of Subcarriers	Total RBs in 0,5 ms
1,4 MHz	1,08 MHz	72	6
3 MHz	2,7 MHz	180	15
5 MHz	4,5 MHz	300	25
10 MHz	9 MHz	600	50
15 MHz	13,5 MHz	900	75
20 MHz	18 MHz	1200	100

Source: NOGUEIRA, 2016

The Table 2.1 present the channel bandwidth, the useful part of each channel bandwidth, the total of subcarriers offers for each usable channel bandwidth, and finally the total of RBs available for each channel during 0,5 ms. With this information, we can conclude, for example, that a channel of 20 MHz can transmit a total of 2000 RBs during a second. These RBs are then distributed by the network cells to the network UEs, to meet user data demand. The total of RBs assignment/distribution will determine the downlink capacity of each user.

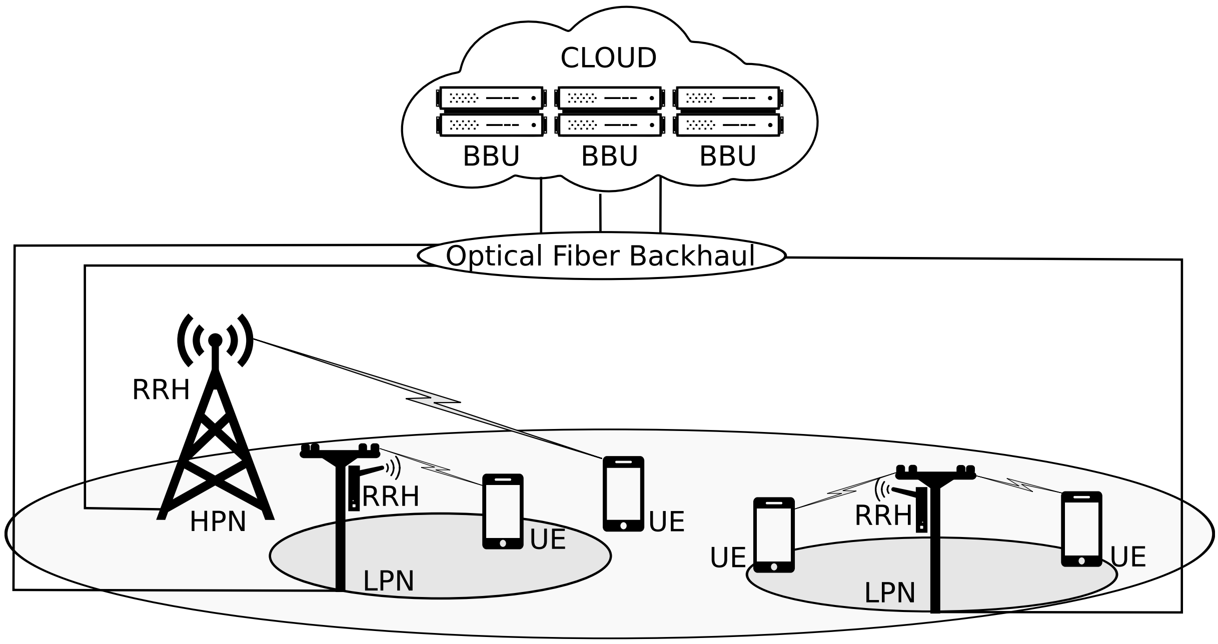
2.3 Next Generation Cellular Network (5G)

In this section, we present some newly proposed architectures for the next generation of the cellular network. Proposed architectures are motivated in the current cellular network issues, such as: (*i*) high deployment cost, (*ii*) high energy consumption, (*iii*) difficult and expensive upgrades, and (*iv*) restricted inter-cell collaboration. The high deployment cost, or the CAPital Expenditure (CAPEX), is originated by the expenditure to site planning and acquisition of BSs, RF, and BBU hardware. The energy consumption is part of the Operating Expenditure (OPEX), that represents the cost of maintaining a cell site after the deployment, and is usually high due to the uninterrupted service transmissions and the use of high processing capacity hardware that demands a considerable amount of power to operate. The difficult and expensive upgrade is due to the geospatial distribution of BBUs, which requires a human intervention for each network cell in case of failures or updates. Finally, the restricted inter-cell collaboration is given by the individual and uncoordinated operation of each cell, that limits the use of cooperative transmission such as Coordinated Multi-Point (CoMP) and Multiple Input Multiple Output (MIMO), as well as failure to deal with channel interference. To solve these traditional architecture issues, cloud-based architectures are the first bets for future 5G networks.

2.3.1 Cloud Radio Access Network (C-RAN)

C-RAN is one, and the primary, of the proposed architectures for the future 5G networks. The C-RAN architecture for 5G was initially introduced in (CHANCLOU et al., 2013) and subsequently described in detail in (DEMESTICHAS et al., 2013) and (WU et al., 2015). The C-RAN has potential to answer the mentioned issues. The central principle of C-RAN is the physical separation of BBU and BS. Where a centralized BBU pool with all BBUs for all BSs will be created to joint process the RF signal of the entire network. A representative of this new network architecture a show in the Figure 2.6.

Figure 2.6: Proposed C-RAN architecture for future 5G cellular network.



Source: the author (2017)

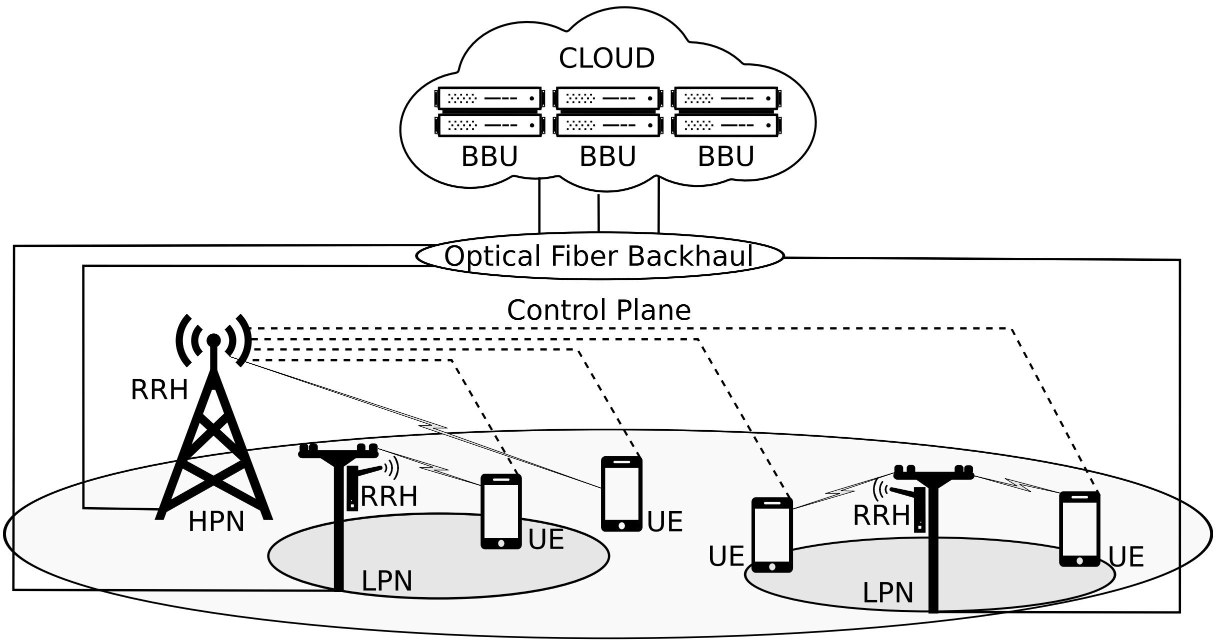
The C-RAN architecture inserts RRH, a BBU pool, and a powerful optical backhaul. The RRHs are simple radio units with provides a RF interface and power amplification for the analog signal. The BBU pool joins all BBUs in a single poll, like a remote data center, enabling the sharing of the processing hardware between different cells. The optical fiber backhaul is a hard constraint to interconnect the RRHs to the BBU pool through low latency requirements, enabling the remote processing of the signal. C-RAN also imports the familiar concepts of HetNets with the insertion of small cells (*i.e.* LPNs), within the coverage areas of traditional BS (*i.e.* HPN). The HPNs coverage long distance and guarantee signal to everywhere, while LPNs that have low coverage range and high data rate can meet high UE demands with high SNR at short distances.

LPNs was mainly designed to be deployed at hot spot zones with high UE demands, such as shopping malls and downtown cities. Dense scenarios are expected to be created, due to the low cost of these LPNs, in the future 5G architecture to absorb the growing demand of UEs. C-RAN has a potential to reduce CAPEX, through low cost RRHs and virtualized BBUs, and the OPEX, by dynamically decoupling BBUs and LPNs according to the presence of UEs in the network. As well, the remote control allows frequent and fast updates, and the centralization enables an inter-cell coordination.

2.3.2 Heterogeneous Cloud Radio Access Network (H-CRAN)

With the high acceptability of C-RAN architecture by academia and industry, a heterogeneous version of this architecture was proposed in (PENG et al., 2014). The H-CRAN made use of the new concept of cloud-based networks and introduced the decoupling of the data and control plan. With the separation of these planes, LPNs only have the data plane *i.e.* LPNs are used only to meet the data user requirements, while the overhead of the control plane (*e.g.* resource management) are unified at the HPNs. The reformulated architecture of an H-CRAN is discretized in the Figure 2.7.

Figure 2.7: Proposed H-CRAN architecture for future 5G cellular network.



Source: the author (2017)

With the centralized coordination of the resources by the HPNs, the LPNs can reuse the same RF resources of the HPNs with controlled interference. This mean, for example, that HPNs and LPNs can operate on the same channel, using different RBs, and the LPNs can re-utilized the same RB in various spatial zones. This new architecture and technologies expand the transmission data capacity expected for the 5G. However, the granularity of radio resources becomes very wide. The complexity in the distribution of the RBs overcomes many times the complexity of a simple allocation of channel or subcarriers. Optimizing the distribution of these radio resources becomes essential in exploring the main potentialities of future 5G cellular networks.

2.4 Chapter Summary

This chapter presents a review of RF and wireless transmissions, more specifically for the current 4G cellular network and the future 5G network. Initially, the LTE, channel, sub carrier, RB, RE are defined. The channel bandwidth, usable bandwidth, subcarriers, and total of RBs for LTE are also discussed. Also, the Shannon theorem is present as a theoretical transmission capacity. Secondly, a traditional cellular network is utilized as the baseline state to show the expectations for the 5G networks. The cloud-based systems concept and its benefits are presented. Further, the cellular heterogeneity is also inserted to the cloud concept. In the next chapter, the leading research efforts that aimed to improve the 5G cellular network are presented.

3 RELATED WORK

This chapter presents the leading researchers to improve the efficient use of the radio resources. Section 3.1 depicts the radio resource allocation enforced for C-RAN and H-CRAN architectures. Subsection 3.1.1 presents an overview of related works on RRH association with minimal power. Subsection 3.1.2 describes some related works on RBs distribution to EE maximization. In the end, Section 2.2 the summary of the related works are presented.

3.1 Radio Resource Allocation

Traditionally, the radio resources of the cellular network are allocated focusing on ubiquitous access and large data capacity exploring as maximum as possible the SE of the network. With the SE, the primary objective is to provide the maximum as possible of the data rate for the users. However, the energy saving becomes a global demand, changing the original focus of wireless researchers to an EE oriented. For EE researchers the primary objective is to minimize the energy waste at the network, without compromise the network service. For 2020 are expected a reduction of the power consumption on cellular networks to archive a 1000 times improvement in EE, as its data rate capacity needs to grow to support the UE demands (PENG et al., 2014).

Fortunately, the presented cloud-based networks offer key possibilities to expand the data capacity and reduce the energy consumption. The important gains on SE of the cloud-based architectures come from the low cost RRHs and massive insertion of new LPNs, and the centralized coordination to explore new transmission techniques, such as Massive-MIMO. The important gains on EE of the cloud-based networks come from the dynamic use of the infrastructure, where LPNs can be dynamically allocated to comport the current network traffic *i.e.* turned-off when network traffic is low. However, to manage the allocation of the radio resources, the UE-RRH association with the new transmission technologies and dynamically use LPNs turn hard, and complex problem to join arrange all of these variables.

3.1.1 RRH association

Considering as radio resources all the network equipment (*e.g.* UE, LPN, and HPN) and the transmission metrics (*e.g.* RB) the optimized use of some of this resources stars to be explored in C-RAN by (SHI; ZHANG; LETAIEF, 2014). The proposal of Shi *et al.* is minimized the network power consumption, including the cost of transport optical backhaul and the radio access network through a restricted UE service requirement. The authors formulate the problem to join considering the UE-RRH association and transmission power minimization per RB. As the UE-RRH association is found, a subset of active RRH, belonging to the entire RRH set, is modeled to indicated operational RRHs. The active RRH set is considerate to distinguish the cost of operational RRHs and standby RRHs, turning the previously ossified association in a dynamic association between UE-RRH.

Shi *et al.* also keeps a well studied OFDMA RB-UE assignment with minimal power assignment per RB. The OFDMA RB distribution has many previews studies, being some of the main (XIONG et al., 2011) and (WANG; ZHANG; ZHANG, 2014). The network energy consumption model, presented by Shi *et al.* consider the total power waste in the transmission plus the radio communication circuit and the energy cost of optical backhaul of each active RRH. The physical maximum transmission power, RB reuse interference, RRH maximum transmission power and UE minimal demand restriction are also modeled. The minimization of the modeled total power consumption with the several constraints turns a Mixed-Integer Non-Linear Programming (MINLP) problem, which is Non-Deterministic Polynomial-time hard (NP-hard) problem.

Due to the NP-hard complexity of the problem, the authors of (SHI; ZHANG; LETAIEF, 2014) focus on low complexity algorithms for practical implementations. The authors present two different greedy based selection of RRHs to solve the problem with yields optimal or near-optimal solutions. Although the authors present a detailed power model, the heterogeneous cells are not considerate, and the minimization energy model limits spectral exploration to the minimal UE request. The subsequent work focuses primarily on the throughput and the H-CRAN architecture.

Recently, the authors Po-Han Huang, Hsu Kao and Wanjium Liao investigate a similar problem for H-CRAN(HUANG; KAO; LIAO, 2016). In this work, the ineffectiveness of the traditional resource allocation model and UE-RRH association, denoted by the use of fractional frequency reuse for OFDMA systems is demonstrated. To solve this problem, the authors proposed the use of cooperation between HPNs and LPNs to realize the resource allocation. The authors then presented a cluster solution for cross-tier collaboration and proposed a hierarchical cooperation strategy to improve the system throughput in H-CRAN. The problem was modeled using linear programming and its approximate solution using a greedy algorithm due to the high complexity of the presented model.

The main contributions of the work were the identification and verification of the ineffectiveness of traditional methods in the allocation of radio resources from traditional networks to networks with cloud processing and the UE-RRH cooperative association model for H-CRANs. The work presents some conditions regarding objectives drawn for H-CRAN. One of the most important aspects disregarded by the work was the demands in EE expected for these networks. The authors focused on maximizing the transmission capacity of the network, neglecting the energy requirements. Also, the proposed solution to solve the complex model is extremely simple and its efficiency in the full allocation of resources, considering the association with the UE-RRH and the distribution of RB is not evaluated. Considering the limitations of the work presented for an association of UE-RRH in the process of allocation of radio resources, the next section presents work focused primarily on the distribution of RBs for heterogeneous networks.

3.1.2 RB distribution

Focusing on the association of UEs, antennas, and the distribution of RBs, Malandrino *et al.* proposed to solve the problem of assaying radio resources in HetNets (MALANDRINO et al., 2015). Its primary focus is to solve the problem of interference in the distribution of radio resources to maximize the SE of the heterogeneous networks, where the different (*i.e.* HPN and LPN) transmission paradigms can generate high levels of interference. In the distribution process of RBs, the authors considered the device-to-device communication, making the investigated problem even more complex. To solve the problem, the authors consider it in two parts. Initially, the UEs are associated with the antennas and then the RBs are distributed among the UEs connected to each antenna to minimize the interference between them.

The solution and model presented by the authors are evaluated considering large scale and realistic scenarios described by 3GPP for an LTE network. One of the main positive aspects of the solution submitted by the authors is the low complexity due to the division of the problem. In spite of being inefficient, the solution proposed by the authors disregards RAN's energy consumption, focusing solely on reducing resource interference. The work still presents a decoupled and ideal solution for HetNet, and not for H-CRAN networks, which have global and integrated control capability to orchestrate the distribution of radio resources, associating the RBs appropriately.

To take full advantage of the H-CRAN *i.e.* the decoupling of the data and control plane, Peng *et al.* study in (PENG et al., 2015) the EE in the RB distribution in a centralized and heterogeneous cells canary. The authors argue that in the H-CRAN architecture the radio resource allocation problem is distinct to the traditional problem due the significant difference in the transmission power between HPNs and LPNs and due to HPNs resources management control for the inter-RRH interference. This coordination enables the HPN and the several LPNs to share the same spectral channel, as well, the heterogeneity turns the UE-RRH association to the neighbor RRH not always efficient. To deal with these differences, the authors propose a joint optimization of RB assignment and power allocation to maximize the EE performances in an

OFDMA downlink system.

The authors consider the constraints at the physical RB-UE assignment, the RRH interference at RB reuse, the UEs minimum data rate and a different maximum transmission power model for HPNs and LPNs. The proposed model turns a mixed-integer programming problem with a fractional EE objective function, which turns a careful optimization to be solved with classical optimization methods. Peng *et al.* proposed a problem reformulation and a feasible interactive solution to enhance the algorithm to assign the RBs with power control. The authors present simulations results to demonstrate that the EE performance gains of the H-CRAN architecture over the C-RAN and the traditional HetNet. Peng *et al.* also proves that the proposed iterative solution is converged and that its EE performance increases over the baseline algorithms.

However, the proposed solution optimization was significantly affected by some simplifications assumptions, turning the solution on locally optimal *i.e.* the optimization of the radio resource allocation occurs individually for each LPN cell without a proper interference management. As well, due to the single cell optimization, the proposed solution assumes that the UE-RRH association has already occurred, and its connection is not optimized, and no one LPN is turned off in any rated scenarios. Meantime, the authors raise an important flag for the need of considering fairness in the radio resources allocation, since providing high data rates for nearby UEs is much more EE than serving distant UEs.

3.2 Discussion and Comparison of approaches

The works presented in this chapter expose several novelties for the literature related to the allocation of radio resources. Different models and strategies are described with several focuses (i.e., EE, SE, throughput and minimal power) for various scenarios (i.e., HetNets, C-RAN, H-RAN). However, the various proposals do not cover a solution capable of associating users and allocating RBs to H-RAN in the maintenance of EE requirements. Also, many proposals use distributed concepts, from the current generation of the cellular network (where centralized control is not feasible) to model and solve resource allocation problems. We argue that a global model capable of centrally allocating the resources of HPNs and LPNs simultaneously is necessary. Only through global solutions will it be possible to perform the necessary optimization of the radio resources of an H-RAN. As proof of this, a Table 3.1 summarizes and presents the main characteristics of each work.

Table 3.1: List of related works and main characteristics

Author	Objective	Radio Resource	User Association	Scenario
SHI	Min. power	RB	Nearest BS	C-RAN
HUANG	Throughput	RB	Dynamic	H-CRAN
MALANDRINO	SE	RB	Dynamic	HetNet
PENG	EE	RB	–	Single cell
SCHIMUNECK	EE	RB	Nearest BS	H-CRAN

Source: the author (2017)

In this work, we consider that a global model for allocating radio resources to an H-RAN is necessary. As the primary objective, we emphasize the need to maintain the EE energy of the network. Also, we believe that the allocation of RBs should be followed by the UE-RRH association for full consolidation of the use of radio resources. As observed in the table presented, these characteristics traced were not found in any work during the bibliographic review. Thus, the model of this problem becomes the first step to be achieved.

We also identified a clear lack of all proposed solutions and problem models. Discretized initially in Peng's work, fairness in the distribution of radio resources is fundamental and should be considered during the process of RB allocation. Equity in the distribution of radio resources is essential to achieving equal transmission capacity for the EU. Also, for a process of optimization of radio resources, the neglect of this restriction favors that the UEs that require less transmission power is favored about the UE that need more power due to the distance of the RRH. Solutions that objective EE or minimal power tend to serve specific users, improving their overall efficiency, but leaving UEs with poor quality because they are in places less advantageous to the objectives outlined. However, despite being identified, none of the articles published until the date of this thesis had proposed a model solution considering this restriction. Thus, in addition to a global model for allocation of radio resources, this thesis will consider fairness in the distribution of resources.

3.3 Chapter Summary

Throughout this chapter, the leading research to optimize the allocation of radio resources is presented. Initially, the research efforts of the UE-RRH association were presented, with a focus on methods to turn off some unused RRHs efficiently. Subsequently, the principal researchers the RBs distribution were studied in an OFDMA environment prioritizing the maximization of the EE. Even though the above solutions represent an important step forward, we argue that in EE optimization approaches to H-CRAN environments, the investigation of a global and heterogeneous solution is an exaction. Moreover, due to the heterogeneity of antennas and UEs relative spatial distribution, consideration of a fairness constraint in the optimization problem is necessary, to avoid EE optimal solutions to favor UEs closest to antennas, it being detrimental to the farthest UEs. Considering the discussions above, we described in the next chapter a detailed research problem and main open challenges of this thesis.

4 ADAPTIVE MONTE CARLO TO GLOBAL RADIO RESOURCES OPTIMIZATION IN H-CRAN

In this chapter, we present the system model, mathematical problem formulation and a stochastic algorithm for a global radio resource allocation in an H-CRAN. Section 4.1 presents the system model to address the EE of an H-CRAN, considering the transmission capacity in bits and the total power consumption in milliwatts. Section 4.2 describes the mathematical problem for EE maximization with the physical and system constraints. We propose the AMC algorithm in Section 4.3 to solve the radio resource allocation problem for EE under the limitations. In the end, Section 4.4 the summary of the proposal is presented.

4.1 System Model

In an OFDMA-based downlink system, with a total of K RBs per antenna, L LPNs and H HPNs attending U UEs, the network total data capacity can be written as:

$$C(a,p) = \sum_{t=1}^{L+H} \sum_{u=1}^{U} \sum_{k=1}^{K} a_{t,u,k} \cdot S_{sh}(p_{t,u,k}), \tag{1}$$

where $t \in \{1, ..., L+H\}$ denotes the LPN and HPN cell transmitters, $u \in \{1, ..., U\}$ corresponds to UEs, and $k \in \{1, ..., K\}$ denotes the respective RB in a given antenna (LPNs and HPNs). Matrices $a = [a_{t,u,k}]_{(L+H)\times U \times K}$ and $p = [p_{t,u,k}]_{(L+H)\times U\times K}$ represent a feasible RB and power allocation possibility, respectively. The element $a_{t,u,k} \in \{0, 1\}$, *i.e.* $a_{t,u,k} = 1$ indicates that the kth RB in the tth LPN or HPN is assigned to the uth UE, and the element $p_{t,u,k}$ denotes the respective transmission power. Whenever $a_{t,u,k} = 0$, for a given (t, u, k), *i.e.* kth RB belonging to tth antenna is not assigned to UE u, the corresponding power $p_{t,u,k} = 0$.

For each $a_{t,u,k} = 1$, $p_{t,u,k} > 0$, we have a transmission bit capacity $S_{sh}(p)$, as defined by Shannon's theorem (PENG et al., 2015):

$$S_{sh}(p_{t,u,k}) = B_0 \log_2 \left(1 + \frac{p_{t,u,k} P_{t,u}^{(pl)}}{I_{t,u,k} + N_0}\right), \tag{2}$$

where $I_{t,u,k}$ represents the power interference due to the assignment of the kth RB to other UEs

($u' \neq u$) by adjacent LPNs or HPNs ($t' \neq t$). $P_{t,u}^{(pl)}$ is the system path loss coming from the inverse square law decay of signal intensity with a distance between uth UE receptor and tth transmitter. Finally, B_0 and N_0 are fixed system bandwidth and noise, respectively.

Each cell type has a different total power consumption. Considering transmission, circuit, and front-haul power, the total power consumption of an LPN (HPN) is defined as $P_L(a,p)$ ($P_H(a,p)$) is defined as (PENG et al., 2015):

$$P_L(a,p) = \sum_{l=1}^{L} \left(\varphi_{\text{eff}}^{L} \sum_{u,k} a_{l,u,k}\, p_{l,u,k} + P_{\text{c}}^{L} + P_{bh}^{L} \right), \tag{3}$$

$$P_H(a,p) = \sum_{h=L+1}^{L+H} \left(\varphi_{\text{eff}}^{H} \sum_{u,k} a_{h,u,k}\, p_{h,u,k} + P_{\text{c}}^{\text{H}} + P_{bh}^{\text{H}} \right), \tag{4}$$

where $l \in \{1, ..., L\}$ denotes LPN cells while $h \in \{L+1, ..., L+H\}$ denotes HPN cells. The constants φ_{eff}^{L} (φ_{eff}^{H}), P_{c}^{L} (P_{c}^{H}), and P_{bh}^{L} (P_{bh}^{H}) denote, respectively, the power amplifier efficiency, circuit power, and power consumption of the front-haul link for LPNs (HPNs) (PENG et al., 2015).

The overall EE performance of a system is obtained by the sum of the network transmission capacity (1) divided by the total power consumption, equations (3) and (4). We represent EE by Γ^E, which for an H-CRAN can be written as:

$$\Gamma^E(a,p) = \frac{C(a,p)}{P_L(a,p) + P_H(a,p)}. \tag{5}$$

4.2 Mathematical Problem Formulation

An allocation problem to optimize the overall efficiency of the radio resources can be formulated with constraints on RB assignments, maximum transmission power per cell, the minimum required UE demand, and fairness among UEs data rate. RB assignment and maximum transmission power constraints are hard physical limitations enforced strictly. However, the minimum required UE demand and equity data rate are soft constraints, yet essential for service insurance. The latter are fundamental to guarantee service to all UEs avoiding that UEs closer to antennas is prioritized. The EE constrained maximization problem can be formulated as:

$$\max_{\{\mathbf{a},\mathbf{p}\}} \quad \Gamma^E(a,p) = \max_{\{\mathbf{a},\mathbf{p}\}} \frac{C(a,p)}{P_L(a,p) + P_H(a,p)}, \tag{6}$$

s.t.

$$\sum_u a_{t,u,k} \leq 1, \qquad a_{t,u,k} \in \{0,1\},\ \forall\, k,\ \forall t, \tag{7}$$

$$\sum_t a_{t,u,k} \leq 1, \qquad \forall\, u,\ \forall\, k, \tag{8}$$

$$\sum_{u,\,k} p_{l,u,k} \leq P^L_{\max}, \qquad 1 \leq l \leq L, \tag{9}$$

$$\sum_{u,\,k} p_{h,u,k} \leq P^{\mathrm{H}}_{\max}, \qquad L+1 \leq h \leq L+H, \tag{10}$$

$$\sum_k S_{sh}(p_{t,u,k}) \geq \eta_{\mathrm{u}}, \qquad \forall\, u,\ \ \forall\, t, \tag{11}$$

$$\frac{\left(\sum_{t,\,u,\,k} S_{sh}(p_{t,u,k})\right)^2}{U \cdot \sum_u \left(\sum_{t,\,k} S_{sh}(p_{t,u,k})\right)^2} \geq \eta_f. \tag{12}$$

Constraint (7) expresses the fact that each RB in a given LPN (HPN) can be assigned at most one UE. Constraint (8) limits assignment of the same RB to a given UE by at most one LPN (HPN). Constraints (9) and (10) limit to $P^L_{\max}$ (P^H_{max}), the maximum transmission power of LPNs (HPNs). Constraint (11) sets UE demand η_{u} as a lower limit to uth UE data rate. Lastly, constraint (12) denotes the fairness restriction for the spectral RBs distribution among the UEs considering the variance of the transmission capacity in bits/s, where $\eta_f = \frac{1}{U}$ being the worst case fairness, and $\eta_f = 1$, the best possible fairness.

Given the complexity of the model (6)-(12), we argue that finding the optimal RB assignment and power allocation, $\mathbf{a}^*$ and $\mathbf{p}^*$, constitute a non-convex optimization problem. Classical convex optimization methods are not well suited to explore the complex EE landscape. Stochastic methods, however, appear as a suitable approach to address the problem by sampling the landscape space of solutions.

