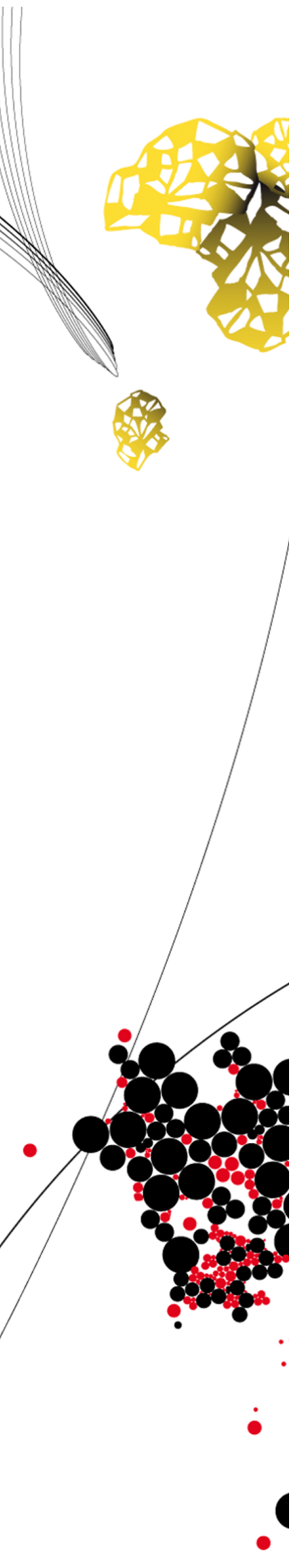


UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science

Leveraging 5G NR-U Bandwidth Parts in Coexistence with Wi-Fi

B.G. Nijenhuis

MSc Thesis

Embedded Systems

April 2024

Supervisors:
prof.dr.ir. G.J. Heijenk
dr. S. Bayhan
dr.ing. Y. Huang

Design and Analysis of Communication Systems
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Abstract

5th Generation New Radio (5G NR) is known for its flexibility in deployment and supports many applications with diverse requirements, e.g. high data rates or low latencies. Because the available spectrum in the licensed bands is very limited, 5G NR traffic can either be offloaded or fully deployed in the unlicensed bands under the same physical layer known as 5G New Radio Unlicensed (5G NR-U). A 5G NR network can have a lot of challenges to meet the Service-Level Agreement (SLA) requirements. These challenges are now also brought into the unlicensed bands. The performance of these applications with diverse requirements deployed in NR-U will experience a degradation compared to the licensed band, due to asynchronous devices and uncertainty of channel access introduced in the unlicensed bands, compared to the uninterrupted and synchronous operation in the licensed band. 5G NR introduces the Bandwidth Part (BWP), which allows the 5G NR network to divide its total channel bandwidth over multiple smaller bandwidths with different configurations to meet specific service requirements. Another feature introduced in 5G NR is Bandwidth Adaptation (BA), which allows a User Equipment (UE) to switch from BWP in order to save energy. Current literature introduces an exploit of the BWP feature used in the unlicensed band, where a multiple BWP Listen Before Talk (LBT) procedure is introduced and a gNB in the NR-U network is able to quickly sense and utilize other BWPs when the channel of the current BWP is busy. However, many challenges are left unaddressed in the research of the deployment of BWP switching in the unlicensed bands. For one, the paper does not discuss the delay overhead at the UE side, because a UE has to tune its radio for the other BWP. Next to this, the choice of configuration of BWPs for the UEs and the BWP time scheduling decisions is not discussed. Therefore, this thesis designs a new and more realistic model for the gNB performing downlink transmission in the unlicensed bands that incorporates the BWP and BA features for the enhancement of the coexistence with a Wi-Fi network. The model includes heuristic solutions for BWP configuration for each UE, BWP time scheduling and incorporates BWP switch delay. We investigate the enhancement of the overall system throughput by comparing different variants of our system model using a simulation methodology in a self made simulator. One baseline model where each UE only has a single BWP and can not switch, com-

pared to variants of a multi-BWP model, where each UE has the possibility to switch between BWPs under different configured delays. Next to this, we also investigate the impact on the Medium Access Latency (MAL) of both basestations and the impact on the throughput of the Wi-Fi network. We find out that the best improvements of the multi-BWP model compared to the baseline model are found under high load conditions of the Wi-Fi Access Point (AP), where the overall throughput and medium access latencies of both networks are improved. The Wi-Fi network benefits most from this, due to its limited available spectrum.

Contents

List of acronyms

OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
5G NR	5th Generation New Radio
5G NR-U	5G New Radio Unlicensed
gNB	Next Generation Node B
UE	User Equipment
BWP	Bandwidth Part
BA	Bandwidth Adaptation
RRC	Radio Resource Control
RF	Radio Frequency
FR-1	Frequency Range 1
LBT	Listen Before Talk
CCA	Clear Channel Assessment
4G LTE	4th Generation Long-Term Evolution
CP	Cyclic prefix
ISI	Intersymbol interference
CSMA/CA	Carrier Sense Multiple Access With Collision Avoidance
RAT	Radio Access Technology
eMMB	Enhanced Mobile Broadband
mMTC	Massive Machine Type Communication
uRLLC	Ultra-Reliable Low Latency Communication
DL	Downlink
UL	Uplink

