Universidade Estadual de Campinas
Instituto de Computação

Alexandre Toshio Hirata

Load Balancing and User Association in HetNets

Balanceamento de Carga e Associação de Usuários em HetNets

CAMPINAS
2017

Alexandre Toshio Hirata

Load Balancing and User Association in HetNets

Balanceamento de Carga e Associação de Usuários em HetNets

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientadora: Profa. Dra. Juliana Freitag Borin
Co-supervisor/Coorientador: Prof. Dr. Eduardo Candido Xavier

Este exemplar corresponde à versão final da Dissertação defendida por Alexandre Toshio Hirata e orientada pela Profa. Dra. Juliana Freitag Borin.

CAMPINAS
2017

Agência(s) de fomento e nº(s) de processo(s): FAPESP, 2013/07064-4

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

H613L

Hirata, Alexandre Toshio, 1984-
Load balancing and user association in HetNets / Alexandre Toshio Hirata. – Campinas, SP : [s.n.], 2017.

Orientador: Juliana Freitag Borin.
Coorientador: Eduardo Candido Xavier.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Balanceamento de carga (Computação). 2. Sistemas de telefonia celular. 3. Redes heterogêneas. I. Borin, Juliana Freitag, 1978-. II. Xavier, Eduardo Candido, 1979-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Balanceamento de carga e associação de usuários em HetNets
Palavras-chave em inglês:
Load balancing (Computing)
Cell phone systems
Heterogeneous networks
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Juliana Freitag Borin [Orientador]
Fabio Luiz Usberti
Paulo Roberto Guardieiro
Data de defesa: 15-12-2017
Programa de Pós-Graduação: Ciência da Computação

Universidade Estadual de Campinas
Instituto de Computação

Alexandre Toshio Hirata

Load Balancing and User Association in HetNets

Balanceamento de Carga e Associação de Usuários em HetNets

Banca Examinadora:

- Profa. Dra. Juliana Freitag Borin
  Universidade Estadual de Campinas
- Prof. Dr. Fabio Luiz Usberti
  Universidade Estadual de Campinas
- Prof. Dr. Paulo Roberto Guardieiro
  Universidade Federal de Uberlândia

A ata da defesa com as respectivas assinaturas dos membros da banca encontra-se no processo de vida acadêmica do aluno.

Campinas, 15 de dezembro de 2017

Aos meus pais, Dominico e Maria, por todo amor e dedicação que tiveram comigo.
Aos meus irmãos, Marcio e Kaleandra, por todo o apoio que me deram

Agradecimentos

A lista de agradecimentos é grande e mesmo que eu não faça justiça a todos que de alguma forma contribuíram para este trabalho, deixo meu obrigado aqui.

À Prof. Dr. Juliana Freitag Borin e ao Prof. Dr. Eduardo Candido Xavier por me orientarem durante todos esses anos. Agradeço-os não só pelo apoio no âmbito técnico, mas também por toda a paciência e compreensão que tiveram comigo.

Ao Prof. Dr. Flávio Keidi Miyazawa que foi meu orientador no programa de iniciação científica num tema relacionado ao deste trabalho e por ter indicado o Eduardo como orientador.

A todos os professores do Instituto de Computação (IC) da Unicamp que contribuíram de forma direta ou indireta para a minha formação tanto durante o programa de pós-graduação como no de graduação. Em especial, gostaria de citar os professores Ricardo da Silva Torres, Rodolfo Jardim de Azevedo, Arthur João Catto, Sandro Rigo, Cid Carvalho de Souza, Ricardo de Oliveira Anido e Cecília Mary Fischer Rubira que de alguma forma me marcaram e me ajudaram.

Aos professores do Instituto de Matemática, Estatística e Computação Científica (IMECC) da Unicamp que muito me ensinaram durante a graduação e, dentre eles, destaco Maria Aparecida Diniz Ehrhardt, Antonio Carlos Moretti e Jayme Vaz Junior que não só foram excelentes professores como também me ajudaram em outros assuntos acadêmicos.

Aos professores do SENAI Prof. Dr. Euryclides de Jesus Zerbini que, durante o curso técnico, apresentaram conceitos fundamentais das áreas de Telecomunicações, Eletrônica e Computação o que foi determinante para a escolha da minha carreira profissional.

À DaitanGroup e todos os seus colaboradores, pelo apoio e suporte durante todos esses anos em que eu conciliava o trabalho com o programa de mestrado. Em especial, gostaria de agradecer a todos que trabalharam e/ou trabalham comigo e principalmente aqueles que acabaram se tornando meus amigos. Destaco, o Hélio Kinoshita que foi o responsável direto por me trazer à Daitan, descanse em paz.

To all teams I've worked with as a partner developer. Among them, I'd like to give special thanks to some British Telecom (BT) teams: BT Softswitch, Ribbit Platform and BT Conferencing teams. I've learned a lot with you guys!

Aos integrantes do Laboratório de Otimização Combinatória (LOCo) e do Laboratório de Redes de Computadores (LRC) pelo apoio técnico.

A todos os meus amigos que dividiram momentos de alegria e tristeza, mas que por mais distantes que possam estar, eles sabem que foram importantes para mim.

Por fim, à minha família. Ao meu pai Dominico, minha mãe Maria, meus irmão Marcio e Kaleandra por todo o amor e suporte que têm me dado.

Science is built up of facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house.
Henri Poincare

Resumo

Redes Heterogêneas (HetNets) apareceram como um modo inteligente de aumentar a capacidade e cobertura de redes de celular nas quais estações rádio-base (BSs) de baixa potência podem dividir a carga de estações de alta potência. Entretanto, essa estratégia também trouxe novos desafios. Por exemplo, técnicas para associação de usuários usadas em redes homogêneas não são eficientes neste tipo de rede quando se considera a quantidade de usuários servidos e o balanço de carga entre as BSs.

Neste trabalho, o problema de associação de usuários em HetNets é modelado como um problema de programação linear inteira (ILP) com o intuito de balancear a carga entre células de curto e longo alcance. Além disso, duas heurísticas são introduzidas: uma solução centralizada baseada em um algoritmo guloso e uma estratégia distribuída, probabilística e ciente de carga. Estas heurísticas produzem bons resultados de balanceamento de carga entre as células e em termos de número de usuários aceitos quando comparados à solução ótima e melhores resultados que algumas das principais estratégias apresentadas na literatura.

Abstract

Heterogeneous Networks (HetNets) come as a clever approach to increase the capacity and the coverage of cellular networks in which low power base stations can share the load of high power ones. However, such strategy also brought new challenges. For instance, user association techniques used on homogeneous networks are no longer efficient when the amount of served users and the load balancing among the BSs are considered.

In this work, the user association in HetNets problem is modeled as an integer linear programming (ILP) problem aiming to balance the traffic load among short and long range cells. In addition, two heuristics are introduced: a centralized solution based on a greedy algorithm and a distributed, probabilistic load-aware solution. These heuristics produce good results of load balancing among the cells and in terms of number of accepted users in comparison with the optimal solution and better results than some of the main strategies presented in the literature.

List of Figures

List of Tables

List of Algorithms

List of Abbreviations and Symbols

1G	First Generation
2G	Second Generation
3G	Third Generation
3GPP	Third Generation Partnership Project
3GPP2	Third Generation Partnership Project 2
4G	Fourth Generation
5G	Fifth Generation
ABS	Almost Blank Subframes
AMPS	Advanced Mobile Phone System
AR	Augmented Reality
BS	Base Station
BT	British Telecom
CA	Carrier Aggregation
CC	Component Carrier
CDMA	Code-Division Multiple Access
CIO	Cell Individual Offset
CN	Core Network
CoMP	Coordinated MultiPoint
CRE	Cell Range Extension
D-AMPS	Digital AMPS
D2D	Device to device
DeNB	Donor eNB
DFT	Discrete Fourier Transform
DL	Downlink

E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
EDGE	Enhanced Data rates for Global Evolution
eNB	E-UTRAN Node B, evolved Node B or eNodeB
EPC	Evolved Packet Core
EPS	Evolved Packet System
FD-MIMO	Full Dimension MIMO
FDD	Frequency-Division Duplexing
FDM	Frequency-Division Multiplexing
GERAN	GSM EDGE Radio Access Network
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
HeNB	Home eNodeB
HetNet	Heterogeneous Network
HetSNet	Heterogeneous and Small-cell Network
HPN	High Power Node
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access
HSS	Home Subscriber Service
HSUPA	High Speed Uplink Packet Access
IC	Instituto de Computação
ICIC	Inter-Cell Interference Coordination
IEEE	Institute of Electrical and Electronics Engineers
ILP	Integer Linear Programming
IMECC	Instituto de Matemática, Estatística e Computação Científica
IMT	International Mobile Telecommunications
IoT	Internet of Things
IP	Internet Protocol
IS	International Standard
ITU	International Telecommunication Union

ITU-R	ITU Radiocommunication Sector
LOCo	Laboratório de Otimização e Combinatória
LP	Linear Programming
LPN	Low Power Node
LRC	Laboratório de Redes de Computadores
LTE	Long Term Evolution
LTE-A	LTE-Advanced
M2M	Machine to machine
MBMS	Multimedia Broadcast Multicast Services
MIMO	Multiple-Input Multiple-Output
MISO	Multiple-Input Single-Output
MME	Mobility Management Entity
MMS	Multimedia Messaging Service
MNO	Mobile Network Operator
MRI	Magnetic Resonance Imaging
MT	Mobile Termination
MTC	Machine-Type Communication
MU-MIMO	Multiple-User MIMO
NB	Node B
NMT	Nordic Mobile Telephone System
NTT	Nippon Telegraph and Telephony
OFDM	Orthogonal Frequency-Division Multiplexing
OFDMA	Orthogonal Frequency-Division Multiple Access
P-GW	Packet Data Network Gateway
P2P	Peer to Peer
PDC	Personal Digital Cellular
PDN	Packet Data Network
PRB	Physical Resource Block
PSTN	Public Switched Telephone Network

QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
RAN	Radio Access Network
RB	Resource Block
RF	Radio Frequency
RN	Relay Node
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RSSI	Received Signal Strength Indicator
Rx	Receiver, Receive or Reception
S-GW	Serving Gateway
SAE	System Architecture Evolution
SC-FDMA	Single Carrier - Frequency-Division Multiple Access
SENAI	Serviço Nacional de Aprendizagem Industrial
SIM	Subscriber Identity Module
SIMO	Single-Input Multiple-Output
SINR	Signal to Interference plus Noise Ratio
SISO	Single-Input Single-Output
SMS	Short Message Service
SON	Self Organized Network
SU-MIMO	Single-User MIMO
TACS	Total Access Communications System
TD-SCDMA	Time Division Synchronous Code Division Multiple Access
TDD	Time-Division Duplexing
TDM	Time-Division Multiplexing
TDMA	Time-Division Multiple Access
TE	Terminal Equipment
TR	Technical Report

Tx	Transmitter, Transmit or Transmission
UE	User Equipment
UHD	Ultra High Definition
UICC	Universal Integrated Circuit Card
UL	Uplink
UMTS	Universal Mobile Telecommunication System
UTRAN	UMTS Terrestrial Radio Access Network
VoIP	Voice Over-IP
VR	Virtual Reality
WAP	Wireless Application Protocol
WCDMA	Wideband Code-Division Multiple Access
WiMax	Worldwide Interoperability for Microwave Access

Table of Contents

Chapter 1

Introduction

Heterogeneous Networks (HetNets) have been proposed to increase the capacity of Long Term Evolution (LTE) networks by introducing small cells, e.g., pico cells and femto cells, into the existing coverage area of the macro cells [11]. Small cells differ mainly in terms of capacity, coverage area, transmit power, signal strength, and cost of deployment [66]. Besides, HetNets play an important role in Long Term Evolution-Advanced (LTE-A) and certainly will do so on the upcoming Fifth Generation Mobile Networks (5G) [42]. Thus, investigating its performance and proposing the required enhancements to it is critical.

Macro cells, created by base stations (BSs) called evolved Node Bs (eNodeBs), are usually deployed by Mobile Network Operators (MNOs), so they are spread in a way to maximize the coverage. Small cells are deployed in high density areas to share the macro cell load or to extend its coverage area. Both pico and femto cells deployment is not necessarily planned [12].

The layered architecture brought by HetNets brings not only advantages but also many challenges to be overcome. For instance, in homogeneous networks, user equipments (UEs), devices used by the end-user to communicate with the system, connect to the BS with the best signal strength in order to provide a better user experience. On HetNets, however, this strategy would make all UEs connect to macro cells while the pico and femto cells would receive little or no load at all, that is, the small cells would work just as a backup when the macro cell is overloaded and the total load would not be evenly shared.

Research on efficient user association strategies that consider the specificities of the HetNets architecture is a hot field since the chosen ones for 5G networks are still not defined. In these networks, the efficiency is bound not only to the number of served users but also how they are distributed among the cells.

From the telecommunications industry perspective, such strategies shall demand as less modifications in terms of hardware as possible which makes distributed strategies more attractive than using a centralized load balancer.

Hence, it is of paramount importance to develop user association strategies to promote load balancing on HetNets [40]. By considering both the signal strength and the load, the association policy is able to assign users to BSs that offer them the best user-perceived rate [12] [20].

In this work, the User Association and Load Balance problems applied to HetNets

were jointly modeled as a single Integer Linear Programming (ILP) Problem. The optimal solution gotten from the ILP problem was used as a baseline to measure the quality of the solutions obtained by the other strategies.

ILP problems are NP-hard in general. Thus, getting an optimal solution usually demands high time complexity. Sometimes good solutions, i.e., non-optimal solutions, are enough and that is why heuristics come for. Although heuristics do not mathematically assure the finding of good solutions, they may empirically get satisfactory solutions.

The first heuristic of this work was relaxing the integrality of the ILP problem to get a Linear Program (LP) problem, which is solvable in polynomial time, and use a greedy strategy over the relaxed obtained solutions in an attempt to construct viable solutions.

In order to be implemented, the previous approach would need a centralized controller that knows the state of the whole system to solve the LP problem. To overcome such limitation and to get an even quicker strategy, this work also proposed a distributed heuristic that is independent of linear programming. In addition, it also performed better than other strategies it was compared to.

1.1 Contributions

The contributions of this dissertation are:

- An Integer Linear Programming Model that aims to solve the load balancing problem on HetNets and provides an efficient user association policy.
- A greedy strategy based on the relaxed version of the linear programming model which can get a possibly good solution in polynomial time.
- A distributed probabilistic heuristic which is independent of both a linear programming solution and a centralized load balancer which makes it very interesting to be deployed in production.

1.2 Scientific Production

The outcomes of this dissertation resulted on the following paper:

- Alexandre Toshio Hirata, Eduardo Candido Xavier, and Juliana Freitag Borin. Load Balance and User Association on HetNets. *IEEE Latin America Transactions,* 14(12):4781–4786, December 2016.

In addition, a paper has been submitted to a conference.

1.3 Dissertation Organization

This dissertation is organized as follows. Chapter 2 introduces basic concepts which are needed for the understanding of this dissertation as well as the research problem. Chapter 3 presents related work. Chapter 4 describes the proposed strategies. Chapter 5

illustrates the performed simulation scenarios and the obtained results. Finally, Chapter 6 shows the final conclusions and possible future work.

Chapter 2

Background

2.1 Telephony: From Landline to 5G

In 1876, landline telephones came with the idea of associating a number to a device called telephone and connecting the telephones through dedicated physical cables. After that, the telephone switch was invented, which allowed the formation of telephone exchanges and, later on, were the building blocks to create the current Public Switched Telephone Network (PSTN). In such network, telephone subscriber lines are able to establish telephone calls among each other.

Through the fixed line telephones, it was possible to associate a telephone to a place but not specifically to a person [24]. Afterwards, many other technologies tried to deliver this person to person association such as the paging systems in the 50's and the car phones in the 70's. However, such goal was truly accomplished with the cellular systems whose first commercial automated cellular network was launched by Nippon Telegraph and Telephony (NTT) in Japan, 1979.

The main idea of the cellular networks is to divide the coverage area in cells of different sizes according to the propagation conditions, density of user equipments (UEs) per kilometer, and traffic. Suppose the areas are initially divided in big hexagonal cells, with a BS in the center of each cell. The basic structure is composed by a set of seven cells as shown in Figure 2.1 [24]. However, in areas with higher density, their cells are subdivided into smaller hexagons and so on. From that basic structure, mobile network operators can plan the growth of the system in terms of quantity of subscribers and traffic.

Although other technologies could connect people similarly, the success of the cell phones can be explained by a combination of freedom, mobility and the increase of the productivity they can provide [24]. Nevertheless, it is causing an increasing demand for mobile broadband services with higher data rates and Quality of Service (QoS) [10]. In Figure 2.2, it is shown how the demand for data and voice traffic have been growing since the fourth quarter of 2011 [22] [23].