4.3 Proposed stochastic algorithm – AMC

The AMC algorithm is an example of a stochastic approach to the radio resource allocation problem. To solve the previous radio resource allocation problem, we propose an AMC algorithm. This stochastic algorithm is suitable to address the problem due to its efficient way to cover the space of solutions with random sampling. The random sampling is an effective way of solving complex constrained optimization through the generation of an ensemble of initial conditions (sample solutions) randomly exploring the solution space (MAHENDRAN et al., 2012). The stochastic dynamics is defined to allow some rate of violation for the target function maximization.

Each sample is evaluated considering model (6)-(12) to find the best possible resource assignment. Algorithm 1 provides an overview of our AMC proposal, which we describe in the remaining of this section.

Algoritmo 1 Adaptive Monte Carlo

1: Set $I_{max}, B, \Lambda, \Upsilon$;
2: **for** $i = 0$ **to** I_{max} **do**
3:     **for** $S = 0$ **to** $K \cdot U$ **do**
4:         **if** $i == 0$ **then** $\beta_S = 1, \lambda_S = 1, \upsilon_S = 1$;
5:         **end if**
6:         **for** $j = 0$ **to** $K \cdot (H + L)$ **do**
7:             $t, u, k = \textit{rand}(S.antennas, S.ues, S.rbs)$;
8:             $B_s = \textit{store}(S)$;
9:             **if** $S.a_{t,:,k}$ **is** *assigned* **then**
10:                 $S.a_{t,:,k}$ = 0; $S.p_{t,:,k}$ = 0;
11:             **else**
12:                 $S.a_{t,u,k}$ = 1; $S.p_{t,u,k}$ = *minpower*(t, u, k);
13:             **end if**
14:             $\Gamma^{[E,D,F]} = W(S.a, S.p, B, \Lambda, \Upsilon)$;
15:             **if** $A(S, B_s, \beta_S, \lambda_S, \upsilon_S) \geq \textit{rand}()$ **then**
16:                 $\Gamma_S^{[E,D,F]} = \Gamma^{[E,D,F]}$;
17:             **else**
18:                 $S = \textit{restore}(B_s)$;
19:             **end if**
20:             *update*$(S, \beta_S, \lambda_S, \upsilon_S)$;
21:         **end for**
22:     **end for**
23:     Γ_i^E = *select*(S^*);
24: **end for**

Parameters $I_{max}, B, \Lambda, \Upsilon$ (line 1) represent, respectively, the number of Monte Carlo Steps (MCSs), AMC weights for EE, data rate, and fairness. The samples' EE is computed after I_{max} MCSs (lines 2-18). The ensemble of sample solutions S contains $K \cdot U$ initial conditions (line 3). For each sample solution S, we define adaptive weights $\beta_S, \lambda_S, \upsilon_S$ that tune the solution design priorities, *i.e.* EE, data rate, and fairness, all initialized to 1 (line 4). Each MCS is represented by a set of movements. We randomly select an antenna t, a random UE u in t's coverage area, and a random RB k (line 6), *i.e.* a triplet t, u, k. A candidate move is an attempt to change the RBs allocation of a sample (*e.g.* vacating it or assigning it to an UE u connected to the antenna t). We store B_s of the original state, to be able to restore it eventually (line 7).

Having randomly selected RB k belonging to antenna t, it will either be vacated with some probability (if initially assigned to some UE u - line 9), or assigned to an UE u' connected to t with some probability (if initially vacant - line 11), whence, constraints (7) and (8) are always satisfied. After assignment, the minimum power $p_{t,u',k}$ to communicate the UE u' with the antenna t through kth RB is obtained considering path loss, received interference, and maximum transmission power. The minimum power thus defined ensures constraints (9) and (10).

The optimal solution S^* must encompass optimal EE, user data rate constraint (11), and fairness constraint (12). These are enforced in a single function W, given as:

$$W(a, p, B, \Lambda, \Upsilon) = \\ B \cdot \Gamma^E(a, p) + \Lambda \cdot \Gamma^D(a, p) + \Upsilon \cdot \Gamma^F(a, p), \tag{13}$$

where B, Λ, and Υ are constant AMC weights, and

$$\Gamma^D = \min\left\{0, \sum_u \left(\sum_{t,\,k} a_{t,u,k} S_{sh}(p_{t,u,k}) - \eta_{\mathrm{u}}\right)\right\}, \tag{14}$$

$$\Gamma^F = \frac{\left(\sum_{t,\,u,\,k} S_{sh}(p_{t,u,k})\right)^2}{U\sum_u \left(\sum_{t,\,k} S_{sh} h(p_{t,u,k})\right)^2} - 1. \tag{15}$$

Function W (line 12) allows one to tune the priority given to each term through the relative size of AMC weights. Once UE data rate and fairness constraints are satisfied by a sample solution S, AMC dynamics is lead by EE maximization. The W definition is inspired by Lagrange Multipliers for constrained variational problems.

Functions Γ^E, Γ^D, and Γ^F provide quality indices of a sample state. Given a candidate move, its acceptance rate (line 13) is:

$$A(S, B_s, \beta_S, \lambda_S, \upsilon_S) \\ = \mathcal{P}_{\beta_S}(\Gamma^E_{B_S} \to \Gamma^E_S) \cdot \mathcal{P}_{\lambda_S}(\Gamma^D_{B_S} \to \Gamma^D_S) \cdot \mathcal{P}_{\upsilon_S}(\Gamma^F_{B_S} \to \Gamma^F_S), \tag{16}$$

where $\mathcal{P}_\xi(\Gamma_i \to \Gamma_f) = \min\{1, \exp(\xi\ (\Gamma_f - \Gamma_i))\}$. Hence, it is possible to accept a move which decreases EE, UE data rate relative to UE demand, and/or fairness with some finite probability, tuned by adaptive weights $\beta_S, \lambda_S, \upsilon_S$. Symbols Γ^E_S ($\Gamma^E_{B_s}$) are the candidate (stored) EE state of the sample solution, Γ^D_S ($\Gamma^D_{B_s}$) are the corresponding candidate (stored) neglected UE demand, and Γ^F_S ($\Gamma^F_{B_s}$) are the corresponding candidate (stored) fairness. Candidate sample solution S is accepted with probability A (line 14), and rejected with probability $1 - A$, in which case S is restored to B_S (line 16). Each MCS amounts to $K \cdot (H + L)$ move attempts.

The adaptive weights $\beta_S, \lambda_S, \upsilon_S$ are updated (line 17) between MCSs according to the current state of the sample solution. We compute the average values $\langle\Gamma^E\rangle$, $\langle\Gamma^D\rangle$, and $\langle\Gamma^F\rangle$ over the ensemble of sample solutions $\{S\}$. For each sample solution S, if $\Gamma^E_S < \langle\Gamma^E\rangle$ (respectively, $\Gamma^D_S < \langle\Gamma^D\rangle$, $\Gamma^F_S < \langle\Gamma^F\rangle$), the corresponding adaptive weight β_S (respectively, λ_S, υ_S) is increased by a fixed factor. Moreover, if Γ^E_S (respectively, Γ^D_S, Γ^F_S) relative fluctuations over a number of MCSs are small, then the respective weight is increased by a larger factor, than in the opposite situation. Conversely, if $\Gamma_S^E > \langle \Gamma^E \rangle$ (respectively, $\Gamma_S^D > \langle \Gamma^D \rangle$, $\Gamma_S^F > \langle \Gamma^F \rangle$), and the relative sample solution fluctuations are small, then the corresponding weight is decreased by a fixed factor. This method is similar to a Simulated Annealing method in Markov Chain Monte Carlo (HAARIO; SAKSMAN; TAMMINEN, 2001).

The AMC dynamics solution is chosen as the best among all sample solutions generated throughout the AMC history (line 18). The best sample will be the solution to the radio resource allocation problem for an H-CRAN architecture. The proposed AMC will find a solution in polynomial time.

4.4 Chapter Summary

This chapter presented the AMC algorithm, proposed in this master thesis to solve the radio resource allocation problem. The chapter began with the system model to evaluate the EE of an H-CRAN. In the sequence, a detailed description of the mathematical problem for EE maximization with the physical and soft constraint was presented. We finalize this chapter with a detailed overview of the AMC algorithm, the Monte Carlo Steps, the Monte Carlo Weights and the adaptive mechanism. In the next chapter, we prototype the proposed algorithm and evaluates it under the 3GPP specifications for a realistic scale scenario.

5 EVALUATION

In this chapter, the evaluation of the model and the proposed solution are carried out. In Section 5.1 we describe a realistic scenario for simulations and tests methodology. In Section 5.2 numerical results are presented, and then, in Section 5.3, the obtained results are analyzed and discussed. For the discussion of the results, we identify three different densities of UE, which will be presented in Subsections 5.3.1, 5.3.2, and 5.3.3. Finally, Section 5.4 summarizes the results and discussion of this chapter.

5.1 H-CRAN Scenery defined by 3GPP

We simulate an H-CRAN scenario assuming 3GPP technical specifications for LTE-A radio access networks with small cell deployment (3GPP TR 36.872, 2013). A dense scenario must contain at least 7 HPNs, each one with three sectors. HPNs are uniformly distributed respecting a minimum distance of 500 meters between each pair of HPNs. A minimum of two Hot Spots ($H_s = 2$) must be deployed inside the HPN coverage area, at a minimum distance of 90 meters from the center of the corresponding HPN. A Hot Spot can be defined as a densely populated area, *e.g.* a city downtown or a shopping mall. Each Hot Spot has a set of four LPNs randomly deployed with a minimum distance of 10 meters between each pair of LPNs. Moreover, the UEs located in a Hot Spot area are randomly placed with minimum distances of 30 meters to the HPN and 5 meters to the nearest LPN. Two-thirds of UEs are forcibly placed within the Hot Spot area, the remaining UEs being randomly placed in the HPN coverage area. We validate the simulation considering the obtained geographical dispersion in relation to the proposed scenario by the LTE-A recommendation. A typical scenario following the 3GPP specification is exemplified in Fig. 1.

We also define minimal power and interference models. We compute the interference caused by the sum of the adjacent cell powers at the kth RB assigned to UE u as:

$$I_{t,u,k} = \sum_{u' \neq u} \sum_{t' \neq t} a_{t',u',k} (p_{t',u',k} P^{(pl)}_{t',u}).$$

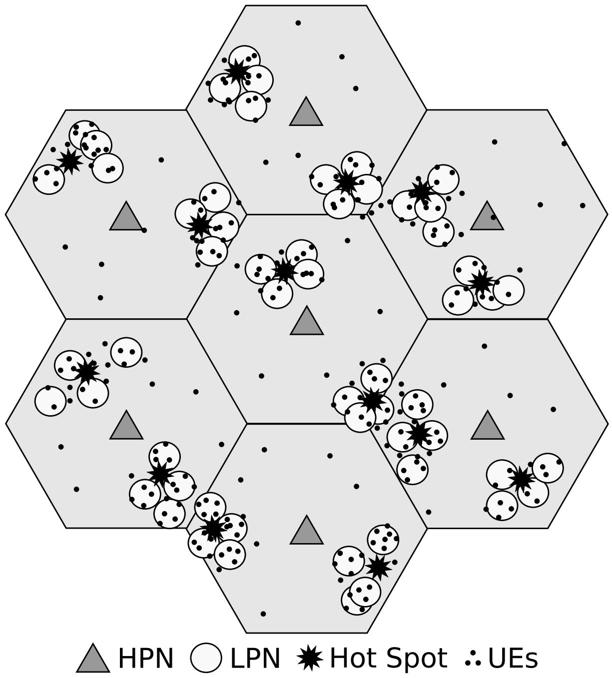


Figure 5.1: H-CRAN scenario with 7 HPNs, 2 Hot Spots per HPN, 4 LPNs per Hot Spot and 30 UEs per Hot Spot.

Source: the author (2017)

We assume the following definition for the minimal power of LPNs (HPNs):

$$p_{t,u,k} = \min\{R_t, (I_{t,u,k} + N_0) \cdot T_{sinr} + P^{(pl)}_{t,u}\},$$

where

$$R_t = \begin{cases} P^{L}_{max} - \sum_{u,\,k} p_{t,u,k}, & \text{if } 1 \leq t \leq L, \\ P^{H}_{max} - \sum_{u,\,k} p_{t,u,k}, & \text{if } L+1 \leq t \leq L+H. \end{cases}$$

T_{sinr} is a fixed target SINR adopted for all system transmissions. The path loss $P^{(pl)}_{t,u}$ is given by the 3GPP TR 36.814 definition (for more details see (3GPP TR 36.814, 2010)). Finally, our UEs are stationary and the transmission is restricted to a single-input downlink system. We compare the performance of the proposed AMC radio resource allocation against a Locally Optimal (LO) algorithm, and two algorithms initially presented by Peng et al., namely Fixed Power and Sequential (PENG et al., 2015). The LO algorithm assigns RBs to UEs, aiming at local interference mitigation and minimal power assignment. By local interference, we mean that quality assessment of an allocation solution is done considering only received interference, which strongly hinders solution robustness. The Fixed Power algorithm equally distributes the transmission power among all K RBs of LPNs and HPNs. The Sequential algorithm relies on RBs assignment to UEs in a sequential manner with minimal power. The criteria for these choices are twofold: (i) to make a parallel between our global proposal and a local solution; (ii) considering commonly presented algorithms in the literature. The three algorithms are restricted to constraints (7)-(12) and execute the same number I_{max} of iterations as our proposal. Simulation parameters for an H-CRAN are enumerated in Table 5.1.

Table 5.1: Simulation parameters for an H-CRAN environment

Parameter	Symbol	Value
Total of HPN	H	7
Total of Hot spot	Hs	2
Total of LPN	L	$4 \cdot (H \times Hs)$
Total of UE	U	$[10, 100 : 10] \cdot (H \times Hs)$
Total of RB	K	100 (20 MHz, 0.5 ms)
RB bandwidth	B_0	180 Hz
Thermal noise	N_0	-110 dB
Transmission amplifiers of HPN	φ^{H}_{eff}	4 W
Transmission amplifiers of LPN	φ^{L}_{eff}	2 W
Maximum transmission powers of HPN	P^{H}_{max}	46 dB
Maximum transmission powers of LPN	P^{L}_{max}	23 dB
Circuit power of HPN	P^{H}_{c}	6.8 W
Circuit power of LPN	P^{L}_{c}	6.8 W
Front-haul consumption of HPN	P^{H}_{bh}	3.85 W
Front-haul consumption of LPN	P^{L}_{bh}	3.85 W
Target SINR	T_{sinr}	16.5

Source: the author (2017)

In this table, the simulation parameters, taken from the recommendation (3GPP TR 36.814, 2010) and related works (SHI; ZHANG; LETAIEF, 2014) are summarized. To the size of the scenario 7 HPNs are defined in the hexagonal form. For each of the 7 HPNs, 2 HotSpots, each one containing 4 LPNs, are overlapping. A range of 10 to a 100 UEs of density are utilized for a fixed total of 100 RBs relating to 0.5 ms in a 20 MHz channel. The fixed band of each RBs is set to 180 Hz. Thermal noise is also set to -110 dB for all transmissions. The power of the transmission amplifiers for HPN and LPN are respectively 4 and 2 W. The maximum transmission powers of HPN and LPN is 46 and 23 dB, respectively. The circuit power of each antenna is 6.8 W, while the cost of front-haul is 3.85 W for both LPNs and HPNs.The quality of the SINR, purposed for each transmission, is 16.5. Finally, Table 5.2 summarizes the parameters used to execute the algorithms.

Table 5.2: Algorithm parameters defined for the evaluation analysis

Parameter	Symbol	Value
Max of interactions	I_{max}	10
EE Weight B	1	
UE demand Weight	Λ	10
Fairness Weight	Υ	1
Objective UE demand	η_{u}	1 Mbit/s
Objective Fairness	η_f	1

Source: the author (2017)

Table 5.1 enumerate the algorithm parameters defined for the evaluation analysis. The execution parameters of the algorithm may vary according to the purpose and complexity of the scenario. For the evaluation of this thesis, it was used a total of 10 interactions, defined initially in (PENG et al., 2015). For Monte Carlos Weights, a high priority was used to meet the minimum user demand (10 times higher than the other objectives of EE and Fairness). The users' demand was fixed as 1 Mbit/s while the fairness function aimed at 100% equality in the distribution of resources. Considering the established parameters, the following section presents the numerical results obtained from the evaluation.

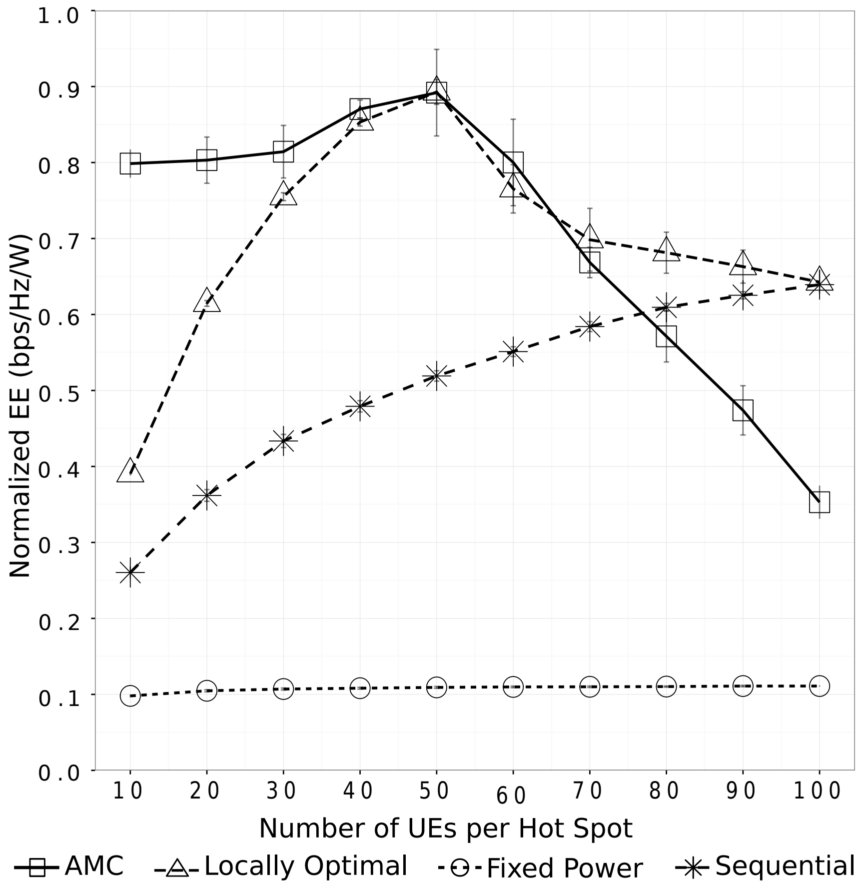
5.2 Obtained Numerical Results

We present numerical results considering system EE and SE, a fraction of served UEs (*i.e.* ratio of minimal data rate constraint fulfillment), and fairness in the resource distribution. We considered a fixed 7 HPNs with two hot spots and 4 LPNs per hot spot. We gradually increase the UE density from 10 to 100 UEs per Hot Spot. The figures below show the results of a total of 30 repetitions with 95% confidence intervals. For all evaluations the value of 1.0 represents the best possible results, as well, the value of 0.0 accounts for the worst case result.

In Fig. 5.2 we present comparative results regarding normalized EE as a function of the density of UEs. We observe a low, though stable, EE (0.1 Mbit/Hz/W) for the fixed power algorithm, confirming the importance of dynamic assignment of the transmission power in an H-CRAN. The Sequential algorithm presents a logarithmic increase of the EE with the UEs increase in density, obtaining the maximal EE of 0.6 Mbit/Hz/W with 100 UEs. The LO algorithm also presents a logarithmic evolution until achieving its maximal EE (0.88 Mbit/Hz/W) point with 50 UEs; above 50 UEs, this algorithm presents a logarithmic decrease in EE, until it achieves the same EE of the Sequential solution at 100 UEs. Our proposed AMC solution shows a higher EE for initial step (0.8 Mbit/Hz/W) and an exponential evolution until the same

50 UEs mark, with a maximal EE point upper to 0.9 Mbit/Hz/W. For more than 50 users, AMC follows the same logarithmic decrease in EE as the LO.

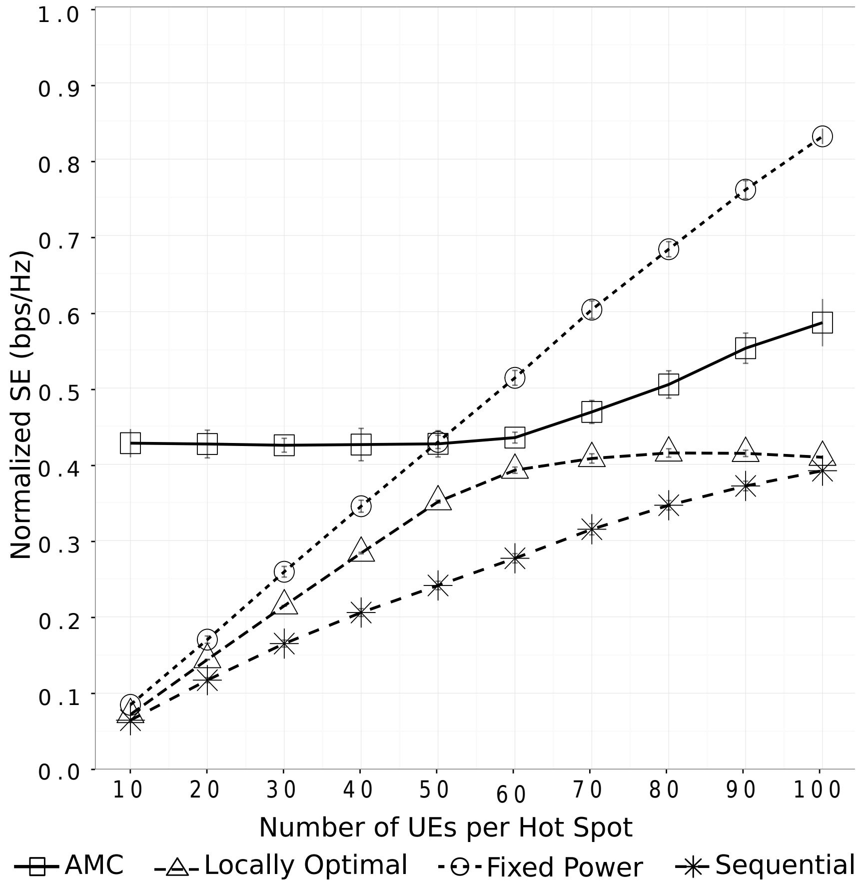
Figure 5.2: Numerical results in Energy Efficiency for the proposed algorithms in the evaluation process



As we can observe, AMC solution achieves the best EE results among the four compared models up to 60 UEs per Hot Spot. Beyond 50 UEs per Hot Spot, the increasing behavior of the normalized EE of the AMC proposal is replaced by a steep decreasing. For densities equal to 70 UEs per Hot Spot and higher, AMC proposal achieves lower standardized EE than the LO solution. For 80 UEs per Hot Spot and beyond, AMC proposal normalized EE also lies below normalized EE performed by the Sequential algorithm.

In Fig. 5.3 we show a similar comparison regarding normalized SE as a function of UE density. We can observe a linear growth of the SE for the Fixed Power solution, with a peak up to 0.8 bps/Hz for 100 UEs. In contrast, the Sequential and LO algorithms, which dynamically adapts the transmission power, achieved a SE of 0.4 Mbit/Hz in the best case, limited by the interference congestion of the scenario. However, each algorithm reached this maximum with different UE's density, *i.e.* 90 UEs for the Sequential and 60 UEs for the Locally Optimal. Exclusively, the AMC proposed solution reached a sequential and top SE for scenarios with 50 UEs or less density. For more than 60 UEs per hot spot, our AMC presents a linear grow of the SE to nearly 0.6 Mbit/Hz.

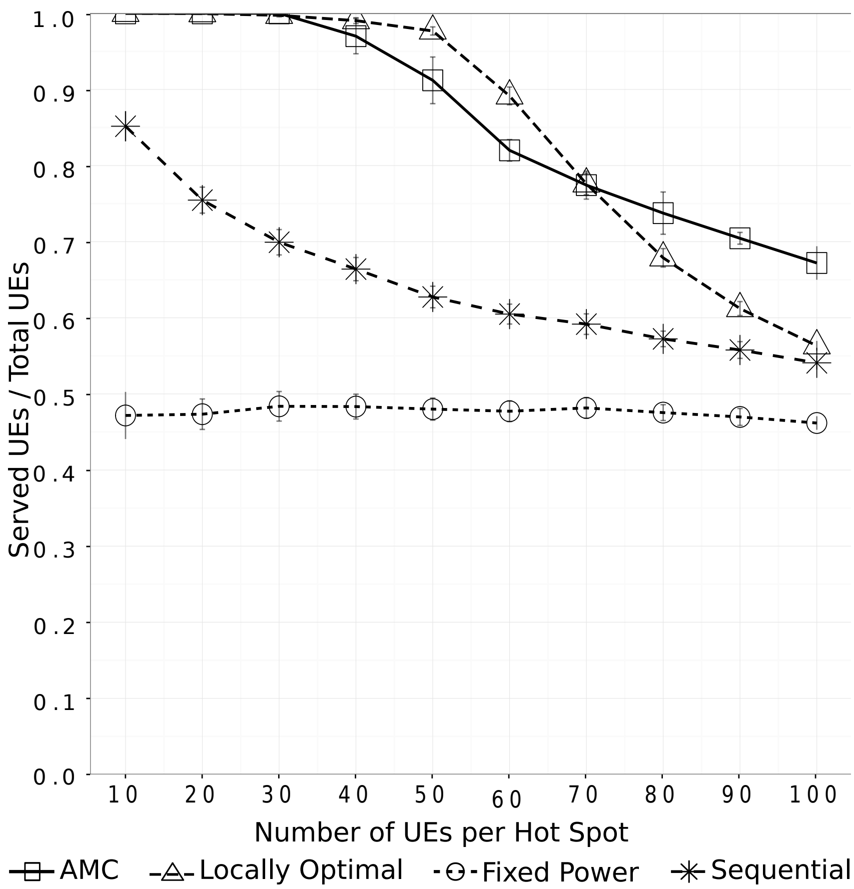
Figure 5.3: Numerical results in Spectral Efficiency for the proposed algorithms in the evaluation process



Again, at low densities (50 UEs per Hot Spot and below) our proposal achieves the highest normalized SE when compared to any of the other three models. At higher UE densities, AMC's normalized SE is surpassed by Fixed Power model, which achieves the highest normalized SE in this density regime. This behavior is explained when considering that Fixed Power model attains very low transmission power, as shown in Fig. 5.6 (a), resulting in a scenario with minimal interference. Above 50 UEs per Hot Spot, AMC's normalized SE attains the second best results among the compared models. We also observe that the normalized SE of the LO model saturates at a density of 60 UEs per Hot Spot.

Fig. 5.4 regards fraction of served UEs as a function of UE density. The Fixed Power algorithm presents the worst performance, satisfying approximately 50% of the UEs in all evaluated UE densities. The Sequential algorithm presents a reasonable solution (attending up to 85% of the UEs when the density is very low). However, the UEs satisfied decrease down to 60% when the density of UEs is high. The others two solutions, LO and AMC, satisfied all the UEs for scenarios with less than 40 UEs per hot spot. As the density increase, both algorithms are unable to assignment 100% of the UEs demand. Our proposed AMC solution met approximately 5% fewer UEs then the LO for a density up to 60 UEs per Hs. Meantime, the AMC solutions archive the highest ratio of four algorithmic to high dense scenarios *i.e.* more than 70 UEs.