PRB Physical Resource Block

MAC Medium Access Control

DCI Downlink Control Information

PDSCH Physical Downlink Shared Channel

PUSCH Physical Uplink Shared Channel

SLA Service-Level Agreement

MAL Medium Access Latency

AP Access Point

DCF Distributed Coordination Function

DIFS DCF Interframe Space

RTS Request-To-Send

CTS Clear-To-Send

TXOP Transmission Opportunity

HARQ Hybrid Automatic Repeat Request

CSAT Carrier Sense Adaptive Transmission

UMi Urban Micro

EIRP Equivalent Isotropic Radiated Power

PRB Physical Resource Block

MCS Modulation Code Scheme

SNR Signal-To-Noise Ratio

FIFO First-In First-Out

USS Unlicensed Spectrum Simulator

LOS Line-of-sight

NLOS Non-line-of-sight

RU Resource Unit

FTP File Transfer Protocol

PC Priority Classes

CA Carrier Aggregation

LBE Load Based Equipment

LEV Largest Extreme Value

List of Symbols

Symbol	Description	Unit
Δf	The subcarrier spacing.	*kHz*
μ	The numerology $\{0, 1, 2\}$.	
T_{switch}	Time delay between the switch of two BWPs.	*slots*
T_{CCA}	Time duration of the CCA procedure.	μs
T_{bs}	Time duration of a single backoff slot.	μs
CW	Fixed contention window.	
CW_{min}, CW_{max}	Minimum and maximum contention window.	
N_{UE}, N_{STA}	Number of UEs and STAs respectively.	
BW_{AP}, BW_{gNB}	The bandwidth size of the AP and gNB respectively.	*MHz*
P_{tx}	The transmission power of the basestations.	*dBm*
BW_{total}	The total observed bandwidth of the unlicensed band.	*MHz*
BW_{ch}	The minimum channel bandwidth.	*MHz*
b, B	The BWP and a set of BWPs respectively.	
f, BW	The start frequency and bandwidth respectively.	*MHz*
s	The SLA requirements of a service.	
u, U	A UE and the set of UEs respectively.	
R_{min}, R_{max}	The minimum acceptable datarate and maximum data rate.	*Mbps*
T_{slot}^{max}	The maximum acceptable slot duration.	*ms*
l, L	An LBT block and the set of LBT blocks respectively.	
T_{pp}	The period for the pre-processing time of the gNB.	*ms*
b_a, B_u	The active BWP and the set of BWPs that belong to a UE.	
B_μ	The set of bandwidth options for numerology μ.	
BW_g	The guard bandwidth.	*MHz*
N_{PRB}	The number of PRBs.	

Symbol	Description	Unit
v_{layers}	The number of transmission streams.	
Q_m, R_c	The modulation order and code rate respectively.	
SF	The scaling factor.	
OH	The overhead in the downlink transmission.	
T_s^μ	The OFDM symbol period of numerology μ.	μs
r, R	A BWP requirement and set of BWP requirements of all UEs.	
r_u	The BWP requirements of UE u.	
b_u	The BWP of UE u.	
c, C	An occupied channel and the set of occupied channels.	
BW_{max}	The maximum bandwidth requirement of a UE.	*MHz*
L_A	The set of all possible LBT blocks.	
BW_l, BW_c	The bandwidth of l, r_u and c respectively.	*MHz*
BW_{r_u}, BW_r	The bandwidth of r and r_u respectively.	*MHz*
f_l, f_c	The start frequency of l and c respectively.	*MHz*
f_e	The end frequency of l.	*MHz*
f_{AP}	The start frequency of the Wi-Fi AP	*MHz*
h_{UT}, h_{BS}	The height of the user and basestation respectively	*m*
σ_{SF}	The standard deviation for shadow fading in pathloss.	
T_{CP}	The time duration of the cyclic prefix.	μs
BW_{max}^l	The bandwidth of the l with the largest bandwidth.	*MHz*
d'_{BP}	The breakpoint distance used in the pathloss model.	*m*
d_{2D}, d_{3D}	The distance of transmission path in 2D and 3D.	*m*
d_{2D-in}, d_{2D-out}	The 2D distance of tx path for inside and outside.	*m*
d_{3D-in}, d_{3D-out}	The 3D distance of tx path for inside and outside.	*m*
f_C	The centre frequency used in the pathloss model.	*Hz*
PL_b	The basic pathloss from either LOS or NLOS.	*dB*
PL_{tw}, PL_{in}	The building penetration loss and inside loss, respectively.	*dB*
PL	The total pathloss from transmitter to receiver.	*dB*
σ_P	The standard deviation for the penetration loss model.	
N_T	The thermal noise over the transmission.	*dBm*
R_w	The data rate for a Wi-Fi STA.	*Mbps*
N_t	The number of tones (subcarriers) in a Wi-Fi RU.	
T_{dft}, T_{gi}	The durations for an OFDM symbol and guard interval.	μs
P_a, P_b	Packet size location a and scale b parameter for LEV.	*bytes*
t_a, t_b	Packet arrival time location a and scale b parameter.	*ms*

Chapter 1

Introduction

5G NR has a wide support for different application requirements. In 5G NR's predecessor named 4th Generation Long-Term Evolution (4G LTE), the waveform had a fixed structure, optimized for high data rate applications, limiting the support for other application types. In contrast to 4G LTE, 5G NR is more focused on flexibility on the physical layer to support these different application requirements. The physical layer of 5G NR is built to meet the difficult and different requirements that these applications introduce through its flexible waveform, but is limited by the available spectrum [1]. 3GPP Release 16 introduces the use of the 5G NR physical layer in the unlicensed spectrum under the name 5G NR-U, or NR-U in short. The deployment of NR-U alleviates the limited licensed cellular spectrum [2] and allows for the offloading of traffic or even fully standalone deployment in the unlicensed spectrum. In the unlicensed spectrum, a channel access method based on the Carrier Sense Multiple Access With Collision Avoidance (CSMA/CA) procedure is mandatory, where the channel is first sensed before it is accessed. 5G NR-U implements this requirement with the LBT mechanism. Unlike the licensed spectrum, the unlicensed spectrum is open for use by anyone and can therefore become crowded. Next to this, the unlicensed bands introduce uncertainty of channel access for all devices, because a node can only transmit if the medium has been determined idle by a channel access method. The switch from the licensed to unlicensed bands is therefore not convenient and will most likely degrade the throughput and latency [3]. This can especially be problematic for 5G NR services with strict service requirements. The main challenge of deploying NR-U in the unlicensed bands is to maintain these service requirements by enhancing the performance of NR-U, while also providing a fair coexistence with other Radio Access Technologies (RATs) operating in the same spectrum.

The BWP is a feature introduced in 5G NR Release 15 [4] and allows a Next Generation Node B (gNB) to divide its total channel bandwidth into multiple smaller

bandwidths that can be configured differently in its waveform parameters. These smaller bandwidths can be assigned to different UEs, to meet specific service requirements that a UE has. In the same release, another feature introduced is BA. This allows a UE to swiftly switch between BWPs that have been assigned to that UE. This is mainly introduced such that a UE can switch from its larger BWP to a smaller BWP to save energy.
Various literature discuss the problems introduced in BWPs and BWP switching within 5G NR in the licensed spectrum and the state-of-the-art provides insightful solutions to the unlicensed coexistence issues, but not many provide a combination of the two. Haghshenas et al.'s research [3] exploits the BWP feature, together with BA to enhance the coexistence in the unlicensed spectrum and mainly provide a higher throughput for the NR-U network. They introduce a new LBT procedure for the gNB on the downlink, where a longer LBT procedure (LBT Category 3) is performed over a defined primary BWP, and a shorter LBT procedure (LBT Category 2) over a secondary BWP when the channel of the primary BWP is busy. However, this research is very limited, unrealistic and many challenges are left unaddressed. The BWP switch delay requirement at the UE side as defined in 3GPP TS 38.133 [5] for the BA feature is not considered at all, which could heavily impact the throughput and latency of the NR-U network. The BWP is designed to allow unique waveform requirements simultaneously over the multiple UEs from a single gNB, Haghshenas et al.'s paper only considers a 20 MHz bandwidth with a single waveform type for all BWPs. This misses the sole purpose of the BWP and the choice of configuration and allocation of many different BWPs must be included. In addition to this, with BWPs in different sizes, time scheduling decisions issues are introduced and must be investigated.