In order to satisfy the expectation of the users to have more advanced wireless access even in the mobile environments, cellular infrastructure has evolved leading to different cellular generations.

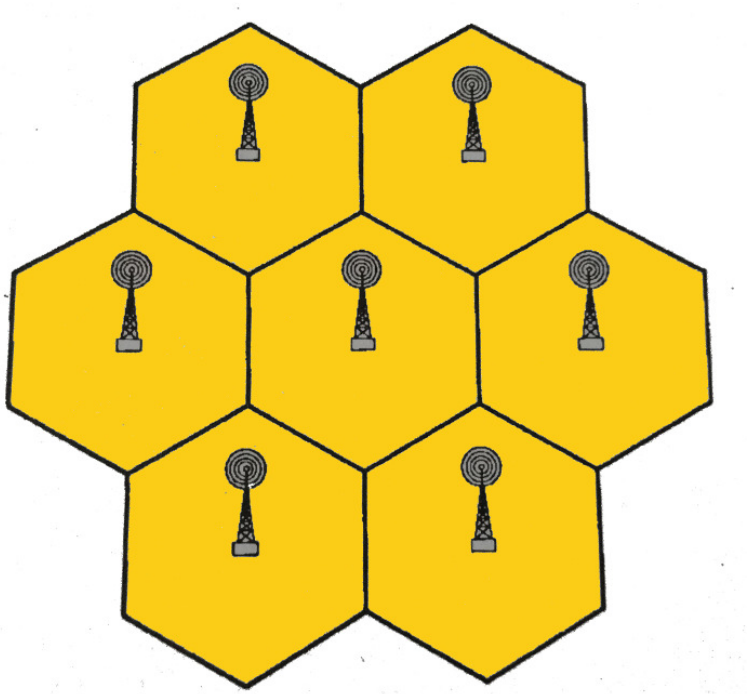

Figure 2.1: Basic structure of seven cells (Adapted from [24])

2.1.1 First Generation Mobile Networks (1G)

In the 80's, the first generation (1G) systems came to deliver voice-only mobile-telephony [19]. These systems used analogue communications techniques and they did not use the available radio spectrum efficiently [17].

Examples of 1G systems include:

- Nordic Mobile Telephone System (NMT)
- Advanced Mobile Phone System (AMPS)
- Total Access Communications System (TACS)

2.1.2 Second Generation Mobile Networks (2G)

In the 90's, the second generation (2G) systems were developed based on digital technology which provided significantly higher capacity than 1G systems [17].

Examples of 2G systems include:

- Personal Digital Cellular (PDC)
- Digital AMPS (D-AMPS) or IS-136, also referred as Time-Division Multiple Access (TDMA)
- cdmaOne or IS-95, which uses Code-Division Multiple Access (CDMA)
- Global System for Mobile Communications (GSM), which introduced the Short Message Service (SMS) later on.

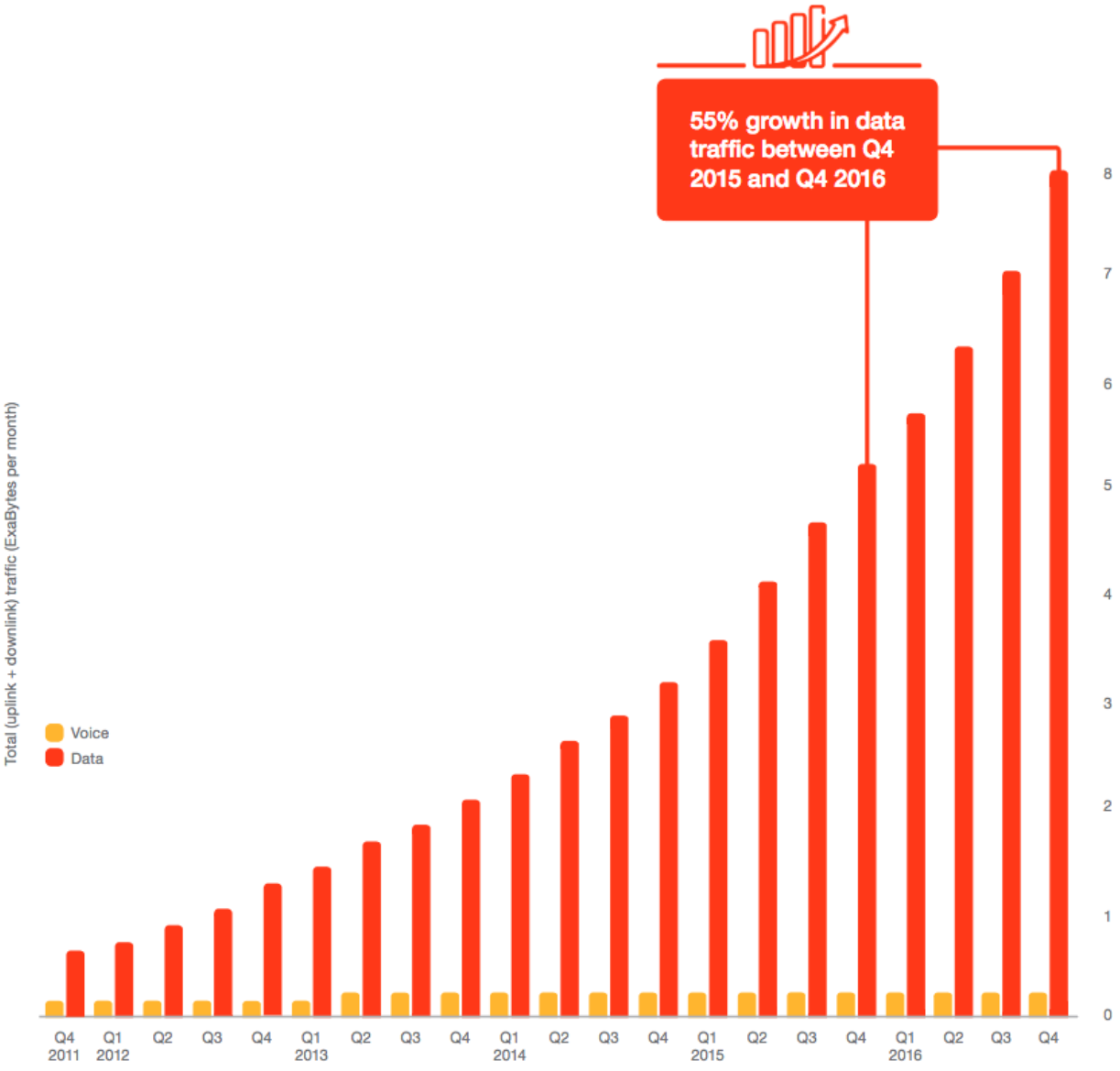

Figure 2.2: Global monthly data and voice traffic [23]

With the introduction of the General Packet Radio Service (GPRS) and the Enhanced Data rates for Global Evolution (EDGE) or Enhanced GPRS, the 2G systems were extended with primitive packet data services. Such services include web browsing via Wireless Application Protocol (WAP) and the ability to send photos and videos via Multimedia Messaging Service (MMS).

2.1.3 Third Generation Mobile Networks (3G)

In 2000's, the third generation (3G) came up using different techniques for radio transmission and reception from their 2G predecessors. In order to have its development carried out in a global basis, the Third Generation Partnership Project (3GPP) was formed to develop the Universal Mobile Telecommunications System (UMTS) based on GSM. On the other hand, the parallel organization 3GPP2 was formed to develop the competing cdma2000 technology, an evolution of cdmaOne [19].

The UMTS air interface has two different implementations: the Time Division Synchronous Code Division Multiple Access (TD-SCDMA) and the Wideband Code Division Multiple Access (WCDMA).

This system was later enhanced for data applications by introducing the High Speed Downlink Packet Access (HSDPA) and High Speed Uplink Packet Access (HSUPA), which are collectively known as High Speed Packet Access (HSPA). HSPA allowed for a "true" mobile-broadband experience with data rates of several MBit/s which was the foundation for the rapid uptake of smart phones such as the iPhone and a wide range of Android devices [17] [19].

2.1.4 Fourth Generation Mobile Networks (4G)

In the mid-2000's, users were using more and more mobile services, so their expectations for a new technology with higher data rate and lower latency ignited the discussions about the fourth generation (4G).

Both Worldwide Interoperability for Microwave Access (WiMax) and Long Term Evolution (LTE) communications systems were submitted and approved as 4G standards. However, LTE is by far the most dominating technology.

The 4G LTE technology was from the beginning developed for packet-data support and has no-support for circuit-switched voice. In addition, since its commercial introduction in 2009, it has been evolving such that, in Release 13, it supports multi-GBit/s peak data rates. Furthermore, it also widened the use cases beyond mobile broadband by, for instance, improving support for massive machine-type communication and introducing direct device-to-device (D2D) communication [17] [19] [40] [52] [53].

4G systems are supposed to meet the requirements issued by International Telecommunication Union Radiocommunication Sector (ITU-R) of an International Mobile Telecommunications - Advanced (IMT-Advanced) system [19].

Figure 2.3: Overall vision of 5G [64]

2.1.5 Fifth Generation Mobile Networks (5G)

Mobile Internet and Internet of Things (IoT) are the two main market drivers for the upcoming fifth generation (5G) which is expected to be deployed beyond 2020. In essence, 5G should be seen as a platform that will enable wireless connectivity to all kinds of services, that is, it aims to provide connectivity anywhere, anytime to anyone and anything. Such scenario is also known as *networked society*.

5G will implement a smart interconnection between people and all things. Hence, as can be seen in Figure 2.3, it will result in a massive number of use cases such as, but not limited to, Ultra High Definition (UHD) video services, Augmented Reality (AR), Virtual Reality (VR), remote computing, collaborative robots. The idea is to connect not only cell phones to people but also a variety of devices such as wearable devices, intelligent home appliances (smart home), medical monitoring devices, and even vehicles. Such connectivity will affect major fields like Industry, Education, Finance, among others [19] [31] [63] [64].

5G is expected to address some challenges not effectively addressed by 4G like higher capacity, higher data rate, lower end-to-end latency, massive device connectivity, and reduced cost. In addition, it also aims to enhance scalability, connectivity, and energy efficiency of the network [16] [21] [31].

Several research groups are working on different technical and probable standardization aspects of 5G. A few examples of such groups are: Mobile and Wireless Communications Enablers for the Twenty–twenty Information Society (METIS), 5th Generation Non-Orthogonal Waveforms for Asynchronous Signaling (5GNOW), Enhanced Multicarrier Technology for Professional Ad-Hoc and Cell-Based Communications (EM-PhAtiC)

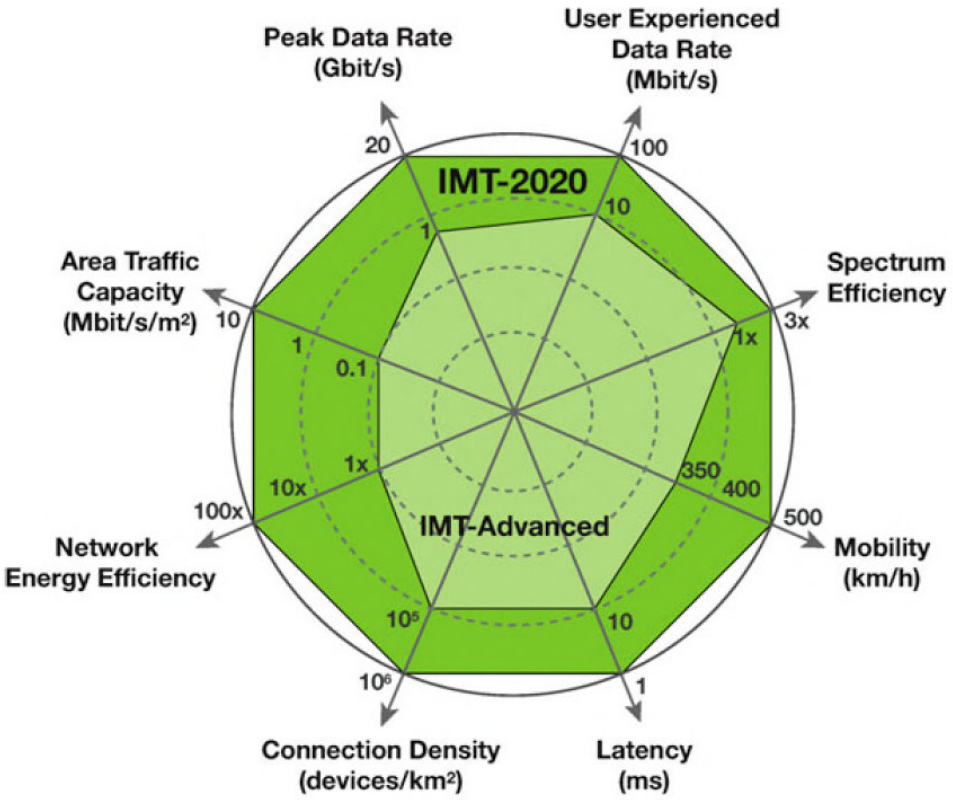

Figure 2.4: Key capabilities of IMT-2020 [64]

and 5G Infrastructure Public Private Partnership (5G-PPP) [46].

The expected key capabilities needed for an IMT-2020 technology to support 5G are described on Table 2.1. In Figure 2.4, they are compared to IMT-Advanced, the original requirements for 4G systems.

Key capabilities	Values
Peek data rate	20 Gbps
User experienced data rate	0.1-1 Gbps
Latency	1 ms over-the-air
Mobility	500 km/h
Connection density	10^6/km^2
Energy efficiency	100 times compared with IMT-Advanced
Spectrum efficiency	3-5 times compared with IMT-Advanced
Area traffic capacity	10 Mbit/s/m^2

Table 2.1: 5G key capabilities and values from ITU-R [64]

2.1.6 Summary

Cellular networks started as analog voice-only mobile networks that did not even allowed text messaging and evolved to the current digital mobile networks which offer

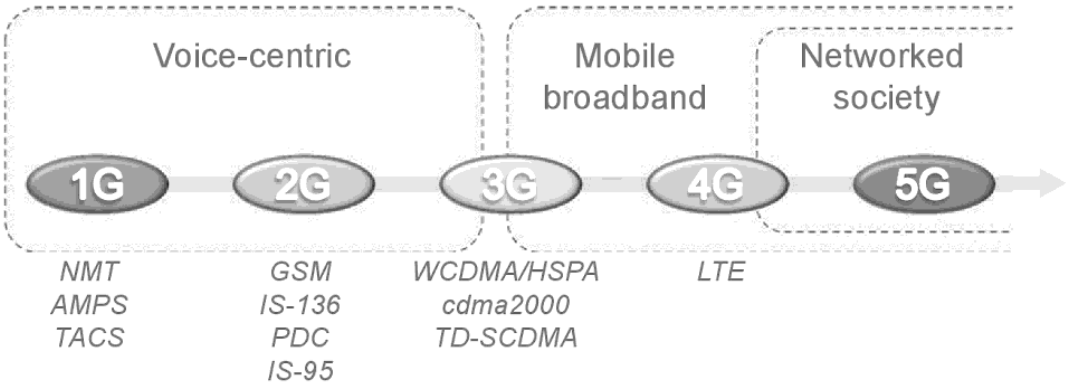

Figure 2.5: Cellular generations [19]

broadband internet access and serve not only cell phones but also a wide number of different devices.

The number of subscribers keep increasing along with the number of different services provided through such networks. Hence, the technologies to supply such demand has also been evolving. Figure 2.5 illustrates a summary of the 5 generations of mobile networks, their main characteristics and the technologies they were composed of.

2.2 Long Term Evolution (LTE)

Evolved Packet System (EPS) represents the evolution of the UMTS. Long Term Evolution (LTE) to UMTS or the E-UTRAN (Evolved Universal Terrestrial Access Network), is the radio access part of the EPS. The main requirements for the new access network are high spectral efficiency, high peak data rates, short round trip time as well as flexibility in frequency and bandwidth [39] [44].

Originally, the term 4G should only refer to systems that met the requirements of IMT-Advanced, for instance:

- All-Internet Protocol (IP) packet-switched network
- Nominal data rate of 100 MBit/s while UEs move at high speeds (e.g. within a train)
- Nominal data rate of 1 GBit/s while UEs are in relatively fixed position.

Although neither LTE nor mobile WiMAX 1.0 (802.16e) did so, they were categorized as 4G systems by ITU in 2010 [17].

2.2.1 Releases

The specifications for LTE are produced by the 3GPP and are organized into *releases*.