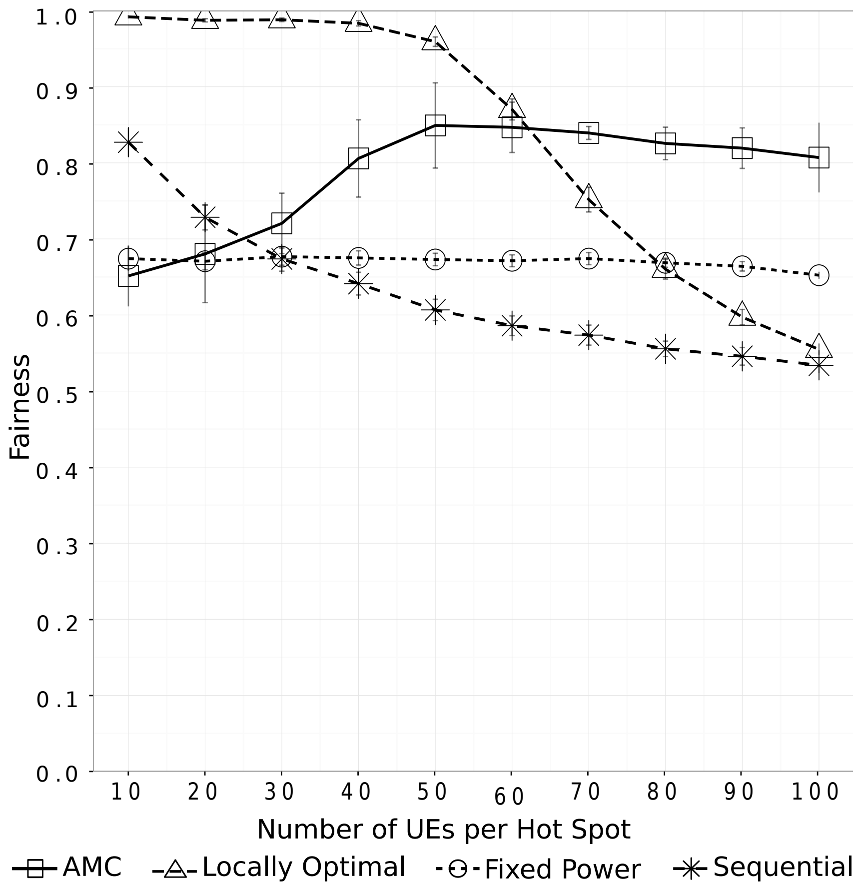
Figure 5.4: Numerical results in Ratio of served UEs for the proposed algorithms in the evaluation process



We can observe that at low densities (up to 40 UEs per Hot Spot), both AMC and LO can attend all UEs. For increasing densities, a fraction of served UEs decreases for both algorithms, though decrease for the LO is steeper. At very high densities (for 70 UEs per Hot Spot and beyond), a fraction of served UEs attained by AMC is the highest among the four models

In the last analysis, presented in Fig. 5.5, we present fairness results. Considering the Fixed Power solution, we observe the same comportment as we observe in fist EE analysis *i.e.* a stable fairness around 0.7. The Sequential solution presents a higher fairness when the UE density is low; however, the fairness decreases as the number of UEs increase. The Sequential solution presents a decrease in fairness, starting up to 0.8 for a low number of UEs and close to 0.5 for a UEs density of 100. The same behavior occurs in the LO algorithm presents, thought the fundamental fairness is close to 1.0 when the UE density is low. Our proposed AMC present an initial equity of 0.7 for low densities, which increases until it surpasses all other algorithms when the network density pass of 60 UEs per hot spot.

Figure 5.5: Numerical results in Fairness for the proposed algorithms in the evaluation process

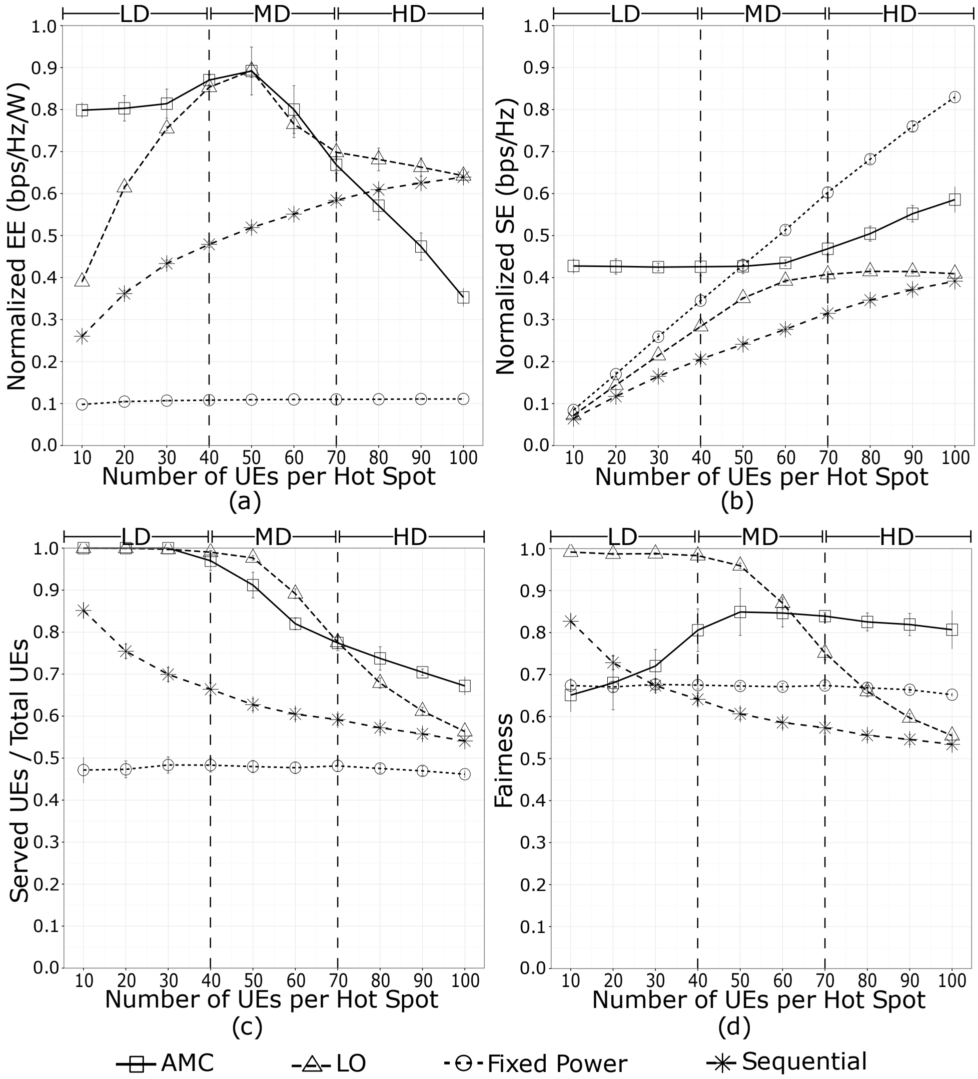


In this analysis, we see that at low UE densities, *i.e.* where available resources are abundant, AMC solution tends to an unequal resource assignment among UEs. In this density regime, LO model is the fairest among the compared solutions. Beyond 60 UEs per Hot Spot, however, AMC solution becomes the most appropriate among solutions, and remains almost constant throughout the high-density regime, while LO steeply decreases attaining a baseline comparison with the worst performing algorithm, considering fairness, *i.e.* Sequential algorithm. In the next section, we will present the evaluation analysis and result in the discussion of this thesis.

5.3 Evaluation Analysis and Discussions

In the following we present a correlated analysis, considering all aspects of the allocation solution, namely, EE, SE, a fraction of served UEs, and fairness. We identify three distinct regimes, namely, Low Density (LD), Moderate Density (MD), and High Density (HD), based on solution characteristics, as highlighted on top of Fig. 5.6.

Figure 5.6: Correlated analysis considering all aspects of the resource allocation problem



5.3.1 Low Density regime

Corresponds to 40 UEs per Hot Spot or less. Sequential and Fixed Power algorithms perform poorly in this regime. We show that the EE of our proposal (Fig. 5.6 (a)) is highest among all algorithms, LO's EE being the second best. Both algorithms can serve the total UE demand. Nonetheless, LO's solution is EE inefficient, due to its local knowledge of solution space. Fig. 5.6 (b) shows that LO's SE is only slightly above the worst performing algorithm's. Indeed, LO has only access to received interference, thus limits its solutions to the minimum data rate. Meanwhile, LO achieves the highest fairness (Fig. 5.6 (d)), but only because it is unable to explore data rates beyond the minimum.

In this abundant resource scenario, AMC achieves both highest EE and SE. This means that, on the one hand, given the abundance of resources in the LD regime, AMC can grant data rate beyond the minimum UE demand. On the contrary, given the global knowledge of the algorithm, it is also able to mitigate interference with intelligent resource allocation. This raises transmission capacity more than the corresponding energetic cost, whence achieving the highest EE. AMC's fairness (Fig. 5.6 (d)) degrades when the UE density is low, *i.e.* despite all UEs having their minimum demand attended (Fig. 5.6 (c)), some UEs receive more data than others.

5.3.2 Moderate Density regime

Between 40 and 70 UEs per Hot Spot, we reach a crossover system, which goes from resource abundant to scarce. Mitigating interference becomes difficult, transmission cost grows to attain the target SINR, hindering EE. AMC can find solutions with increasing SE, to supply UE demand and fairness, up to a point. Radio resources are not enough to service all UEs, so AMC solutions divide resources prioritizing fairness. On the other hand, LO, given its limited knowledge on interference, is unable to locally reconfigure its network allocation in this saturation regime, the interference due to RB reuse increases and its SE saturates. This means the number of UEs LO solutions can serve saturates as well, and a fraction of served UEs decreases steeply with increasing density.

5.3.3 High Density regime

Beyond 70 UEs per Hot Spot, the steep decrease in EE (Fig. 5.6 (a)) for the AMC solution is explained by a corresponding increase in SE (Fig. 5.6 (b)), with a resulting increase in energy cost. When the number of unattended UEs grows, the AMC solution sacrifices EE (including below Sequential's EE) to achieve higher SE, to serve more UEs with also higher fairness (Fig. 5.6 (c), (d)). Indeed, our solution provides the best fairness among all models. In contrast, the LO solution decreases equity and ratio of served UEs, due to its limited local knowledge. This means that solutions limited to local knowledge are unable to explore all the network reconfiguration capacity (limiting SE), as opposed to global solutions that can promote greater SE to the network.

High EE is easy to achieve, when considered alone, as can be seen from the fact that the Sequential algorithm can provide the same normalized EE as the LO. Had we considered EE only, we would have mistakenly inferred the AMC solution to provide a bad resource allocation strategy at high UE densities. However, this behavior is justified given our solution prioritizing served UEs and fairness, as forecast in 5G networks.

5.4 Chapter Summary

In this chapter, the prototype of the proposed AMC algorithm was developed and evaluated following 3GPP recommendations. The details of the scenarios of an H-CRAN were described, as well as the final signal propagation models required for the simulation. The parameters considered were summarized and described in detail in tables. Subsequently, the numerical results were described and analyzed. Finally, a discussion is presented considering the different levels of UE density. The arguments relevant to prove the efficiency of our proposal was introduced, indicating their superiority about the other solutions.

6 CONCLUSION AND FUTURE WORK

Despite the benefits of an H-CRAN, where the decoupling of BBU and RRH enable a dense deployment of LPNs with high SE, the energy consumption of such networks radically increase due to the massive insertion of new antennas. In fact, the control of the energy consumption of an H-CRAN is a challenging task. To tackle this problem, an optimized use of the radio resources can be a viable tool to reduce energy waste, minimizing the interference and the transmission power of HPNs and LPNs. The benefits of the optimized allocation of the radio resources are enormous. However, the problem complexity is also huge. To reduce the energy consumption of a network requires rigorous care not to degrade the quality of service provided to UEs. An unplanned energy optimization to minimal power can limit the overall transmission capacity of the network.

In this thesis, we proposed a radio resource allocation model for an H-RAN. We model this problem considering the EE maximization, under the UE service and fairness constraints. The service and fairness constraints are essential to guarantee minimal quality and equality to UEs. In this chapter, we will conclude this work, presenting the main contributions, obtained results, and future works.

6.1 Summary of Contributions

Our contributions of this master thesis are threefold: (*i*) the proposal of a global approach to the radio resource assignment problem in H-CRAN architecture, whose stochastic character ensures an overall solution space sampling; (*ii*) a critical comparison between our global solution and a local model; (*iii*) the demonstration that for high density scenarios only Energy Efficiency is not a well suited metric for efficient allocation, considering data rate capacity, fairness, and served users.

As the main contribution of this thesis, we present an AMC algorithm to perform global EE resource allocation for H-CRAN architectures, which are forecast as future 5G networks. We argue that our global proposal offers an efficient solution to the resource allocation for the 5G, due it stochastic characteristics which provides an effective way of sampling highly complex solution spaces, as is the case for the EE landscape in the resource allocation problem for H-

CRAN. Our stochastic AMC approach enables the global optimization of the radio resource by jointly assign of RBs for LPNs and HPNs in polynomial time.

As the second contribution, we compared our solution against three state-of-the-art resource allocation algorithms for 5G networks, following the guidelines defined by the 3GPP. The results of our evaluation show that the global solution improves the efficiency of the radio access network when compared to a local model. We demonstrated that our proposal achieves the highest EE when the UE density is low; whereas, for high-density networks, our solution achieves an EE similar or lower to LO algorithms. However, our solution can guarantee a higher fraction of served UEs with the highest fairness. We also proved that our algorithm provides the best SE due to overall knowledge and global control of the network resources.

Finally, as the third contribution, we demonstrate that for high density scenarios EE is not a well suited metric for efficient allocation. We showed that in highly dense scenarios, EE maximization does not provide the best allocation solution, considering a fraction of served UEs and fairness. Thus, EE alone is not a well-suited metric for an efficiency of resource allocation. Moreover, we compare the AMC proposal against three state-of-the-art resource allocation algorithms and conclude that our proposal is the best suited to meet 2020 cellular network requirements.

6.2 Final Remarks and Future Work

As future work, we believe that the dynamic association of UE-RRH can further optimize the proposed solution. Also, through the dynamic association of UE-RRH, a vertical handover model can be defined as part of the objective function of the resource allocation problem. For this, the mobility of the user should be investigated and considered. We also encourage the modeling of different user profiles for minimum demand as well as multiple service contracts for several users priority. Finally, we believe that more sophisticated and realistic models of signal propagation can be used in the communication of RBs. Different characteristics pertinent to this design, such as Modulation and Coding Schemes and Channel Quality Index, currently used in LTE, could be adopted in the future.

REFERENCES

3GPP TR 36.814. Further advancements for e-utra - physical layer aspects. *3rd Generation Partnership Project*, 2010.

3GPP TR 36.872. Small cell enhancements for e-utra and e-utran - physical layer aspects. *3rd Generation Partnership Project*, 2013.

ADACHI, F. Wireless past and future–evolving mobile communications systems–. **IEICE transactions on fundamentals of electronics, communications and computer sciences**, The Institute of Electronics, Information and Communication Engineers, v. 84, n. 1, p. 55–60, 2001.

BALACHANDRAN, K.; KADABA, S. R.; NANDA, S. Channel quality estimation and rate adaptation for cellular mobile radio. **IEEE Journal on Selected Areas in Communications**, IEEE, v. 17, n. 7, p. 1244–1256, 1999.

BENNETT, C. H. et al. Entanglement-assisted capacity of a quantum channel and the reverse shannon theorem. **IEEE Transactions on Information Theory**, IEEE, v. 48, n. 10, p. 2637–2655, 2002.

CHANCLOU, P. et al. Optical fiber solution for mobile fronthaul to achieve cloud radio access network. In: IEEE. **Future Network and Mobile Summit (FutureNetworkSummit), 2013**. [S.l.], 2013. p. 1–11.

DAHROUJ, H. et al. Resource allocation in heterogeneous cloud radio access networks: advances and challenges. **IEEE Wireless Communications**, v. 22, n. 3, p. 66–73, June 2015. ISSN 1536-1284.

DAI, B.; YU, W. Energy efficiency of downlink transmission strategies for cloud radio access networks. **IEEE Journal on Selected Areas in Communications**, IEEE, v. 34, n. 4, p. 1037–1050, 2016.

DEMESTICHAS, P. et al. 5g on the horizon: key challenges for the radio-access network. **IEEE Vehicular Technology Magazine**, IEEE, v. 8, n. 3, p. 47–53, 2013.

EKSTROM, H. et al. Technical solutions for the 3g long-term evolution. **IEEE Communications Magazine**, IEEE, v. 44, n. 3, p. 38–45, 2006.

ERICSSON. **Ericsson mobility report: On the pulse of the networked society**. [S.l.], 2013.

FRENCH, R. C. The effect of fading and shadowing on channel reuse in mobile radio. **IEEE transactions on vehicular technology**, IEEE, v. 28, n. 3, p. 171–181, 1979.

HAARIO, H.; SAKSMAN, E.; TAMMINEN, J. An adaptive metropolis algorithm. **Bernoulli**, Bernoulli Society for Mathematical Statistics and Probability, v. 7, n. 2, p. 223–242, April 2001.

HUANG, P.-H.; KAO, H.; LIAO, W. Hierarchical cooperation in heterogeneous cloud radio access networks. In: IEEE. **Communications (ICC), 2016 IEEE International Conference on**. [S.l.], 2016. p. 1–6.

LAFORTUNE, J.-F.; LECOURS, M. Measurement and modeling of propagation losses in a building at 900 mhz. **IEEE Transactions on Vehicular Technology**, IEEE, v. 39, n. 2, p. 101–108, 1990.

LIU, D. et al. User association in 5g networks: A survey and an outlook. **IEEE Communications Surveys & Tutorials**, IEEE, v. 18, n. 2, p. 1018–1044, 2016.

LUO, J. et al. Investigation of radio resource scheduling in wlans coupled with 3g cellular network. **IEEE Communications Magazine**, IEEE, v. 41, n. 6, p. 108–115, 2003.

MAHENDRAN, N. et al. Adaptive MCMC with bayesian optimization. In: **Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS**. [S.l.: s.n.], 2012. p. 751–760.

MALANDRINO, F. et al. Interference-aware downlink and uplink resource allocation in hetnets with d2d support. **IEEE Transactions on Wireless Communications**, IEEE, v. 14, n. 5, p. 2729–2741, 2015.

MAROTTA, M. A. et al. Resource sharing in heterogeneous cloud radio access networks. **IEEE Wireless Communications**, v. 22, n. 3, p. 74–82, June 2015. ISSN 1536-1284.

MORELLI, M.; KUO, C.-C. J.; PUN, M.-O. Synchronization techniques for orthogonal frequency division multiple access (ofdma): A tutorial review. **Proceedings of the IEEE**, IEEE, v. 95, n. 7, p. 1394–1427, 2007.

NOGUEIRA, M. C.; ROCHOL, J. Estratégias de escalonamento ofdma dl para redes móveis. 2016.

PAHLAVAN, K.; LEVESQUE, A. H. Wireless data communications. **Proceedings of the IEEE**, IEEE, v. 82, n. 9, p. 1398–1430, 1994.

PENG, M. et al. Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies. **IEEE Wireless Communications**, IEEE, v. 21, n. 6, p. 126–135, 2014.

PENG, M. et al. Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks. **IEEE Transactions on Vehicular Technology**, v. 64, n. 11, p. 5275–5287, November 2015. ISSN 0018-9545.

QIAN, F. et al. Top: Tail optimization protocol for cellular radio resource allocation. In: IEEE. **Network Protocols (ICNP), 2010 18th IEEE International Conference on**. [S.l.], 2010. p. 285–294.

ROBERTS, M. L. et al. Evolution of the air interface of cellular communications systems toward 4g realization. **IEEE Communications Surveys and Tutorials**, v. 8, n. 1-4, p. 2–23, 2006.

ROCHOL, J. Comunicação de dados. **Porto Alegre: Instituto de Informática da UFRGS: Bookman**, 2012.

SAMPATH, A.; KUMAR, P. S.; HOLTZMAN, J. M. Power control and resource management for a multimedia cdma wireless system. In: IEEE. **Personal, Indoor and Mobile Radio Communications, 1995. PIMRC'95. Wireless: Merging onto the Information Superhighway., Sixth IEEE International Symposium on**. [S.l.], 1995. v. 1, p. 21–25.

SHI, Y.; ZHANG, J.; LETAIEF, K. B. Group sparse beamforming for green cloud-ran. **IEEE Transactions on Wireless Communications**, v. 13, n. 5, p. 2809–2823, May 2014. ISSN 1536-1276.

TANG, J. et al. Resource efficiency: A new paradigm on energy efficiency and spectral efficiency tradeoff. **IEEE Transactions on Wireless Communications**, v. 13, n. 8, p. 4656–4669, August 2014. ISSN 1536-1276.

TAROKH, V.; SESHADRI, N.; CALDERBANK, A. R. Space-time codes for high data rate wireless communication: Performance criterion and code construction. **IEEE transactions on information theory**, IEEE, v. 44, n. 2, p. 744–765, 1998.

TRIPATHI, N.; REED, J. H. **Cellular communications: a comprehensive and practical guide**. [S.l.]: John Wiley & Sons, 2014.

WANG, Y.; ZHANG, J.; ZHANG, P. Energy-efficient power and subcarrier allocation in multiuser ofdma networks. In: **2014 IEEE International Conference on Communications (ICC)**. [S.l.: s.n.], 2014. p. 5492–5496. ISSN 1550-3607.

WU, J. et al. Cloud radio access network (c-ran): a primer. **IEEE Network**, IEEE, v. 29, n. 1, p. 35–41, 2015.

XIONG, C. et al. Energy- and spectral-efficiency tradeoff in downlink ofdma networks. **IEEE Trans. on Wireless Communications**, v. 10, n. 11, p. 3874–3886, Nov. 2011. ISSN 1536-1276.

YANG, C. et al. When icn meets c-ran for hetnets: an sdn approach. **IEEE Communications Magazine**, v. 53, n. 11, p. 118–125, November 2015. ISSN 0163-6804.

AppendixA PUBLISHED PAPER – ICC 2017

Matias A. K. Schimuneck, Maicon Kist, Ana Carolina Ribeiro-Teixeira, Cristiano Bonato Both, Juergen Rochol. **Adaptive Monte Carlo Algorithm to Global Radio Resources Optimization in H-CRAN**. 2017 IEEE ICC 2017 Wireless Communications Symposium, Paris, 2017, pp. 2894-2899. DOI: 10.1109/ICC.2017.7996788

- **Title:** *Adaptive Monte Carlo Algorithm to Global Radio Resources Optimization in H-CRAN.*
- **Contribution:** A proposal of an Adaptive Monte Carlo (AMC) algorithm to the optimal radio resources allocation model.
- **Abstract:** We propose an Adaptive Monte Carlo algorithm to perform global energy efficient resource allocation for Heterogeneous Cloud Radio Access Network (H-CRAN) architectures, which are forecast as future fifth-generation (5G) networks. We argue that our global proposal offers an efficient solution to the resource allocation for both high and low density scenarios. Our contributions are threefold: (*i*) the proposal of a global approach to the radio resource assignment problem in H-CRAN architecture, whose stochastic character ensures an overall solution space sampling; (*ii*) a critical comparison between our global solution and a local model; (*iii*) the demonstration that for high density scenarios Energy Efficiency is not a well suited metric for efficient allocation, considering data rate capacity, fairness, and served users. Moreover, we compare our proposal against three state-of-the-art resource allocation algorithms for 5G networks.
- **Status:** Published.
- **Qualis:** A1.
- **Conference:** IEEE International Conference on Communications (ICC).
- **Date:** May 21 - May 25, 2017.
- **Local:** Paris, France.
- **URL:** <http://icc2017.ieee-icc.org/>.

- **Digital Object Identifier (DOI):** <10.1109/ICC.2017.7996788>.

Adaptive Monte Carlo Algorithm to Global Radio Resources Optimization in H-CRAN

Matias A. K. Schimuneck,
Maicon Kist, and Juergen Rochol
Institute of Informatics
Federal University of Rio Grande do Sul
Av. Bento Gonçalves, 9500 – Porto Alegre, Brazil
Email:{makschimuneck, mkist, juergen}@inf.ufrgs.br

Ana Carolina Ribeiro-Teixeira
and Cristiano Bonato Both
Institute of Physics and Informatics
Federal University of Health Sciences of Porto Alegre
R. Sarmento Leite, 245 – Porto Alegre, Brazil
Email:{anacarolinart, cbboth}@ufcspa.edu.br

***Abstract*—We propose an Adaptive Monte Carlo algorithm to perform global energy efficient resource allocation for Heterogeneous Cloud Radio Access Network (H-CRAN) architectures, which are forecast as future fifth-generation (5G) networks. We argue that our global proposal offers an efficient solution to the resource allocation for both high and low density scenarios. Our contributions are threefold: (*i*) the proposal of a global approach to the radio resource assignment problem in H-CRAN architecture, whose stochastic character ensures an overall solution space sampling; (*ii*) a critical comparison between our global solution and a local model; (*iii*) the demonstration that for high density scenarios Energy Efficiency is not a well suited metric for efficient allocation, considering data rate capacity, fairness, and served users. Moreover, we compare our proposal against three state-of-the-art resource allocation algorithms for 5G networks.**

I. Introduction

Projections on cellular network service requirements foretell a 10-fold expansion in coverage area, a 100-fold increase in User's Equipment (UEs), as well as a 1000-fold raise in data rate transmission capacity by 2020 in comparison with current cellular networks specifications [1][2]. Current network architectures are unable to meet the preconized requirements. Luckily, the dense deployment of small cells is reckoned as a promising solution to reach such stringent improvements since it moves the antennas closer to UEs, thus achieving higher data rates by taking profit of better signal quality at short distances [3]. However, operating a massive number of antennas can significantly increase the energy consumption of the network [4]. Moreover, insertion of a large number of new radios means a buildup on spectral interference between the cells, potentially challenging the gain in spectral capacity [5].

Fortunately, Energy Efficiency (EE), *i.e.,* the relation between spectral capacity and energy consumption, can be managed through a radio resource allocation algorithm that controls transmission power and distribution of spectral Resource Blocks (RBs), *e.g.*, in a Generalized Frequency Division Multiple Access Allocation Scheme [6]. Nonetheless, granting service to UEs through the optimal management of radio resources is a challenging task. For instance, while low transmission powers may preclude UE's connection, high transmission powers may increase interference, thus hindering Spectral Efficiency (SE), *i.e.,* the ratio between transmission capacity and channel bandwidth [7]. Similarly, unplanned reuse of RBs may also intensify interference, resulting in low data transmission per RB, while under-reuse of RBs will limit the overall data transmission capacity. These competing effects are at the core of the high complexity of the management of radio resources, and motivate the design of a solution that encompasses control of transmission power, assignment of the spectral RBs, while ensuring service to the UEs.

The radio resource allocation problem has been investigated in the context of Orthogonal Frequency-Division Multiple Access (OFDMA) based systems for traditional and scattered networks [6][8]. Recently, this problem has attracted renewed attention within the community in connection with the trends on cloud-based centralized cellular architectures, *i.e.,* Cloud Radio Access Network (C-RAN) and Heterogeneous C-RAN (H-CRAN). These architectures allow for new control possibilities and crucially depend on global network knowledge [3][4]. Shi, Zhang, and Letaief [4] presented a global optimization model for C-RAN aiming to reduce the energy consumption of the network through a joint selection of active cell and minimization of the transmit power. Peng et al. [3] investigated the H-CRAN architecture problem aiming at single cell (*i.e.,* local) EE optimization through iterative RBs distribution with minimal power assignment.

The above solutions represent an important step forward. Yet, given resource competition between the cellular radios, EE landscape is highly complex. In view of this competition, we argue that in EE optimization approaches to H-CRAN environments, the investigation of a global and heterogeneous solution is an exaction. Moreover, due to the heterogeneity of antennas and UEs relative spatial distribution, consideration of a fairness constraint in the optimization problem is necessary, to avoid EE optimal solutions to favor UEs closest to antennas, being detrimental to the farthest UEs. Evidently, service is guaranteed by a UE demand constraint.

We propose a global approach to the radio resource allocation problem in a multi-cell H-CRAN model. As the main contribution of this paper, we present an Adaptive Monte Carlo (AMC) algorithm to the optimal radio resources allocation model. Stochastic approaches, such as AMC, provide an efficient way of sampling highly complex solution spaces, as is the case for the EE landscape in the resource allocation problem for H-CRAN. Moreover, we also demonstrate that for high density scenarios EE is not a well suited metric for efficient allocation, considering data rate capacity, fairness, and served users. We evaluate our solution in a scenario defined by the 3rd Generation Partnership Project (3GPP) for the latest definition of Long-Term Evolution (LTE-A) with small cell deployment [9]. We compare the AMC proposal against three state-of-the-art resource allocation algorithms. The results of our evaluation show that the global solution improves the efficiency of the radio access network when compared to a local model.

The remainder of the paper is organized as follows. The system model is analyzed in Section 2 while our proposed solution is presented in Section 3. Our solution is then evaluated in Section 4. Finally, conclusions are presented in Section 5.