To address the missing challenges, this thesis designs a new and more realistic model for the gNB that implements both the BWP and BA features of the 5G NR physical layer in the unlicensed spectrum. We then implement this model in a simulation methodology to investigate the performance improvement that NR-U can gain from this model and the impact that it has on the coexistence with a Wi-Fi network. We introduce two variants of the model, with the goal of measuring the impact on the overall network throughput and latency when UEs are allowed to BWP switch, compared to a baseline model:

- Single-BWP model: each UE is assigned a single BWP;
- Multi-BWP model: each UE is assigned two or more BWPs.

The single-BWP model is considered the baseline model and is used to compare the possible improvements of the multi-BWP model. For these models, the following

research questions are formulated:

- Does the multi-BWP model enhance the coexistence of NR-U and Wi-Fi in the unlicensed bands?
  - To what extent is the throughput improved of the gNB in the multi-BWP model compared to the baseline model?
  - What is the impact on the latency of the gNB in the multi-BWP model compared to the baseline model?
  - What is the impact on the throughput of the AP in the multi-BWP model compared to the baseline model?
  - What is the impact on the latency of the AP in the multi-BWP model compared to the baseline model?

The rest of this thesis is organized as follows. Chapter 2 provides the relevant background information that gives a better understanding of the technical features in the context of this thesis. Chapter 3 discusses the state-of-the-art literature and provides a literature review of Haghshenas et al.'s research [3]. In Chapter 4, the system model for the gNB is explained in detail. Chapter 5 presents the simulator framework that has been additionally designed and developed to provide quantifiable results for the given research questions. The simulation configuration and results are evaluated in Chapter 6. Chapter 7 discusses the performance evaluation of Chapter 6 and explains the missing pieces of this thesis, providing inspiration for possible future directions. The thesis concludes the results of the research in Chapter 8.

Chapter 2

Background

2.1 Waveform principles

2.1.1 OFDM

Orthogonal Frequency Division Multiplexing (OFDM) is a waveform design principle that is very popular in the wideband digital communication, e.g. wireless networks (IEEE 802.11 Wi-Fi) and 4/5G mobile communication networks. In most modulation techniques, information is modulated onto a single carrier frequency. In OFDM, a single information stream is split among several closely-spaced parallel orthogonal subcarrier frequencies. Each of this partial signal in these subcarriers is modulated with a conventional modulation scheme (QAM or PSK) at a low symbol rate (thus increased symbol duration). This increased symbol duration improves the robustness of OFDM to channel delay spread [6]. OFDM uses Cyclic prefix (CP) as a guard interval to make sure the subcarriers do not interfere with one another, which can completely eliminate Intersymbol interference (ISI), as long as the CP duration is longer than the channel delay spread. The subcarriers of OFDM overlap each other, but due to orthogonality, they do not interfere. Orthogonality suggest that the subcarriers are independent of each other. This independence is created by overlapping the peak of a subcarrier with the zero of the neighbouring subcarriers, resulting in no interference. This is also known as the subcarrier spacing, denoted by Δf. Figure 2.1 shows an example of an OFDM waveform in the frequency domain. In the frequency domain, these subcarriers are sinc functions, which corresponds to a rectangular function in the time domain. In OFDM, all of the subcarriers are assigned to a single user at a time, as shown in Figure 2.2a.

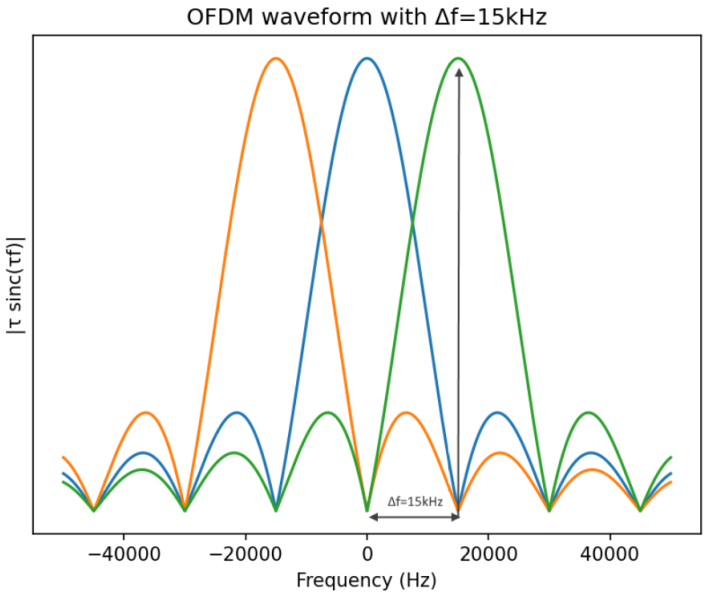
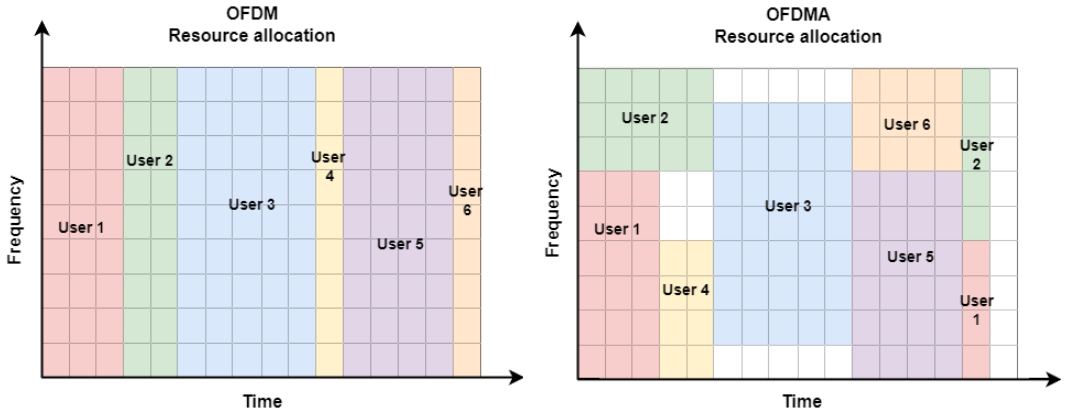


Figure 2.1: OFDM waveform showing three subcarriers in the frequency domain with a subcarrier spacing $\Delta f = 15$ kHz. The top of each subcarrier aligns with the zero of the adjacent subcarrier, providing orthogonality and no interference.