LTE was first introduced in Release 8 and some minor enhancements were presented in Release 9. Both form the foundation of LTE.

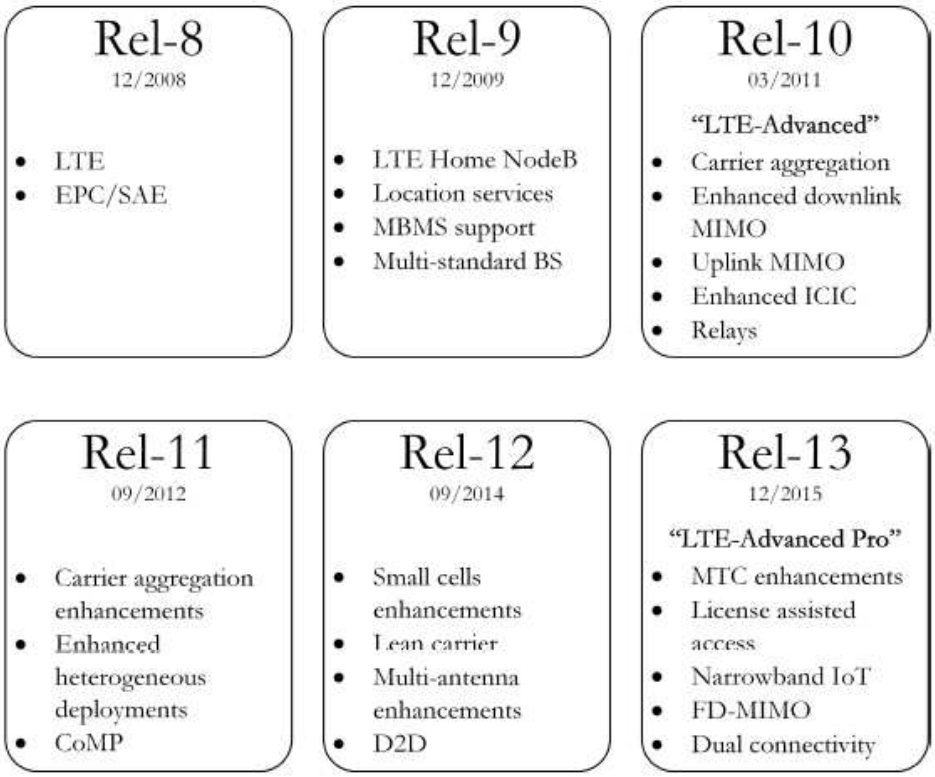

Figure 2.6: LTE releases (Adapted from [19])

Releases 10 and 11 specified LTE's evolution, known as LTE-Advanced (LTE-A). LTE-A is backward compatible to LTE and it formally satisfies the IMT-Advanced requirements [19] [51].

Release 12 focused on the use of LTE technology for emergency and security services. In addition, other important features were completed such as small cells and network densification, D2D and Carrier Aggregation (CA) [3].

A further evolution was introduced in Release 13 and it is also known as LTE-Advanced Pro.

Figure 2.6 presents the main features of each LTE Release.

2.2.2 Network Architecture

The LTE system architecture is basically composed of the Core Network (CN) and the Radio-Access Network (RAN). Both networks, including the split of functionality between these two, is known as System Architecture Evolution (SAE).

As mentioned earlier, LTE refers to the RAN and the new CN architecture is referred to as the Evolved Packet Core (EPC). Together, EPC and LTE RAN are referred to as the Evolved Packet System (EPS) [18]. However, LTE has become the colloquial name for the whole system even for 3GPP [17].

Figure 2.7 shows a comparison of the simplified network architecture starting from GSM until LTE.

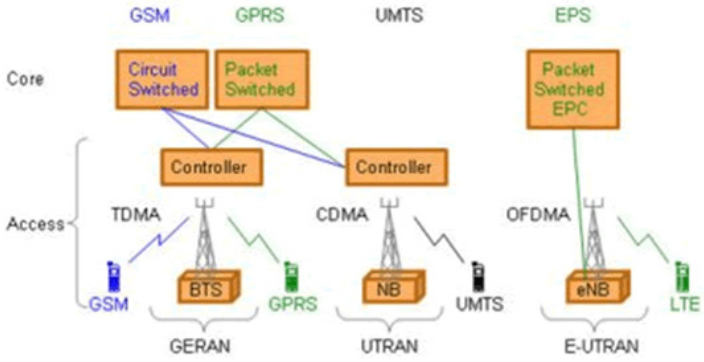

Figure 2.7: Network Solutions from GSM to LTE [44]

Core Network (CN)

The EPC is responsible for tasks not related to radio access such as authentication, billing and end-to-end connections. In addition, since EPS is purely IP based, all the services that used to be handled by the circuit switched systems, e.g. phone calls, are now handled by the packet switched EPC [19]. Figure 2.8 shows the CN architecture.

The EPC consists of several types of nodes [19] [56]:

- Mobility Management Entity (MME) is the control node that processes the signaling between the UEs and the CN. Its responsibilities include connection/release of bearers to a UE, handling of IDLE to ACTIVE transitions, and handling of security keys.
- Serving Gateway (S-GW) serves as a mobility anchor for both when UEs move between eNBs and for inter-working with other 3GPP technologies (GPRS and UMTS). In addition, it can collect information for billing and legal interception.
- Packet Data Network Gateway (P-GW) connects the EPC to the Internet and it is responsible for IP address allocation, as well as QoS enforcement according to the policy controlled by the Policy and Charging Rules Function (PCRF). Besides, it is also the mobility anchor for non-3GPP radio-access technologies connected to the EPC, e.g. cdma2000.
- Home Subscriber Service (HSS) contains users' SAE subscription data, information about which Packet Data Network (PDN) the user can connect, and some dynamic information like the identity of the MME to which the user is currently attached or registered.

Radio-Access Network (RAN)

The RAN is responsible for all radio-related functionality of the overall network such as scheduling, radio-resource handling, coding and multi-antenna schemes.

In Release 8, the LTE access network is simply a network of eNBs without any intelligent centralized controller, generating a flat architecture. Additional node types were

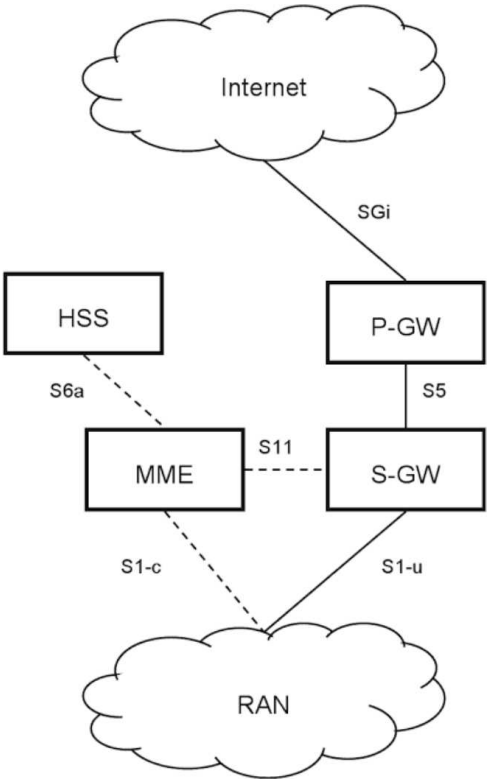

Figure 2.8: Core Network Architecture [19]

introduced with Multimedia Broadcast Multicast Services (MBMS) in Release 9 and the Relay Nodes in Release 10 [19].

The eNB has two main functions. Firstly, sending radio transmissions to all its UEs on the Downlink (DL) and receiving transmissions from them on the Uplink (UL). Secondly, sending signalling messages to control the UE's low-level operation.

Each eNB is a BS that controls the UEs in one or more cells. Prior to Release 12, a UE was able to connect to a single eNB and an IP address was allocated when the UE was switched on and released when switched off [44]. The Dual Connectivity introduced in Release 12 allowed a UE to be connected to two simultaneous eNBs at the same time [3]. Such feature is not going to be considered in this thesis.

UE's architecture is identical to the one used by UMTS and GSM, it can be divided into Mobile Termination (MT), which handles all the communication functions and the Terminal Equipment (TE) which contains the end-to-end application. For instance, the MT can be a plug-in LTE card for laptop and the TE would be the laptop itself. Besides, it also has a Universal Integrated Circuit Card (UICC), colloquially known as Subscriber Identity Module (SIM) card which stores user-specific data [17].

2.2.3 Multiple Access Schemes

In order to achieve high radio spectral efficiency, 3GPP chose the OFDMA (Orthogonal Frequency Division Multiple Access) for the downlink communication and the SC-FDMA (Single Carrier - Frequency Division Multiple Access), also known as Discrete Fourier

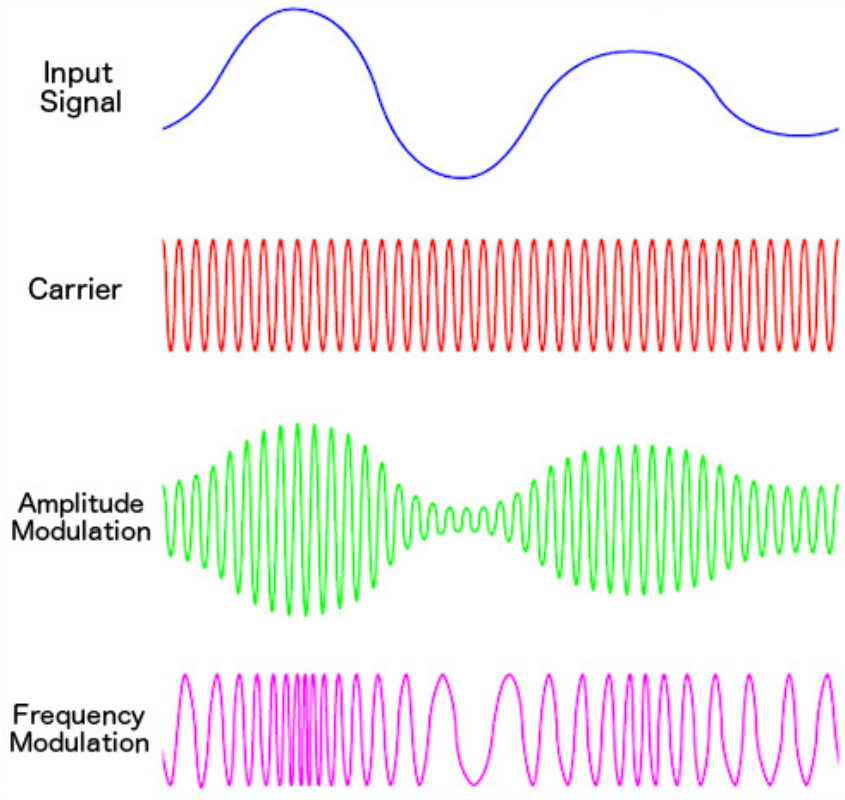

Figure 2.9: Input signal in AM and FM (Adapted from [5])

Transform (DFT) spread OFDMA, for the uplink communication.

Orthogonal Frequency Division Multiple Access (OFDMA)

Typically, most of the signals, in the way they originally are, cannot be physically transmitted over communication media. They need to be encoded in a way they can be transmitted and also decoded on the receiving endpoint.

Carrier signal, or just carrier, is a waveform (usually sinusoidal) that is going to be modulated in order to represent the information (input signal) that needs to be transmitted. Usually, the carrier's frequency is higher than the input signal's.

Modulation is the process of modifying some properties of a carrier signal like amplitude, frequency or phase such that there is a one-to-one correspondence between these properties and the input signal. Demodulation is the inverse of the modulation. Figure 2.9 shows an example of an input signal modulated using Amplitude Modulation (AM) and Frequency Modulation (FM). In AM, the amplitude of the carrier changes according to the amplitude of the input signal while the the frequency and the phase of the carrier remain unchanged. In FM, the property that changes is the frequency [9] [24] [28].

Multiplexing is the method in which multiple signals are combined into one complex signal over a communication link.

In Frequency-Division Multiplexing (FDM), the available frequency spectrum is divided into a series of non-overlapping frequency subcarriers. Each subcarrier can carry a different signal [24] [37].

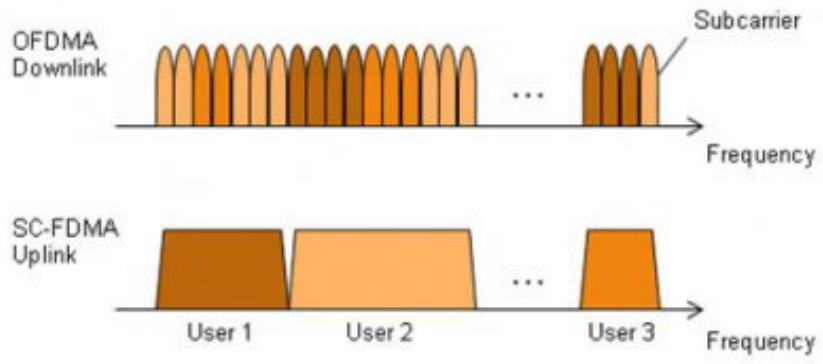

Figure 2.10: OFDMA and SC-FDMA [1]

Orthogonal Frequency-Division Multiplexing (OFDM) is a special case of FDM in which subcarriers are overlapping but orthogonal. This avoids the need to separate the carriers by means of guard-bands, and therefore makes OFDM highly spectrally efficient [56].

The OFDMA solution leads to high Peak-to-Average Power Ratio (PAPR) requiring expensive power amplifiers with high requirements on linearity, increasing the power consumption for the sender. This is no problem in the eNB, but would lead to very expensive handsets [1]. Hence a different solution was selected for the UL.

OFDMA can achieve high data rates due to the combination of higher order modulation (up to 256-Quadrature Amplitude Modulation - QAM), large bandwidths (up to 20 MHz) and spatial multiplexing (up to 8x8 Multiple-Input Multiple-Output (MIMO) in the DL and 4x4 MIMO in the UL). The highest theoretical peak data rate on the transport channel , using MIMO and 20MHz channel, is 300 Mbps in the DL and 170 Mbps in the UL [4] [44].

Single Carrier - Frequency Division Multiple Access (SC-FDMA)

The Single Carrier - Frequency Division Multiple Access (SC-FDMA) solution generates a signal with single carrier characteristics and, hence, with lower PAPR in comparison to OFDMA [49]. So, it is used to optimize the range and power consumption in the UL while the OFDMA is used in the DL to minimize receiver complexity, especially with large bandwidths, and to enable frequency domain scheduling with flexibility in resource allocation [34].

Figure 2.10 compares OFDMA and SC-FDMA in terms of carrier usage. Note that OFDMA uses many subcarriers for a single user, while SC-FDMA just uses a single carrier per user [1].

2.2.4 Generic Frame Structure

In OFDMA, users are allocated a specific number of subcarriers for a predetermined amount of time. These are referred to as Physical Resource Blocks (PRBs) in the LTE specifications. Thus, PRBs have both time and frequency dimensions.

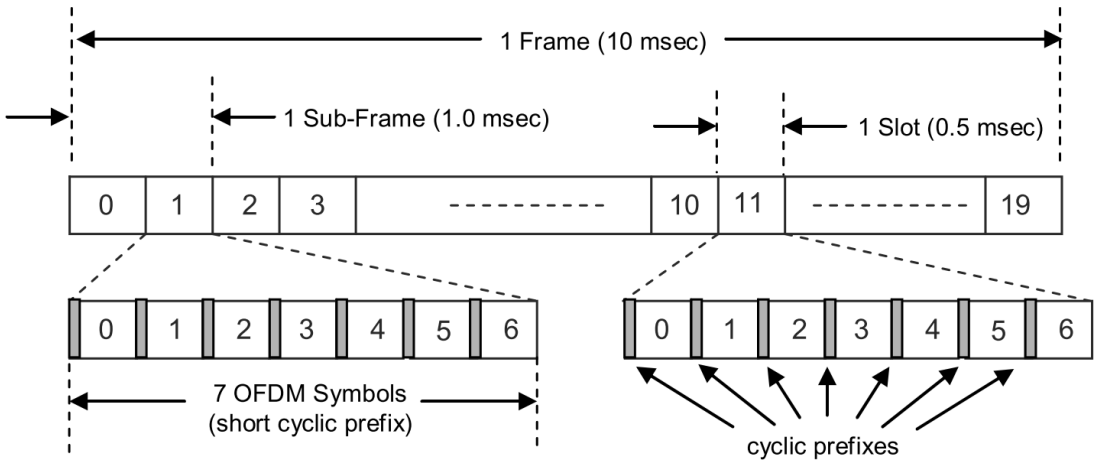

Figure 2.11: LTE Generic Frame Structure [67]

PRBs are the minimum unit of resource that can be allocated to a UE to receive data and their association is handled by a scheduling function at the eNB.