II. System Model

H-CRAN architecture adopts Remote Radio Heads (RRH) that digitize and forward the signal samples to be processed at a centralized Base Band Unit (BBU) pool [3]. RRHs are simple antennas with a lower cost of deployment and operation when compared to traditional base stations. In practice, several RRHs acting as Low Power Nodes (LPNs), will be deployed in the coverage area of High Power Nodes (HPNs) to increase the network bit rate in hot spot areas, *i.e.,* holding a large number of UEs and high traffic demand [4]. Fortunately, the coupling of RRHs to a centralized BBU enables the use of cloud computing concepts to centralize decisions and achieve global optimization of resources [10].

Despite the advantages offered by RRH malleability, studies indicate that the densification of RRHs can significantly increase the energy consumption of the network [4]. Beside linearly scaling with the number of RRHs, energy consumption also depends on energy demands of a high-performance optical network connecting RRHs to remote BBUs [3]. Reducing energy consumption is possible by dynamically decreasing the RRHs transmission power through global cloud control solutions [4]. Nevertheless, energy consumption alone cannot be taken as a metric for network efficiency. Indeed, limiting the transmission power reduces Signal-to-Noise Ratio (SNR), possibly restraining the data transmission capacity, thus demanding more spectral RBs to meet a target UE demand [4][11]. The network EE is pointed as an important metric to evaluate the system performance, balancing the energy consumption and ensuring service to the UEs.

In an OFDMA-based downlink system, with a total of K RBs per antenna, L LPNs and H HPNs attending U UEs, the network total data capacity can be written as:

$$C(a,p) = \sum_{t=1}^{L+H} \sum_{u=1}^{U} \sum_{k=1}^{K} a_{t,u,k} \cdot S_{sh}(p_{t,u,k}), \qquad (1)$$

where $t \in \{1,...,L+H\}$ denotes the LPN and HPN cell transmitters, $u \in \{1,...,U\}$ corresponds to UEs, and $k \in \{1,...,K\}$ denotes the respective RB in a given antenna (LPNs and HPNs). Matrices $a = [a_{t,u,k}]_{(L+H)\times U\times K}$ and $p = [p_{t,u,k}]_{(L+H)\times U\times K}$ represent a feasible RB and power allocation possibility, respectively. The element $a_{t,u,k} \in \{0,1\}$, *i.e.,* $a_{t,u,k} = 1$ indicates that the kth RB in the tth LPN or HPN is assigned to the uth UE, and the element $p_{t,u,k}$ denotes the respective transmition power. Whenever $a_{t,u,k} = 0$, for a given (t,u,k), *i.e.,* kth RB belonging to tth antenna is not assigned to UE u, the corresponding power $p_{t,u,k} = 0$.

For each $a_{t,u,k} = 1$, $p_{t,u,k} > 0$, and we have a transmission bit capacity $S_{sh}(p)$, as defined by Shannon's theorem [3]:

$$S_{sh}(p_{t,u,k}) = B_0 \log_2\left(1 + \frac{p_{t,u,k} P_{t,u}^{(pl)}}{I_{t,u,k} + N_0}\right), \qquad (2)$$

where $I_{t,u,k}$ represents the power interference due to the assignment of the kth RB to other UEs ($u' \neq u$) by adjacent LPNs or HPNs ($t' \neq t$). $P_{t,u}^{(pl)}$ is the system path loss coming from the inverse square law decay of signal intensity with distance between uth UE receptor and tth transmitter. Finally, B_0 and N_0 are fixed system bandwidth and noise, respectively.

Each cell type has a different total power consumption. Considering transmission, circuit, and front-haul power, the total power consumption of an LPN (HPN) is defined as $P_L(a,p)$ ($P_H(a,p)$) is defined as [3]:

$$P_L(a,p) = \sum_{l=1}^{L} \left(\varphi_{\text{eff}}^{L} \sum_{u,k} a_{l,u,k}\, p_{l,u,k} + P_{\text{c}}^{L} + P_{bh}^{L}\right), \qquad (3)$$

$$P_H(a,p) = \sum_{h=L+1}^{L+H} \left(\varphi_{\text{eff}}^{H} \sum_{u,k} a_{h,u,k}\, p_{h,u,k} + P_{\text{c}}^{\text{H}} + P_{bh}^{\text{H}}\right), \qquad (4)$$

where $l \in \{1,...,L\}$ denotes LPN cells while $h \in \{L+1,...,L+H\}$ denotes HPN cells. The constants φ_{eff}^{L} (φ_{eff}^{H}), P_{c}^{L} (P_{c}^{H}), and P_{bh}^{L} (P_{bh}^{H}) denote, respectively, the power amplifier efficiency, circuit power, and power consumption of the front-haul link for LPNs (HPNs) [3].

The overall EE performance of a system is obtained by the sum of the network transmission capacity (1) divided by the total power consumption, equations (3) and (4). We represent EE by Γ^{E}, which for an H-CRAN can be written as:

$$\Gamma^{E}(a,p) = \frac{C(a,p)}{P_L(a,p) + P_H(a,p)}. \qquad (5)$$

An allocation problem to optimize the overall efficiency of the radio resources can be formulated with constraints on RB assignments, maximum transmission power per cell, minimum required UE demand, and fairness among UEs data rate. RB assignment and maximum transmission power constraints are hard physical limitations enforced strictly. However, the minimum required UE demand and fairness data rate are soft constraints, yet essential for service ensurance. The latter constraints are fundamental to guarantee service to all UEs avoiding that UEs closer to antennas be prioritized. The EE constrained maximization problem can be formulated as:

$$\max_{\{\mathbf{a},\mathbf{p}\}} \quad \Gamma^E(a,p) = \max_{\{\mathbf{a},\mathbf{p}\}} \frac{C(a,p)}{P_L(a,p)+P_H(a,p)}, \tag{6}$$

s.t.

$$\sum_u a_{t,u,k} \le 1, \qquad a_{t,u,k} \in \{0,1\},\ \forall k,\ \forall t, \tag{7}$$

$$\sum_t a_{t,u,k} \le 1, \qquad \forall u,\ \forall k, \tag{8}$$

$$\sum_{u,\,k} p_{l,u,k} \le P^L_{\max}, \qquad 1 \le l \le L, \tag{9}$$

$$\sum_{u,\,k} p_{h,u,k} \le P^{\mathrm{H}}_{\max}, \qquad L+1 \le h \le L+H, \tag{10}$$

$$\sum_k S_{sh}(p_{t,u,k}) \ge \eta_{\mathrm{u}}, \qquad \forall u,\ \forall t, \tag{11}$$

$$\frac{\left(\sum_{t,\,u,\,k} S_{sh}(p_{t,u,k})\right)^2}{U \cdot \sum_u \left(\sum_{t,\,k} S_{sh}(p_{t,u,k})\right)^2} \ge \eta_f. \tag{12}$$

Constraint (7) expresses that each RB in a given LPN (HPN) can be assigned to at most one UE. Constraint (8) limits assignment of the same RB to a given UE by at most one LPN (HPN). Constraints (9) and (10) limit to $P^L_{\max}$ (P^H_{max}), the maximum transmission power of LPNs (HPNs). Constraint (11) sets UE demand η_{u} as a lower limit to uth UE data rate. Lastly, constraint (12) denotes the fairness restriction for the spectral RBs distribution among the UEs, $\eta_f = \frac{1}{U}$ being the worst case fairness, and $\eta_f = 1$ being the best case fairness.

In view of the complexity of the model (6)-(12), we argue that finding the optimal RB assignment and power allocation, $\mathbf{a}^*$ and $\mathbf{p}^*$, constitute a non-convex optimization problem. Classical convex optimization methods are not well suited to explore the complex EE landscape. Stochastic methods, however, appear as a suitable approach to address the solution space landscape sampling problem.

III. Adaptive Monte Carlo

The AMC algorithm is an example of a stochastic approach to the radio resource allocation problem. Random sampling is an efficient way of solving complex constrained optimization through the generation of an ensemble of initial conditions (sample solutions) randomly exploring the solution space [12]. A stochastic dynamics is then defined that allows for some rate of violation of the target function maximization. Each sample is evaluated considering model (6)-(12) to find the best possible resource assignment. Algorithm 1 provides an overview of our AMC proposal, which we describe in the remaining of this section.

Parameters $I_{max}, B, \Lambda, \Upsilon$ (line 1) represent, respectively, the number of Monte Carlo Steps (MCSs), AMC weights for EE, data rate, and fairness. The samples' EE is computed after I_{max} MCSs (lines 2-18). The ensemble of sample solutions S contains $K \cdot U$ initial conditions (line 3). For each sample solution S, we define adaptive weights $\beta_S, \lambda_S, \upsilon_S$ that tune the solution design priorities, *i.e.,* EE, data rate, and fairness, all initialized to 1 (line 4). Each MCS is represented by a set of moves. We randomly select an antenna t, a random UE

Algorithm 1 Adaptive Monte Carlo

1: Set $I_{max}, B, \Lambda, \Upsilon$;
2: **for** $i = 0$ **to** I_{max} **do**
3: **for** $S = 0$ **to** $K \cdot U$ **do**
4: **if** $i == 0$ **then** $\beta_S = 1, \lambda_S = 1, \upsilon_S = 1$;
5: **for** $j = 0$ **to** $K \cdot (H + L)$ **do**
6: $t, u, k = \mathit{rand}(S.antennas, S.ues, S.rbs)$;
7: $B_s = \mathit{store}(S)$;
8: **if** $S.a_{t,:,k}$ **is** *assigned* **then**
9: $S.a_{t,:,k} = 0$; $S.p_{t,:,k} = 0$;
10: **else**
11: $S.a_{t,u,k} = 1$; $S.p_{t,u,k} = \mathit{minpower}(t,u,k)$;
12: $\Gamma^{[E,D,F]} = W(S.a, S.p, B, \Lambda, \Upsilon)$;
13: **if** $A(S, B_s, \beta_S, \lambda_S, \upsilon_S) \ge \mathit{rand}()$ **then**
14: $\Gamma_S^{[E,D,F]} = \Gamma^{[E,D,F]}$;
15: **else**
16: $S = \mathit{restore}(B_s)$;
17: $\mathit{update}(S, \beta_S, \lambda_S, \upsilon_S)$;
18: $\Gamma_i^E = \mathit{select}(S^*)$;

u in t's coverage area, and a random RB k (line 6), *i.e.,* a triplet t, u, k. A candidate move is an attempt to change the RBs allocation of a sample (*e.g.,* vacating it or assigning it to a UE u connected to the antenna t). We store B_s of the original state, to be able to restore it eventually (line 7).

Having randomly selected RB k belonging to antenna t, it will either be vacated with some probability (if initially assigned to some UE u - line 9), or assigned to a UE u' connected to t with some probability (if initially vacant - line 11), whence, constraints (7) and (8) are always satisfied. After assignment, the minimum power $p_{t,u',k}$ to communicate the UE u' with the antenna t through kth RB is obtained considering path loss, received interference, and maximum transmission power. The minimum power thus defined ensures constraints (9) and (10).

The optimal solution S^* must encompass optimal EE, user data rate constraint (11), and fairness constraint (12). These are enforced in a single function W, given as:

$$W(a,p,B,\Lambda,\Upsilon) = \\ B \cdot \Gamma^E(a,p) + \Lambda \cdot \Gamma^D(a,p) + \Upsilon \cdot \Gamma^F(a,p), \tag{13}$$

where B, Λ, and Υ are constant AMC weights, and

$$\Gamma^D = \min\left\{0, \sum_u \left(\sum_{t,\,k} a_{t,u,k} S_{sh}(p_{t,u,k}) - \eta_{\mathrm{u}}\right)\right\}, \tag{14}$$

$$\Gamma^F = \frac{\left(\sum_{t,\,u,\,k} S_{sh}(p_{t,u,k})\right)^2}{U \sum_u \left(\sum_{t,\,k} S_{sh} h(p_{t,u,k})\right)^2} - 1. \tag{15}$$

Function W (line 12) allows one to tune the priority given to each term through the relative size of AMC weights. Once UE data rate and fairness constraints are satisfied by a sample solution S, AMC dynamics is lead by EE maximization. The W definition is inspired by Lagrange Multipliers for constrained variational problems.

Functions Γ^E, Γ^D, and Γ^F provide quality indices of a sample state. Given a candidate move, its acceptance rate (line 13) is:

$$A(S, B_s, \beta_S, \lambda_S, \upsilon_S) \quad (16)$$
$$= \mathcal{P}_{\beta_S}(\Gamma^E_{B_S} \to \Gamma^E_S) \cdot \mathcal{P}_{\lambda_S}(\Gamma^D_{B_S} \to \Gamma^D_S) \cdot \mathcal{P}_{\upsilon_S}(\Gamma^F_{B_S} \to \Gamma^F_S),$$

where $\mathcal{P}_\xi(\Gamma_i \to \Gamma_f) = \min\{1, \exp(\xi\ (\Gamma_f - \Gamma_i))\}$. Hence, it is possible to accept a move which decreases EE, UE data rate relative to UE demand, and/or fairness with some finite probability, tuned by adaptive weights $\beta_S, \lambda_S, \upsilon_S$. Symbols Γ^E_S ($\Gamma^E_{B_s}$) are the candidate (stored) EE state of the sample solution, Γ^D_S ($\Gamma^D_{B_s}$) are the corresponding candidate (stored) neglected UE demand, and Γ^F_S ($\Gamma^F_{B_s}$) are the corresponding candidate (stored) fairness. Candidate sample solution S is accepted with probability A (line 14), and rejected with probability $1 - A$, in which case S is restored to B_S (line 16). Each MCS amounts to $K \cdot (H + L)$ move attempts.

The adaptive weights $\beta_S, \lambda_S, \upsilon_S$ are updated (line 17) between MCSs according to the current state of the sample solution. We compute the average values $\langle\Gamma^E\rangle$, $\langle\Gamma^D\rangle$, and $\langle\Gamma^F\rangle$ over the ensemble of sample solutions $\{S\}$. For each sample solution S, if $\Gamma^E_S < \langle\Gamma^E\rangle$ (respectively, $\Gamma^D_S < \langle\Gamma^D\rangle$, $\Gamma^F_S < \langle\Gamma^F\rangle$), the corresponding adaptive weight β_S (respectively, λ_S, υ_S) is increased by a fixed factor. Moreover, if Γ^E_S (respectively, Γ^D_S, Γ^F_S) relative fluctuations over a number of MCSs are small, then the respective weight is increased by a larger factor, than in the opposite situation. Conversely, if $\Gamma^E_S > \langle\Gamma^E\rangle$ (respectively, $\Gamma^D_S > \langle\Gamma^D\rangle$, $\Gamma^F_S > \langle\Gamma^F\rangle$), and the relative sample solution fluctuations are small, then the corresponding weight is decreased by a fixed factor. This method is similar to a Simulated Annealing method in Markov Chain Monte Carlo [13].

The AMC dynamics solution is chosen as the best among all sample solutions generated throughout the AMC history (line 18). The best sample will be the solution to the radio resource allocation problem for an H-CRAN architecture. The proposed AMC will find a solution in polynomial time.

IV. Evaluation

In the following, we describe a realistic scenario for simulations and tests methodology. In the sequence, numerical results are presented and then analyzed and discussed.

A. Scenario

We simulate an H-CRAN scenario assuming 3GPP technical specifications for LTE-A radio access networks with small cell deployment [9]. A dense scenario must contain at least 7 HPNs, each one with three sectors. HPNs are uniformly distributed respecting a minimum distance of 500 meters between each pair of HPNs. A minimum of two Hot Spots ($H_s = 2$) must be deployed inside the HPN coverage area, at a minimum distance of 90 meters from the center of the corresponding HPN. A Hot Spot can be defined as a densely populated area, *e.g.,* a city downtown or a shopping mall. Each Hot Spot has a set of four LPNs randomly deployed with a minimum distance of 10 meters between each pair of LPNs. Moreover, the UEs located in a Hot Spot area are randomly placed with minimum distances of 30 meters to the HPN and 5 meters to the nearest LPN. Two thirds of UEs are forcibly placed within the Hot Spot area, the remaining UEs being randomly placed in the HPN coverage area. A typical scenario following the 3GPP specification is exemplified in Fig. 1.

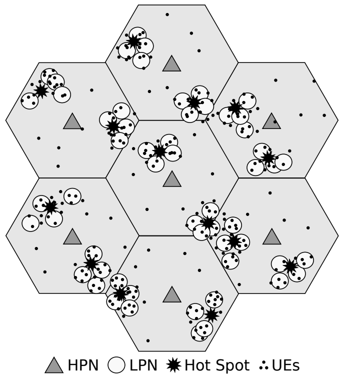


Fig. 1. H-CRAN scenario.

We also define minimal power and interference models. We compute the interference caused by the sum of the adjacent cell powers at the kth RB assigned to UE u as:

$$I_{t,u,k} = \sum_{u' \neq u} \sum_{t' \neq t} a_{t',u',k} (p_{t',u',k} P^{(pl)}_{t',u}).$$

We assume the following definition for the minimal power of LPNs (HPNs):

$$p_{t,u,k} = \min\{R_t, (I_{t,u,k} + N_0) \cdot T_{sinr} + P^{(pl)}_{t,u}\},$$

where

$$R_t = \begin{cases} P^L_{max} - \sum_{u,\,k} p_{t,u,k}, & \text{if } \ 1 \le t \le L, \\ P^H_{max} - \sum_{u,\,k} p_{t,u,k}, & \text{if } \ L+1 \le t \le L+H. \end{cases}$$

T_{sinr} is a fixed target Signal-to-Interference-plus-Noise Ratio (SINR) adopted for all system transmissions. The path loss $P^{(pl)}_{t,u}$ is given by the 3GPP TR 36.814 definition (for more details see [14]). Finally, our UEs are stationary and the transmission is restricted to a single-input downlink system.

We compare the performance of the proposed AMC radio resource allocation against a Locally Optimal (LO) algorithm, and two algorithms initially presented by Peng et al., namely Fixed Power and Sequential [3]. The LO algorithm assigns RBs to UEs, aiming at local interference mitigation and minimal power assignment. By local interference we mean that quality assessment of an allocation solution is done considering only received interference, which strongly hinders solution robustness. The Fixed Power algorithm equally distributes the transmission power among all K RBs of LPNs and HPNs. The Sequential algorithm relies on RBs assignment to UEs in a sequential manner with minimal power. The criteria for these choices are twofold: (*i*) to make a parallel between our global proposal and a local solution; (*ii*) considering commonly presented algorithms in the literature. The three algorithms are restricted to constraints (7)-(12) and execute the same number I_{max} of iterations as our proposal. Simulation and algorithm parameters are enumerated in Table I.

TABLE I
SIMULATION AND ALGORITHM PARAMETERS

Parameter	Value	Parameter	Value
H/Hs	7 / 2	B_0	180 Hz
L	$4 \cdot (H \times Hs)$	N_0	-110 dB
U	$[10, 100 : 10] \cdot (H \times Hs)$	$\varphi_{\mathrm{eff}}^{H} / \varphi_{\mathrm{eff}}^{L}$	4 / 2 W
$P_{max}^{H} / P_{max}^{L}$	46 / 23 dB	$P_{\mathrm{c}}^{\mathrm{H}}$ / P_{c}^{L}	10 / 6.8 W
T_{sinr}	16.5	P_{bh}^{H} and P_{bh}^{L}	3.85 W
η_{u} / η_f	1 Mbit/s / 1	I_{max}	10
K	100 (20 MHz, 0.5 ms)	$[B, \Lambda, \Upsilon]$	[1, 10, 1]

B. Numerical Results

We present numerical results considering system EE and SE, fraction of served UEs (*i.e.,* ratio of minimal data rate constraint fulfillment), and fairness in the resource distribution. We gradually increase the UE density from 10 to 100 UEs per Hot Spot. Fig. 2 shows the results of a total of 30 repetitions with 95% confidence intervals.

In Fig. 2 (a) we present comparative results regarding normalized EE as a function of density of UEs. AMC solution achieves the best EE results among the four compared models up to 60 UEs per Hot Spot. Beyond 50 UEs per Hot Spot, the increasing behavior of the normalized EE of the AMC proposal is replaced by a steep decreasing. For densities equal to 70 UEs per Hot Spot and higher, AMC proposal achieves lower normalized EE than the LO solution. For 80 UEs per Hot Spot and beyond, AMC proposal normalized EE also lies below normalized EE achieved by the Sequential algorithm.

In Fig. 2 (b) we show a similar comparison regarding normalized SE as a function of UE density. Again, at low densities (50 UEs per Hot Spot and below) our proposal achieves the highest normalized SE, when compared to any of the other three models. At higher UE densities, AMC's normalized SE is surpassed by Fixed Power model, which achieves the highest normalized SE in this density regime. This behavior is explained when considering that Fixed Power model attains very low transmission power, as shown in Fig. 2 (a), resulting in a scenario with minimal interference. Yet, above 50 UEs per Hot Spot, AMC's normalized SE attains the second best results among the compared models. We also observe that the normalized SE of the LO model saturates at a density of 60 UEs per Hot Spot.

Fig. 2 (c) regards fraction of served UEs as a function of UE density. At low densities (up to 40 UEs per Hot Spot), both AMC and LO are able to attend all UEs. For increasing densities, fraction of served UEs decreases for both algorithms, though decrease for the LO is steeper. At very high densities (for 70 UEs per Hot Spot and beyond), fraction of served UEs attained by AMC is the highest among the four models.

In the last analysis, presented in Fig. 2 (d), we present fairness results. At low UE densities, *i.e.,* where available resources are abundant, AMC solution tends to an unequal resource assignment among UEs. In this density regime, LO model is the fairest among the compared solutions. Beyond 60 UEs per Hot Spot, however, AMC solution becomes the fairest among solutions, and remains almost constant throughout the high density regime, while LO steeply decreases attaining a baseline comparable with the worst performing algorithm, considering fairness, *i.e.,* Sequential algorithm.

C. Analysis

In the following we present a correlated analysis, considering all aspects of the allocation solution. We identify three distinct regimes, namely, Low Density (LD), Moderate Density (MD), and High Density (HD), based on solution characteristics, as highlighted on top of Fig. 2.

• LD regime: Corresponds to 40 UEs per Hot Spot or less. Sequential and Fixed Power algorithms perform poorly in this regime. We show that the EE of our proposal (Fig. 2 (a)) is highest among all algorithms, LO's EE being the second best. Both algorithms are able to serve the total UE demand. Nonetheless, LO's solution is EE inefficient, due to its local knowledge of solution space. Fig. 2 (b) shows that LO's SE is only slightly above the worst performing algorithm's. Indeed, LO has only access to received interference, thus limits its solutions to the minimum data rate. Meanwhile, LO achieves the highest fairness (Fig. 2 (d)), but only because it is unable to explore data rates beyond the minimum.

In this resource abundant scenario, AMC achieves both highest EE and SE. This means that, on the one hand, given the abundance of resources in the LD regime, AMC is able to grant data rate beyond the minimum UE demand. On the other hand, given the global knowledge of the algorithm, it is also able to mitigate interference with intelligent resource allocation. This raises transmission capacity more than the corresponding energetic cost, whence achieving the highest EE. AMC's fairness (Fig. 2 (d)) degrades when the UE density is low, *i.e.,* despite all UEs having their minimum demand attended (Fig. 2 (c)), some UEs receive more data than others.

• MD regime: Between 40 and 70 UEs per Hot Spot, we reach a crossover regime, which goes from resource abundant to scarce. Mitigating interference becomes difficult, transmission cost grows to attain the target SINR, hindering EE. AMC is able to find solutions with increasing SE, to supply UE demand and fairness, up to a point. Radio resources are not enough to service all UEs, so AMC solutions divide resources prioritizing fairness. On the other hand, LO, given its limited knowledge on interference, is unable to locally reconfigure its network allocation in this saturation regime, the interference due to RB reuse increases and its SE saturates. This means the number of UEs LO solutions are able to serve saturates as well, and fraction of served UEs decreases steeply with increasing density.

• HD regime: Beyond 70 UEs per Hot Spot, the steep decrease in EE (Fig. 2 (a)) for the AMC solution is explained

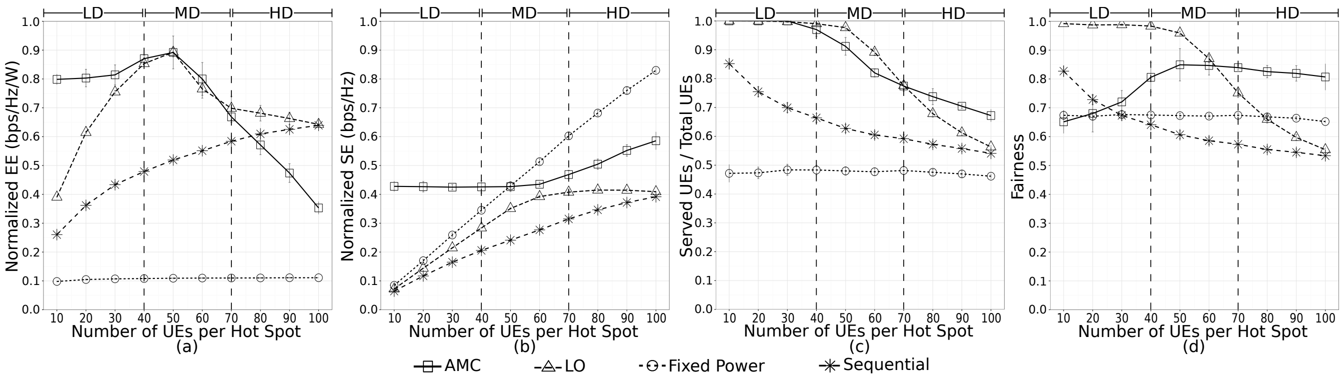


Fig. 2. Quality assessment for UEs densification in a fixed H-CRAN scenario with 7 HPN and 4 LPNs per Hot Spot

by a corresponding increase in SE (Fig. 2 (b)), with a resulting increase in energy cost. When the number of unattended UEs grows, the AMC solution sacrifices EE (including below Sequential's EE) to achieve higher SE, to serve more UEs with also higher fairness (Fig. 2 (c), (d)). Indeed, our solution provides the best fairness among all models. In contrast, the LO solution decreases the fairness and ratio of served UEs, due to its limited local knowledge. This means that solutions limited to local knowledge are unable to explore all the network reconfiguration capacity (limiting SE), as opposed to global solutions that can promote greater SE to the network.

High EE is easy to achieve, when considered alone, as can be seen from the fact that the Sequential algorithm is able to provide the same normalized EE as the LO. Had we considered EE only, we would have mistakenly inferred the AMC solution to provide a bad resource allocation strategy at high UE densities. However, this behavior is justified in view of our solution prioritizing served UEs and fairness, as forecast in 5G networks.

V. Conclusion and Future Works

In this paper, we proposed an AMC algorithm to perform global EE resource allocation for *e.g.,* H-CRAN architectures. We compared our solution against three state-of-the-art resource allocation algorithms for 5G networks, following the guidelines defined by the 3GPP. We showed that our proposal achieves the highest EE when the UE density is low; whereas, for high density networks, our proposal achieves an EE similar or lower to LO algorithms. However, our solution is able to guarantee higher fraction of served UEs with the highest fairness. We also proved that our algorithm provides the best SE due to overall knowledge and control of the network. We showed that in highly dense scenarios, EE maximization does not provide the best allocation solution, considering fraction of served UEs and fairness. Thus, EE alone is not a well suited metric for efficiency of resource allocation. We conclude that our proposal is the best suited, among all models considered, to meet 2020 cellular network requirements. For future works, we envisage the application of the AMC solution to a scenario with user mobility and the modeling of the horizontal handover as a function of the radio resource allocation problem.

Acknowledgment

The research leading to these results received funding from the European Commission H2020 programme under grant agreement no. 688941 (FUTEBOL), as well from the Brazilian Ministry of Science, Technology, Innovation, and Communication (MCTIC) through RNP and CTIC. The authors are also grateful to Marcelo M. Marotta reviews.

References

[1] Ericsson, "Ericsson mobility report: On the pulse of the networked society," Ericsson, White Paper, Tech. Rep., 2013.

[2] C. Yang, Z. Chen, B. Xia, and J. Wang, "When icn meets c-ran for hetnets: an sdn approach," *IEEE Communications Magazine*, vol. 53, no. 11, pp. 118–125, November 2015.

[3] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 11, pp. 5275–5287, November 2015.

[4] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-ran," *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2809–2823, May 2014.

[5] J. Tang, D. K. C. So, E. Alsusa, and K. A. Hamdi, "Resource efficiency: A new paradigm on energy efficiency and spectral efficiency tradeoff," *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4656–4669, August 2014.

[6] C. Xiong, G. Y. Li, S. Zhang, Y. Chen, and S. Xu, "Energy- and spectral-efficiency tradeoff in downlink ofdma networks," *IEEE Trans. on Wireless Communications*, vol. 10, no. 11, pp. 3874–3886, Nov. 2011.