2.1.2 OFDMA

Orthogonal Frequency Division Multiple Access (OFDMA) allows multiple access in OFDM systems by dividing the available subcarriers over multiple users. Resources are now available to users in both time and frequency. In the time domain, they are available as OFDM symbols, while in the frequency domain as subcarriers. The time and frequency resources can be organized into subchannels, called resource blocks, for allocation to individual users, as shown in Figure 2.2b. OFDMA is a multiple-access/multiplexing scheme that provides multiplexing operation of user data streams onto the downlink subchannels and uplink multiple access by means of uplink subchannels [6]. Especially the multiplexing operation onto the downlink is important in this thesis and later discussed in Section 2.2.2. OFDMA was applied in the older 4G LTE already, but only in downlink, because uplink OFDMA requires good coordination and timing between the basestation and interacting users. Also for Wi-Fi, OFDMA was difficult to implement due to already existing multiple access schemes, such as CSMA/CA, in which devices would also contend for the channel and use the whole bandwidth when access is granted. The newest Wi-Fi generation 802.11ax has the option to use OFDMA in both uplink and downlink. Also in 5G NR, OFDMA is supported in both uplink and downlink. The key advantages of OFDMA over other traditional access technologies, are the Multiple Input, Multiple Output (MIMO) friendliness, uplink orthogonality, channel frequency selectivity and

scalability. Next to this, since the bandwidth is simultaneously divided over multiple users, it is possible to have different power levels over parts of the bandwidth.



(a): OFDM: Users are allocated all subcarriers sequentially in time. Allowing transmission to only a single user at a time.

(b): OFDMA: Users can be assigned resources in both time and frequency. Allowing transmission to multiple users at a time.

Figure 2.2: Resource allocation over a number of users for both OFDM and OFDMA. Where OFDM can only allocate all the subcarriers to a single user at at time and OFDMA provides much more flexibility.

2.1.3 Numerology

The scalability that OFDM and OFDMA provides, due to its flexible time-frequency grid, can also be referred to as numerology in the context of 3GPP 5G standardization. It provides the configuration of waveform parameters. For OFDM(A), this numerology set consist of the number of subcarriers, subcarrier spacing, slot duration and CP duration. In the rest of this thesis, when the term numerology is used, it refers to OFDM numerology. Different numerologies are possible due to the flexibility of OFDM, which supports different services under various channel conditions. The OFDM parameters allow a flexible adaptation to channel conditions and user bit rate requirements. In simple terms, a numerology provides a set of predefined settings for OFDM. The implemented flexibility of OFDM in numerology is different for 5G NR and Wi-Fi, where 5G NR provides much more flexibility.

2.2 5G New Radio features

5G NR is a Radio Access Technology (RAT) used in the fifth generation mobile network. In contrast to its predecessor 4G LTE, 5G NR is more focused on flexibility. LTE waveform has a fixed structure, optimized for high data rate applications and limited support for other applications. The physical layer of 5G NR is designed for better flexibility to support diverse services and user requirements, for three essential use case classes known as Enhanced Mobile Broadband (eMMB), Massive Machine Type Communication (mMTC) and Ultra-Reliable Low Latency Communication (uRLLC). Both LTE and 5G NR are based on OFDM and support OFDMA. However, OFDMA in LTE was only limited to Downlink (DL). In 5G NR, it is possible to use OFDMA in both Uplink (UL) and DL. The flexibility of 5G NR is provided through flexible time-frequency grid of OFDM and OFDMA enabling multi-numerology structure. In the case of OFDM and OFDMA, this numerology set consist of the number of subcarriers, subcarrier spacing, slot duration and CP duration. The possible numerology configuration for the sub-7 GHz Frequency Range 1 (FR-1) is shown in Table 2.1. In LTE, a single numerology was used with a subcarrier spacing of 15 kHz. 5G NR can have subcarrier spacings that are multiples of 15 kHz. Also, 5G NR can support a maximum bandwidth of 100 MHz in FR-1 [7].

Frequency Range (FR)	μ	Δf (kHz)	T_{CP} (μs)	Slots/subframe	Slot Duration (ms)	Max BW (MHz)
FR-1	0	15	4.76	1	1	50
	1	30	2.38	2	0.5	100
	2	60	1.19	4	0.25	100

Table 2.1: Possible numerologies for the FR-1 range (sub-7 GHz bands). Where μ is the numerology, Δf the subcarrier spacing and T_{CP} the duration of the CP.

2.2.1 Radio Resource Control (RRC)

Before a UE can send and receive data in a 5G NR network, both the gNB and UE must be configured such that the two can communicate with each other under the correct waveform configurations. This control mechanism is called Radio Resource Control (RRC) in 5G NR. Next to initial configuration, the common RRC protocol also dynamically updates the configuration between the UE and gNB and is responsible for managing radio resources and establishing, configuring and releasing logical channels that carry user and control data. There are three RRC states that the UE can have.

1. **RRC Connected**: The state in which the UE performs data transmission in the network.

2. **RRC Idle**: In this state, the UE saves battery, reduces signalling and there is no data transmission. A new connection to both the radio and core network must be established before data delivery, which can take a matter of seconds.

3. **RRC Inactive**: A new RRC state introduced in NR Release 15, where the radio connection is suspended, whereas the core connectivity is maintained [8]. The UE context is saved, including the RRC configuration. Meaning the UE can quickly resume data transmission in the network on the previous configuration.

If an RRC connection has been established between the gNB and UE, the UE must be either in the RRC Connected or RRC Inactive state. If no RRC connection is established, the UE is in the RRC Idle state. This paper only focuses on the RRC Idle and RRC Connected states, where RRC is used in 5G NR for BWP configuration and can be used in the BA feature. See 3GPP TS 38.331 for a further read on the RRC protocol [9].