The LTE physical layer generic frame structure considered here is used with Frequency-Division Duplexing (FDD). Alternative frame structures are defined for use with Time-Division Duplexing (TDD). As shown in Figure 2.11, LTE frames are 10 msec in duration. They are divided into 10 subframes of 1.0 msec and each subframe is further divided into two slots of 0.5 msec each. Each slot consists of either 6 or 7 OFDM symbols, depending on whether the normal or extended cyclic prefix is employed.

The total number of available subcarriers depends on the overall transmission bandwidth of the system. The LTE specifications define parameters for system bandwidths from 1.25 MHz to 20 MHz as shown in Table 2.2. A PRB is defined as consisting of 12 consecutive subcarriers for one slot in duration.

Bandwidth (MHz)	1.25	2.5	5.0	10.0	15.0	20.0
Subcarrier bandwidth (kHz)	15					
Physical resource block (PRB) bandwidth (kHz)	180					
Number of available PRBs	6	12	25	50	75	100

Table 2.2: Available downlink bandwidth is divided into PRBs [67]

Figure 2.12 illustrates the DL resource grid which consists of N_{BW} subcarriers for a duration of the DL slot T_{slot}. A resource element, that is, each box within the grid, represents a single subcarrier for one symbol period. Note that in case of MIMO applications, see Section 2.2.5 for more details, there is a resource grid for each transmitting antenna [67].

2.2.5 LTE-Advanced (LTE-A)

LTE-Advanced (LTE-A) focus was to provide higher bitrates in a cost efficient way and, at the same time, completely fulfill the requirements set by ITU for IMT-Advanced,

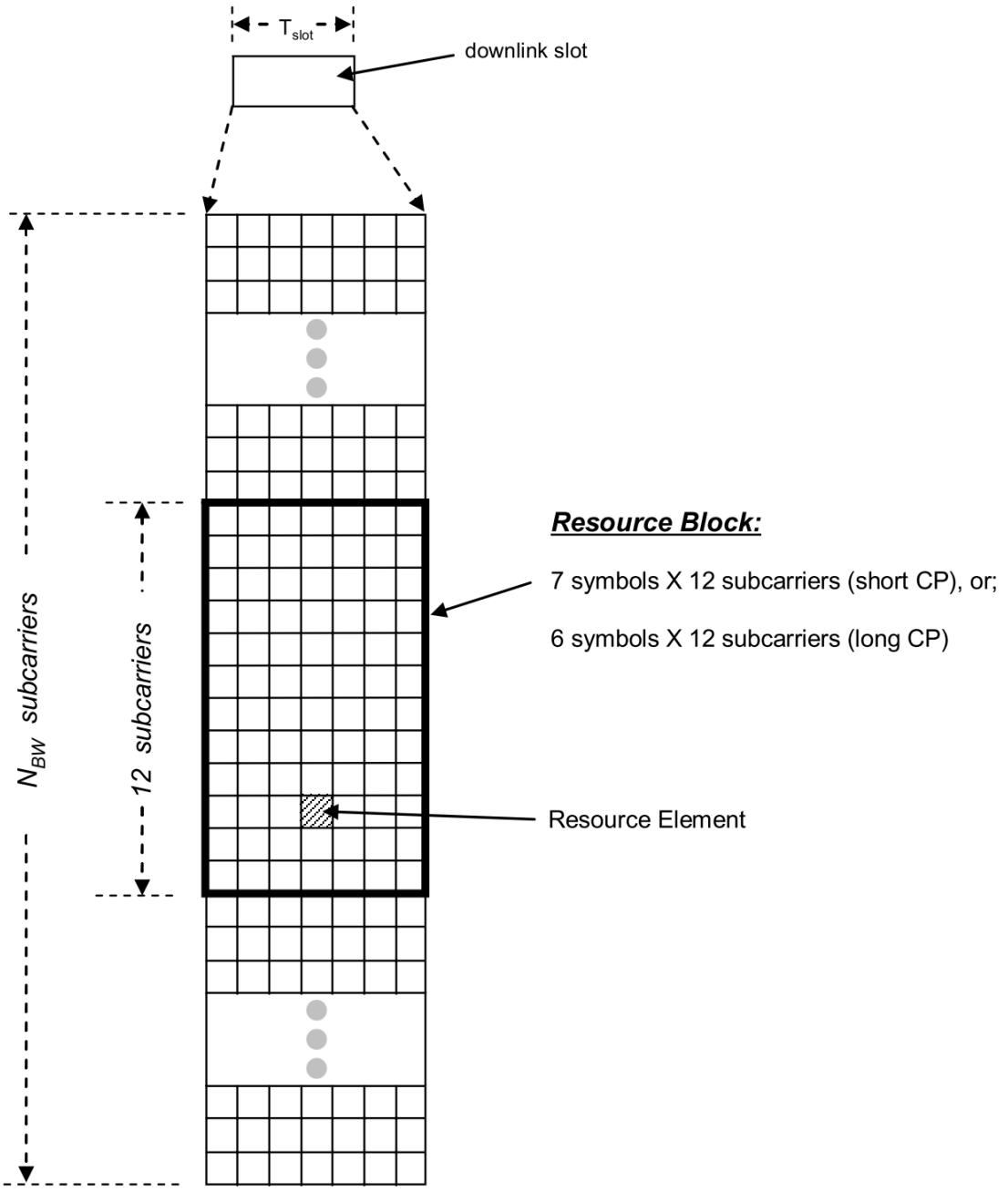

Figure 2.12: Downlink Resource Grid [67]

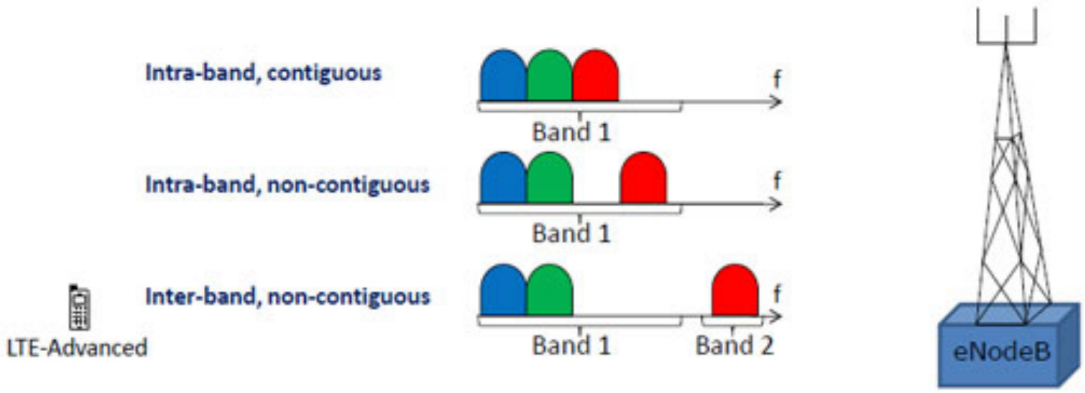

Figure 2.13: Carrier Aggregation [2]

also referred to as 4G [2].

The main features introduced in LTE-Advanced are Carrier Aggregation (CA), enhanced use of multi-antenna techniques and support for Relay Nodes (RNs). The latter will be further detailed in Section 2.3.3.

Carrier Aggregation (CA)

LTE-Advanced aims to support peak data rates of 1 Gbps in the DL and 500 Mbps in the UL. To achieve this, a transmission bandwidth of up to 100 MHz is required but the availability of such large portions of contiguous spectrum is rare in practice. LTE-A uses Carrier Aggregation (CA) of multiple Component Carriers (CCs) to achieve high-bandwidth transmission [56].

The component carriers can have a bandwidth of 1.4, 3, 5, 10, 15 or 20 MHz and a maximum of five component carriers can be aggregated. Hence the maximum bandwidth is 100 MHz. The number of aggregated carriers can be different in DL and UL, however the number of UL component carriers is never larger than the number of DL component carriers. The individual component carriers can also be of different bandwidths [2].

The easiest way to arrange aggregation is to use contiguous component carriers within the same operating frequency band, so called intra-band contiguous. For non-contiguous allocation, it could either be intra-band, i.e. the component carriers belong to the same operating frequency band, but are separated by a frequency gap, or it could be inter-band, in which case the component carriers belong to different operating frequency bands, see Figure 2.13.

So, CA facilitates the efficient use of fragmented spectrum. In addition, it supports HetNets (see Section 2.3).

Multiple-Input Multiple-Output (MIMO)

Multiple-Input Multiple-Output (MIMO) is used to increase the overall bitrate through transmission of two (or more) different data streams on two (or more) different antennas - using the same resources in both frequency and time, separated only through use of different reference signals - to be received by two or more antennas [2].

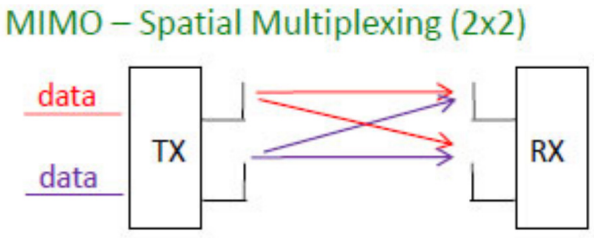

Figure 2.14: Simplified illustration of 2x2 MIMO [2]

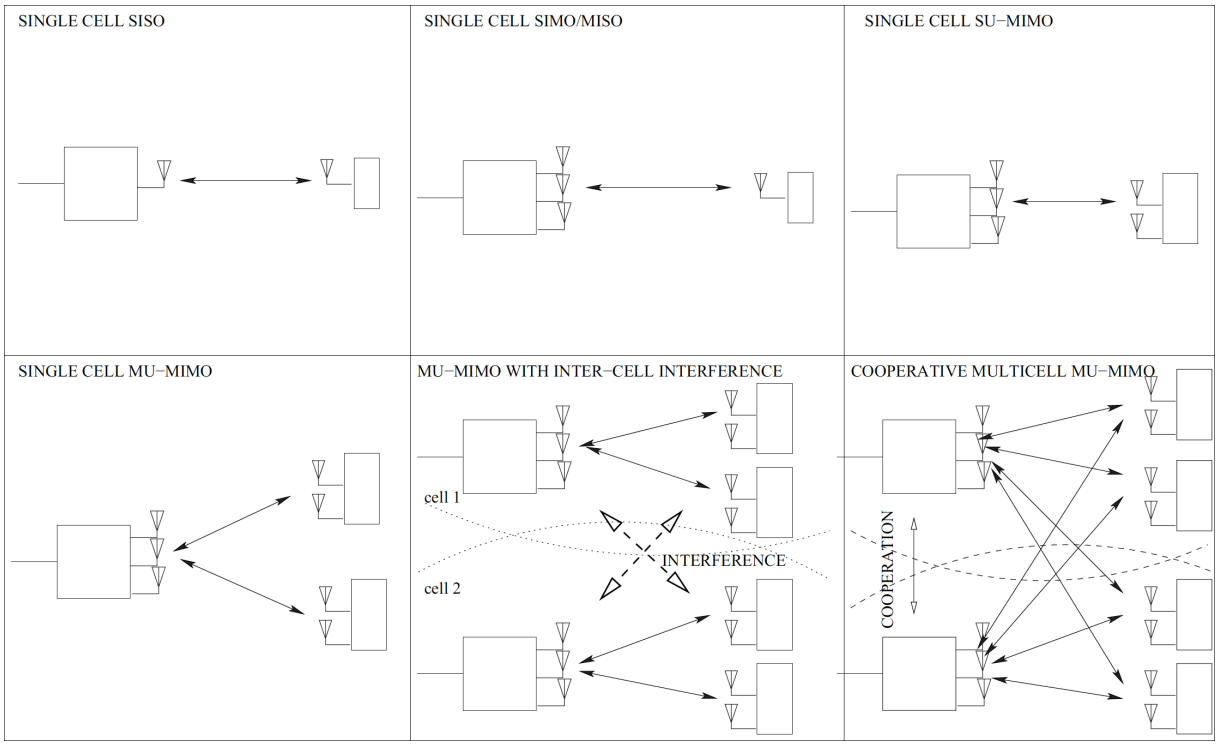

Figure 2.15: The evolution of MIMO technology [56]

An example of 2x2 MIMO can be seen in Figure 2.14, in which two different data streams are transmitted on two Tx (transmitter) antennas and received by two Rx (receiver) antennas, using the same frequency and time, separated only by the use of different reference signals.

Figure 2.15 shows the evolution of MIMO technologies from traditional single antenna communication to multi-User MIMO scenarios: Single Cell Single-Input Single-Output (SISO), Single Cell Single-Input Multiple-Output (SIMO) / Multiple-Input Single-Output (MISO), Single Cell Single-User MIMO (SU-MIMO), Single Cell Multiple-User MIMO (MU-MIMO), MU-MIMO with Inter-Cell Interference and Cooperative Multicell MU-MIMO.

A major change in LTE-Advanced is the introduction of 8x8 MIMO in the DL and 4x4 in the UL.

Coordinated MultiPoint (CoMP)

Coordinated MultiPoint is a type of cooperative MIMO (see Figure 2.15) whose introduction aims to improve network performance at cell edges. In CoMP a number of Tx points provide coordinated transmission in the DL, and a number of Rx points provide coordinated reception in the UL. A Tx/Rx-point constitutes of a set of co-located Tx/Rx antennas providing coverage in the same sector. The set of Tx/Rx-points used in CoMP can either be at different locations, or co-sited but providing coverage in different sectors, they can also belong to the same or different eNBs. CoMP can be done in a number of ways, and the coordination can be done for both homogeneous networks as well as HetNets [2].

2.3 Heterogeneous Networks (HetNets)

Heterogeneous Networks (HetNets) have been proposed to increase the capacity of LTE networks by introducing low power nodes (LPNs), e.g. pico and femto cells, overlaid into the existing coverage area of the macro cells, high power nodes (HPNs). These networks are *heterogeneous* due to this mix of macro cells and LPNs, in the deployment sense [11] [20].

HetNets came trying to overcome the theoretical limit that other advanced technologies like CA, MIMO and CoMP are reaching. Under low Signal to Interference plus Noise Ratio (SINR) conditions, where received powers are low due to attenuation and/or interference might be high, such techniques may not work well whereas HetNets can [30] [35].

LPNs are usually deployed in high density areas, *hotspots*, to share the macro cell load or to extend its coverage area and their deployment are not necessarily planned [12].

Macro cells, pico cells and femto cells (tiers 1, 2, and 3 respectively) compose the building blocks of a HetNet. Tier 1 cells have more capacity, bigger coverage area, greater power transmission, higher signal strength and are more expensive to deploy in comparison to tier 2 and so on [32] [66].

Figure 2.16 shows a typical architecture of a traditional cellular network which can be considered an homogeneous network. Likewise, Figure 2.17 displays the architecture of HetNets.

2.3.1 Macro and Micro cells

Macro cells have transmission range of several kilometers, while the transmission range for micro cells is of hundreds of meters. Both are costly to deploy and maintain [15].

Macro cells, created by base stations called eNodeBs (eNBs), are usually deployed by the operators, so they are spread in a way to maximize the coverage.

2.3.2 Pico and Femto cells

Both, pico and femto cells aim to cover small areas, with a transmission range of tens of meters and are cheaper to maintain and easier to deploy than macro and micro cells.