[7] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 126–135, December 2014.

[8] Y. Wang, J. Zhang, and P. Zhang, "Energy-efficient power and subcarrier allocation in multiuser ofdma networks," in *2014 IEEE International Conference on Communications (ICC)*, June 2014, pp. 5492–5496.

[9] T. S. G. R. A. Network, "Small cell enhancements for e-utra and e-utran - physical layer aspects," 3GPP TR 36.872, Tech. Rep., 2013.

[10] M. A. Marotta, N. Kaminski, I. Gomez-Miguelez, L. Z. Granville, J. Rochol, L. DaSilva, and C. B. Both, "Resource sharing in heterogeneous cloud radio access networks," *IEEE Wireless Communications*, vol. 22, no. 3, pp. 74–82, June 2015.

[11] H. Dahrouj, A. Douik, O. Dhifallah, T. Y. Al-Naffouri, and M. S. Alouini, "Resource allocation in heterogeneous cloud radio access networks: advances and challenges," *IEEE Wireless Communications*, vol. 22, no. 3, pp. 66–73, June 2015.

[12] N. Mahendran, Z. Wang, F. Hamze, and N. de Freitas, "Adaptive MCMC with bayesian optimization," in *Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS*, April 2012, pp. 751–760.

[13] H. Haario, E. Saksman, and J. Tamminen, "An adaptive metropolis algorithm," *Bernoulli*, vol. 7, no. 2, pp. 223–242, April 2001.

[14] E. U. T. R. Access, "Further advancements for e-utra physical layer aspects," 3GPP TR 36.814, Tech. Rep., 2010.

AppendixB PUBLISHED PAPER – SBRC 2017

Ariel Galante Dalla-Costa, Matias A. K. Schimuneck, Juliano Araujo Wickboldt, Cristiano Bonato Both, Luciano Paschoal Gaspary, Lisandro Zambenedetti Granville. **NFV em Redes 5G: Avaliando o Desempenho de Composição de Funções Virtualizadas via Maestro**. 2017 SBRC 2017 Brazilian Symposium on Computer Networks and Distributed Systems, Brazil, Belem-Pará, 2017.

- **Title:** *NFV em Redes 5G: Avaliando o Desempenho de Composição de Funções Virtualizadas via Maestro*.
- **Contribution:** An evaluate of the effectiveness and orchestration overhead for deciding the best place to VNF components.
- **Abstract:** Dynamic Cloud Radio Access Network (Dynamic C-RAN) is an architecture wireless that aims benefits such as flexibility and agility to new five generation (5G) network. In Dynamic C-RAN, the radio functions processing is distributed along of a hierarchical of clouds. Network Functions Virtualization has been investigated the facilities to employ this wireless functions. NFV offers orchestration of split of radio functions into Virtualized Network Functions (VNF). The orchestrator needs to consider strict requirements of C-RAN between distributed components of VNF to get real earnings. In this work, we propose a detailed evaluation of Maestro - an NFV orchestrator for wireless environments aware VNF composition. We evaluate Maestro through of linear model programming, considering the effectiveness and orchestration overhead for deciding the best place to VNF components. In addition, we evaluate the trade-off between network data rate and time to generate the placement of components. Moreover, we compare our proposal against three state-of-the-art resource allocation algorithms for 5G networks.
- **Status:** Published.
- **Qualis:** B2.
- **Conference:** Brazilian Symposium on Computer Networks and Distributed Systems (SBRC).

- **Date:** May 15 - May 19, 2017.
- **Local:** Belém, Pará.
- **URL:** <http://sbrc2017.ufpa.br/en/>.

NFV em Redes 5G: Avaliando o Desempenho de Composição de Funções Virtualizadas via Maestro

Ariel Galante Dalla-Costa[1], Matias A. K. Schimuneck[1], Juliano Araujo Wickboldt[1], Cristiano Bonato Both[2], Luciano Paschoal Gaspary[1], Lisandro Zambenedetti Granville[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul

[2]Departamento de Matemática Aplicada e Ciências Sociais
Universidade Federal das Ciências da Saúde de Porto Alegre

{agdcosta, makschimuneck, jwickboldt}@inf.ufrgs.br

cbboth@ufcspa.edu.br, {paschoal, granville}@inf.ufrgs.br

Abstract. *Dynamic Cloud Radio Access Network (Dynamic C-RAN) is an architecture wireless that aims benefits such as flexibility and agility to new five generation (5G) network. In Dynamic C-RAN, the radio functions processing is distributed along of a hierarchical of clouds. Network Functions Virtualization has been investigated the facilities to employ this wireless functions. NFV offers orchestration of split of radio functions into Virtualized Network Functions (VNF). The orchestrator needs to consider strict requirements of C-RAN between distributed components of VNF to get real earnings. In this work, we propose a detailed evaluation of Maestro - an NFV orchestrator for wireless environments aware VNF composition. We evaluate Maestro through of linear model programming, considering the effectiveness and orchestration overhead for deciding the best place to VNF components. In addition, we evaluate the trade-off between network data rate and time to generate the placement of components.*

Resumo. *Dynamic Cloud Radio Access Network (Dynamic C-RAN) é uma arquitetura de redes sem fio emergente que objetiva trazer benefícios como flexibilidade e agilidade para redes móveis de quinta geração (5G). Em Dynamic C-RAN, a carga de processamento das funções de rádio é distribuída ao longo de uma hierarquia de nuvens e, para tal, tem-se investigado o uso de Network Function Virtualization (NFV). NFV tem potencial para oferecer a orquestração das funções de rádio, através da sua subdivisão em componentes de Virtualized Network Function (VNF). Neste trabalho propõe-se uma avaliação detalhada do Maestro – um orquestrador de NFV para ambientes sem fio orientado a composição de VNFs, a fim de estimar os potenciais ganhos ao se aplicar os conceitos de NFV em Dynamic C-RAN. Através de um modelo baseado em programação linear, avaliou-se a efetividade e custo de orquestração em termos de aderência aos requisitos de comunicação entre as funções de rádio e os posicionamentos dos componentes das VNFs na rede. Além disso, avaliou-se a relação entre tempo para realizar esses posicionamentos e a taxa de dados transmitida, considerando cenários de ocupação variável em Dynamic C-RAN.*

1. Introdução

Cloud Radio Access Network (C-RAN) é uma arquitetura de redes sem fio que separa os componentes das *Base Stations* (BS) em *Radio Remote Head* (RRH) e *Baseband Unit* (BBU). Uma RRH é uma unidade de rádio, enquanto uma BBU é uma unidade de processamento de sinais [Checko et al. 2015]. As RRHs estão localizadas na RAN e as amostras de sinais são transportadas através de enlaces de fibra óptica até as BBUs, que são centralizadas em uma infraestrutura de nuvem, também conhecidas como BBU *pool*. *Dynamic* C-RAN é uma evolução do conceito original de C-RAN, que prevê a distribuição das funções de rádio sobre uma infraestrutura de nuvem, como por exemplo, uma nuvem de borda que processa parte das funções de rádio, e uma nuvem regional que realiza o processamento das demais funções [5G-PPP-Working-Group 2016]. Em *Dynamic* C-RAN, o processamento das funções de rádio pode ser dividido de acordo com as suas funcionalidades, sendo distribuídas em diferentes locais da infraestrutura da rede, resultando em benefícios, tais como adaptabilidade, balanceamento de carga, flexibilidade na implantação de serviços e eficiência energética [Bartelt et al. 2015].

Apesar dos benefícios de *Dynamic* C-RAN, coordenar as funções de rádio não é uma tarefa trivial, por exemplo distribuir as funções de rádio ao longo de diferentes níveis da rede necessita de um algoritmo especializado. Esse algoritmo de distribuição deve respeitar requisitos restritos, tais como taxa mínima de dados transmitidos e atraso mínimo. Por esta razão, tanto academia quanto indústria vêm explorando conceitos de *Network Functions Virtualization* (NFV), na coordenação e distribuição das funções de rede. Utilizar NFV em *Dynamic* C-RAN, torna possível mover as funções de rádio para uma nuvem distribuída, virtualizando essas funções, chamadas de *Virtualized Network Functions* (VNF). Desta forma, as características de NFV, como por exemplo, escalabilidade, migração e orquestração, podem ser aproveitadas em cenários dinâmicos, como em *Dymamic* C-RAN [Heideker e Kamienski 2016], [Abdelwahab et al. 2016].

Os conceitos de NFV devem ser adaptados para gerenciar e orquestrar as funcionalidades de rádio virtualizadas na próxima geração de redes sem fio [Mijumbi et al. 2016], permitindo adaptação e dinamicidade ao ambiente 5G. Para tal, faz-se necessário calcular o posicionamento das funções, atendendo aos requisitos restritos do ambiente *Dynamic* C-RAN. Mais especificamente, um orquestrador deve determinar a localização dos componentes de funções virtualizadas, chamados de *Virtualized Network Functions Components* (VNFCs) ao longo das *Virtualized Deployment Units* (VDUs), considerando as restrições temporais das aplicações. Esse processo, conhecido como posicionamento de funções, é de responsabilidade do orquestrador, pois uma mesma função pode ser dividida em VNFs de diferentes formas pré-estabelecidas.

O orquestrador, chamado *Maestro* [Dalla-Costa et al. 2017], foi proposto em um trabalho anterior, que considera a composição de funções virtualizadas na implantação de VNFCs ao longo da rede. Naquele trabalho, conduziu-se uma prova de conceito focando nas características funcionais e na simulação de ambientes *Dynamic* C-RAN. Neste artigo, realiza-se uma avaliação detalhada do *Maestro*, utilizando diferentes composições de VNFs, buscando caracterizar a capacidade deste orquestrador em satisfazer requisitos temporais restritos, impostos pelo ambiente dinâmico de rede de acesso sem fio. Ademais, avalia-se a relação entre o tempo consumido para o cálculo de posicionamento dos VNFCs e a redução da quantidade de dados transmitidos (ganho de até 83,68%) no *fronthaul*. Assim, a principal contribuição deste artigo é mostrar o posicionamento dos VNFCs em um ambiente *Dynamic* C-RAN, devido a ocupação das BSs, observando quais cenários e quais características de ocupação influenciam no tempo para realizar tal posicionamento.

O restante deste artigo está organizado da seguinte maneira. Na Seção 2 são discutidos os trabalhos relacionados. Na Seção 3 é descrito e exemplificado um cenário de *Dynamic* C-RAN. Na Seção 4, propõe-se um estudo de caso, descrevendo e discutindo o desempenho do orquestrador. Finalmente, na Seção 5 são apresentadas as conclusões e os trabalhos futuros.

2. Trabalhos relacionados

Nessa seção são apresentados os trabalhos relacionados sobre *Dynamic* C-RAN e, posteriormente, discute-se o estado da arte sobre orquestração e posicionamento de funções virtualizadas, através de NFV.

2.1. *Dynamic* C-RAN

Diferente de C-RAN, em que o processamento de funções de rádio é movido para a nuvem centralizada, em *Dynamic* C-RAN funções de rádio podem ser distribuídas entre nuvens organizadas hierarquicamente. Essa distribuição possui diversas vantagens potenciais quando comparada a C-RAN, como maior eficiência energética, diminuição da latência, maior escalabilidade, flexibilidade e adaptabilidade de serviços. Por essa razão, redes 5G estão sendo idealizadas em ambientes *Dynamic* C-RAN [5G-PPP-Working-Group 2016]. Trabalhos encontrados na literatura especializada em redes sem fio [Bartelt et al. 2015, Liu et al. 2015] demonstram quais são as possibilidades e as vantagens de dividir as funções de rádio, por exemplo através da divisão do processamento local e remoto em diferentes funções, tais como MAC, Soft-Bit, RX Data, *Subframe* e *In-phase & Quadrature* (I/Q). Além disso, estudos mostram como essas subdivisões de funções de rádio implicam na taxa de dados transmitida no *fronthaul* [Wubben et al. 2014].

Coordenar *Dynamic* C-RAN por meio de funções de rádio virtualizadas é um tema de pesquisa que está sendo bastante investigado [Abdelwahab et al. 2016, Liu et al. 2015]. Existem trabalhos na literatura que advogam que NFV pode desempenhar um papel importante no gerenciamento e na orquestração de funções de rede sem fio virtualizadas [Mijumbi et al. 2016, Riggio et al. 2015]. Desta forma, *Dynamic* C-RAN pode considerar a virtualização da BBU, utilizando componentes para cada camada de rádio. Entretanto, ao empregar NFV nesse tipo de rede, depara-se com outros dois principais desafios de pesquisa. O primeiro refere-se a como posicionar e orquestrar funções virtualizadas em uma infraestrutura do tipo *Dynamic* C-RAN. O segundo consiste em assegurar desempenho satisfatório para o cálculo do posicionamento de VNFs, de modo a não prejudicar a dinamicidade do ambiente como um todo. Por exemplo, as funções de rádio de retransmissão da camada *Subframe* não suportam latências superiores a 1ms, enquanto na cada MAC o tempo do *Hybrid Automatic Repeat Request* (HARQ) deve ser menor ou igual a 8ms [Ortín et al. 2014]. Outro fator que pode alterar o funcionamento adequado do ambiente *Dynamic* C-RAN é a mobilidade dos *User Equipments* (UEs), ao longo da infraestrutura, exigindo que o orquestrador gerencie a rede de acordo com as variações das demandas do usuário. Na Subseção 2.2 são apresentados os trabalhos relacionados sobre orquestração em um ambiente *Dynamic* C-RAN utilizando NFV.

2.2. Orquestração baseada em NFV

Dando um passo adiante na área de *Dynamic* C-RAN, foi proposta uma arquitetura para NFV em redes 5G [Abdelwahab et al. 2016], discutindo a implantação de funções de rádio virtualizadas. Essa proposta se assemelha a arquitetura *Management and Orchestration* (MANO) do *European Telecommunications Standards Institute* (ETSI) [Quittek et al. 2014]. A arquitetura do ETSI visa padronizar as implementações, fornecer suporte à indústria e prover gerenciamento e orquestração de funções virtualizadas, definindo um *Network Function Virtualization Orchestrator* (NFVO). Segundo a especificação do ETSI, uma VNF é composta por VNFCs, que são os menores elementos da VNF e são implantados em *containers* virtualizados, ou VDUs. Todas essas VNFs são descritas por um VNF *Descriptor* (VNFD), que representa como os VNFCs são distribuídos e implantados ao longo dos VDUs.

Na área de orquestração em NFV, destacam-se, inicialmente, dois trabalhos relacionados a solução de orquestração. O primeiro refere-se a uma plataforma aberta para NFV, baseado em nuvem, que suporta uma API e outros componentes provenientes do *framework* ETSI MANO [Makaya et al. 2015]. O segundo propõe uma arquitetura modular para NFV, que permite o gerenciamento baseado em políticas de VNFs [Giotis et al. 2015]. Além disso, pode-se destacar outros trabalhos que abordam o posicionamento de funções virtualizadas, tais como T-NOVA [Xilouris et al. 2014] e Cloud4NFV [Soares et al. 2014]. Entretanto, nenhum desses trabalhos considera a composição das funções para a tomada de decisão em um ambiente sem fio, sendo assim, não atendem os requisitos necessários para orquestrar ambientes *Dynamic* C-RAN.

Um orquestrador específico para WLAN foi proposto [Riggio et al. 2015], sem considerar a divisão de componentes de funções, ou seja, não sendo adaptável ao cenário de *Dynamic* C-RAN. Além disso, uma avaliação formal de posicionamento de recursos de funções virtualizadas foi proposta [Moens e Turck 2014], focando em cenários híbridos entre VNFs e funções físicas legadas. Em relação as propostas de posicionamento automático dos nós virtuais e alocação de serviços [Clayman et al. 2014], pode-se destacar a abordagem baseada em heurísticas de *Service Function Chaining* (SFC) [Luizelli et al. 2015, Li e Qian 2016]. Essa abordagem, apesar de considerar o posicionamento de VNFs, não leva em consideração os componentes na tomada de decisão, assim, os tempos para efetuar o posicionamento não são diretamente comparáveis. Desta forma, com a finalidade de avaliar a dinamicidade da orquestração em *Dynamic* C-RAN, neste trabalho avalia-se a arquitetura do orquestrador Maestro, proposto anteriormente [Dalla-Costa et al. 2017].

3. Orquestração em *Dynamic* C-RAN

O cenário *Dynamic* C-RAN é composto de uma hierarquia em nuvens, aproveitando a distribuição geográfica das RANs como em ambientes FOG [Bonomi et al. 2012]. A nuvem central geralmente possui a maior quantidade de recursos disponíveis, enquanto as nuvens regionais e de borda são posicionadas próximas ao *front-end*, possuindo geralmente menos recursos. Essa organização hierárquica permite a implantação de funções virtualizadas de forma dinâmica, tornando o provisionamento do serviço mais flexível [Peng et al. 2014]. Por exemplo, uma nova unidade de processamento de sinal virtualizado pode ser implantada em uma nuvem regional para suprir uma determinada demanda, devido ao aumento do número de usuários na RAN. Além disso, algumas técnicas de balanceamento de carga e de economia de energia tornam-se facilmente aplicáveis nesses cenários, uma vez que existe flexibilidade na migração de VNFs entre os diferentes níveis de nuvem. Assim, dividir e posicionar as funções de rádio que demandam maior transmissão de dados, faz-se necessário através de um orquestrador. Desta forma, na Subseção 3.1 apresenta-se um modelo de posicionamento de VNFs e sua implementação, no orquestrador *Maestro* através de programação linear, é discutida na Subseção 3.2.

3.1. Modelo de posicionamento de VNFs em *Dynamic* C-RAN

Considerando um ambiente *Dynamic* C-RAN, contendo um conjunto de A BSs com um total de N nuvens (borda, regionais e centrais) e capacidade de F funções de rádios virtualizadas, o total de dados transmitidos no *fronthaul* pode ser escrito como:

$$O_B(d) = \sum_{f=0}^{F} \sum_{n=0}^{N} \sum_{a=0}^{A} d_{f,n,a} D_B(f,n,a) \tag{1}$$

onde, $f \in \{0, ..., F\}$ corresponde as funções de rádio virtualizadas, $n \in \{0, ..., N\}$ representa as nuvens disponíveis para processamento dessas funções e $a \in \{0, ..., A\}$ simboliza as BSs correspondentes a uma *Dynamic* C-RAN. A matriz tridimensional $d = [d_{f,n,a}]_{F \times N \times A}$ representa um posicionamento factível para as funções de rádio virtualizadas em cada BSs entre as nuvens. A matriz de posicionamento $d_{f,n,a} \in \{0, 1\}$, *i.e.,* $d_{f,n,a} = 1$ indica que f foi posicionada sobre n para processar o sinal de a. Caso contrário, *i.e.,* $d_{f,n,a} = 0$, representa que f da BS a não está sendo processada em n. Finalmente, $D_B(f,n,a)$ determina a taxa de dados necessária para processar f da BS a em uma determinada nuvem n. A equação $D_B(f,n,a)$ considera o tipo de f a ser processada, a distância da nuvem n (*i.e.,* borda, regionais ou centrais) e a demanda de UEs na BS a, para determinar a quantidade de dados transmitidos no *fronthaul*. Maiores informações sobre o cálculo da quantidade de dados transmitidos, podem ser encontrados em Wubben, et al. [Wubben et al. 2014].

Baseado nas definições acima, o problema de minimização do total de dados transmitidos no *fronthaul*, através do posicionamento dinâmico das funções de rádio virtualizadas e suas restrições pode ser formulada como:

$$\min_{\{\mathbf{d}\}} \quad O_B(d) = \min_{\{\mathbf{d}\}} \sum_{f,n,a} d_{f,n,a} D_B(f,n,a) \tag{2}$$

s.t.

$$\sum_{n} d_{f,n,a} = 1, \qquad \forall f, \forall a, \tag{3}$$

$$\sum_{f,a} d_{f,n,a} D_V(f,n,a) \leq C_V^n, \qquad \forall n, \tag{4}$$

$$\sum_{f,a} d_{f,n,a} D_B(f,n,a) \leq C_B^n, \qquad \forall n, \tag{5}$$

onde, $D_V(f,n,a)$ determina o consumo dos recursos de nuvem necessários para processar f da BS a em uma determinada nuvem n. C_V^n e C_B^n representam os limites de VDUs e de banda do *fronthaul* da nuvem n, respectivamente.

As restrições (3), (4) e (5) garantem a factibilidade da solução perante o ambiente de uma *Dynamic* C-RAN. A restrição (3) garante que para toda e qualquer BS $a \in A$ exista obrigatoriamente uma função de rádio virtualizada $f \in F$. Além disso, a restrição garante que nenhuma f de uma mesma a seja duplicada dentre todas as nuvens possíveis $n \in N$. A restrição (4) garante que o posicionamento de uma função de rádio virtualizada f de uma BS a seja realizada somente em uma nuvem n, com capacidade de processamento, *i.e.,* a nuvem deve conter VDUs suficientes para atender todas as funções atribuídas. Por fim, a restrição (5) garante que o posicionamento de uma função f de uma BS a ocorra obrigatoriamente em uma nuvem n com capacidade de transmissão, *i.e.,* uma nuvem só pode receber f caso o *fronthaul* possua banda suficiente para atender os requisitos de transmissão de todas as funções virtualizadas nessa nuvem.

3.2. Programação linear dos recursos em *Dynamic* C-RAN utilizando Maestro

O diferencial proposto pelo orquestrador *Maestro* em relação aos trabalhos anteriores é a capacidade de posicionar VNFCs em diferentes VDUs, sobre diferentes nuvens, quando as funções virtualizadas possuem mais que uma opção de divisão. Sendo assim, o *Maestro* pode posicionar os VNFCs de várias maneiras diferentes, possibilitando ganhos na transmissão de dados e na maximização do uso dos recursos computacionais nas nuvens de borda. Desta forma, o orquestrador recebe um conjunto de informações sobre a infraestrutura de rede, definidas pelo operador, e sua utilização, para determinar o posicionamento de cada função virtualizada para todas as BSs dentre as nuvens existentes.

A principal função do *Maestro* é minimizar a taxa de dados transmitidos no *fronthaul*, adotando o modelo descrito na Subseção 3.1. Para acelerar a resolução do problema linear, um novo elemento objetivo é inserido em (2), para maximizar a quantidade de recursos utilizados nas nuvens de borda, uma vez que estas nuvens não transmitem dados no *fronthaul* e devem ser priorizas pelas soluções. Desta maneira, insere-se uma ponderação na função objetivo, afim de atingir os objetivos propostos. Tal minimização ponderada é descrita como:

$$\begin{aligned} \min_{\{\mathbf{d}\}} \quad & \omega \cdot \sum_{f,n,a} d_{f,n,a} D_B(f,n,a) + (\omega - 1) \cdot \sum_{f,a} d_{f,0,a} D_V(f,n,a) \\ \text{s.t.} \quad & (3)(4)(5) \end{aligned} \tag{6}$$

onde, $0 \leq \omega \leq 1$ realiza a ponderação das funções objetivos para resolução do problema e $d_{f,0,a} D_V(f,n,a)$ da prioridade das nuvens de borda no posicionamento das funções virtualizadas. A segunda parte de (6) tem por finalidade maximizar a ocupação de recursos nas nuvens de borda e acelerar a convergência para a solução. A maximização das nuvens de borda e a minimização da taxa de dados transmitidos no *fronthaul* fazem com que os algoritmos consigam obter um melhor desempenho quanto ao tempo, através do direcionamento das possibilidades de posicionamento das VNFs, durante sua execução.

O *Maestro* garante que os componentes de funções de rádio não sejam implantados na mesma nuvem para a mesma BS, que não se exceda a capacidade computacional das nuvens e a capacidade de transmissão disponível no *fronthaul*, respeitando as restrições descritas anteriormente. O *Maestro* foi implementado utilizando o *IBM CPLEX* [IBM Software 2010], com a função objetivo descrita em (6). Na Seção 4 é apresentado o estudo de caso e os experimentos utilizados para avaliação do *Maestro*.

4. Estudo de Caso e Experimentos

O estudo de caso e os experimentos apresentados nesta seção visam demonstrar o desempenho do orquestrador *Maestro* em posicionar as composições de funções mais apropriadas para a implantação dos componentes de VNFs. Baseado em uma infraestrutura hierárquica, modelou-se um conjunto de VDUs, juntamente com um conjunto de BSs, que são descritos a seguir.

4.1. Estudo de Caso

Este estudo de caso é baseado nas opções de divisão de funções de rádio, descritas por Wubben, et al. [Wubben et al. 2014] e aplicadas no ambiente de *Dynamic* C-RAN. A abordagem proposta por Wubben, et al. prevê a execução distribuída de cinco funções de rádio, a saber: *MAC*, *Soft-Bit*, *RX Data*, *Subframe* e *I/Q*. Essa abordagem objetiva permitir um posicionamento mais flexível dos recursos de computação e de rede necessários para processar o sinal de rádio de forma distribuída, através do uso de virtualização. Maiores detalhes sobre a operação de cada divisão e das próprias funções de rádio são descritas em Wubben, et al. [Wubben et al. 2014]. A decisão acerca da utilização de cada opção de divisão gera uma variabilidade na taxa de dados transmitidos no *fronthaul* e de recursos de computação nas nuvens, o que torna essa abordagem interessante para este estudo de caso. Considerando essa abordagem, é possível mostrar como a orquestração de componentes de VNFs causa impacto no desempenho da rede e na utilização de recursos.

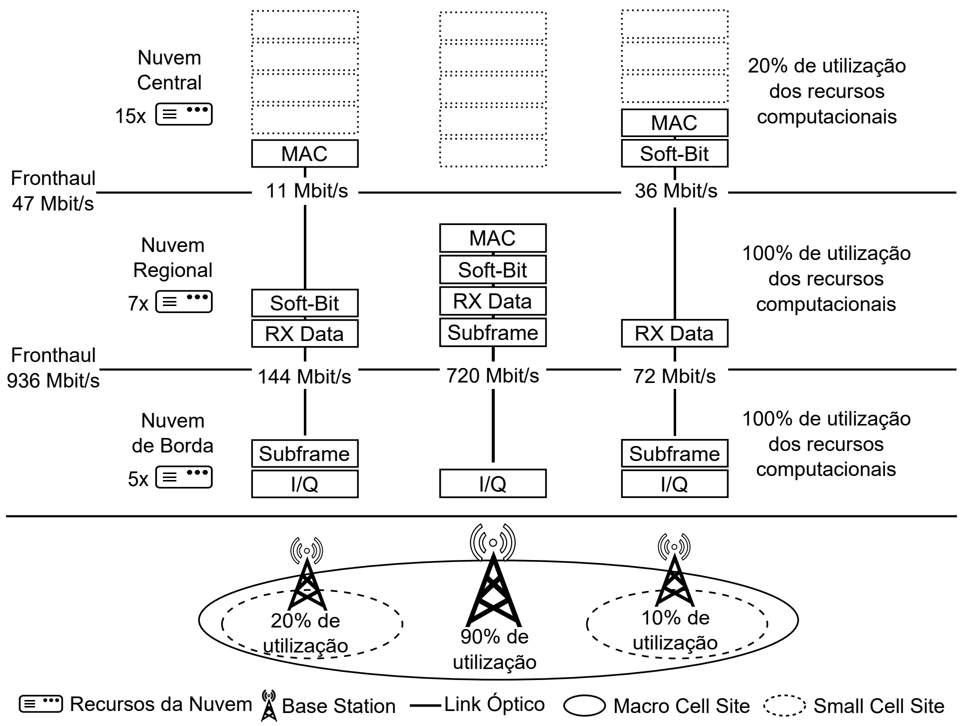


Figura 1. Exemplo de divisões de funções de rádio implantadas como VNFs

Na Figura 1 pode-se observar um exemplo de posicionamento das funções de rádio em componentes de VNFs e as suas respectivas distribuições ao longo dos níveis de nuvens disponíveis (*i.e.*, borda, regional e central). No exemplo, cada VNF implementa uma BS, sendo que cada uma das cinco funções de rádio são encapsuladas em um VNFC, que por sua vez são implantados em até 3 VDUs. As BSs podem ser visualizadas na parte inferior da Figura 1, onde cada uma delas possui uma porcentagem de ocupação diferente em função da capacidade total de transmissão. Essa variação de ocupação influencia o impacto das opções de divisão selecionadas em termos de dados transmitidos no *fronthaul*. No exemplo, a *macro cell* na parte inferior central, tendo ocupados 90% de sua capacidade, foi implantada com 2 VDUs posicionados na borda e na nuvem regional, consumindo uma taxa fixa de 720Mbit/s para comunicação entre as funções de rádio I/Q e *Subframe*. Já para as *small cells*, tendo ocupações de apenas 10% e 20%, optou-se por utilizar 3 VDUs posicionando as funções acima do RX Data – que têm seus requisitos de banda minimizados pelas baixas taxa de ocupação – nas nuvens regional e central.