2.2.2 Bandwidth Part

5G NR defines a system that divides up the total channel bandwidth of a cell into smaller bandwidths with a given numerology, called Bandwidth Part (BWP). It acts as a bridge between the numerology and scheduling mechanism [10]. A BWP is defined as a contigious set of physical resource blocks, selected from a contiguous subset of the common resource blocks for a given numerology on a given carrier. BWPs are controlled at the gNB based on UE needs and network requirements. A big advantage is that the UEs do not need to monitor the whole gNB downlink transmission bandwidth, but only have to scan the BWPs that are assigned to themselves, saving a lot of energy at the UE side. Also, some UEs do not support the large bandwidth used by the gNB and the BWP allows the gNB to still provide services for those UEs. For example, a gNB has a full channel bandwidth of 100 MHz, while the UE can only support up to 20 MHz of bandwidth [7]. BWPs are allowed to overlap in the frequency domain. Figure 2.3 shows an example of a gNB BWP configuration using the maximum total channel bandwidth of 100 MHz. Every color indicates a different BWP. It might be possible that different BWPs contain the same BWP configuration, including numerology and bandwidth, but are on a different carrier frequency. Thus, a BWP is defined by its numerology, bandwidth and carrier frequency. A single UE can be configured via RRC with up to four BWPs. However, only one BWP can be active at a time.

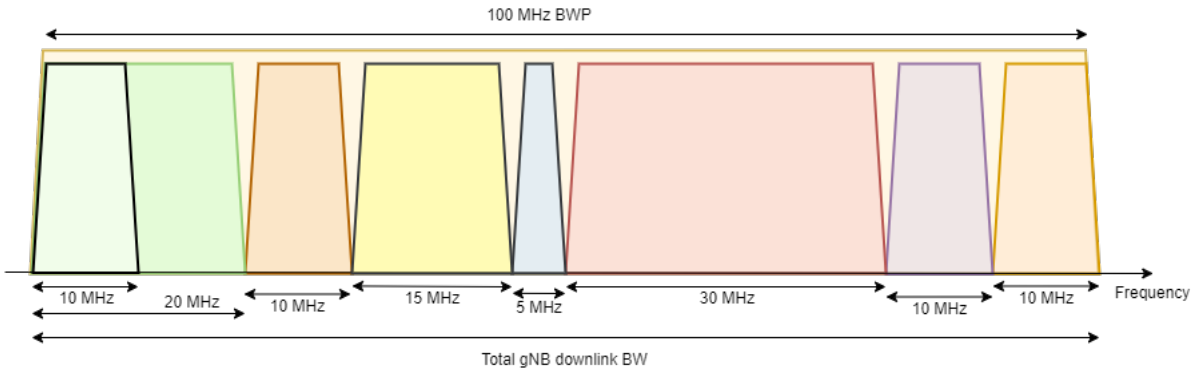


Figure 2.3: BWP configuration example in the frequency domain. The total transmission bandwidth of the gNB equals 100 MHz. A number of BWPs with different bandwidth sizes and different numerologies can be placed adjacent, partly overlapped or fully overlapped with each other.

BW (MHz) / Δf	5	10	15	20	25	30	40	50	60	80	90	100
15 kHz	25	52	79	106	133	160	216	270	NA	NA	NA	NA
30 kHz	11	24	38	51	65	78	106	133	162	217	245	273
60 kHz	NA	11	18	24	31	38	51	65	79	107	121	135

Table 2.2: Number of PRBs for each Δf/ bandwidth combination [11].

2.2.3 Bandwidth Adaptation

5G NR networks support a variety of services with different traffic conditions. However, services do not have static traffic rates and BWPs assigned to each UE are configured to support the maximum traffic conditions. Even though UEs only have to monitor the control channel of their own BWP, they still often use too many resources for their current received data rates. 5G NR Release 15 introduces the BA feature, that allows a UE to quickly switch from BWP to adapt to the current traffic conditions for that UE. This heavily decreases the energy consumption of a UE.

There are four types of BWPs configured by RRC, that are used in two UE modes named Idle and Connected, that correspond to RRC Idle and RRC Connected respectively from Section 2.2.1:

- **Idle Mode**: The UE is not connected to the 5G NR network and no data to be send or received by the UE. This mode and BWP type below can be ignored for the rest of this thesis, the focus is on the Active mode.
  - **Initial BWP**: The BWP that performs the initial access, after Synchroniza-

tion Signal Block (SSB) decoding. This BWP is common to all UEs in the network and the possible sizes are 24, 48 or 96 PRBs [4]. It receives the information to configure the first Active BWP, which includes RMSI (Requested Minimum System Information), CORESET (Control Resource Set) and RMSI Frequency location / bandwidth / subcarrier spacing.

- **Connected Mode**: The UE is connected with the 5G NR network and has data to send or receive.
  - **First Active BWP**: The BWP that is activated at first after initial attach is completed;
  - **Default BWP**: The BWP that the UE automatically switches to when there is no activity in the current BWP. The default BWP must be smaller than the active BWPs. If no default BWP is set, the PRBs of the Initial BWP are used as the default BWP;
  - **Regular** / **Active BWPs**: BWPs for specific traffic conditions.

The BA feature includes four switching mechanisms:

- **DCI-based switch**: Downlink Control Information (DCI) is a special set of information which schedules downlink or uplink data channel. The BWP indicator field [12] is a 2-bit field to indicate any of the four RRC-configured BWPs to quickly switch to. DCI is the fastest way to switch between Active BWPs, with a maximum of 3 ms for FR-1, as defined in Table 2.3. The downside of DCI is that it requires extra error handling, due to UEs possibly failing to decode the DCI with BWP activation/deactivation command.

- **Timer-based switch**: An inactivity timer that is used to switch any active BWP to the default BWP. This is introduced to automatically reduce power consumption at the UE side. A timer can have a value configured from 2 to 2560 ms and has a granularity of 1 ms (1 subframe), the timer is decremented at the end of each subframe for FR-1 [13]. The timer is started at the beginning of each active BWP, and restarted when a DCI with downlink assignment or uplink grant is decoded, indicating that there is still activity on the current BWP.

- **RRC-based switch**: Next to initially configuring BWPs to a UE, it can also be used to reconfigure or switch from BWP. Unlike DCI-based switching, the delays for RRC-based switching are not yet defined, but will be a minimum of 10 ms. This is the time to process the long RRC message. In addition to this, a BWP switch delay longer than the delay defined in Table 2.3 must be added to RRC processing delay [4]. The RRC-based switching is therefore at least 3 times as slow as DCI-based switching.