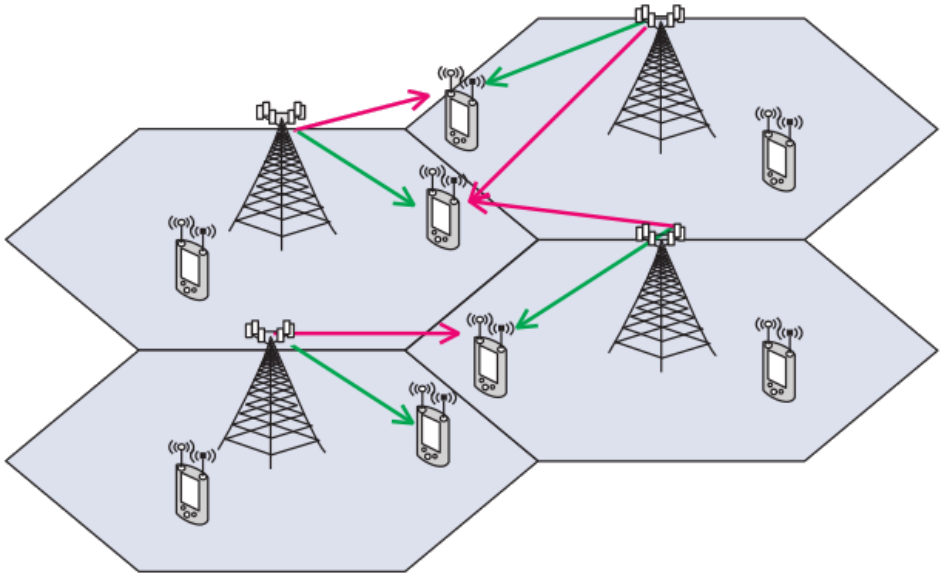

Figure 2.16: Homogeneous Networks Architecture [38]

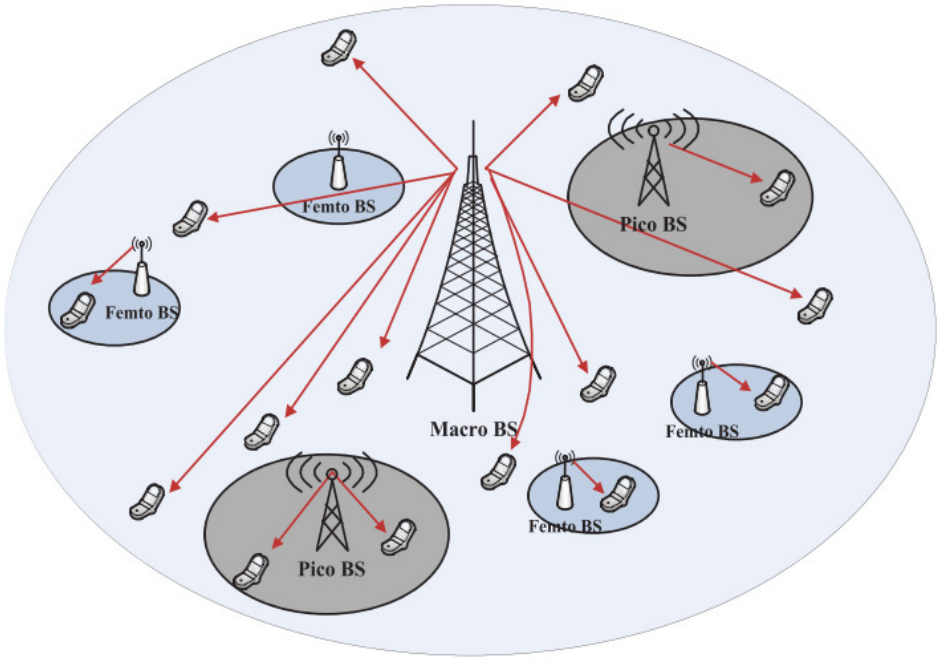

Figure 2.17: Heterogeneous Networks Architecture [66]

Pico cells are fully functioning eNBs with reduced size while femto cells, also known as Home eNBs (HeNBs), are usually used for connecting personal equipment and their transmission range is limited to a home or office area, for instance [15].

2.3.3 Relay Nodes (RNs)

The Relay Nodes (RNs) are low power base stations that will provide enhanced coverage and capacity at cell edges, and hot-spot areas and it can also be used to connect to remote areas without fibre connection. Figure 2.18 shows how the relay node, through

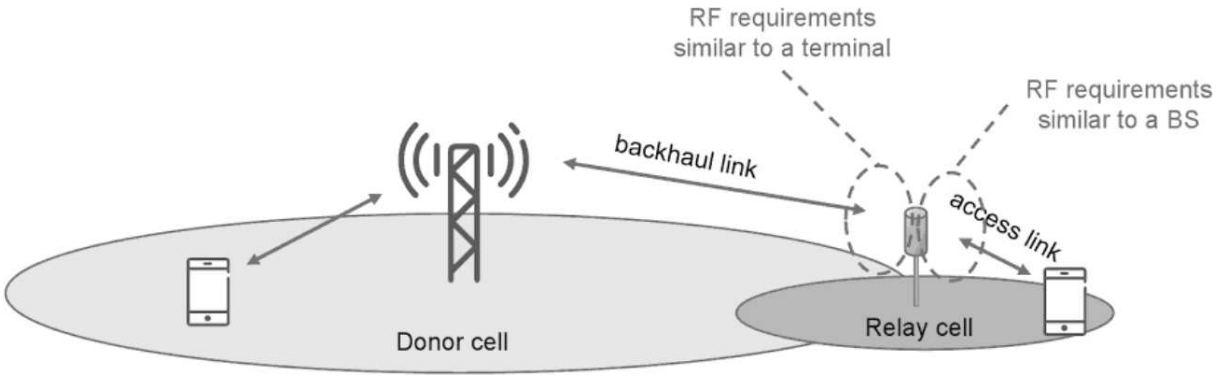

Figure 2.18: LTE-Advanced relaying [19]

its relay cell, extends the coverage area of the donor cell created by the eNB.

The Relay Node is connected to the Donor eNB (DeNB) via a radio interface. So, in the Donor cell the radio resources are shared between UEs served directly by the DeNB and the Relay Nodes.

2.3.4 Challenges

HetNets differ from the so-called traditional cellular network in many ways, so, a summary of such differences can be observed on Table 2.3.

The major challenges for HetNets are interference mitigation, user association and resource allocation problems. Unlike classical wireless networks, in HetNets, the number of choices or configurations increases exponentially with the number of deployed LPNs. Thus, centralized resource management algorithms are not cut out for the massive overhead in computation and signaling required by the LPNs [50].

2.4 Summary

In this chapter, following a top-down approach, an overview about the evolution of mobile networks and their main characteristics was presented. Afterwards, an introduction about LTE and its main concepts were shown. At last, details about HetNets were explained along with some new challenges such approach brings up. In the next chapter, some relevant strategies to deal with these challenges will be presented.

Aspect	Traditional Cellular	HetNet
Performance Metric	Outage/coverage probability distribution (in terms of SINR) or spectral efficiency (bps/Hz).	Outage/coverage probability distribution (in terms of rate) or area spectral efficiency (bps/Hz/m^2).
Topology	BSs spaced out, have distinct coverage areas. Hexagonal grid is an ubiquitous model for BS locations.	Nested cells (pico/femto) inside macrocells. BSs are placed opportunistically and their locations are better modeled as a random process.
Cell Association	Usually connect to strongest BS, or perhaps two strongest during soft handover	Connect to BS(s) able to provide the highest data rate, rather than signal strength. Use biasing for small BSs.
Downlink vs. Uplink	Downlink and uplink to a given BS have approximately the same SINR. The best DL BS is usually the best in UL too.	Downlink and uplink can have very different SINRs; should not necessarily use the same BS in each direction.
Mobility	Handoff to a stronger BS when entering its coverage area, involves signaling over wired core network.	Handoffs and dropped calls may be too frequent if use small cells when highly mobile, overhead a major concern.
Backhaul	BSs have heavy-duty wired backhaul, are connected into the core network. BS to UE connection is the bottleneck.	BSs often will not have high speed wired connections. BS to core network (backhaul) link is often the bottleneck in terms of performance and cost.
Interference Management	Employ (fractional) frequency reuse and/or simply tolerate very poor cell edge rates. All BSs are available for connection, i.e. "open access".	Manage closed access interference through resource allocation; users may be "in" one cell while communicating with a different BS; interference management hard due to irregular backhaul and sheer number of BSs.

Table 2.3: Differences between Traditional Cellular and HetNet [12]

Chapter 3

Literature Review

3.1 Load Balance in HetNets

The Load Balancing Problem consists of distributing tasks ("loads"), as evenly as possible among the entities that can process them according to their "capacity" [11]. In HetNets, the UEs request some kind of service from the BSs such as calls, video streaming, instant text messaging, etc. The aim of the load balancer is to distribute the UEs among all available BSs (LPNs and HPNs) in a way no BS becomes either overloaded or underutilized.

Load balancing strategies can be either centralized or distributed. Such strategies can rely on a load balancer, a centralized controller, which effectively performs the balance by assigning or moving resources from one service provider to another. Additionally, the entities that compose the network can adopt policies that lead towards a load balancing, for instance, BSs can have a cooperative user association policy in which they communicate with each other when deciding to serve or not a user.

With a centralized controller, it has all the information about the logical state of the system and can possibly make the best decision in terms of allocation to keep the system as evenly balanced as possible. Nevertheless, it becomes another hop in the network and a critical component too, that is, if it goes down all the routing is compromised.

Strategies that distribute the logic of this central controller make the coordination of the entities that compete for resources a real challenge. Using algorithimic game theory, it would be the same as making selfish players, in a possibly non-cooperative game, cooperate in such way that each of them receive their profits and all the combined strategies turn out to be an optimal solution.

Andrews et al [11] provided a good overview about load balancing in HetNets by pointing out open challenges such as:

- Mobility - Handovers among various types of cells in a HetNet must be seamless but it is known that such procedures are complicated and have a costly overhead.
- Asymetric Downlink and Uplink - the optimal downlink association is not necessarily optimal for uplink transmission. Hence, downlink load balancing strategies must be extended to cover uplink scenarios.

- Regulatory Issues - the regulatory focus should be on making it easier to deploy and use small cell infrastructure which could also include legal means to encourage such usage. In addition, LTE-A and WiFi networks should be able to hand off users seamlessly among them.

3.1.1 Centralized Load Balancing

Prasad, Arslan, and Rangarajan [54] considered the problem of maximizing the proportional fairness utility over HetNets by jointly exploiting cell dormancy and load balancing. A low complexity greedy algorithm for the load balancing sub-problem and an algorithm for the joint optimization problem were presented and nearly optimal results were reached. However, the proposed solution requires a central controller to provide all the information the algorithms need. In addition, its approximation algorithm relies on cell dormancy to maximize proportional fairness which may work very well for underloaded situations by turning off some of the BSs but not for the overloaded scenarios proposed on our simulations.

Anedda, Muntean and Murroni [13] considered the problem of balancing the real-time load of heterogeneous wireless networks using a centralized controller. The proposed mechanism ranks the networks and suggests the best network to the UE according to the traffic load and the application type running by the UE. It is a framework that depends on applications running on the UEs, BSs, and also on a load balancer. In our simulations, this strategy would behave alike Pico Cell First (see Section 5.1) due to the parameters we used, e.g., the UEs are uniformly distributed among 3 different services.

3.1.2 Distributed Load Balancing

Fotiadis et al [27] proposed a distributed Mobility Load Balancing scheme that adjusts Cell Individual Offsets (CIOs) in favor of under-utilized cells, relying on the exchanged load information. CIO is a parameter used in Cell Range Extension (CRE) technique. The higher the CIO is, the more attractive a cell is. This strategy relies on cooperation among cells and also transferring the load from one cell to the other for the sake of load balancing. It does not perform any kind of user association, therefore, it is not suitable to be compared with our strategies. Nevertheless, it can be used along with them to improve overall load balancing.

Hong et al [50] formulated the problem of using of HetNets to offload the traffic of existing cellular systems, traffic offloading, as a joint optimization problem of interference mitigation, user association and resource allocation. Then, they proposed a self-organizing algorithm based on Markov approximation and log-linear learning which was able to offload more than 90% of the traffic from the macro cell BSs to small cell BSs, in simulation results. This work focuses on offloading the traffic from high power nodes to low power nodes regardless the user admission policy. So, similarly to Fotiadis et al [27], it could be a partner that can work alongside our final distributed probabilistic heuristic.

3.2 User Association in HetNets

The objective of the user association problem is to find an optimal set of UE-BS associations such that the number of served UEs is maximized.

As previously pointed out in Section 2.3.4, user association policies in HetNets cannot rely on the same strategies used in homogeneous networks [12][48].

In addition, likewise load balancing schemes, user association strategies can be categorized as centralized or distributed according to the way the entities that decide the association works.

3.2.1 Centralized User Association

Fooladivanda, Daoud, and Rosenberg [25] formulated a joint association, channel allocation, and inter-cell interference management problem as a Non-linear NP-hard problem and compared it against SINR based strategies. Fooladivanda and Rosenberg [26] also proposed an association rule that combines both user and resource allocation with significant performance gains in throughput. Ye *et al* [66] presented a distributed algorithm for cell association and resource allocation whose complexity time is linear to the number of UEs and the number of BSs. The main goal of these works is to achieve user association and they analyze load balancing as a side effect.

Mishra, Rangineni, and Murthy [45] formulated a joint optimization problem of UE-BS association and Almost Blank Subframe (ABS) density as an NP-Hard problem. A relaxed version of the problem was reduced to a minimum weighted bipartite matching problem and it was solved using the Hungarian algorithm for a given ABS density. This approach is used iteratively on the set of candidate ABS densities to identify the optimal ABS density and minimize the number of blocked users. It tries to increase the attendance, but load balancing is not considered and it also needs a centralized controller.

Trabelsi et al [61] used game theory to present a framework to optimize user association and dynamic radio resource sharing, and to coordinate the inter-cell interference for enhancing overall LTE network utility. Such framework relied on a central controller to gather information from BSs and the UEs in the system to determine the optimal parameters.

Note that these works rely on frameworks whose characteristics make it difficult to compare to our strategies, for instance, some of them focus on cell interference and energy efficiency.

3.2.2 Distributed User Association

The CRE technique is part of 3GPP standardization efforts. This technique uses an association bias to artificially extend the coverage area of small cells by increasing their CIOs, hence, offloading users from the macro cell [11] [60]. Siomina and Yuan [58] proposed an optimization framework for load balancing in LTE HetNets by means of cell range assignment using cell-specific offsets. Choosing the optimal biases is not straightforward [66] and this technique cannot respond or adapt dynamically to real-time distribution of traffic among multi-tier cells [43].

Kim et al [36] proposed a framework for user association and presented an iterative distributed user association algorithm based on the BSs' loads and signal strength. In HetNets, it would privilege high power nodes over low power nodes. In addition, admission control policies were proposed when the system cannot stabilize due to excessive traffic loads. Due to its iterative characteristic, an active load balancing strategy reallocate UEs over time which is not well comparable with our user association strategies that already aim to balance the load.

Mlika, Driouch and Ajib [47] proposed a fully distributed algorithm based on a learning mechanism to solve the user association problem on HetNets. Such algorithm achieved near-to-optimal performance on simulations. They modeled the problem of associating users to BSs and allocating frequency channels using non-cooperative game theory and despite proving this game may not always admit pure Nash Equilibra, they are efficient when they exist. The success of this strategy is very dependent on what the UEs learn when playing, that is, the amount of iterations that are performed. Besides, it also requests more information from the BSs. In our dynamic simulations, the UEs just perform a single request and they leave afterwards, so, such strategy would not be suited to be compared to ours.

3.3 Joint User Association and Load Balance in HetNets

Arani et al [14] and Sohn et al [59] tackled the load balancing and user association problems in a distributed way. In Arani's work, a UE association policy and a resource allocation mechanism attempt to balance the load in a self-organizing fashion. This policy was based on BSs' estimated load and SINR at the location of the UE. In Sohn's work, the distributed load balancing algorithm relies on message passing and aims to maximize the network-wide sum rate of all users. Both works use quite different frameworks in comparison to ours.

Recent works are handling massive Multiple Input-Multiple Output (MIMO) HetNets. In such scenario, the BS is equipped with a massive number of antennas (hundreds, thousands or even tens of thousands). Xu and Mao [65] used Algorithmic Game Theory to solve the problem and developed distributed user association algorithms that converged to the Nash Equilibrium.

3.4 Summary

In this chapter, we presented many related works and it can be noted that there is a big effort towards trying to develop new techniques to handle both user association and load balancing in HetNets.

For some works, the focus is just to maximize the number of served users and the load balancing is considered just as another quality parameter for such user association problems. On the other hand, there are some works that just focus on getting the best load balancing regardless the user association policy.

In addition, there are other works that consider them jointly but they either are not distributed or they rely on machine learning to solve the problem. If the strategies are not distributed, they usually rely on a centralized load balancer which may easily become the bottleneck of the whole architecture and, if it goes down, it would compromise the whole network. If the strategies rely on machine learning, data sets for training would be needed and, perhaps, a massive quantity of data must be collected until it starts performing well enough.

Finally, some works propose frameworks to address either one of them or even both the user association and load balancing problems altogether. For such works, the main question is to know how simple and feasible is to implement such solutions because the industry is looking for solutions that demand less modifications with the biggest revenue as possible. Our final distributed heuristic basically demands that the BSs provide how loaded they are and, with that, UEs can selfishly decide what is the best BS to connect to, ending up with solutions very close to the optimal ones.

In the next chapter, the proposed solution for the UE association and load balancing is going to be further explained. It evolved from a ILP solver-dependent strategy to a fully distributed load-aware probabilistic heuristic.

Chapter 4

Load Balance and User Association in HetNets

In this chapter, our proposed solution for the Load Balance and User Association in HetNets problem will be presented. Section 4.1 introduces an abstraction of the system including some assumptions and simplifications adopted to narrow down the system model.