A lógica por trás do exemplo da Figura 1 é de tentar posicionar os VDUs de forma a minimizar o total de dados transmitidos no *fronthaul*. Consequentemente, a nuvem de borda, que fica mais próxima da RAN, é prioritariamente ocupada com funções de rádio de nível mais baixo que são aquelas que apresentam requisitos mais estritos de comunicação. A nuvem regional, por sua vez, possui maior quantidade de recursos computacionais em relação a nuvem de borda, porém encontra-se mais distante da RAN. Sendo assim, as funções de rádio implantadas na nuvem regional enfrentam atrasos de comunicação para transporte dos dados até esse local. Finalmente, a nuvem central é a nuvem que possui maior quantidade de recursos disponíveis, porém é também aquela que está mais distante da RAN. Essa nuvem se torna mais atraente para o processamento de funções de nível mais alto (*e.g.*, Soft-Bit ou MAC), principalmente para BSs com menor ocupação. Vale salientar que, por medida de simplificação, nesse exemplo considera-se que cada função de rádio encapsulada em um VNFC consome uma unidade de processamento de uma nuvem onde é posicionada. Em um ambiente real, a demanda de processamento potencialmente seria variável em função de fatores como a ocupação das BSs, por exemplo.

Sabendo que os dados transmitidos no *fronthaul* variam de acordo com a ocupação das BSs, para realizar os experimentos foi necessário estimar os percentuais de ocupação com base em uma simulação utilizando as recomendações do *3rd Generation Partnership Project* (3GPP), do documento *Further advancements for E-UTRA physical layer aspects* [3GPP 2010]. Esse percentual de ocupação foi calculado através do método de *Monte Carlo*, em uma simulação do tráfego gerado por UEs associados a uma infraestrutura com 7 *macro cells* e 15 *small cells* distribuídas dentro da área de cobertura de cada *macro cell*, totalizando 112 BSs. Em média, 60 usuários foram distribuídos aleatoriamente dentro da área de cobertura de cada *macro cell* e associados a BS mais próxima de cada usuário (seja uma *macro* ou *small cell*). A distribuição das BSs e UEs foi realizada em uma área de 2000x2000 metros, seguindo as normas para simulação definidas no documento *Small cell enhancements for E-UTRA and E-UTRAN - Physical layer aspects* do 3GPP [3GPP 2013] para uma rede celular, considerando a existência de áreas com diferentes densidades de BS e UEs *i.e.,* zonas rurais e urbanas.

Na Tabela 1 pode-se encontrar a ocupação das BSs utilizadas na simulação, expressa através de uma distribuição discreta de probabilidades, servindo como entrada para o modelo linear detalhado na Seção 3. A ocupação das BSs influencia a taxa de dados transmitidos para as funções de *MAC, Soft-Bit, RX Data*, tal como é descrito por Wubben et al. [Wubben et al. 2014]. A quantidade de *macro cells* consideradas na simulação cor-

Tabela 1. Distribuição da Ocupação das *macro* e *small cells*

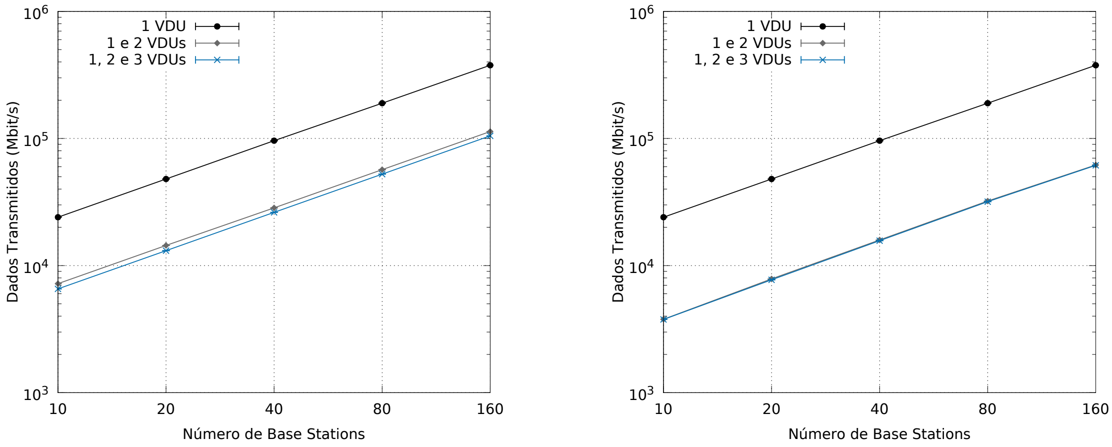
	Macro cells		*Small cells*	
Índice de Ocupação	Carga de Ocupação	Probabilidade de Ocupação	Carga de Ocupação	Probabilidade de Ocupação
0	15%	0,42%	2,5%	56,97%
1	25%	1,42%	12,5%	31,59%
2	35%	5,14%	22,5%	9,28%
3	45%	10,28%	37,5%	1,88%
4	65%	15,71%	47,5%	0,28%
5	75%	19,17%	-	-
6	85%	17,29%	-	-
7	95%	30,57%	-	-
	Total	100%	-	100%

responde a 33,3% da infraestrutura, enquanto que as *small cells* representam 66,7%. Essa medida se deve à quantidade de UEs distribuídos ao longo da região geográfica, sendo maior para as *macro cells* e menor para as *small cells*. Os cenários de avaliação possuem 10, 20, 40, 80 e 160 BSs ao longo da infraestrutura. Para esse trabalho, considerou-se que cada um dos três níveis hierárquicos de nuvem possui respectivamente 5, 7 e 15 unidades de processamento, como pode ser observado na Figura 1, aumentando essas unidades proporcionalmente à quantidade de BSs em cada cenário. Baseado os cenários de simulação apresentados, na Subseção 4.2 os resultados obtidos são descritos e analisados.

4.2. Resultados

O modelo de programação linear foi implementado com *IBM CPLEX Optimizer* [IBM Software 2010], utilizando 8 *threads* e o mesmo algoritmo *Dual Simplex* para todos os cenários. A simulação foi executada em um computador com um processador Intel i7 de 3.1 GHz, com 16 GB de memória RAM, executando sobre um sistema operacional Linux de 64 *bits*. Para todos os cenários, considerou-se três nuvens (*i.e.*, de borda, regional e central), com capacidade proporcional para acomodar a carga de trabalho nos diferentes cenários, tal como descrito na Subseção 4.1. Todos os experimentos foram repetidos 50 vezes e o tempo para posicionar os VNFCs ao longo dos VDUs, foi medido pelo *IBM CPLEX*. Neste artigo, considerou-se tanto a ocupação fixa das BSs no posicionamento das funções de rádio, quanto ocupação variável (conforme a Tabela 1), diferentemente da avaliação realizada no trabalho anterior [Dalla-Costa et al. 2017]. Desta forma, é possível comparar a quantidade de tempo necessária para efetuar o posicionamento das funções de rádio em ambos os casos.

Pode-se observar na Figura 2(a) a variação total de dados transmitidos no *fronthaul* para diferentes opções de divisão de funções de rádio, considerando que as BSs tenham ocupação percentual fixa em 50%. A Figura 2(b) apresenta o mesmo resultado, porém considerando ocupação variável das BSs (aleatoriamente atribuídos conforme a Tabela 1). O eixo x denota a quantidade de BSs em cada experimento, enquanto o eixo y (em escala logaritmica) denota os dados transmitidos no *fronthaul* em Mbit/s ocasionado pelas opções de divisão selecionadas pela função de posicionamento. A curva *1 VDU* representa a transmissão dos dados considerando uma VNF atômica, utilizando apenas um VDU contendo todos os VNFCs. Já a curva *1 e 2 VDUs* expressa a transmissão dos dados



(a) Percentual de ocupação fixa das BSs

(b) Percentual de ocupação variável das BSs

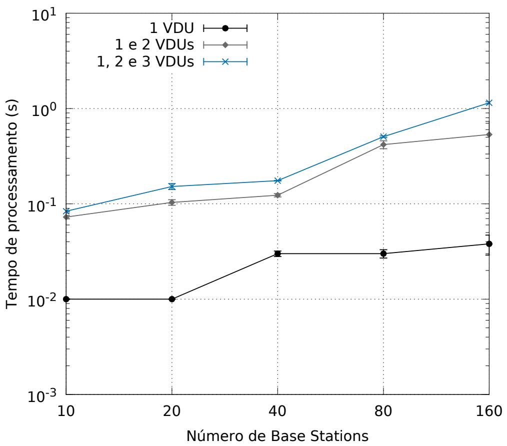
Figura 2. Transmissão de dados no *fronthaul*

quando VNFs são posicionadas em um ou dois VDUs (*i.e.*, podendo se utilizar uma opção de divisão para separar VNFCs em até dois VDUs). Por fim, a curva *1, 2 e 3 VDUs* representa os dados transmitidos de uma VNF implantada em até 3 VDUs, similar ao que é demonstrado na Figura 1 (*i.e.*, podendo usar simultaneamente até duas opções de divisão para separar os VNFCs de uma mesma VNF).

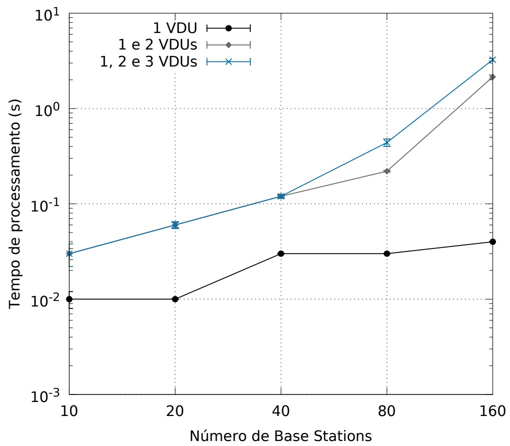
Observando o gráfico de dados transmitidos com ocupação fixa, para 10 BSs a quantidade de dados transmitido utilizando apenas 1 VDU foi de 24 Gbit/s, enquanto que até 2 e até 3 VDUs esse valor é reduzido para 7,2 Gbit/s e 6,5 Gbit/s, respectivamente. Percentualmente, houve uma redução de 70% entre os cenários utilizando 1 VDU versus 2 VDUs e de 72,7% comparando 1 e 3 VDUs. Entre os cenários de até 2 e até 3 VDUs a diferença foi apenas de 9%. Já no caso mais extremo, considerando 160 BSs, a banda consumida utilizando 1 VDU atingiu 376,8 Gbit/s, enquanto que para até 2 VDUs foi de 113,4 Gbit/s e para até 3 VDUs foi de 104,83 Gbit/s. Assim, manteve-se aproximadamente os mesmos preceituais de redução entre as curvas, enfatizando a importância da composição de VNFs no uso racional dos recursos do *fronthaul*.

Ao observar o gráfico de dados transmitidos com ocupação variável para 10 BSs, a quantidade de dados transmitidos para 1 VDU foi dos mesmos 24 Gbit/s obtidos com ocupação fixa. Isso se deve ao fato de que no posicionamento de 1 VDU pode-se utilizar apenas a opção de divisão da função de rádio I/Q, que requer uma taxa de transferência fixa independente da ocupação da BS. Ainda considerando 10 BSs, as curvas de até 2 VDUs e até 3 VDUs obtiveram taxas de transmissão de dados na ordem de 3,8 Gbit/s e 3,7 Gbit/s, respectivamente, atingindo reduções de aproximadamente 83% em comparação com a curva de 1 VDU. Na comparação dos resultados entre as curvas de 2 VDUs e 3 VDUS, os valores se mantiveram sempre muito similares para todas as quantidade de BSs simuladas com diferenças sempre abaixo de 1%.

Neste trabalho, diferentemente dos trabalhos anteriores, compara-se o desempenho do posicionamento de BSs considerando ocupação fixa e variável em ambientes de *Dynamic* C-RAN. Pode-se observar que tanto na ocupação fixa como na variável, o orquestrador é capaz de minimizar a transmissão de dados no *fronthaul*, quando existir a possibilidade de selecionar as divisões de funções de rádio com 2 ou 3 VDUs. Essa redução é ainda mais evidente quando há demanda variável nas BSs, o que se deve principalmente a dois fatores: (*i*) existe uma diferença de ocupação média significativa entre *macro* e *small cells*, o que fornece ao orquestrador uma gama de opções de VDUs com transmissão de dados diferentes para posicionar; e (*ii*) apesar da alta ocupação nas *macro cells* (em média 75%), há um grande número de *small cells* com baixa ocupação (em média 8%), o que reduz a ocupação média por BS para a casa dos 30%. A utilização de até 3 VDUs, ainda que acarrete ganhos em relação a até 2 VDUs, traz vantagens muito tímidas no cenário de ocupação variável, o que coloca em dúvida se de fato o vale esforço extra para calcular posicionamentos mais complexos para obter ganhos pouco significativos.



(a) Solução na ocupação fixa



(b) Solução na ocupação variável

Figura 3. Tempo necessário para encontrar uma solução

As Figuras 3(a) e 3(b) ilustram o tempo médio para computar o posicionamento de todos os VNFCs sobre a infraestrutura de nuvens disponível, considerando as BSs com ocupação fixa e variável. As barras de erro representam um intervalo de confiança de 99%, obtido por meio de 50 repetições para cada experimento. O eixo x expressa o número de BSs no cenário, enquanto o eixo y denota o tempo médio em segundos necessário para calcular o posicionamento dos VNFCs.

Os gráficos de tempo demonstram que para até 80 BSs, o tempo necessário para realizar o posicionamento com ocupação fixa é superior ao tempo de ocupação variável. Isso se justifica pois no cenário de ocupação fixa os recursos se esgotam mais rapidamente, de forma que o orquestrador testa mais possibilidades para alocar os VDUs de forma eficiente. Entretanto, para 160 BSs o cenário de ocupação variável consome mais tempo (3,19s para até 3 VDUs, contra 1,12s da ocupação fixa) do que a ocupação fixa. Fato este que pode ser explicado pela grande quantidade de combinações de VDUs com demandas de banda diferentes disponíveis para posicionamento pelo orquestrador.

No cenário de ocupação variável, para realizar o posicionamento de até 40 BSs utilizando 2 ou 3 VDUs, os tempos de cálculo são extremamente similares. Isso torna viável a utilização de divisões de funções de rádio em até 3 VDUs, uma vez que isso leva a uma redução, ainda que pequena, no consumo de recursos do *fronthaul*. Já para 80 e 160

BSs, o tempo necessário para realizar o posicionamento utilizando até 3 VDUs, ocasiona um aumento de mais de aproximadamente 40% no tempo médio de calculo em relação a até 2 VDUs. Desta forma, pode-se concluir que em um ambiente com uma grande quantidade de BSs, onde a variação da ocupação for frequente, a utilização de até 2 VDUs pode ser mais adequada para permitir que o orquestrador responda mais rapidamente às mudanças do ambiente sem prejudicar significativamente o consumo de recursos.

Finalmente, vale ressaltar que não está sendo considerado neste trabalho o tempo necessário para a implantação/migração dos componentes de VNFs entre as nuvens, o que, em uma infraestrutura real, deverá influenciar no tempo total necessário para que a nova configuração calculada pelo orquestrador entre em vigor. Ademais, os tempos de resposta da modelagem linear sempre são dados até a execução total do posicionamento dos VDUs, de acordo com a função objetivo definida na Seção 3.2. Nesse caso, se essa função for modificada, os tempos de posicionamento serão diferentes.

5. Conclusão e Trabalhos Futuros

Neste trabalho, foi realizada uma avaliação do *Maestro*, um orquestrador para ambientes *Dynamic* C-RAN, que baseia-se nos componentes de funções virtualizadas para realizar o posicionamento das VNFs. Seu mecanismo de decisão foi baseado na modelagem de posicionamento de VNFs e em programação linear dos recursos em *Dynamic* C-RAN. A avaliação foi realizada baseada em simulação de um caso de uso, considerando as especificações de uma rede celular do 3GPP. O *Maestro* cumpre com o objetivo proposto, reduzindo a taxa de dados transmitida no *fronthaul* e maximizando a quantidade de recursos utilizados nas nuvens de borda e regionais, tentando maximizar o uso dos recursos computacionais nas nuvens de borda. Ademais, o *Maestro* fornece uma visão geral dos ganhos que podem ser adquiridos através da divisão de funções de rádio em diferentes nuvens e VDUs.

Além disso, pode-se concluir que se o mecanismo de decisão necessita ser executado a cada 3,19 segundos (pior tempo para 160 BSs), consegue-se atender os requisitos do ambiente de uma rede *Dynamic* C-RAN, encontrando uma solução de posicionamento de todos os componentes. Vale ressaltar que quando implantado em uma rede física, faz-se necessário considerar outros parâmetros, tal como a quantidade de tempo para a migração de uma VNF de um VDU para outro, e desta forma, o mecanismo de decisão pode ter seu tempo para calcular o posicionamento, alterado.

Como trabalho futuro pretende-se implantar o *Maestro* em uma infraestrutura real, conduzindo experimentação de forma real, utilizando elementos do *Maestro* para realizar o posicionamento dos componentes ao longo da rede. Além disso, pretende-se adaptar o orquestrador *Maestro* para outros tipos de rede, como por exemplo, para redes ópticas. Outro caminho a seguir, é que o *Maestro* possa considerar fatores (*i.e.* políticas) durante o processo de decisão, de acordo com regras pré-definidas por operadores de rede.

Agradecimentos

Este trabalho foi financiado pelo programa Horizon 2020 da União Europeia para pesquisa, desenvolvimento tecnológico e demonstração no âmbito do acordo nº. 688941 (FUTEBOL), bem como pelo Ministério da Ciência, Tecnologia, Inovação e Comunicação (MCTIC) através da RNP/CTIC.

Referências

3GPP (2010). Further advancements for E-UTRA physical layer aspects; Evolved Universal Terrestrial Radio Access - TR 36.814. Technical report. Disponível em: $www.qtc.jp/3GPP/Specs/36814-900.pdf$.

3GPP (2013). Small cell enhancements for E-UTRA and E-UTRAN - Physical layer aspects; Technical Specification Group Radio Access Network - TR 36.872. Technical report. Disponível em: $www.3gpp.org/ftp/Specs/archive/36\_series/36.872/36872-c10.zip$.

5G-PPP-Working-Group (2016). View on 5G Architecture. Technical report, 5G PPP.

Abdelwahab, S., Hamdaoui, B., Guizani, M., e Znati, T. (2016). Network function virtualization in 5G. *IEEE Communications Magazine*, 54(4):84–91.

Bartelt, J., Rost, P., Wubben, D., Lessmann, J., Melis, B., e Fettweis, G. (2015). Fronthaul and backhaul requirements of flexibly centralized radio access networks. *IEEE Wireless Communications*, 22(5):105–111.

Bonomi, F., Milito, R., Zhu, J., e Addepalli, S. (2012). Fog Computing and Its Role in the Internet of Things. In *SIGCOMM Workshop on Mobile Cloud Computing (MCC)*, MCC '12, pages 13–16, New York. ACM.

Checko, A., Christiansen, H. L., Yan, Y., Scolari, L., Kardaras, G., Berger, M. S., e Dittmann, L. (2015). Cloud RAN for Mobile Networks - A Technology Overview. *IEEE Communications Surveys Tutorials*, 17(1):405–426.

Clayman, S., Maini, E., Galis, A., Manzalini, A., e Mazzocca, N. (2014). The dynamic placement of virtual network functions. In *IEEE Network Operations and Management Symposium (NOMS)*, pages 1–9.

Dalla-Costa, A. G., Bondan, L., Wickboldt, J. A., Both, C. B., e Granville, L. Z. (2017). Maestro: An nfv orchestrator for wireless environments aware of vnf internal compositions (to appear). *IEEE International Conference on. Advanced Information Networking and Applications*.

Giotis, K., Kryftis, Y., e Maglaris, V. (2015). Policy-based orchestration of NFV services in Software-Defined Networks. In *Network Softwarization (NetSoft), 2015 1st IEEE Conference on*, pages 1–5.

Heideker, A. e Kamienski, C. (2016). Gerenciamento flexível de infraestrutura de acesso público à internet com nfv. In *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*.

IBM Software (2010). Efficient modeling with the IBM ILOG CPLEX Optimization Studio. Technical Report WSW14059-USEN-02, ILOG Optimization and Analytical Decision Support Solutions, Somers, NY. White Paper.

Li, X. e Qian, C. (2016). A survey of network function placement. In *2016 13th IEEE Annual Consumer Communications Networking Conference (CCNC)*, pages 948–953.

Liu, J., Zhou, S., Gong, J., Niu, Z., e Xu, S. (2015). Graph-based framework for flexible baseband function splitting and placement in C-RAN. In *IEEE International Conference on Communications (ICC)*, pages 1958–1963.

Luizelli, M. C., Bays, L. R., Buriol, L. S., Barcellos, M. P., e Gaspary, L. P. (2015). Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions. In *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 98–106.

Makaya, C., Freimuth, D., Wood, D., e Calo, S. (2015). Policy-based NFV management and orchestration. In *Network Function Virtualization and Software Defined Network (NFV-SDN), 2015 IEEE Conference on*, pages 128–134.

Mijumbi, R., Serrat, J., l. Gorricho, J., Latre, S., Charalambides, M., e Lopez, D. (2016). Management and orchestration challenges in network functions virtualization. *IEEE Communications Magazine*, 54(1):98–105.

Moens, H. e Turck, F. D. (2014). VNF-P: A model for efficient placement of virtualized network functions. In *10th International Conference on Network and Service Management (CNSM) and Workshop*, pages 418–423.

Ortín, J., Caballero, P., IMDEA, Rost, P., e NEC (2014). D5.2 final definition of ijoin requirements and scenarios. Technical report, INFSO-ICT-317941 iJOIN, http://www.ict-ijoin.eu/wp-content/uploads/2012/10/D5.2.pdf.

Peng, M., Li, Y., Jiang, J., Li, J., e Wang, C. (2014). Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies. *IEEE Wireless Communications*, 21(6):126–135.

Quittek, J. et al. (2014). Network Functions Virtualisation (NFV) - Management and Orchestration. White paper, ETSI NFV ISG.

Riggio, R., Rasheed, T., e Narayanan, R. (2015). Virtual network functions orchestration in enterprise WLANs. In *IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 1220–1225.

Soares, J., Dias, M., Carapinha, J., Parreira, B., e Sargento, S. (2014). Cloud4NFV: A platform for Virtual Network Functions. In *IEEE International Conference on Cloud Networking (CloudNet)*, pages 288–293.

Wubben, D., Rost, P., Bartelt, J. S., Lalam, M., Savin, V., Gorgoglione, M., Dekorsy, A., e Fettweis, G. (2014). Benefits and Impact of Cloud Computing on 5G Signal Processing: Flexible centralization through cloud-RAN. *IEEE Signal Processing Magazine*, 31(6):35–44.

Xilouris, G., Trouva, E., Lobillo, F., Soares, J. M., Carapinha, J., McGrath, M. J., Gardikis, G., Paglierani, P., Pallis, E., Zuccaro, L., Rebahi, Y., e Kourtis, A. (2014). T-NOVA: A marketplace for virtualized network functions. In *European Conference on Networks and Communications (EuCNC)*, pages 1–5.

AppendixC PUBLISHED JOURNAL – COMPUTER NETWORKS 2015

Marcelo Antonio Marotta, Leonardo Roveda Faganello, Matias Artur Klafke Schimuneck Lisandro Zambenedetti Granville, Juergen Rochol, Cristiano Bonato Both. **Managing mobile cloud computing considering objective and subjective perspectives**. Computer Networks - Cloud Networking and Communications II, Volume 93, Part 3, 24 December 2015, Pages 531–542.

- **Title:** *Managing mobile cloud computing considering objective and subjective perspectives.*
- **Contribution:** An management model and architecture for mobile cloud computing, exploiting both objective and subjective perspectives.
- **Abstract:** Mobile Cloud Computing enables mobile devices to augment constrained resources such as processing, storage, and battery autonomy by using the cloud infrastructure. As the network is a key element in integrating mobile devices to the cloud, a proper management of the mobile cloud computing environment is necessary. Such a management must take into account two main perspectives: administrator's and end-user's perspectives. The administrator is usually concerned with a more objective perspective based on Quality of Service parameters, such as throughput, delay, and jitter. On the other hand, the end-user has a more subjective perspective, observing his/her Quality of Experience when using a mobile cloud application or service. In this article, we introduce a management model and architecture for mobile cloud computing, exploiting both objective and subjective perspectives. As a proof of concept, we prototyped the architecture in a management system called CoLisEU, which allowed us to investigate this architecture and we discuss the benefits of the combined objective and subjective perspectives in our management architecture.
- **Status:** Published.
- **Qualis:** A1.
- **Jornal:** Computer Networks.

- **Date:** 24 December 2015.
- **URL:** <http://dl.acm.org/citation.cfm?id=2870023>.
- **Digital Object Identifier (DOI):** <10.1016/j.comnet.2015.09.040>.

Computer Networks 93 (2015) 531–542

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Managing mobile cloud computing considering objective and subjective perspectives

Marcelo Antonio Marotta[a], Leonardo Roveda Faganello[a], Matias Artur Klafke Schimuneck[a], Lisandro Zambenedetti Granville[a], Juergen Rochol[a], Cristiano Bonato Both[b,*]

[a] Institute of Informatics – Federal University of Rio Grande do Sul Av. Bento Gonçalves, 9500 – Porto Alegre, RS - Brazil
[b] Federal University of Health Sciences of Porto Alegre, Brazil Av. Independência, 2293 – Santa Cruz do Sul - RS / Brasil

ARTICLE INFO

Article history:
Received 16 March 2015
Revised 10 August 2015
Accepted 19 September 2015
Available online 24 October 2015

Keywords:
Mobile cloud computing
Network management
QoE
QoS

ABSTRACT

Mobile Cloud Computing enables mobile devices to augment constrained resources such as processing, storage, and battery autonomy by using the cloud infrastructure. As the network is a key element in integrating mobile devices to the cloud, a proper management of the mobile cloud computing environment is necessary. Such a management must take into account two main perspectives: administrator's and end-user's perspectives. The administrator is usually concerned with a more objective perspective based on Quality of Service parameters, such as throughput, delay, and jitter. On the other hand, the end-user has a more subjective perspective, observing his/her Quality of Experience when using a mobile cloud application or service. In this article, we introduce a management model and architecture for mobile cloud computing, exploiting both objective and subjective perspectives. As a proof of concept, we prototyped the architecture in a management system called CoLisEU, which allowed us to investigate this architecture and we discuss the benefits of the combined objective and subjective perspectives in our management architecture.

1. Introduction

Mobile Cloud Computing (MCC) is the paradigm that integrates mobile devices and cloud computing. MCC allows the augmentation of mobile devices' constrained resources, such as processing, storage, and battery autonomy by using cloud services [1]. Counting on augmented resources, mobile devices can execute more sophisticated versions of key applications and services, such as mobile learning, e-commerce, and healthcare [2]. This is possible because tasks in MCC are partially performed in mobile devices and partially computed in the cloud. The network that interconnects mobile devices and the cloud largely impacts the proper execution of MCC applications and services. Besides, this network also allows MCC to perform a second augmentation, now in terms of coverage, extending the cloud towards the edge of the network, where the mobile devices and End-users lie [3]. In this article, we address the important issue of managing the MCC environment, focusing on the networking aspects. Such a management must take into account two main perspectives: the network Administrator's perspective and the End-user's perspective.

The Administrator is usually concerned with Quality of Service (QoS) parameters (*e.g.*, throughput, delay, and jitter [4]), which provide an *objective* view about the quality of the network. By observing QoS-related measurements, the Administrator can tune the operation of underlying

* Corresponding author. Tel.: +55 5193080875.
E-mail addresses: mamarotta@inf.ufrgs.br (M.A. Marotta), lrfaganello@inf.ufrgs.br (L.R. Faganello), makschimuneck@inf.ufrgs.br (M.A.K. Schimuneck), granville@inf.ufrgs.br (L.Z. Granville), juergen@inf.ufrgs.br (J. Rochol), cbboth@ufcspa.edu.br, cbboth@inf.ufrgs.br (C.B. Both).

http://dx.doi.org/10.1016/j.comnet.2015.09.040
1389-1286/

infrastructures, including wireless infrastructure, backhaul, backbone, provider infrastructure, and cloud infrastructure [1]. The End-user, in turn, tends to neglect QoS parameters and focus instead on his/her experience when using an MCC application or service. This experience can be measured by observing Quality of Experience (QoE) parameters, such as satisfaction level with application navigation, response time, and cloud ubiquity [5]. Since each End-user has a personal experience, QoE parameters provide a *subjective* view about MCC applications and services. Although different to the objective view obtained from QoS observation, the subjective, QoE-based observation of MCC is also important because it is directly related to the applications and services consumed by the End-user.