- **MAC-based switch**: MAC layer switch operation from the current BWP to the initial BWP, for cases when random access occasions are not configured [7].

Δf	Slot Duration (ms)	BWP switch delay T_{switch} (slots)	
		Type 1	Type 2
15	1	1	3
30	0.5	2	5
60	0.25	3	9

Table 2.3: Required BWP switch delay in slots for DCI- and timer-based BWP switching, type 1 and 2 are two levels of requirements and is dependent on the UE capability.

For DCI- and timer-based switching, Table 2.3 shows the required delay in slots that should happen between switching of BWPs [4]. This is required to accommodate Radio Frequency (RF) tuning and modem warm-up time for the UE [7]. We refer to this delay as T_{switch}, and is defined as the offset between the slot of the DCI switch request and first slot the UE is able to receive Physical Downlink Shared Channel (PDSCH) for DCI-based switching. For timer-based switching, this is the offset between the end slot of a subframe where a timer is expired and the first slot the UE is able to receive PDSCH, as shown in Figure 2.4. T_{switch} depends on the performance of the UE. When a switch is made between BWPs with a different subcarrier spacing, T_{switch} is determined by the smaller subcarrier spacing. Thus, the delay between a BWP switch is dependent on the numerology and UE capability. This UE capability information is reported by the UE to the gNB via RRC during its setup in the network. More specifically, the report tells the gNB how fast the UE can process the PDSCH data, or prepare the Physical Uplink Shared Channel (PUSCH) data. For both PDSCH and PUSCH, it is subdivided into two capability tables, showing the requirements in number of symbols for decoding time per numerology that the UE should meet [14]. Every UE should meet the capability 1 requirements by default and it is optional to meet the capability 2 table. When the UE performance is able to meet the requirements of the second capability table, it is a Type 1 UE and can use shorter delays, otherwise it is a type 2 UE. This thesis only focuses on DCI switching between active BWPs.

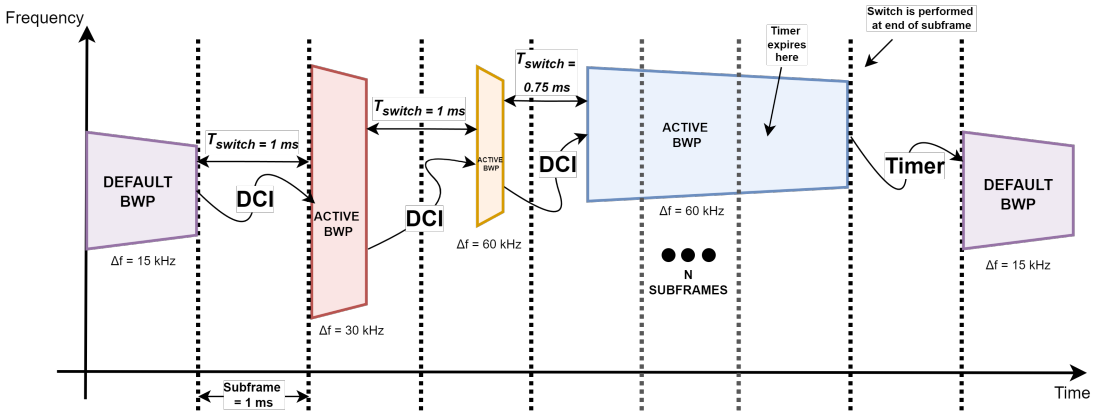


Figure 2.4: An example of BWP adaptation using the DCI- and timer-based switching mechanism of a Type 1 UE in Connected Mode.

2.3 Channel sensing procedures

2.3.1 CSMA/CA

Wi-Fi systems always operate in the unlicensed spectrum and are designed to coexist with other RATs. Therefore, the IEEE 802.11 Medium Access Control (MAC) layer is based on the CSMA/CA mechanism and is also known as the Distributed Coordination Function (DCF) [15]. DCF is the basic medium access protocol that allows for automatic medium sharing between compatible physical layer nodes through the use of CSMA/CA and the binary exponential algorithm when a transmission has failed. CSMA/CA minimizes the collision probability between these physical layer nodes by sensing the medium (e.g. energy detection) for a standard time duration known as a DCF Interframe Space (DIFS) and goes into a backoff period if the medium is sensed as busy. This backoff period is defined by a random amount of backoff slots followed by a DIFS [16]. A backoff counter is used to make sure the node waits for the random generated amount of backoff slots. The counter is decremented by one when there is no activity sensed in the medium during a backoff slot. If there is activity detected during a backoff slot, the backoff counter will not be decremented. The countdown will continue when the medium is determined idle for the duration of a DIFS followed by backoff slot. After the backoff counter reaches zero, the node is able to transmit. The random backoff decreases the probability of collisions, but does not prevent them. Collisions can still occur due to the hidden node problem. Therefore, Wi-Fi can use an additional (virtual) collision avoidance mechanism after the transmitter senses the medium as idle, by transmitting and receiving the short control frames: Request-To-Send (RTS) and Clear-To-Send (CTS) respectively.
In addition, after transmission, an acknowledge frame (ACK) is sent by the receiver

to notify the sender that the transmission was not interfered. When no ACK frame is received, the sender goes into a binary exponential backoff, before it tries to re-transmit. See Figure 2.5 for the flowchart of the CSMA/CA procedure.

2.3.2 Listen Before Talk

Similar to Wi-Fi, the deployment of an NR-U network in the unlicensed bands requires the network to use a spectrum sharing mechanism, where a device senses the channel using a Clear Channel Assessment (CCA) and only access the channel when no other nodes are transmitting on that channel, to prevent collisions. The use of LBT is mandatory in Europe and Japan and NR-U adopts the LBT protocol defined in LTE-LAA. Prior literature has studied the coexistence in the 5 GHz bands and commonly accepted that LAA has a better fairness relation with Wi-Fi then LTE-U does, because LAA uses the LBT mechanism [17]. LBT is derived from the Wi-Fi CSMA/CA mechanism. There are four categories defined for LBT:

1. CAT1-LBT (Type 2C): A gNB can immediately access the channel without any performed sensing, the maximum Transmission Opportunity (TXOP) times are therefore short and can be up to 548 μs.

2. CAT2-LBT (Type 2A / 2B): Any device in the NR-U network must sense the channel for a fixed duration T_{CCA}. For Type 2A, $T_{CCA} = 25\ \mu s$, for type 2B $T_{CCA} = 16\ \mu s$.