Section 4.2 presents three strategies to solve this problem. Firstly, in 4.2.1, the problem is formulated as an integer linear programming (ILP) problem. Secondly, in 4.2.2, a greedy heuristic over solutions of a relaxed version of the proposed ILP is presented as a faster alternative. Finally, in 4.2.3, a distributed probabilistic load-aware heuristic is proposed as an even faster strategy to solve this problem.

4.1 System Model

We consider a two-tier cellular network consisting of a set of Pico and Macro Base Stations distributed as shown in Figure 4.1. The values in this figure indicate the percentage of the total number of UEs whose position is uniformly distributed within each cell.

Despite being represented as hexagons, the BSs' coverage area are circles and they actually overlap their neighbor's coverage area. Such fact is illustrated in Figure 4.2, the green area is the intersection of coverage areas of neighbor cells.

Downlink and uplink communication do not behave similarly on HetNets. The focus of this work, therefore, is on the downlink problem.

The downlink capacity of a given BS is given by the number of Physical Resource Blocks (PRBs) that are available to transport data in a timeframe of 1 second considering that Orthogonal Frequency-Division Multiple Access (OFDMA) is used and a bandwidth of 20 MHz. The PRBs used to control the communication are not being considered.

Each UE can connect to a single BS and it requests a minimum data rate in Mbps. However, the same number of PRBs does not necessarily delivers the same data rate for two distinct UEs due to different levels of interference.

In order to compute the number of needed PRBs a BS b has to provide to satisfy the

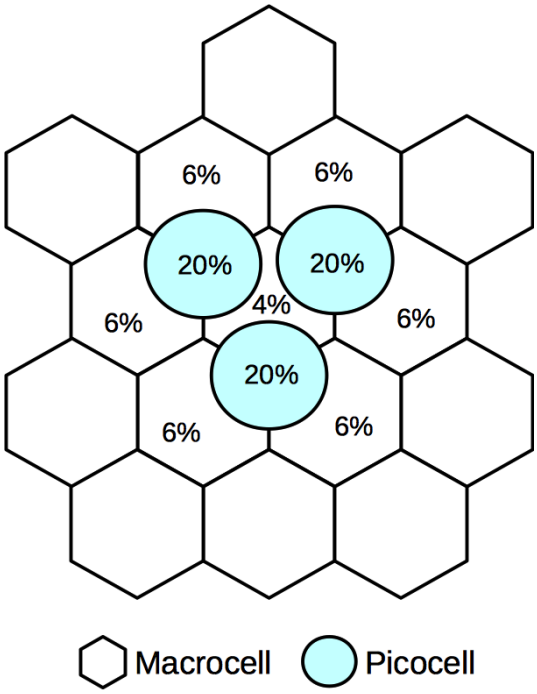

Figure 4.1: Simulation scenario (Adapted from [41])

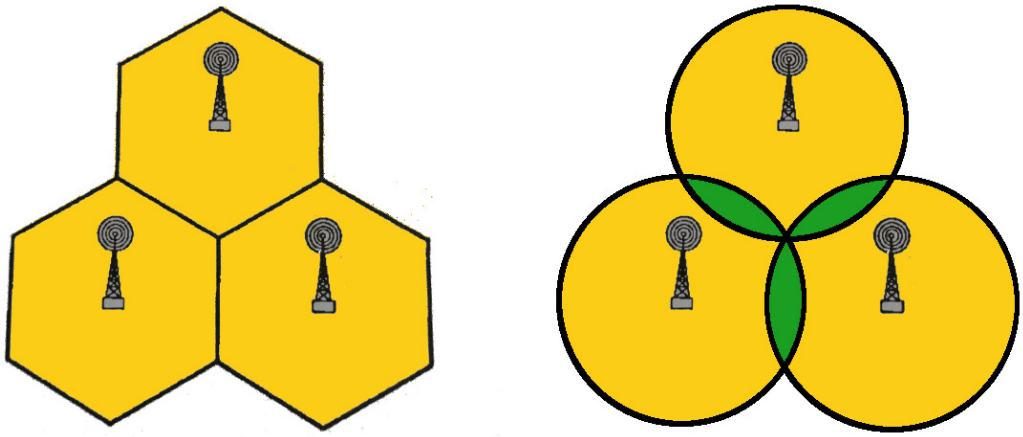

Figure 4.2: Base Stations' coverage area (Adapted from [24])

desired data rate for a given UE m, we use the Shannon equation that gives the maximum data rate bm supports under the presence of interference.

In Shannon's equation, the maximum data rate of the channel, denoted by C_{bm}, depends on the bandwidth W and on the $SINR_{bm}$ of the channel and is given by:

$$C_{bm} = W \cdot \log_2(1 + SINR_{bm}) \tag{4.1}$$

By equation 4.1 we need to compute the $SINR_{bm}$ which can be defined as the ratio of the signal power to the summation of the average interference power from the other cells and the background noise power [8] [53] [62]:

$$SINR_{bm} = \frac{P_b \cdot A_{mb}(\vec{q_m}, \theta_b)}{N + \sum_{b' \in B \setminus \{b\}} \rho_{b'} \cdot P_{b'} \cdot A_{mb'}(\vec{q_m}, \theta_{b'})} \tag{4.2}$$

where B is the set of BSs, P_b is the transmit power of the BS $b \in B$, θ_b is the network wide downtilt angle which is constant in this simulation. A_{mb} is the overall attenuation incurred by m if connected to b; $\vec{q_m}$ is the position of m which is constant; N is the thermal noise power; $P_{b'}$ is the maximum transmit power of each one of the other BSs; and $\rho_{b'}$ is the cell availability, i.e., the fraction of available PRBs in b'.

After calculating the $SINR_{bm}$, we can compute the maximum data rate C_{bm} supported by the channel between b and m. The number of PRBs in a timeframe of the LTE network is N_W for each bandwidth W. The effective data rate supported by one PRB in the channel between b and m is $d = C_{bm}/N_W$. Given that the user m requires a data rate of l Mbps, we then compute the number of PRBs required to achieve this data rate: $l_{mb} = l/d = N_W/(l \cdot C_{bm})$.

LTE supports peak downlink speeds of up to 300 Mbps with maximum channel bandwidth of 20 MHz, which corresponds to 100 PRBs [57] [67].

We consider that UEs request different services, that is, different minimum data rates. In addition, note that the UEs are not susceptible to exactly the same interference, which affects the minimum number of required PRBs.

Although the services requested by users differ by the minimum rate they demand, priority among them is not being considered in this work.

Furthermore, we assume the interference, the position of the UE, and the minimum number of PRBs required by a UE are constant in the considered timeframe.

4.2 Problem Formulation

There are many association strategies that aim to associate the UEs with BSs to maximize the number of accepted UEs. In these strategies, load balancing is considered as another parameter of quality and not the final goal [25] [26] [66].

This section introduces an ILP model, a greedy linear programming heuristic, denoted by LP_GREEDY, and a distributed probabilistic strategy, called PROB_MIN_LOAD, for the problem of UE association considering load balancing as an objective. The goal of the introduced strategies is to perform user association taking into account the minimum data rate requested by each UE as well as the available downlink capacity of the BSs in

such a way that the total load is evenly distributed among macro and small cells.

4.2.1 Integer Linear Programming Model

Objective:

Minimize the difference in terms of load between the most overloaded BS (L_{max}) and the least overloaded BS (L_{min}).

Definitions:

M = set of UEs
B = set of BSs
$M_b = \{m \in M; m \text{ is in coverage area of } b \in B\}$
c_b = capacity of $b \in B$ in PRBs
l_{mb} = minimum load of PRBs that $m \in M_b$ demands from $b \in B$ (this number includes the losses caused by interference)
α = minimum percentage of UEs that must be served

Decision variables:

$x_{mb} \in \{0, 1\}$ indicates whether $m \in M_b$ should be served by $b \in B$ or not
$L_{max} \in [0, 1]$ is the fraction of used PRBs of the most overloaded BS in the found solution
$L_{min} \in [0, 1]$ is the fraction of used PRBs of the least overloaded BS

Model:

(P.1)

$$\min L_{max} - L_{min}$$

$$\text{s.t.} \quad L_{max} \geq \frac{1}{c_b} \sum_{m \in M_b} l_{mb} x_{mb} \qquad \forall b \in B \tag{4.3}$$

$$L_{min} \leq \frac{1}{c_b} \sum_{m \in M_b} l_{mb} x_{mb} \qquad \forall b \in B \tag{4.4}$$

$$\sum_{b \in B} x_{mb} \leq 1 \qquad \forall m \in M \tag{4.5}$$

$$\sum_{m \in M_b} l_{mb} x_{mb} \leq c_b \qquad \forall b \in B \tag{4.6}$$

$$\sum_{b \in B} \sum_{m \in M_b} x_{mb} \geq \alpha |M| \tag{4.7}$$

$$x_{mb} \in \{0, 1\} \qquad \forall m \in M, \forall b \in B \tag{4.8}$$

$$L_{min}, L_{max} \in [0, 1] \tag{4.9}$$

Objective Function: Minimize the difference between the most overloaded BS and the least overloaded one (load balancing). The constraints are:

(4.3) L_{max} must be the biggest among all loads.

(4.4) L_{min} must be the smallest among all loads.

(4.5) Each UE must be served by at most one BS.

(4.6) The sum of the loads of all served UEs in a BS cannot be greater than its capacity.

(4.7) The number of served UEs must be greater than α of all UEs. Without this constraint, $\{x_{mb} = 0, \forall m \in M, \forall b \in B\}$ would be an optimal solution.

We want the largest possible value of α, i.e., to serve the largest number of UEs, and for this value the ILP optimizes the load balancing. This value of α is found via an iterative process described below.

For each fixed value of α in the finite set $\{1.00, 0.99, 0.98, \ldots, 0.00\}$, in decreasing order of value, we solve the above ILP and check if a feasible solution could be found. As soon as a feasible solution is found for the ILP, the process stops. This way we are trying to find the solution that gives the best possible load balancing and serves the largest number of UEs.

A simple implementation of this method results in an algorithm that in the worst case has to solve 100 ILPs. In our implementation, instead of this linear search of the greatest value of α inducing a valid ILP, we perform a binary search in the finite set $\{1.00, 0.99, 0.98, \ldots, 0.00\}$ which results in at most 7 ILPs to be solved.

In addition, the partition problem can be reduced to this load balancing problem, so optimal solutions can only be found with exponential time algorithms, assuming $P \neq NP$ [29].

An instance of the partition problem is a set of natural numbers S and it needs to be decided whether there exists a partition of S in two sets A and $S \setminus A$ such that

$$\sum_{e \in A} e = \sum_{e' \in S \setminus A} e' = \frac{1}{2} \sum_{\hat{e} \in S} \hat{e}$$

Similarly, a version of the decision problem of P.1 is to decide whether it is possible to allocate all UEs such that $L_{max} - L_{min} \leq k$. In order to make the reduction in polynomial time, for each $e \in S$, create a UE, $m_e \in M$, that demands the load e. In addition, create two BSs $b_1, b_2 \in B$ with infinity capacity such that m_e is served by b_i if and only if $e \in B_i$, where $i = \{1, 2\}$. Thus, there is a solution for the partition problem if and only if there is a solution for our problem with $L_{max} - L_{min} \leq 0$.

4.2.2 Greedy Linear Programming Heuristic

The ILP model described previously gives an optimal solution but it is computationally expensive as shown in Section 4.2.1. Thus, in order to get a possibly good solution in polynomial time, the integrality constraints of the variables x_{mb} were relaxed and a greedy strategy was applied over the relaxed solution.

In Algorithm 1, given the solution of the relaxed version of the linear program, LP_SOL, all decision variables x_{mb} are enqueued in descending order of their values.

Algorithm 1 Greedy strategy over LP Solution

```
1: procedure LP_GREEDY(LP_SOL)
2:     Let x = (x_11, x_12, ..., x_|M||B|) be the fractional solution of LP_SOL
3:     Sort x in descending order of values
4:     for each x_mb in this order do
5:         if m is not associated then
6:             associate m with b if b has enough capacity
7:         end if
8:     end for
9: end procedure
```

Afterwards, this queue is consumed in this order, and for each x_{mb}, if the UE m was not previously associated, it is then associated to the corresponding BS b in case it has available capacity to serve m. If b is overloaded, the UE is not associated to any BS in this iteration but it will possibly be in a future one while the queue is being consumed.

4.2.3 Distributed Probabilistic Minimum Load Heuristic

The heuristic LP_GREEDY is fast but it is relatively complex to be implemented since it requires the resolution of an LP program. Moreover, it requires a stateful central controller. To overcome these limitations, a distributed probabilistic heuristic is proposed in this section.

The probabilistic heuristic requires that each UE receives information about the load of its reachable BSs, which can be easily done if each BS transmits this information. Then the UE probabilistically decides to which BS it is going to connect based on the received load information.

Algorithm 2 Probabilistic Minimum Load Strategy

```
1: procedure UE.PROB_MIN_LOAD(ReachableBS)
2:     total = 0.0
3:     for each b in ReachableBS do
4:         available_b = (capacity_b − load_b)/capacity_b
5:         total = total + available_b
6:     end for
7:     for each b in ReachableBS do
8:         Pr(b) = available_b/total
9:     end for
10:    Assign m to one of the reachable BSs considering the computed probabilities
11: end procedure
```

We assume that each UE independently runs Algorithm 2. Each UE independently calculates the percentage of available capacity for each of its reachable BSs. These values are used to define the probability distribution of the chance of connection to each reachable BS. Note that the higher the available capacity in some BS, the higher the probability that some UE connects to it, but since associations are done randomly, the load tends to be evenly distributed among the BSs.

However, such heuristic does not assure the constraint 4.7 will be respected, that is, a minimum number of served UEs is very likely to be greater than zero but it also may be below than *alpha* of all UEs.

In addition, note that the LP-based strategies depend on a centralized controller, that is, a load balancer on the system. Since this distributed probabilistic heuristic should be implemented as part of the UE's business logic, it could work along with load balancers by trying to make potentially good associations. So, it is likely even better results could be obtained by using a hybrid approach in which load balancers work along with user association policies implemented on the UEs.

4.3 Summary

In this chapter, we introduced the system model that is being considered and proposed an Integer Linear Programming (ILP) model to jointly solve the Load Balance and User Association problems in HetNets. Afterwards, two heuristics are proposed: a greedy one that depends of a solution a relaxed version of the ILP model and a distributed and probabilistic one which is independent of linear programming. The next chapter presents the performance evaluation of the proposed strategies and the comparison with state-of-the-art solutions.

Chapter 5

Simulations and Results

5.1 General Considerations

To evaluate the performance of the proposed heuristics, the scenario described on [41] and showed in Figure 4.1 was used, since it illustrates one of the most common situations in HetNets [12]: the deployment of picocells on high demanding service areas to share the macrocells' loads. In addition, a simulator[1] was implemented and the simulation parameters adopted in the proposed scenario were based on a 3GPP recommendation [7] and are presented in Table 5.1.

The ILP model and the proposed heuristics were compared to the following strategies that were proposed by Fooladivanda [25] and Ye et al [66]:

- Best Downlink: UEs are associated with the BS which presents the best downlink bit rate.
- Best SINR: UEs are associated with the BS which presents the best SINR (Signal to Interference plus Noise Ratio).
- Pico Cell First: UEs are associated with the pico cell as long as its SINR is better by a given factor in comparison with the macro cell.
- Range Expansion: it is similar to Best SINR strategy but a multiplicative SINR bias is assigned to each BS tier.
- Rate Bias: alike Range Expansion but the biasing factor is in the exponential term of the SINR instead.

Our algorithms are denoted by ILP_SIMPLE_MODEL (Section 4.2.1), LP_GREEDY (Section 4.2.2) and PROB_MIN_LOAD (Section 4.2.3). The comparison is performed considering static and dynamic scenarios.

A framework to perform the simulations and all the strategies were implemented in Java 7. This framework included tools to generate the simulation scenarios, classes to model HetNet entities, a common interface to ease the implementation of the several strategies, a CPLEX LP file generator and utilities to aid the analysis of the results.

[1] `https://bitbucket.org/alehirata_team/load-balancer-simulator`

In addition, Bash scripts were written to automate the execution of multiple test scenarios and the generation of results report. Besides, the ILP problems were solved using IBM ILOG CPLEX Optimization Studio 12.5.1.

The source code of the simulator is publicly available at `https://bitbucket.org/alehirata_team/load-balancer-simulator`.

In order to run the simulations, a Macbook Pro Mid-2014 was used. In summary, it runs the macOS with a 2.5 GHz Quad-Core Intel Core i7, 16 GB of DDR3 RAM and a 512 GB SSD.