The combined observation of QoS and QoE parameters are fundamental to the management of MCC. However, such management is still underexploited by the research community. Some preliminary investigations point out that QoS and QoE can serve as guiding paradigms [6,7] for improving MCC management [8]. Other investigations focus on proposing architectural models for MCC management that are centered on the End-user [9,10] or on cloud applications [2], but neglect QoE parameters. Therefore, there is a need for concrete architectural models that explicitly take into account both QoS and QoE together.

In this article, we investigate the management of MCC through an architecture that considers both objective (QoS-based) and subjective (QoE-based) perspectives. In Section 2, we review MCC basic concepts, as well as the state-of-the-art on MCC management. Afterwards, we highlight five key requirements for managing the MCC environment. Next, a mapping of traditional management entities (*i.e.,* agent, gateway, and management application) to the MCC environment is introduced. In Section 3, we describe a management architecture that explicitly takes into account both QoS and QoE. As a proof of concept, in Section 4, we prototyped the proposed architecture in a management system called CoLiSeU,[1] whose details are presented in the article as well. To evaluate our prototype, in Section 5, we analyze one thousand samples provided by End-users about their satisfaction using a wireless infrastructure to request cloud services. Finally, the benefits of the combined analysis of QoS and QoE in our MCC management architecture are summarized in the concluding section.

2. Mobile Cloud Computing

In this section, we present the components of an MCC infrastructure. Next, we present the state-of-the-art of the MCC management. Afterwards, we detail key management requirements considering the MCC environment.

2.1. MCC infrastructure

MCC enables the augmentation of computing resources of mobile devices in an environment composed of the following infrastructures: (*i*) *wireless infrastructure*, (*ii*) *backhaul*, (*iii*) *backbone*, (*iv*) *provider infrastructure*, and (*v*) *cloud infrastructure*. These components are depicted in Fig. 1. The *End-user*, through his/her *Mobile Device* establishes communication with a *Base Station* or an *Access Point*, requesting a resource (*e.g.,* storage or processing) augmentation from the cloud. This request may be based on different wireless technologies, such as LTE or WiFi. After reaching the *Base Station/Access Point*, the request is forwarded through the *Backhaul* to an Internet Service Provider (ISP).

The *Backbone* routes the request along different ISPs. When the request reaches the destination ISP, it accesses the *Provider Infrastructure*. This infrastructure is similar to the *Backhaul*, only with typically larger capacity links. The target Cloud then receives the request and allocates resources inside the *Cloud Infrastructure*. In the Cloud, virtual nodes communicate with one another through virtual links. These links are created by the Cloud administrators as an abstraction of the real network links with specific characteristics, such as capacity, routing protocol, and virtual node endpoints. Finally, the Cloud provides the requested resources, replying to the *Mobile Device* across the five infrastructures.

Each infrastructure presents different problems, such as intermittent wireless signal, insufficient coverage areas, traffic overhead, non-optimal service level agreements, and virtual link misconfiguration [11]. These problems may degrade both QoS and QoE. To try to avoid this degradation, MCC management must monitor each infrastructure to detect these problems. When problems are detected, MCC management reconfigures and tunes the internal components of the five infrastructures, in order to lead the network to a state that again satisfies the *End-user* expectations, thus keeping QoE at satisfactory levels.

2.2. MCC management

MCC extends from the *End-user* to the *Cloud Infrastructure*. Therefore it is important that the MCC management strategy takes into account the perspectives of both the *Administrator* and *End-user*. The *Administrator* has a more objective view of the managed network, typically observing QoS parameters to detect specific problems, such as link bottlenecks or network node inactivity [4]. However, solely relying on QoS parameters may cause Administrators to be unaware of whether the *End-user* expectations are fulfilled (*i.e.,* QoE). As opposed to Administrators, *End-users* have a more subjective perspective of the network, based on QoE. As such, aspects that escape the Administrator's objective analysis (*e.g.,* satisfaction level, network efficiency, and Cloud ubiquity), end up being confined to the End-user's point-of-view. That, however, is inadequate and QoE must be observed too. On the opposite side, QoE evaluation alone is not able to accurately point out specific problems in any of the five MCC infrastructures. We thus argue that MCC management must be based on a joint observation of both QoS and QoE.

In reviewing the state-of-the-art on the management of MCC, we divide the literature work into two groups: (*i*) works that combine the use of QoS and QoE and (*ii*) works that focus on architectural models for MCC management. The investigations from the first group indicate that the traditional QoS-based management alone is insufficient to guarantee *End-user's* satisfaction [6–8,12]. Meanwhile, management architectures (from the second group) do not consider QoE, as described in the surveys of Rahimi, Ren, and Liu [2], and

[1] CoLisEU: https://coliseu.rnp.br .

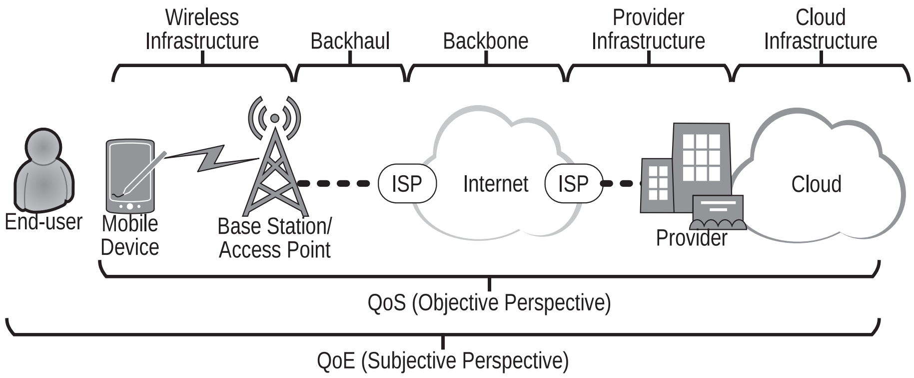


Fig. 1. Mobile cloud computing environment.

Abolfazli *et al.*, [1]. Although most current MCC management architectures represent new proposals, they are not based on the quite large and solid management principles established by the network and services management research field. For example, management entities, such as agents and management stations, are weakly discussed and presented, remaining unmapped in today's MCC management environments.

In the network management research field, two management entities are well known and explored: (*i*) agents and (*ii*) managers. The agent is a software module placed inside an infrastructure component and is responsible for monitoring and controlling local parameters, such as available memory and Maximum Transmission Unit (MTU). The manager, in turn, is a role assumed by a network node, *e.g.*, computers, set-top-boxes, or routers, and is responsible for retrieving information from agents and managing a network slice or domain. The main difference between other investigations on the management of MCC and our research is that we provide a mapping of the management entities to the MCC environment, considering the joint observation of QoS and QoE. In this context, we summarize the main management requirements considering the five infrastructures of an MCC environment and the observation of QoS and QoE in the Table 1.

In Table 1, the five functional areas Fault, Configuration, Accounting, Performance, and Security (FCAPS) [13] are mapped into key requirements for any MCC management systems. The MCC requirements are not limited to the presented Table 1, for example, exploring the heterogeneity or scalability of the MCC environment would lead to other requirements. However, FCAPS requirements are broader, being able to include other requirements.

3. MCC management architecture

In this section, we first show an MCC environment in terms of management entities. Next, we propose an architecture for the MCC management system, considering the perspectives of End-users and Administrators. In addition, we detail how the requirements listed in the previous section are addressed.

As can be seen in Fig. 2, *Agent* is an MCC application that can be remotely programmed or triggered by other applications within mobile devices to measure End-users' QoE and network QoS. *Gateways* are intermediary nodes installed at the edge of the network that bridge communications between *Management Applications* and *Agents*. *Management Application* corresponds to software installed in a cloud virtual node, handled by the cloud infrastructure, to provide management tools for Administrators and End-users. Moreover, cloud components, such as the *AAA Server*, as well as other *Cloud Resources*, are presented inside the cloud. *Base Stations* and *Access Points* are shown as intermediate equipments between *Mobile Devices* and logical links to *Gateways*. Finally, the Administrator and End-users that participate in the MCC environment are also shown using a desktop and mobile devices, respectively.

In such an environment, we propose a management architecture where *Agent*, *Gateway*, and *Management Application* are placed inside different MCC components, respectively: *mobile device*, *virtual network node*, and *virtual cloud node*. These architecture components are shown in Fig. 3, and according to each management entity, are better described in the next subsections.

3.1. Agent

The MCC infrastructure must be properly managed through the monitoring of its evaluation status, which is composed of several QoS parameters (*e.g.*, throughput, *Mobile Device* Received Signal Strength Indicator - RSSI, and Round Trip Time - RTT), as well as QoE measurements about the MCC infrastructure. Currently, the cloud, the provider, and the core infrastructures can have their state partially exposed with QoS measurements using, for example, traditional SNMP or NETCONF agents. In turn, the backhaul and the wireless infrastructures – where Administrators have loose/partial control and cannot freely deploy agents – have their states defined based on some statistical estimate of link usage, *e.g.*, bit rate of an entire domain. Whereas, the gathering of QoE measurements for each of the five infrastructures is a metric that is difficult to obtain, mainly because End-users have indirect contact with most of these infrastructures and are unable to give a proper review of a particular one. It means that QoE and QoS measurements are gathered partially preventing the retrieving of a complete evaluation status of the MCC environment. In this sense, the cooperation of End-users that can use their *Mobile Devices* to gather QoS and QoE measurements involving MCC infrastructures becomes fundamental.

Table 1
MCC management requirements.

Requirement	MCC environment
Fault [14]	The management system must be aware of the five infrastructure faults, *e.g.*, nodes failures, to avoid QoS and QoE degradation.
Configuration [15]	The management system must reconfigure the five infrastructures to achieve correctness and autonomy based on QoS and QoE.
Accounting [16]	The management system must monitor and measure the usage of the five infrastructure through QoS and QoE for billing correctness and auditing purposes.
Performance [17]	The management system must support a large number of mobile devices performing asynchronous communications to avoid compromising the proper operation of the MCC environment.
Security [18]	The management system must Authenticate, Authorize, and Account (AAA), *End-users* actions inside the MCC environment, avoiding impersonation attacks as well as providing a stronger auditing.

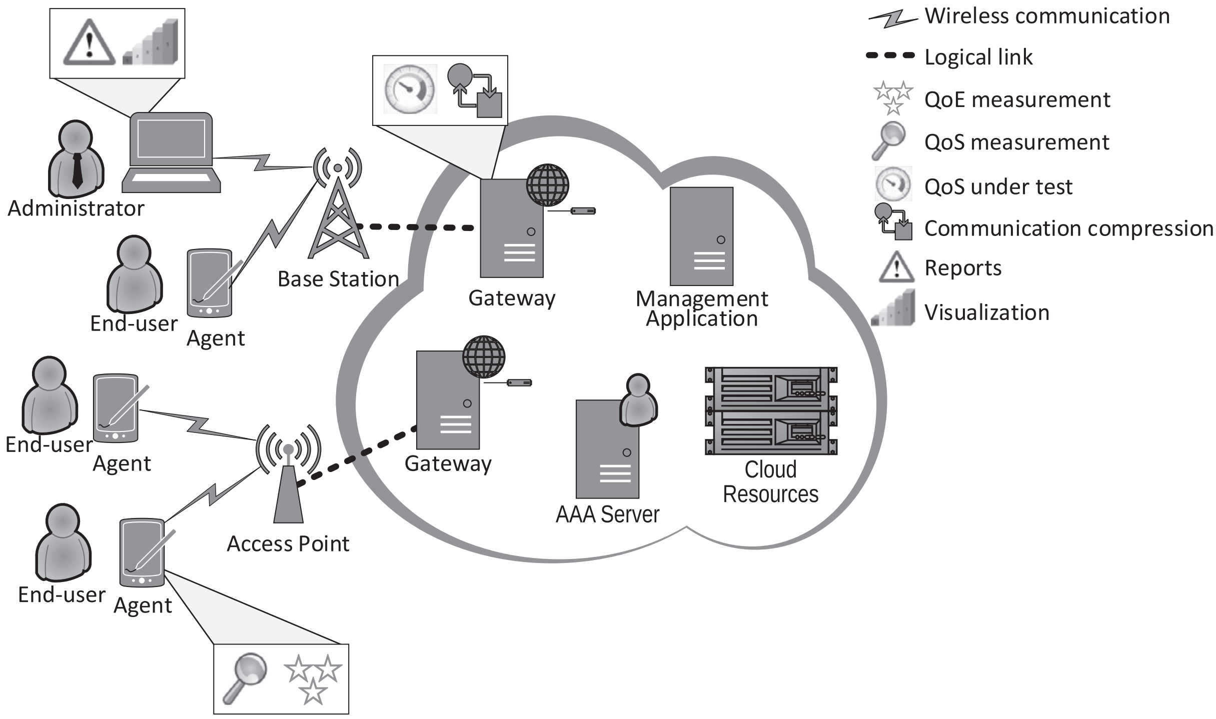


Fig. 2. MCC management environment.

To allow End-users to participate actively in the management of MCC environments, a new *Agent* placed inside the *Mobile Device* is proposed. The *Agent* is responsible for gathering QoS measurements from the wireless and backhaul infrastructures. In addition, the *Agent* provides capabilities for End-users to report their QoE about the entire MCC infrastructure. *Agent* entities must cope with the specific hardware and software characteristics of a *Mobile Device*, such as *Wireless Network Interfaces*, *e.g.*, WiFi or LTE cards, and *Operating Systems* (OS), *e.g.*, Android, iOS, or Windows Phone, since the performance of QoS and QoE gathering is largely influenced by these characteristics.

Inside the *Agent* we propose a main module called *Agent Controller* that communicates with four lower level modules, *QoS Meter*, *System Extractor*, *AAA client*, and *Service Interface*, as well as four upper level modules, *QoE Meter*, *Graphic Viewer*, *Cache*, and *MCC Applications*, described below.

- *QoS Meter* is responsible for collecting QoS measurements from the wireless and backhaul infrastructures of the MCC environment. To account for different QoS metrics, the *QoS Meter* must use a set of well established network measurement tools, such as perfSonar from Internet2. In terms of wireless infrastructure metrics, the *QoS Meter* must account for downlink throughput, delay, and RSSI. Whereas, for the backhaul infrastructure, the *Gateway* must be used with the *Agent* to gather other metrics, such as RTT, jitter, and uplink throughput. It is important to notice that the *QoS meter* only accounts for backhaul and wireless infrastructure, whereas traditional SNMP

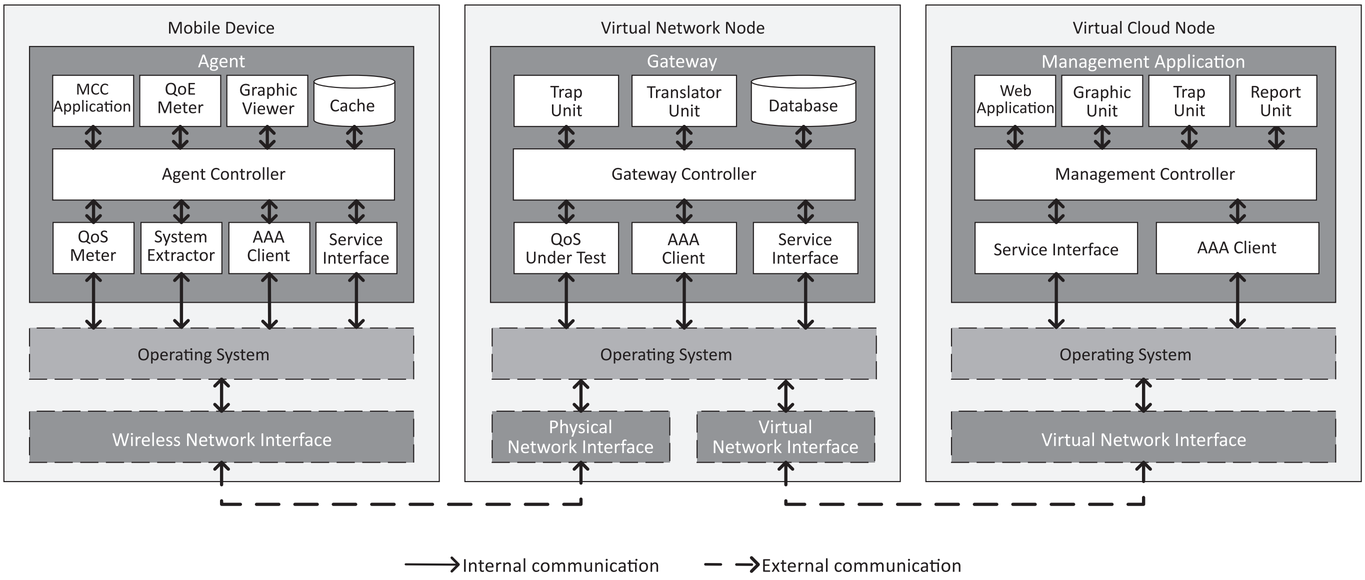


Fig. 3. MCC management architecture.

and NETCONF agents are used for gathering QoS metrics in the core, provider, and cloud infrastructures.

- *System Extractor* retrieves meaningful information about the software and hardware details of the *Mobile Device*. Examples of gathered information are the OS branch and version, device manufacturer, and battery level. In addition, the system extractor is responsible for control system parameters, for example, IP Maximum Transmission Unit (MTU) and TCP Maximum Segment Size (MSS).
- *AAA client* accounts for the authentication of mobile devices, authorizing other modules to perform their actions, and audit each action performed by the *Agent*. This means that each measurement gathering will be secured according to the AAA solution employed, for example, OAuth or RADIUS.
- *Service Interface* homogenizes the communication among *Agents* and *Gateways*, providing a set of services that can be remotely consumed by *Gateways*. In addition, this module defines how each other modules can consume different services provided by *Gateways*, *e.g.*, store QoE measurement or store QoS measurement.
- *QoE Meter* is proposed to be used by other *MCC Applications* to gather QoE from End-users as a default tool to *Mobile Devices*. It means that the MCC management can receive QoE measurements in a homogeneous manner that measures the user satisfaction accounting for different types of *MCC Applications*. In addition, *QoE Meter* provides graphical input tools to interact with End-users to gather their experience about the MCC environment.
- *Graphic Viewer* enables different *MCC Applications* to retrieve chart visualizations of QoS and QoE measurements gathered from the End-User mobile device. These visualizations can be later displayed to the End-users according to the *MCC Application* usage.
- The *Agent Controller* is responsible for managing the other modules' processing requests from the *Virtual Cloud Node* or providing communication among higher and lower level modules.
- *Cache* allows offloading operations, *i.e.*, it stores temporary offline information to be used later [19]. These offloading operations are particularly important for MCC infrastructures, mainly when End-users are facing disconnections and for compressing information before sending to *Gateways*.
- *MCC Applications* are mobile applications that can take advantage of all the other modules from the *Agent* to perform gathering of QoE and QoS. These applications can also use underlying functions for standardization purposes. For example, *MCC Applications* can use the *Agent* as a platform for AAA as much as for the retrieving of visualizations of the network state.

Considering the design of the *Agent* component, we can improve the *Accounting* requirement by monitoring End-users' usage of MCC, such as described in Section 2.2. In addition, the *Security* requirement is provided by the *AAA client*, using the *Mobile as a Representer* approach, where each End-user can be represented by a virtualized entity in the cloud through his/her mobile device [10]. The *Fault* requirement is met using the System Extractor to send reports about malfunctions as well as QoS and QoE degradation to be analyzed by the *Gateways*. *Configuration* is assured by the Service Interface by standardizing the communication protocol among components, as well as the messages format.

3.2. Gateway

To intermediate the communication between mobile devices and the cloud environment, *Gateways* placed inside *Virtual Network Nodes* between the wireless and cloud infrastructures are used. These gateways can temporarily store measured data from *Agents* about the MCC environment. Periodically, the stored data is compressed and sent to a centralized server in the cloud infrastructure. This approach avoids flooding the network with several asynchronous messages, greatly improving the scalability of the management system. *Gateways* also can assume the role of intermediary managers, executing tasks, such as performing polling on *Mobile Devices* or gathering information from the cloud, provider, and core infrastructure through traditional agents. *Gateways* aim to shorten the time taken to exchange management messages between *Agents* and the cloud infrastructure.

Similar to the *Agent*, the *Gateway* must be designed considering that a *Virtual Network Node* may present different OS and Network Interfaces. However, these interfaces are classified in Physical and Virtual. The first is connected to the backbone. The second is connected to an internal network within the core or the cloud infrastructure. Inside the *Gateway*, we specify seven modules, as below:

- *QoS Under Test* is responsible for answering the requests made by the *QoS Meter* in the *Agent*, composing a Client-Server paradigm. In this sense, the *Gateway* can be considered as a measurement point, where *Agents* can measure their network through different tools, *e.g.,* iPerf, OWAMP, or OWPING.
- *AAA client* provides a bridge service for *Mobile Devices* to perform AAA through the *AAA Server*, as depicted in Fig. 2. Authentication requests allow *Gateways* to secure the information provided by each *Agent* by associating it to a trusted and unique ID. It is important to notice that centralized AAA solutions, such as RADIUS, can also be used, excluding *Gateways* from the AAA process. However, a distributed AAA approach favors security and scalability. It means that *Gateways* can distribute content inside the MCC environment to End-users taking advantage of distributed content network concepts in a secure manner.
- *Service Interface*, in the same manner as in the *Agent*, standardizes the communication among *Agent-Gateway* and *Gateway-Manager*. In contrast to *Agents*, the *Gateway Service Interface* module communicates with the *Translator Unit* for compressing and decompressing of messages.
- *Translator Unit* performs all the compressing, decompressing, and translation of messages. Before being sent, any information must be grouped and processed by the *Translator Unit*. This module defines which information will be compressed to be sent to the *Management application*. In addition, received compressed information must be decompressed by the *Translation Unit* using a previous established encryption key for security purposes. Finally, the *Translation Unit* module may also organize information in different formats, *e.g.,* JavaScript Object Notation (JSON) to eXtensible Markup Language (XML).
- A *Database* stores temporary information received from *Agents*, to be later processed by other modules and forwarded to the *Virtual Cloud Node*. Also, the *Gateway* database contains stored actions of each *Agent* for auditing purposes.
- *Trap Unit* handles the notification of undesired events that may occur within the MCC infrastructures, such as signal interference, low throughput, gateway redirections due to a node inactivity, among others. Since all these events are usually unexpected, the *Trap Unit* of each *Gateway* must be always active to receive trap messages. As soon as a trap occurs, depending on its severity, the *Gateway* should perform at least one of three actions: (*i*) forward the trap immediately to an *Alert Unit* on a *Management Application*, (*ii*) cache the trap to be sent later, or (*iii*) perform an extra task. In the first action, the *Gateway* sends the trap message containing a severity level indicator, informing that a drastic event occurred. In the second, a minor severity problem is cached to be sent later. In the third, the *Gateway* may perform any other task, according to the architecture implementation. For example, the *Gateway* may be programmed to execute a custom script in the case of a network interface going down.
- *Gateway controller* is responsible for intermediating management operations, such as performing polling on the *Agents* or monitoring and forwarding traps. In addition, according to where the *Virtual Network Node* is placed, *Gateways* may control other infrastructure's components directly or indirectly through the network. For example, switches installed with a *Virtual Network Node* can have their internal parameters changed by the *Gateway controller* locally, *e.g.,* changing the MTU and the routing protocol, or indirectly, by receiving remote requests through the network.

In summary, *Gateways* are important components to address the four MCC requirements. By intermediating communication with *Agents*, *Gateways* minimize network traffic and response time by being placed closer to the wireless infrastructure, partially tackling the *Performance* requirement. Also, when *Gateways* perform management operations, this alleviates the overall management task of the MCC environment that otherwise would have to be completely performed by *Management Applications*. Like in the Agent, *Security* is provided by the *AAA client*, while the *Configuration* requirement is met by the Service Interface. Finally, the *Fault* requirement is addressed by the *Gateway controller*, which monitors and processes the information sent by the *Agent System Extractor* to detect malfunctions as well as unavailabilities that degrades QoS and QoE.

3.3. Management application

An Administrator is not able to fully understand all the raw data gathered from the MCC environment without proper summarization, followed by the creation of comprehensive charts. Moreover, the Administrator cannot perform distribution of tasks inside an MCC environment without employing a distribution solution, *e.g.,* using management by delegation [20] or network function virtualization [21]. In addition, there are several management tasks that must be processed without the need for waiting for Administrators to make a proper decision. For example, when a critical trap is sent to inform about a link failure, the management system is supposed to perform an action to try to overcome this link failure, by preventing *Agents* from using the defective links. In this sense, data summarization and visualization , as well as distribution and processing of management tasks, are some of the responsibilities that a *Management Application* must perform in an MCC environment.

A *Management Application* is cloud software that can be placed in a *Virtual Cloud Node*, which has an OS and *Virtual Network Interface* that allow the communication with *Gateways* in an internal network. To perform all the aforementioned responsibilities, we propose a *Management Application* composed of seven modules, described below

- *Management Controller* is responsible for processing and distributing all the management tasks of the MCC environment. The controller can take advantage of the huge availability of *Cloud Resources* to employ different management algorithms for optimization purposes, *e.g.,* policy mechanisms, autonomous machine learning algorithms, and fuzzy logic. According to the employed algorithm, the MCC environment can be reorganized dynamically aiming for different objectives, for example, fairness or overall throughput improvement.
- *Service Interface* presents all the services available to be consumed for *Gateways* and other *Management Applications* in a standardized manner. In addition, all the distribution of management tasks among *Management Applications* is based on network services to provide decoupling and easy integration.
- *Graphic Unit* must continuously process all the data gathered by *Gateways* and *Agents* to create rich summaries, composing comprehensive charts or visual graphs. Through the visualizations created by this module, the *Management Application* allows Administrators to audit and monitor their MCC domains.
- *Report Unit* creates documents that summarize all the management performed inside the MCC domain. These documents can be periodically sent to Administrators for auditing or monitoring purposes. Also, inside a report, different management tasks are summarized in terms of traps received and what actions were performed by different entities.
- *Alert Unit* receives messages from the *Traps Unit* from *Gateways* and are responsible for warning Administrators through notifications, mainly when a critical event happened inside the MCC environment. In addition, the *Alert Unit* is responsible for summarizing the traps received and actions performed to be used later by the *Report* and *Visualization Units*.
- *AAA client* assures that all the gathered measurements are correctly binded to the End-users that collected it, for report generation purposes. Moreover, each role, *i.e.,* End-user or Administrator, which a user can assume when accessing the *Management Application* is granted by the *AAA client* module.
- *Web Application* is a web tool for Administrators and End-users to analyze reports and charts generated by the *Report* and *Graphic Units*, respectively. Also, Administrators can manage their MCC domains by auditing reports generated by the *Alert Unit*, and send management tasks to be performed by different *Gateways* or *Agents* manually if needed.

Management Applications play a key role in addressing three of the MCC management requirements. *Security*, addressed by *AAA client*, is the most important requirement for this component, since it allows grouping measurements binded to the End-user that collected it. In addition, it provides user access control in the *Web Application* module, granting protection for private data and different levels of permission for End-users and Administrators. The *Fault* requirement is met by analyzing charts and reports that include data collected from mobile devices. An Administrator can analyze this information to verify whether or not a given problem reported by an End-user is related to his/her mobile device. Similar to other components, *Configuration* is assured by the *Service Interface* module that guarantees interoperability among other components exposing configurable services to be used remotely.

4. CoLisEU

In this section, we describe the developed prototype, called CoLisEU, that was based on our architecture as a proof of concept. The scenario used for deployment and testing of the prototype is also detailed. Finally, metrics are defined for assessing the performance of CoLisEU.

4.1. Prototype and scenario

We developed the CoLisEU prototype considering the Agent, Gateway, and Management Application entities. The Agent was developed using Java and the software development kit provided by Google^{inc} for mobile devices with Android OS. A Gateway was developed as a network application over an Apache Web Server using PHP combined with a MySQL database. A virtual node was used to deploy the Management Application, which is an Asynchronous JavaScript and XML (AJAX)-based Web system. In such an application, reports and visualizations are processed using Web Service composition, *e.g.,* Google MAPs and Google Big Tables services.