3. CAT3-LBT: If an NR-U device has sensed the medium to be busy during CCA, a random backoff of sensing must be performed before it can access the medium. This random backoff is sampled from $[0,\ CW]$, where CW is the fixed contention window.

4. CAT4-LBT (Type 1): Similar to CAT3-LBT, but the contention window is increased upon a collision, which is detected by the Hybrid Automatic Repeat Request (HARQ) [18]. This LBT type is similar to the CSMA/CA procedure without the RTS/CTS Exchange, see Figure 2.5.

CAT4-LBT is the type that is adopted by LTE-LAA and is also considered to be the baseline for shared spectrum access of Load Based Equipment (LBE) [1]. In both CSMA/CA and CAT4-LBT, a device can transmit into the medium for a given time period known as the COT in the 3GPP context and TXOP in the IEEE 802.11 standard. In the rest of this thesis, we use TXOP for both Wi-Fi and NR-U. The TXOP of a device is based on a set of parameters. This is were Wi-Fi and NR-U typically differ from each other in the implementation of the channel access methods. Table 2.4 shows the available set of parameters that can be used in the an NR-U

PC	T_{CCA}	CW_{min}	CW_{max}	Max TXOP
P_1	1,2/ 25, 34 μs	4	8	2 ms
P_2	1,2/ 25, 34 μs	8	16	3 or 4 ms
P_3	3/ 34 μs	16	64	6, 8 or 10 ms
P_4	7/ 79 μs	16	1024	6, 8, or 10 ms

Table 2.4: The set of channel access parameters to be used in CAT4-LBT for different PCs [19].

device with CAT4-LBT, for different types of Priority Classes (PC) defined in 3GPP TR 36.213 [19].

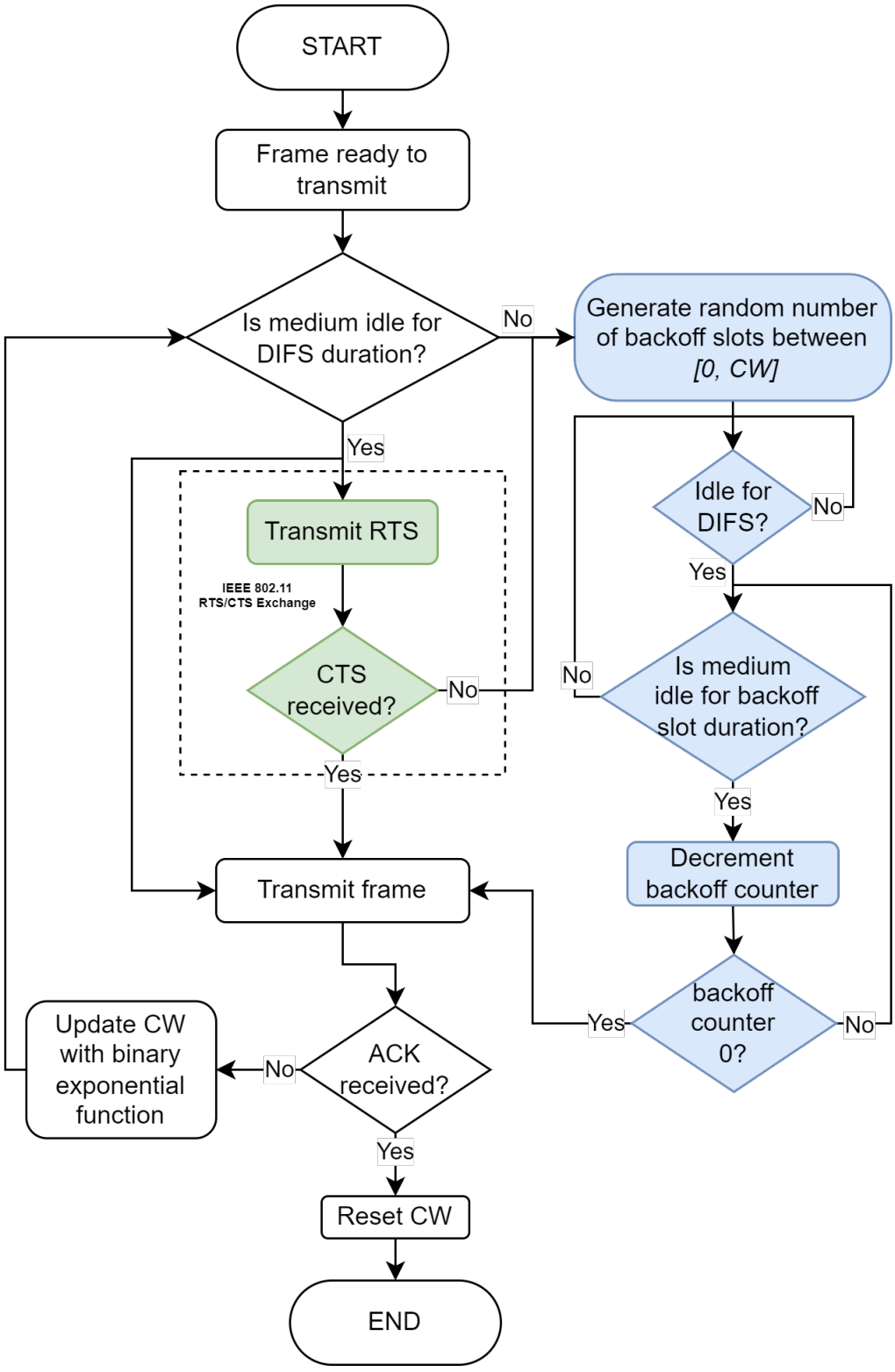


Figure 2.5: Flowchart of the CSMA/CA procedure. The blue and green components indicate the backoff procedure and optional RTS/CTS Exchange respectively.

Chapter 3

Related Work

This chapter presents the state-of-the-art literature in the context of NR-U and Wi-Fi coexisting in the unlicensed bands in general, to obtain maximum total throughput and maintain fairness. We will then present literature that evaluates the overhead caused by BWP switching with the BA feature in the 5G NR network and finally discuss the research from Haghshenas et al. [3], that leverages the BWP and BA feature in NR-U. In the end, we present the new direction that this thesis will take.