Parameters	Values/assumptions
Total number of active UEs ($\|M\|$)	100, 200, 300, ..., 1000
Number of cells	Seven macro cells and three pico cells
Cells layout (as shown in Figure 4.1)	Seven hexagonal-grid loaded macro cells with wrap-around. Three loaded pico cells are deployed at the boundary of the central macro cell with wrap-around.
Inter-site distance of macro cells	500m
UE deployment (Fig. 4.1)	Each pico cell is deployed by 20% of UEs among $\|M\|$. These UEs attach to the BS of the pico cell. Each macro cell (except the central) is deployed by 6% of UEs among $\|M\|$. These UEs attach to BSs of corresponding macro cells. 4% of UEs among $\|B\|$ are deployed to and attach to the central macro cell.
The downlink Tx power of the BS in the macro cell	46 dBm
The downlink Tx power of the BS in each pico cell	30 dBm

Table 5.1: Parameters used in the simulation scenarios.

5.2 Static Scenarios

5.2.1 Simulations

To create a static scenario, the positions of the UEs were calculated randomly but respecting the distribution illustrated on Figure 4.1. Furthermore, the UEs request bandwidth for three different services according to the parameters shown in Table 5.2. The services are evenly distributed among the UEs and, for each fixed value of the total number of UEs, 30 randomly different test scenarios were generated.

In order to avoid too long runs of CPLEX when trying to find optimal solutions of ILPs and LPs, it was configured not to exceed 5 minutes of run time per problem.

Telecommunication Service	Minimum Bandwidth
Voice over IP (VoIP) Call	100 kbps
Video Call (HD)	1.5 Mbps
FullHD Video Streaming or Group video call (+7 people)	8 Mbps

Table 5.2: Minimum bandwidth required by the services used in the simulations [6].

5.2.2 Results

The figures presented in this section show mean values with confidence intervals of 95% derived using the independent replication method. Figures 5.1, 5.2, and 5.3 present the results for the static scenarios.

Figure 5.1 presents the percentage of associated UEs by each one of the implemented algorithms. We can see that LP_GREEDY accepts as much UEs as the ILP-based strategy, ILP_SIMPLE_MODEL. Besides, PROB_MIN_LOAD is the best strategy which does not rely on linear programming. Moreover, the proposed strategies accept more UEs than the competing strategies of the literature specially for overloaded scenarios (more than 200 UEs).

Figure 5.2 shows the load balancing results. Both proposed strategies, LP_GREEDY and ILP_SIMPLE_MODEL, balance the load more evenly than the other strategies. Furthermore, PROB_MIN_LOAD is also the best among the ones that do not use linear programming. In addition, note that all the other strategies leave at least one cell with 0% of load (points in the figure where $L_{max} - L_{min} = 100\%$) because they do not consider the load of the cells when deciding the assignment of a UE. Thus, for UEs in the boundary of the coverage area of a loaded cell, the chance of choosing the loaded cell or the unloaded one is the same.

Figure 5.3 presents the average load of the BSs. The strategies proposed in this work provide a better usage of the available network resources than the other strategies. This can be partially explained by the fact that the proposed algorithms were able to serve more UEs than the other strategies. As explained before, the higher acceptance rate is possible due to the load awareness because the UEs in the boundary of loaded cells choose the unloaded cells in the neighborhood instead of the loaded ones.

In terms of run time, while LP-based strategies were taking about 3.5 seconds to solve an instance of the problem, the other strategies were taking around 350 milliseconds on the 100 UEs scenario. Going up to 500 UEs, LP-based strategies were taking more than 10 minutes while other strategies were not taking more than 400 milliseconds. Finally, on the 1000 UEs scenario, LP-based strategies were taking more 15 minutes while the others were not reaching 500 milliseconds.

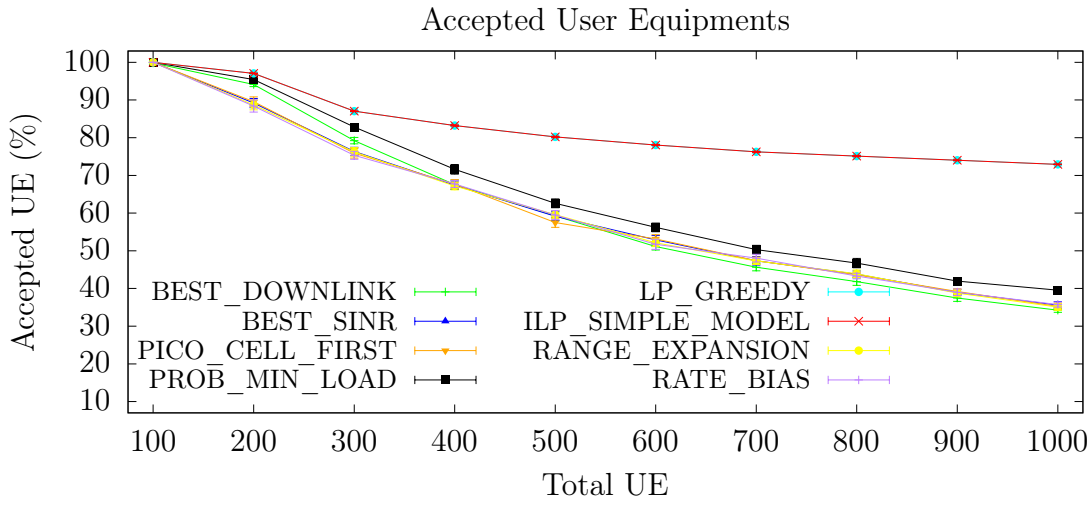

Figure 5.1: Static Scenarios - Accepted Users

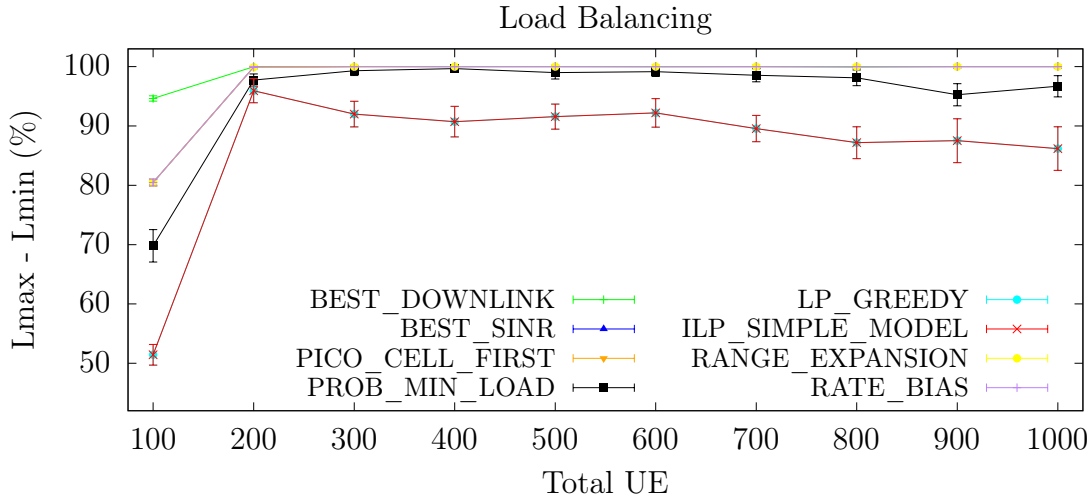

Figure 5.2: Static Scenarios - Load balancing

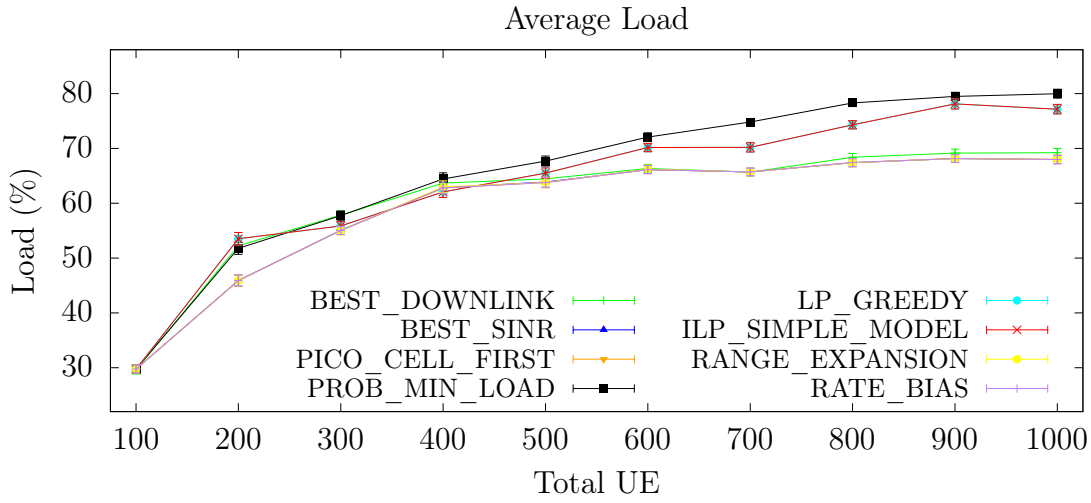

Figure 5.3: Static Scenarios - Average Load

Considering the metric of load balancing, we can see that both proposed strategies that are independent of ILP, LP_GREEDY and PROB_MIN_LOAD produce as good results as the other strategies from the literature for underloaded cells scenarios, and present better results on overloaded cells scenarios. It is also interesting to note that while our algorithms main objective is to improve load balancing, LP_GREEDY and PROB_MIN_LOAD not only achieved better results on load balancing but also accepted more users even in comparison with strategies that aim for user attendance.

5.3 Dynamic Scenarios

5.3.1 Simulations

The static scenarios show how the strategies would behave when they need to match all the UEs with all the BSs as much as possible but in a single shot. However, what happens in real life is that UEs come one by one or even in batches but not all of them at once. That is, it is very unlikely all UEs happen to turn on at the exact same time.

In order to evaluate the proposed solutions in dynamic scenarios, we also performed simulations with UEs randomly joining the system one by one respecting a fixed arrival rate for 60 minutes. Likewise the previous simulation, the values used in Table 5.1 are still valid but with the addition of the values in Table 5.3. Moreover, the same strategies from the literature used for comparison in the static scenario are also going to be considered here.

Parameters	Values/assumptions
UEs joining per minute	10, 20, 30, . . . , 120
Call duration	100s
Video Call duration	300s
FullHD Video Streaming duration	600s
Test duration	60min

Table 5.3: Parameters used in the dynamic simulation scenarios

5.3.2 Results

The figures presented in this section show mean values with confidence intervals of 95% derived using the independent replication method. Figures 5.4, 5.5, 5.6 present the results for the dynamic scenarios. Note that, in these charts, BEST_SINR, PICO_CELL_FIRST and RANGE_EXTENSION show almost identical results, that is why their curves are overlapped.

Figure 5.4 presents the percentage of associated UEs by each one of the implemented algorithms when scenarios with different UE arrival rates were tested. It can be observed that PROB_MIN_LOAD is the strategy that accepts more UEs. In addition note that BEST_DOWNLINK is close to PROB_MIN_LOAD but it is always accepting less users than the first one. Furthermore, note that the other strategies behave similarly but they are always outperformed in this criterion.

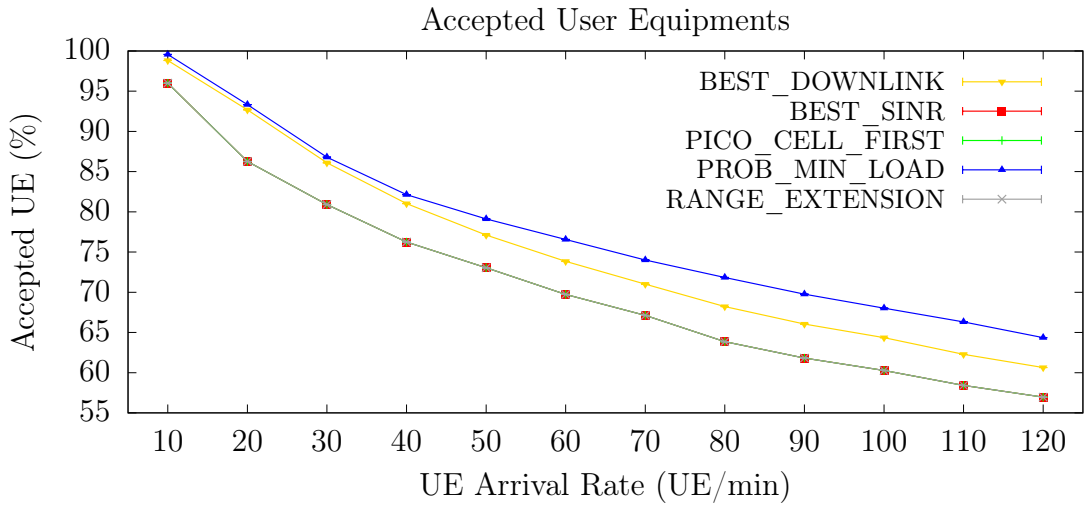

Figure 5.4: Dynamic Scenarios - Accepted Users

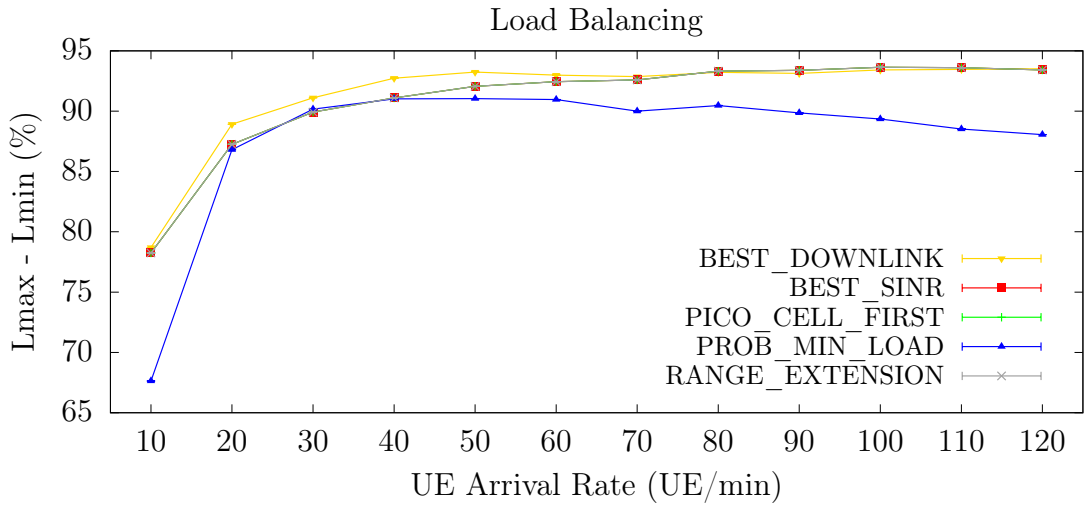

Figure 5.5: Dynamic Scenarios - Load balancing

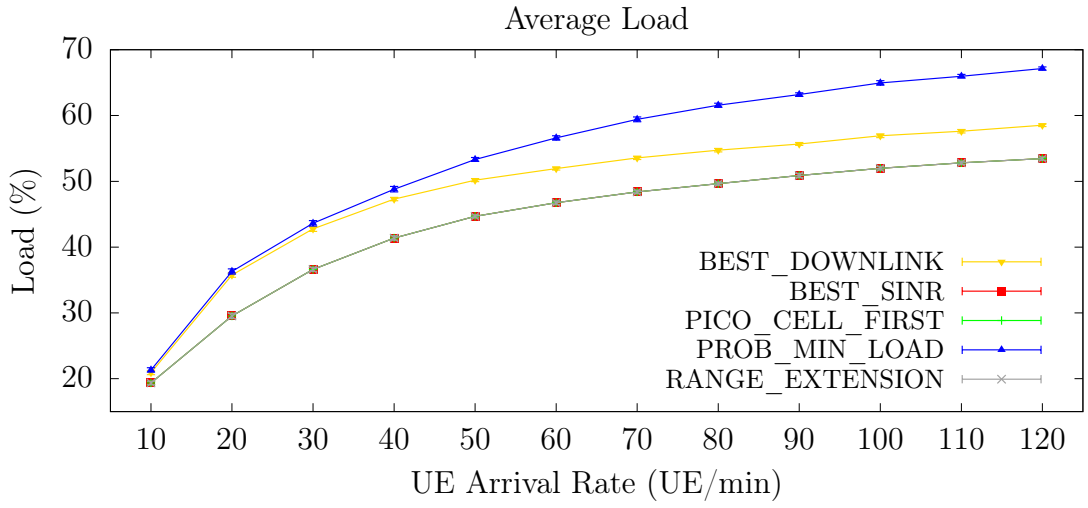

Figure 5.6: Dynamic Scenarios - Average Load

Figure 5.5 shows that the PROB_MIN_LOAD strategy provides the best load balancing results. However, in this criterion, BEST_DOWNLINK is not always better than the other strategies like in the previous one, i.e., it is accepting more UEs than the others but its load is less balanced among the BSs.