The AAA server was developed as a database placed inside a virtual node on the cloud infrastructure. The Cloud Resource was implemented as a database to store QoS and QoE measurements. All the developed virtual nodes were placed over a private cloud infrastructure, based on a Xen server. The main developed modules for evaluating our prototype are QoS Meter, QoS Under Test, and the QoE Meter. The Agent QoS Meter and the Gateway QoS Under Test modules were both developed as part of the performance tools from Internet2 perfSONAR, OWPING, OWAMP, and BWCTL,[2] while the QoE Meter was developed using Java.

To evaluate the prototype, we defined two different test scenarios to understand how CoLisEU must be applied to real MCC environments. The first scenario regards an educational wireless network, *i.e.,* students from the University of Santa Cruz used their mobile devices to contribute QoS and QoE measurements on the evaluation of the university wireless network. On the other hand, the second scenario is composed of residential wireless networks, where End-users are at their own home. Both scenarios differ from each other, since the behavior and the expectation of the End-Users are different, affecting the subjective evaluations.

4.2. Evaluation methodology

A team composed of 68 volunteers was responsible for collecting QoS and QoE measurements for the two scenarios, educational and residential, using mobile devices such as

[2] http://www.internet2.edu/products-services/performance-analytics/performance-tools/ .

smartphones and tablets. Using the gathered measurements, we were able to evaluate the combination of QoS and QoE, making two contributions: (*i*) the assessment of the composed visualization and the management cost that CoLisEU inserts into the MCC environment; (*ii*) the establishment of the relationship between QoE and QoS to improve the management of an MCC environment. In the first contribution, three metrics were considered: the consumption of battery, network traffic, and response time for each measurement. These metrics were chosen because they directly impact on the End-user experience when using the MCC application.

As a baseline for comparison, we used a management operation with the minimum information required for a mobile device to communicate with a gateway, *i.e.,* username, password, and OS details. For each experiment performed, we conducted replications of 30 experiments to assess each measurement gathering. For each replication we presented the mean values with a 95% confidence level.

Our second contribution is deriving an equation for each scenario that uses QoS measurements to infer a subjective perception about the wireless infrastructure. Deriving these equations is important for a management system so it can predict the user experience through QoS metrics. This means that the management system can reconfigure the network with fewer QoE measurements. Such derivation was possible by applying a linear regression to the measurements collected by the volunteer teams in both evaluation scenarios. In the next section, we analyze the prototype, using the experimental results collected for the proposed scenarios.

5. Results & discussions

Since CoLisEU is concerned with the management of MCC considering both objective and subjective perspectives, we chose to focus on a qualitative analysis in Section 5.1 and on a quantitative analysis in Section 5.2.

5.1. Qualitative analysis

This analysis intends to explore the benefits acquired from using the combination of QoS and QoE in the management of MCC. To expose these benefits, we chose the main screen of our prototype, as shown in Fig. 4. This visualization composes a heat map showing the experiences that End-users had when using a federated network, in particular, the eduroam WiFi network. The colored amorphous shape in the map represents the estimated geographic coverage area of the analyzed network. In addition, the spreading gradient colors depict the average End-user experience within the selected network. CoLisEU End-users can interact with the visualization to analyze the nearest access point in the selected area, along with meaningful information about it. Examples of such information would be the IP address or geographical position of the access point, or even a throughput chart that allows End-users to realize the average throughput experienced by other users using the same access point. Furthermore, the side menu also enables the End-users to switch between networks and change the measurement time intervals being considered.

Based on the visualization provided by the prototype, we encourage some discussion on the advantages of managing MCC considering objective and subjective perspectives combined with rich visualizations.

- **Detecting End-users behavior** in MCC can potentially influence Administrators' decision-making about network configurations. QoE and QoS measurements can be combined to determine different End-user profiles, *e.g.,* End-users that are interested in games, video chats, or web navigation. Each profile may lead Administrators to configure their infrastructures differently, such as for example, creating policies that prioritize smaller round trip times, instead of improving link capacity.
- **Estimating the wireless network coverage** is essential for Administrators to reduce the cost of the wireless infrastructure, in terms of antennas and network devices. For the End-user, the main advantage could be the possibility of improving the connectivity achieved inside a covered area.
- **Summarizing the average satisfaction** within the network may be interesting for End-users to know what to expect when connecting in a given wireless network. In addition, this satisfaction could be used by an End-user to troubleshoot his/her mobile device.
- **Analyzing the infrastructure performance and the End-users' satisfaction** is possible through detailed reports, which contains data and charts. Synthesizing a huge amount of information in a comprehensive visualization allows the management system to be used either by expert Administrators or lay End-users.

5.2. Quantitative analysis

The developed prototype is a management system that enables the retrieval of QoE in terms of Mean Opinion Score (MOS), which is represented by the values 1 (very unsatisfied), 2 (unsatisfied), 3 (regular), 4 (satisfied), and 5 (very satisfied). QoS values are gathered in terms of RSSI, RTT, and throughput. However, gathering these measurements requires the consumption of resources from device, network, and cloud.

5.2.1. Resource consumption

For the quantitative analysis, we assessed the consumption of battery, network traffic, and the average response time for each measurement. The results are shown in Fig. 5. The battery consumption (*a*) was measured in milliamperes per hour [mA/h], the network traffic (*b*) in bytes, and the response time (*c*) in milliseconds [ms]. The measurements were organized and classified in four groups: Baseline only, Baseline and QoE, Baseline and QoS, and Baseline, QoS, and QoE simultaneously.

The summary presented in Fig. 5 allows us to evaluate the management cost for an MCC environment. From the network consumption point-of-view, gathering QoS measurements with our prototype is, in the worst case, 42% costlier than retrieving a QoE measurement. This difference is mainly due to the technique employed to measure QoS, *i.e.,* the network stress test. Considering the battery consumption, QoS costs 6% more than QoE, while considering the response time, QoS takes 252% more time to gather than QoE. These two results are related, since the QoS stress test takes much more processing time than simply informing a MOS.

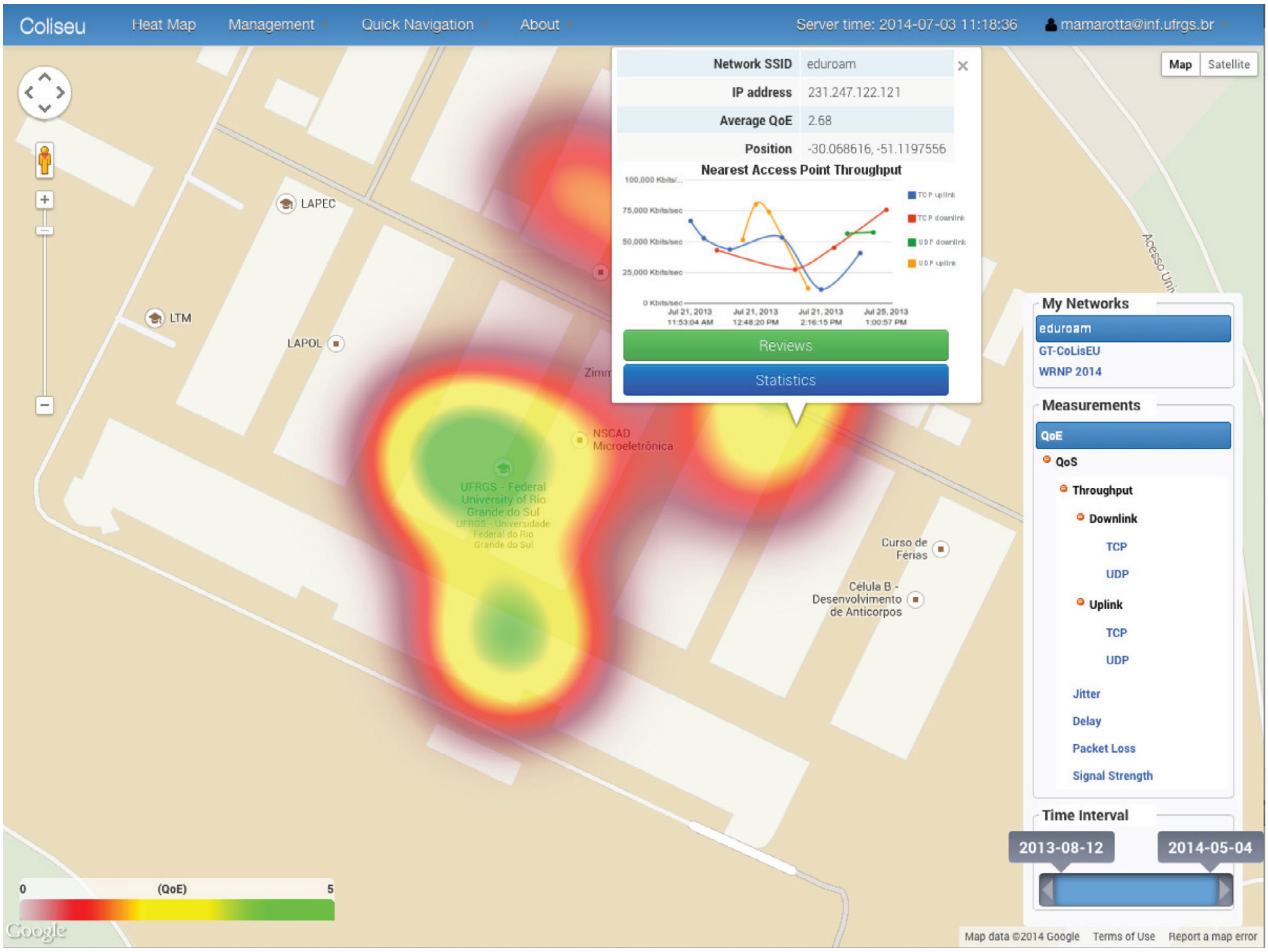
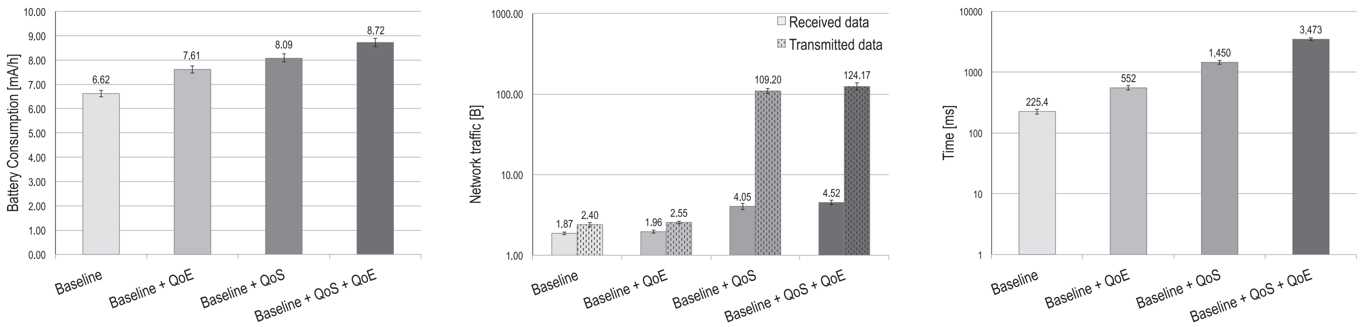


Fig. 4. Management application.



(a) Battery cost (b) Network Traffic (c) Response Time

Fig. 5. Prototype resource consumption evaluation.

Based on these results, it is possible to state that, from the End-user point-of-view, considering QoE measurements is important to avoid resource depletion with management operations. However, considering the Administrator point-of-view, the QoS measurements is the best way to detect specific infrastructure problems. By combining these two methods, the management system should explore the benefits of both, avoiding excessive costs. Therefore, the joint analysis of QoE and QoS improves the awareness of the MCC environment for both End-users and Administrators.

Our prototype showed that through the use of the designed model, it is possible to achieve the derived requirements to properly manage an MCC environment, *i.e.*, scalability, authentication, resource management, interoperability, and decoupling. In addition, charts and innovative visualizations were depicted according to the End-user geographic

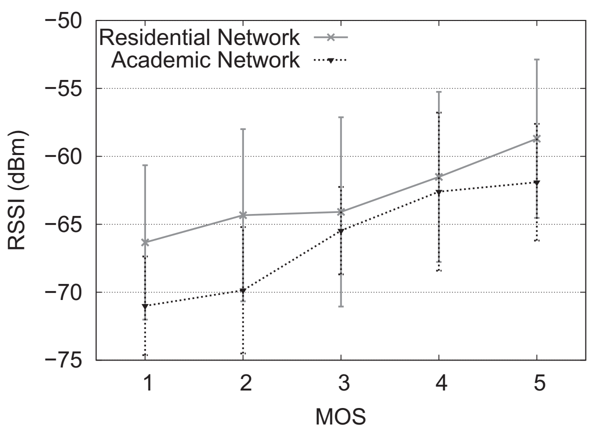


Fig. 6. RSSI versus QoE.

area, considering different authenticated Administrators. Moreover, the distribution of the management tasks, using Gateways, enabled trap detection for a larger number of End-users and network nodes from different infrastructures, without a single point of failure and network overload.

5.2.2. QoE versus QoS

The main idea of this subsection is to establish a relationship between QoE and QoS measurements, such as shown in Eq. 1. In order to establish this relationship, one thousand samples of QoE versus QoS were gathered in academic and residential scenarios. These samples were summarized in Fig. 7, according to three analyses: (*i*) the relationship between the *RSSI* versus the experiences of users (Fig. 6); (*ii*) the variation of *RTT* versus throughput (*TP*) for the experiences of academic users (Fig. 7a); and (*iii*) the same relationship of *RTT* versus throughput (*TP*) for the experiences of residential users (Fig. 7b); All of the analysis was carried out to investigate the importance of considering the satisfactions of users in different contexts for the management of MCC networks.

$$QoE = f(RSSI, TP, RTT) \quad (1)$$

As can be seen in Fig. 7, through QoS and QoE parameters we can establish the following relationships for a management system.

- The intensity of RSSI influences the user satisfaction, regardless of other QoS parameters for both networks.
- In an academic network where the users experience a low RTT, the main parameter that influences their satisfaction is the throughput because they expect a high data transfer rate. This influence becomes clear by the lightening of gradient colors that go along with the improvement of throughput as presented in Fig. 7a.
- In a residential network, which has an Internet access with limited bit rate according to a contracted SLA by users, the main QoS parameter that influences their satisfaction is the RTT. This influence becomes clear by the darkening of colors that go along with the RTT when it gradually gets larger, such as presented in Fig. 7b.

All of these relationships contribute to the definition of the function *f*(*RSSI*, *TP*, *RTT*). To make this definition, we used a multi-variable linear regression with our thousand samples of QoS and QoE, resulting in Eqs. 2 and 3. For each equation we found a significance level under 5%, proving their correctness for statistical purposes.

$$QoE = 5,435 + 0,06(RSSI) + 0,179(TP) - 0,00005(RTT) \quad (2)$$

In the academic network the *TP* is the most relevant QoS metric that influences the users' QoE. Through the Eq. 2, the *TP* presented a positive weight of 0,179 whereas the others presented less influence in the QoE summary. For a management system, this weight relationship among metrics is useful to coordinate each infrastructure of MCC to improve their link capacity, instead of, for example, coordinating flow priorities. In addition, this relationship also helps the University of Santa Cruz to direct their expenditure and, in this case, suggests to invest in wireless equipment in order to improve their link capacity with users.

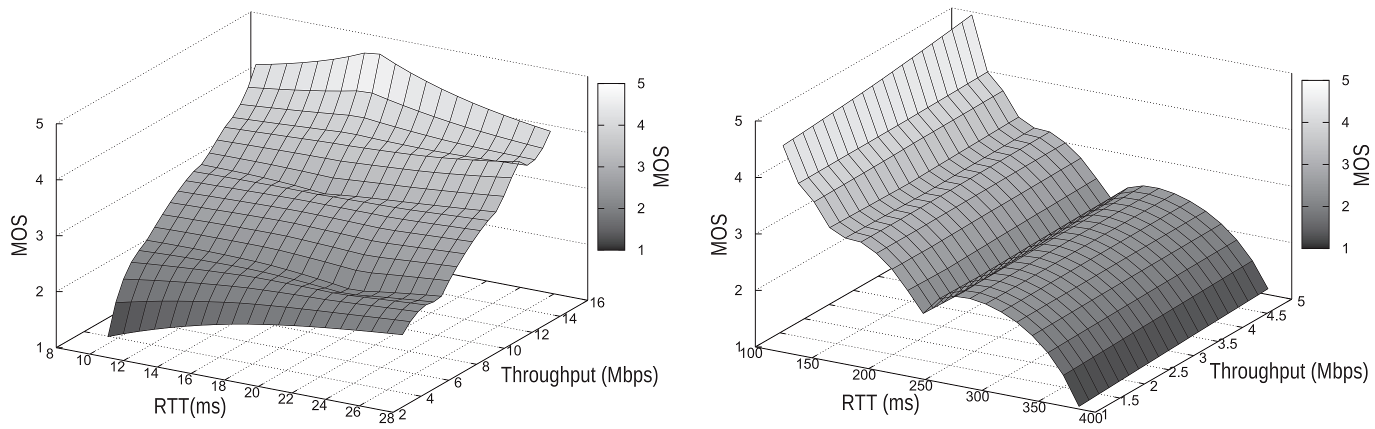
$$QoE = 2,762 + 0,002(RSSI) + 0,424(TP) - 0,018(RTT) \quad (3)$$

In residential networks, in turn, the Eq. 3 showed a higher weight for RTT than in an academic network. Despite the influence of RTT in residential networks, the Eq. 3 also showed that the *TP* also has a great influence in residential networks. The influence of *TP* was masked into the graphical charts, showing how important it is for a management system to calculate the linear regression. Using the Eq. 3, the management system must reconfigure the MCC environments to consider both, RTT and *TP*, as the main indicators for users' experiences. As a consequence, management systems will focus on guaranteeing the SLA for End-users, adding flows prioritizing minimizing RTT.

As can be seen, both equations are very different, which allows us to determine that users from different contexts have different QoS needs. This means that we cannot generalize this function for any context, making the gathering of QoE a neccessity for the management of the MCC environment according to different contexts. For the management system, it means that QoE values must be gathered and later clustered according to different contexts before the establishment of a relationship with QoS values.

Under another perspective, we tried to evaluate the End-user's point of view in academic and residential networks according to their most used type of applications, such as depicted in Fig. 8.

We considered five types of End-user applications; browsers, e-mail, games, chat, and social networks. End-users interested in playing games, using chats, or social networks, are susceptible to become dissatisfied and give a bad review for an academic network. Such a dissatisfaction might be related to the use of firewalls or network configurations that prevent the full usage of these type of applications in an academic institution. End-users in residential networks have their satisfaction less influenced by the type of application in use. Therefore, there will be cases where the same applications will influence the End-user satisfaction differently within the MCC environment. This means that a management system cannot consider just the type of applications in use when managing an entire MCC domain. The management of MCC must be performed through the gathering of QoE, QoS, and information concerning



(a) Academic network (b) Residential network

Fig. 7. QoS versus QoE values in academic and residential scenarios.

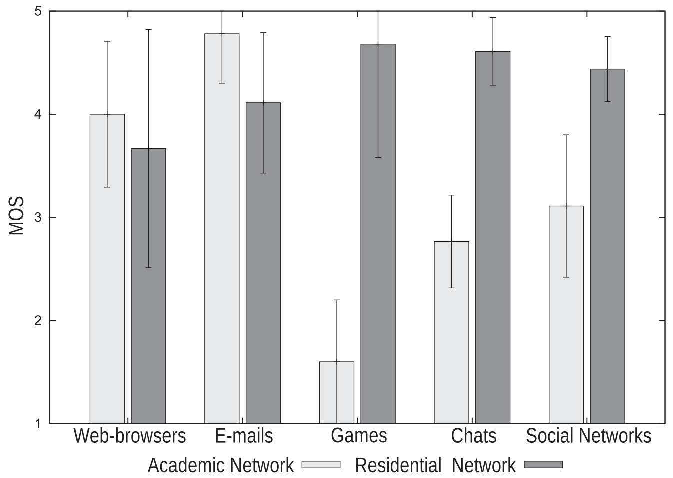


Fig. 8. QoE according to different users interests.

End-users' applications to be later clustered to find the real source of End-users' dissatisfaction managing each part of the MCC environment.

6. Conclusion & future work

In this paper we outlined MCC as the process of extending the cloud for resource augmentation of constrained mobile devices. Managing the infrastructures involved in the process of augmentation is fundamental to the MCC environment to properly operate. Administrators and End-users can collaborate with the MCC management by sharing their experiences through the use of a survey with detailed reviews. In addition, End-users can perform QoS measurements, enabling problem detection in the MCC environment. A model, architecture, and prototype were described to enable the collaboration between Administrators and End-users within the management of MCC, creating rich visualizations to be analyzed with regards to MCC performance.

The model and architecture were designed to meet MCC requirements integrating mobile devices, the MCC infrastructures, and the cloud. Our prototype, called CoLisEU, validates the architecture and is divided into three components. Afterwards, we analyzed the prototype qualitatively and quantitatively, discussing the benefits of using objective and subjective perspectives for the management of the MCC environment. Combining both perspectives improves the awareness of Administrators and End-users about the MCC environment.

Our research is intended to be an initial step toward the introduction of a management system for MCC environments. As MCC develops, there will likely be other interesting research opportunities in this area. For example, semantic Web technologies can be used to improve the extraction of management information from QoE reviews. Finally, a management system could employ machine learning techniques to adapt these requirements to each different infrastructure of an MCC environment.

Acknowledgments

The authors thank to the Rede Nacional de Ensino e Pesquisa (RNP), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Coordenação de Aperfeiçoamento de Nível Superior (CAPES) for the financial support.

References

[1] S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, R. Buyya, Cloud-based augmentation for mobile devices: motivation, taxonomies, and open challenges, IEEE Commun. Surveys Tutorials 16 (1) (2014) 337–368.
[2] M.R. Rahimi, J. Ren, C.H. Liu, A.V. Vasilakios, N. Venkatasubramanian, Mobile cloud computing: a survey, state of art and future directions, Mobile Netw. Appl. 19 (2) (2014) 133–143.
[3] D. Huang, Mobile cloud computing, Multimedia Commun. Techn. Committee IEEE Communications Soc. E-Letter 6 (10) (2011) 27–30.
[4] F. Sardis, G. Mapp, J. Loo, M. Aiash, A. Vinel, On the investigation of cloud-based mobile media environments with service-Populating and QoS-aware mechanisms, IEEE Trans. Multimedia 15 (4) (2013) 769–777.
[5] P. Rengaraju, C.-H. Lung, F. Yu, A. Srinivasan, On qoe monitoring and e2e service assurance in 4g wireless networks, Wirel. Commun. IEEE 19 (4) (2012) 89–96.
[6] T. Hobfeld, R. Schatz, M. Varela, C. Timmerer, Challenges of qoe management for cloud applications, IEEE Commun. Magaz. 50 (4) (2012) 28–36.
[7] P. Casas, R. Schatz, Quality of experience in cloud services: survey and measurements, Comput. Netw. 68 (5) (2014) 149–165.

[8] S. Ickin, K. Wac, M. Fiedler, L. Janowski, J.-H. Hong, A.K. Dey, Factors influencing quality of experience of common;y used mobile applications, IEEE Commun. Magaz. 50 (4) (2012) 48–56.
[9] J. Wan, Z. Liu, K. Zhou, R. Lu, Mobile cloud computing: application scenarios and service models, in: Proceedings of the 9th International Conference Wireless Communications and Mobile Computing, 2013, pp. 644–648.
[10] D. Huang, T. Xing, H. Wu, Mobile cloud computing service models: a user-centric approach, IEEE Network 27 (5) (2013) 6–11.
[11] Y. Cai, F. Yu, S. Bu, Cloud computing meets mobile wireless communications in next generation cellular networks, IEEE Network 28 (2014) 54–59.
[12] M. Santos, J. Villalon, L. Orozco-Barbosa, A novel QoE-aware multicast mechanism for video communications over IEEE 802.11 WLANs, IEEE J. Select. Areas Commun. 30 (7) (2012) 1205–1214.
[13] R. Boutaba, A. Polyrakis, Projecting fcaps to active networks, in: Proceedings of international Conference on Enterprise Networking, Applications and Services, 2001, pp. 97–104.
[14] A. Khan, M. Othman, S. Madani, S. Khan, A survey of mobile cloud computing application models, IEEE Commun. Surveys Tutorials 16 (1) (2014) 393–413.
[15] A.N. Toosi, R.N. Calheiros, R. Buyya, Interconnected cloud computing environments: challenges, taxonomy, and survey, ACM Comput. Surveys 47 (1) (2014) 7:1–7:47.
[16] B. Parak, Z. Ustr, Challenges in achieving iaaS cloud interoperability across multiple cloud management frameworks, in: Proceedings of the 7th International Conference on Utility and Cloud Computing IEEE/ACM, IEEE, 2014, pp. 404–411.
[17] H.T. Dinh, C. Lee, D. Niyato, P. Wang, A survey of mobile cloud computing: architecture, applications, and approaches, Wirel. Commun. Mobile Comput. 13 (18) (2013) 1587–1611.
[18] N. Fernando, S.W. Loke, W. Rahayu, Mobile cloud computing: a survey, Elsevier Future Generation Comput. Syst. 29 (1) (2013) 84–106.
[19] H. Wu, D. Huang, S. Bouzefrane, Making offloading decisions resistant to network unavailability for mobile cloud collaboration, in: Proceedings of IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, 2013, pp. 168–177.
[20] M.A. Marotta, F.J. Carbone, J.J.C. de Santanna, L.M.R. Tarouco, Through the internet of things – a management by delegation smart object aware system (MbDSAS), in: Proceedings of the 37th Annual Computer Software and Applications Conference, 2013 IEEE Kyoto, Japan, 2013, pp. 732–741.
[21] M. Peng, Y. Li, Z. Zhao, C. Wang, System architecture and key technologies for 5g heterogeneous cloud radio access networks, IEEE Network 29 (2) (2015) 6–14.

Marcelo A. Marotta is a Ph.D. student in Computer Science at the Institute of Informatics (INF) of the Federal University of Rio Grande do Sul (UFRGS), Brazil. He received his M.Sc. title in (2013) from INF of UFRGS, Brazil. In addition, he holds a B.Sc.in Computer Science from the Federal University of Itajubá (UNIFEI) (2010), Brazil. His research involves Wireless Networks, Next Generation Networks, Internet of Things, Software Defined Radio, Cognitive Radio Networks, and Heterogeneous Cloud-Radio Access Networks.

Leonardo R. Faganello is an M.Sc. student at the Federal University of Rio Grande do Sul (UFRGS), in Brazil. He achieved his B.Sc. degree in Computer Engineering at UFRGS in 2012. His current research interests include wireless networks, cognitive radio networks and spectrum decision.

Matias Artur Klafke Schimuneck is an M.Sc. student at the Federal University of Rio Grande do Sul (UFRGS), in Brazil. He achieved his B.Sc. degree in Computer Science at University of Santa Cruz do Sul in 2014. His current research interests include wireless systems and users quality of experience.

Lisandro Z. Granville is an associate professor at the Institute of Informatics of the Federal University of Rio Grande do Sul (UFRGS), Brazil. He received his M.Sc. and Ph.D. degrees, both in computer science, from UFRGS in 1998 and 2001, respectively. He is member of the Brazilian Internet Committee (CGI.br). He has served as a TPC member for many important events in the area of computer networks (e.g.,IM, NOMS, and CNSM) and was TPC co-chair of DSOM 2007 and NOMS 2010.

Juergen Rochol is an associate professor at the Institute of Informatics of the Federal University of Rio Grande do Sul (UFRGS), Brazil. He received his M.Sc. degree in Physics, and his Ph.D. degree in Computer Science, both from UFRGS in 1972 and 2001, respectively. His research interests include wireless networks, next generation networks, optical networks and traffic control on broadband computer networks.

Cristiano B. Both is a Postdoctoral in Computer Science at the Institute of Informatics of the Federal University of Rio Grande do Sul (UFRGS), Brazil. He received a Ph.D. degree at UFRGS in 2011 and his M.Sc. degree in Computer Science from the Pontifical Catholic University of Rio Grande do Sul, Brazil, in 2003. He is currently an assistant professor at the Federal University of Health Sciences of Porto Alegre, Brazil. His research interests include wireless networks, next generation networks, and software-defined networking.