Figure 5.6 shows the average load of the BSs. Likewise the load results presented for the static scenarios (Figure 5.3), the strategy proposed in this work leads to a better usage of the BSs capacity. This can be partially explained by the fact that the proposed algorithm is able to accept more UEs than the other strategies. Likewise the results presented in Figure 5.5, it happens because the UEs in the boundary of loaded cells choose the unloaded cells instead of the loaded ones due to their load awareness.

Regarding run time, the strategies were taking around 400 to 900 milliseconds to simulate all the events of each proposed scenario.

Likewise the results presented by the static scenario, PROB_MIN_LOAD produces as good results as the other strategies from the literature for underloaded cells scenarios, and present better results on overloaded cells scenarios. It is also interesting to note that while our algorithm main objective is to improve load balancing, PROB_MIN_LOAD not only achieved better results on load balancing but also accepted more users even in comparison with strategies that aim for user acceptance.

5.4 Summary

In this chapter, we presented the performance evaluation of the strategies proposed in this work and compared them with some of the mains solutions proposed in the literature. According to the simulation results, the strategies proposed in this work perform better than the others specially in scenarios of heavy traffic load. The next chapter presents the conclusions as well as future works.

Chapter 6

Conclusions and Future Work

HetNets were introduced as a possible solution to increase the capacity and the coverage area of the existing homogeneous networks and it is part of the specifications of LTE, LTE-A and the emerging 5G.

HetNets presented a layered architecture in which LPNs can share the load of the high power nodes. Nevertheless, they also brought some challenges to be overcome such as the search for good strategies to both associate users and balance the load.

User association and load balancing strategies devised for homogeneous networks are no longer optimal in HetNets because they mostly relied on the signal strength between the UE and BS. In HetNets, such strategies would make the UEs to always try to connect to the high power nodes and the LPNs would be underutilized.

In this work, the user association and load balancing on HetNets problem was formulated as an integer linear program problem whose solution was used as a baseline. Afterwards, a greedy strategy over a relaxed version of this model was presented. Finally, a distributed probabilistic heuristic, that privileges BSs with greater availability was introduced.

According to the presented results, in static scenarios, the proposed strategies perform better than other non-optimal strategies, mainly on overloaded cells scenarios, by accepting more UEs and by sharing the load more evenly.

Note that, for static scenarios, the strategies based on linear programming presented the best results. However, they would require a centralized load balancer in order to be deployed in production. Besides, they are computationally expensive to be calculated since the linear program is modified every time a new UE needs to be served or a UE is no longer served by any BS.

On the other hand, the distributed heuristic would not require a load balancer, the UEs themselves would decide to which BS they would prefer to be connected to. UEs would just need the data about the load state of the BSs they are trying to connect to. For dynamic scenarios, this solution, also performed better than the other strategies on overloaded scenarios. In such situations, the load is more balanced among the BSs, more UEs are accepted and the capacity of the BSs are used in a more efficient way.

The strategies proposed here are not free of limitations; there is a great number of open challenges concerning User Association and Load Balancing in HetNets in which this work can be extended to.

Load Balancing and User Association are two different objectives that may oppose to each other when trying to maximize one or another. Thus, the proposed ILP model could be extended to use multi-objective optimization, so it would also have an objective function to maximize the user association instead of using binary search for *alpha*. An example of possible objective function to be considered is adding up the all x_{mb}, that is, the number of served UEs. Therefore, it is likely better solutions may be found.

This work focused on the downlink, however, there are some works that either consider just the uplink or try to consider both like Sapountzis et al [55]. In HetNets, the downlink and uplink are considerably different such that the decision of considering just downlink, only uplink or both can lead to totally different solutions. So, such scenario can be more thoroughly explored.

Besides, this work considered that there was no priority among the services and there was no fairness among them either. So, the given solutions may lead to starvation for some UEs that demand high bandwidth services. Moreover, by not considering priority among the services, the presented solution may refrain from accepting UEs using emergency services. For telecommunication switches in the industry, emergency services in general (e.g. calls to police, firemen, emergency medical services) must be processed at all costs, even if someone else's call must be dropped to free up some resources.

The current work assumed a UE can be served by just one BS. However, by using massive MIMO in HetNets like Xu and Mao [65], such restriction can be removed and possibly better solutions can be achieved. So, we could adapt our distributed strategy to this scenario and compare it with their strategy.

References

[1] 3GPP LTE. http://www.3gpp.org/technologies/keywords-acronyms/98-lte. Accessed: 2017-06-13.

[2] 3GPP LTE-Advanced. http://www.3gpp.org/technologies/keywords-acronyms/97-lte-advanced. Accessed: 2017-06-14.

[3] 3GPP Releases. http://www.3gpp.org/specifications/releases/. Accessed: 2017-06-12.

[4] LTE and LTE-Advanced capabilities. http://anisimoff.org/eng/lte_performance.html. Accessed: 2017-12-28.

[5] Questions and Answers in MRI. http://mriquestions.com/signal-squiggles.html. Accessed: 2017-10-09.

[6] Skype required bandwidth. https://support.skype.com/en/faq/FA1417/how-much-bandwidth-does-skype-need. Accessed: 2017-09-17.

[7] 3GPP. Study on RAN Improvements for Machine-type Communications. TR 37.868, 3GPP, 2011.

[8] F. Afroz, R. Subramanian, R. Heidary, K. Sandrasegaran, and S. Ahmed. SINR, RSRP, RSSI and RSRQ Measurements in Long Term Evolution Networks. *International Journal of Wireless & Mobile Networks*, 7(4):113–123, 2015.

[9] S. Agbo. *Principles of Modern Communication Systems*. Cambridge University Press, Cambridge, United Kingdom New York, NY, 2017.

[10] I. F. Akyildiz, D. M. Gutierrez-Estevez, and E. C. Reyes. The Evolution to 4G Cellular Systems: LTE-Advanced. *Phys. Commun.*, 3(4):217–244, December 2010.

[11] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon. An overview of load balancing in HetNets: old myths and open problems. *IEEE Wireless Communications*, 21(2):18–25, April 2014.

[12] J.G. Andrews. Seven ways that HetNets are a cellular paradigm shift. *Communications Magazine, IEEE*, 51(3):136–144, March 2013.

[13] M. Anedda, G. Muntean, and M. Murroni. Adaptive real-time multi-user access network selection algorithm for load-balancing over heterogeneous wireless networks. In *2016 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–4. IEEE, jun 2016.

[14] A. H. Arani, A. Mehbodniya, M. J. Omidi, and F. Adachi. Distributed Load Balancing User Association and Self-Organizing Resource Allocation in HetNets. In *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, pages 1–5. IEEE, sep 2016.

[15] E. Bodanese. A brief introduction to heterogeneous networks (hetnets) and its challenges. *IET Conference Proceedings*, pages 605–609(4), January 2011.

[16] S. Chen and J. Zhao. The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication. *IEEE Communications Magazine*, 52(5):36–43, May 2014.

[17] C. Cox. *An introduction to LTE : LTE, LTE-Advanced, SAE, and 4G mobile communications*. John Wiley & Sons, Hoboken, NJ, 2012.

[18] E. Dahlman, S. Parkvall, and J. Skold. *4G: LTE/LTE-advanced for mobile broadband*. Academic press, 2nd edition, 2013.

[19] E. Dahlman, S. Parkvall, and J. Skold. *4G, LTE-Advanced Pro and The Road to 5G, Third Edition*. Academic Press, 3rd edition, 2016.

[20] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi. A survey on 3GPP heterogeneous networks. *IEEE Wireless Communications*, 18(3):10–21, June 2011.

[21] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, and J. Yao. 5G on the Horizon: Key Challenges for the Radio-Access Network. *IEEE Vehicular Technology Magazine*, 8(3):47–53, Sept 2013.

[22] Ericsson. Ericsson mobility report. Tr, Ericsson, November 2016.

[23] Ericsson. Ericsson mobility report - interim update. Tr, Ericsson, February 2017.

[24] A. Ferrari. *Telecomunicaçãoes : evolução & revolução*. Érica, São Paulo, 1998.

[25] D. Fooladivanda. Joint channel allocation and user association for heterogeneous wireless cellular networks. *2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 384–390, September 2011.

[26] D. Fooladivanda and C. Rosenberg. Joint resource allocation and user association for heterogeneous wireless cellular networks. *IEEE Transactions on Wireless Communications*, 12(1):248–257, 2013.

[27] P. Fotiadis, M. Polignano, D. Laselva, B. Vejlgaard, P. Mogensen, R. Irmer, and N. Scully. Multi-Layer Mobility Load Balancing in a Heterogeneous LTE Network. In *2012 IEEE Vehicular Technology Conference (VTC Fall)*, pages 1–5, Sept 2012.

[28] R. Gallager. *Principles of Digital Communication*. Cambridge University Press, Cambridge, 2008.

[29] M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990.

[30] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. A. Thomas, J. G. Andrews, P. Xia, H. S. Jo, H. S. Dhillon, and T. D. Novlan. Heterogeneous cellular networks: From theory to practice. *IEEE Communications Magazine*, 50(6):54–64, June 2012.

[31] A. Gupta and R. K. Jha. A Survey of 5G Network: Architecture and Emerging Technologies. *IEEE Access*, 3:1206–1232, 2015.

[32] S. V. Hanly and P. A. Whiting. On the capacity of HetNets. In *Information Theory and Applications Workshop (ITA), 2014*, pages 1–9, Feb 2014.

[33] A. T. Hirata, E. C. Xavier, and J. F. Borin. Load balance and user association on HetNets. *IEEE Latin America Transactions*, 14(12):4781–4786, December 2016.

[34] H. Holma and A. Toskala. *LTE for UMTS-OFDMA and SC-FDMA based radio access*. John Wiley & Sons, 2009.

[35] M. Joud. Pico Cell Range Expansion toward LTE-Advanced Wireless Heterogeneous Networks. Master's thesis, Universitat Politècnica de Catalunya, 2013.

[36] H. Kim, G. Veciana, X. Yang, and M. Venkatachalam. Distributed α-Optimal User Association and Cell Load Balancing in Wireless Networks. *IEEE/ACM Transactions on Networking (TON)*, 20(1):177–190, 2012.

[37] J. F. Kurose and K. W. Ross. *Computer Networking: A Top-Down Approach*. Pearson, 6th edition, 2012.

[38] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana. Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges. *IEEE Communications Magazine*, 50(2):148–155, February 2012.

[39] P. Lescuyer and T. Lucidarme. *Evolved Packet System (EPS): The LTE and SAE Evolution of 3G UMTS*. Wiley Publishing, 2008.

[40] Y. Li, Z. Gao, L. Huang, X. Du, and M. Guizani. Resource management for future mobile networks: Architecture and technologies. *Computer Networks*, 0:1–7, apr 2017.

[41] S. Lien, T. Liau, C. Kao, and K. Chen. Cooperative access class barring for machine-to-machine communications. *Wireless Communications, IEEE Transactions on*, 11(1):27–32, January 2012.

[42] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. Wong, R. Schober, and L. Hanzo. User Association in 5G Networks: A Survey and an Outlook. *IEEE Communications Surveys & Tutorials*, 18(2):1018–1044, 2016.

[43] J. Liu, Y. Kawamoto, H. Nishiyama, N. Kato, and N. Kadowaki. Device-to-device communications achieve efficient load balancing in lte-advanced networks. *Wireless Communications, IEEE*, 21(2):57–65, April 2014.

[44] F. A. Elmaryami M. A. Emsaed, A. R. Zerek. Network solution from GSM to LTE. *International Conference on Control, Engineering & Information Technology (CEIT'14)*, 2014.

[45] S. Mishra, S. Rangineni, and C. S. R. Murthy. Exploiting an optimal user association strategy for interference management in HetNets. *Communications Letters, IEEE*, 18(10):1799–1802, Oct 2014.

[46] R. N. Mitra and D. P. Agrawal. 5G mobile technology: A survey. *{ICT} Express*, 1(3):132 – 137, 2015. Special Issue on Next Generation (5G/6G) Mobile Communications.

[47] Z. Mlika, E. Driouch, and W. Ajib. A fully distributed algorithm for user-base station association in HetNets. *Computer Communications*, 105:66–78, jun 2017.

[48] Z. Mlika, M. Goonewardena, W. Ajib, and H. Elbiaze. User-Base-Station Association in HetSNets: Complexity and Efficient Algorithms. *IEEE Transactions on Vehicular Technology*, 66(2):1484–1495, Feb 2017.

[49] H. G. Myung. Introduction to single carrier fdma. In *2007 15th European Signal Processing Conference*, pages 2144–2148, Sept 2007.

[50] T. Z. Oo, N. H. Tran, W. Saad, D. Niyato, Z. Han, and C. S. Hong. Offloading in HetNet: A Coordination of Interference Mitigation, User Association and Resource Allocation. *IEEE Transactions on Mobile Computing*, PP(99):1–1, 2016.

[51] S. Parkvall, E. Dahlman, A. Furuskar, Y. Jading, M. Olsson, S. Wanstedt, and K. Zangi. LTE-Advanced - Evolving LTE towards IMT-Advanced. In *2008 IEEE 68th Vehicular Technology Conference*, pages 1–5, Sept 2008.

[52] K. I. Pedersen, T. E. Kolding, F. Frederiksen, I. Z. Kovacs, D. Laselva, and P. E. Mogensen. An overview of downlink radio resource management for UTRAN long-term evolution. *IEEE Communications Magazine*, 47(7):86–93, July 2009.

[53] D. Pernes, D. Neves, and P. Vieira. Análise de Cobertura e Capacidade em Redes Móveis LTE de Quarta Geração (4G). In *URSI Seminar of the Portuguese Committee*, volume 1, pages 1–1, November 2012.

[54] N. Prasad, M. Y. Arslan, and S. Rangarajan. Exploiting cell dormancy and load balancing in LTE HetNets: Optimizing the proportional fairness utility. In *IEEE International Conference on Communications, ICC 2014, Sydney, Australia, June 10-14, 2014*, pages 1916–1921, 2014.

[55] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim. An analytical framework for optimal downlink-uplink user association in hetnets with traffic differentiation. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–7, Dec 2015.

[56] S. Sesia, M. Baker, and I. Toufik. *LTE - The UMTS Long Term Evolution: from Theory to Practice*. John Wiley & Sons, 2nd edition, 2011.

[57] A. Sibille, C. Oestges, and A. Zanella. *MIMO: From Theory to Implementation*. Academic Press, 1st edition, 2010.

[58] I. Siomina and D. Yuan. Load balancing in heterogeneous LTE: Range optimization via cell offset and load-coupling characterization. *2012 IEEE International Conference on Communications (ICC)*, pages 1357–1361, June 2012.

[59] I. Sohn and S. H. Lee. Distributed Load Balancing via Message Passing for Heterogeneous Cellular Networks. *IEEE Transactions on Vehicular Technology*, 65(11):9287–9298, nov 2016.

[60] A. Tall, Z. Altman, and E. Altman. Self organizing strategies for enhanced icic (eicic). In *2014 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pages 318–325, May 2014.

[61] N. Trabelsi, C. S. Chen, R. El Azouzi, L. Roullet, and E. Altman. User Association and Resource Allocation Optimization in LTE Cellular Networks. *IEEE Transactions on Network and Service Management*, 4537(c):1–1, 2017.

[62] I. Viering, M. Dottling, and A. Lobinger. A mathematical perspective of self-optimizing wireless networks. In *Communications, 2009. ICC '09. IEEE International Conference on*, pages 1–6, June 2009.

[63] C. X. Wang, F. Haider, X. Gao, X. H. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir. Cellular architecture and key technologies for 5G wireless communication networks. *IEEE Communications Magazine*, 52(2):122–130, February 2014.

[64] W. Xiang, K. Zheng, and X. Shen, editors. *5G mobile communications*. Springer International Publishing, 2017.

[65] Y. Xu and S. Mao. User Association in Massive MIMO HetNets. *IEEE Systems Journal*, 11(1):7–19, March 2017.

[66] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews. User Association for Load Balancing in Heterogeneous Cellular Networks. *Wireless Communications, IEEE Transactions on*, 12(6):2706–2716, June 2013.

[67] J. Zyren and W. McCoy. Overview of the 3GPP Long Term Evolution Physical Layer. *Freescale Semiconductor, Inc., white paper*, 07/:2–22, 2007.