3.1 Coexistence in the unlicensed bands

There are many papers that discuss the coexistence between Wi-Fi and NR-U in the unlicensed bands and the challenges that arise. Before we go into that, we first discuss the state-of-the-art between Wi-Fi and the unlicensed deployments of LTE, such as LTE-U, LAA and eLAA. Huang et al.'s research presents several coexistence mechanisms to improve the LTE and Wi-Fi coexistence [20]. The mechanisms either utilize the time, frequency or power domain at the LTE system to enhance the coexistence scenario. The fundamental challenge is the centralized characteristics of LTE. For example, in the time domain, LTE-U uses Carrier Sense Adaptive Transmission (CSAT) that periodically turns off the centralized LTE transmission, to allow Wi-Fi adequate access time. The LBT mechanism is deployed in LAA and eLAA and shows fairer coexistence than CSAT, because LBT is based on CSMA/CA used in Wi-Fi systems [17]. In CSAT, the LTE system would be more dominant than Wi-Fi, due to the exponentional backoff system of CSMA/CA. With LBT, the two RATs would behave more similar to each other and cause fairer coexistence.
Therefore, LBT is adopted as the standard for spectrum sharing in NR-U. However, NR-U derives from scheduled and synchronous radio access technology, where transmissions are expected to begin at fixed slot boundaries. LBT is an asynchronous channel access mechanism and the end of an LBT procedure may not

coincide with a slot boundary. To solve this, a gNB must postpone transmission to align the end of the LBT procedure with the slot boundary. Zajac et al. [21] conducts a performance analysis using a discrete-event simulator of two types of solutions to obtain this alignment: gap-based or reservation signal (RS) based access. The gap-based access adds an extra time period of no activity before the LBT backoff procedure, to make sure the number of backoff slots align with the slot boundary. The RS-based access transmits energy into the medium after the backoff procedure and before the slot boundary, to block other nodes from accessing the channel and reserving it for itself. Zajac et al.'s simulator shows that RS-based access allows for a fairer coexistence between basestations, but wastes more energy and radio resources and has a higher collision probability. The gap-based access gives basestations a higher spectral efficiency due to a significantly lower probability of collision.

Furthermore, the introduction of flexible slot durations by scaling numerologies and mini-slots can bring the scheduling granularity down to a single OFDM symbol and can even further improve the fairness in channel access for the coexisting scheduled and distributed technologies [22].

With LBT set as the standard for NR-U to behave similar to Wi-Fi with CSMA/CA, the fairness between the two is not guaranteed, and it often seems that NR-U is more dominant. Hirzallah et al.'s evaluation [1] indicates that under heavy traffic, NR-U often achieves higher throughput and lower latencies than Wi-Fi. The loss of certain critical control messages of Wi-Fi due to collisions with NR-U transmission are the cause of the deterioration of the Wi-Fi network. The tuning of both the parameters of NR-U LBT and the Wi-Fi CSMA/CA are very important for a smooth coexistence. Critical messages or transmission with low latency requirements should be configured with parameters that allow faster medium access. 3GPP proposed a fairness criterion for the coexistence of NR-U and Wi-Fi. The papers of Luo et al. [2] and Kakkad et al. [23] evaluate different NR-U parameters under the 3GPP fairness constraints. Luo et al. [2] states that Wi-Fi networks usually have fixed access parameters, thus the configuration should happen at the NR-U side. Luo et al. tunes the initial backoff window size of the LBT of NR-U nodes and tries to achieve two optimizations, one where the coexistence throughput is maximized and one where the throughput of the NR-U network is maximized, both under 3GPP fairness contraints. Their findings are that NR-U is starved when trying to optimize the total coexistence throughput, since the 3GPP fairness contraints protects the performance of the Wi-Fi network. Kakkad et al. [23] also configures the slot duration and reveals that when the 3GPP fairness is met, the NR-U parameters are significantly larger than that of Wi-Fi, reducing the throughput of the NR-U network. In contradiction to the Hirzallah et al.'s evaluation [1], Luo et al. suggest optimizing the throughput of NR-U in a

practical scenario to achieve fairness.

3.2 State-of-the-Art of Bandwidth Adaptation

The BA feature is introduced as an enhancement on the BWP feature to meet the changing traffic conditions or service requirements of the UE to switch to a matching BWP. For example, a scenario with high delay spread requires the UE to use a BWP with a low numerology to have a better performance, while changing traffic conditions requires the UE to switch to a BWP with a different bandwidth. As a result, this further improves energy effiency, but can cause overhead due to the BWP switching delays. Research by Ramaswamy et al. [7] introduces a model to evaluate the performance of the BA feature in a 5G NR network. The model characterizes the power savings and packet scheduling delays that come with the BA feature. They define a cycle of a UE in the default BWP, that at some point in time receives a grant with cross-slot scheduling, and after a constant number of slots must be able to monitor and decode data on the active BWP. Then, the UE keeps using this active BWP as long as it receives grants with same-slot scheduling. When it does not receive a grant within a defined timer duration, the cycle will end at the expiration of the timer. This cycle is used to observe the energy savings and scheduling delay of the model with the BA feature compared to a baseline model where a UE always operates at a bandwidth corresponding to the active BWP configuration. The energy in the model is calculated on a per slot basis of 1 ms. The slots where data is transmitted consume power for monitoring PDCCH and decoding PDSCH. Slots where no data is transmitted only consume power for monitoring PDCCH. During the switch time, they consider that the UE does not transmit or receive any signals, thus uses no power in this model. The average scheduling delay in this model is mostly dependent on the switch duration in slots, as well on the packet arrival probability (that is equal to the probability of an arriving grant). When increasing the packet arrival probability, the average scheduling delay is decreased, because the UE will incur less frequent BWP switch delays, due to the BWP timer expiring less frequently. This decrease in scheduling delay becomes larger for higher BWP timer values, since this also helps to less frequently switch from the active BWP. However, for low packet arrival probabilities, the impact of the BWP timer is very small. In contrary, for the power savings gain it is better to decrease the packet arrival probability. When the UE will stay longer in the active BWP, it consumes more energy by monitoring PDCCH over a larger bandwidth compared to the default BWP. For higher packet arrival probabilities, decreasing the BWP inactivity timer leads to significant lower power consumption, because the UE will switch back to the default BWP more often.

The paper suggest future research to focus on the packet arrival probabilities and e.g. deploy packet shaping mechanisms, rather than focus on the settings of the BA feature. The key observation of this paper is that transmitting larger packets using wider bandwidth, but less frequently, is a better strategy for reducing UE power consumption than transmitting shorter packets on a narrower bandwidth more frequently. Meaning that a low packet arrival probability should be used. This leads however, according to the observations of the paper, to higher average scheduling latencies.
Ramaswamy et al.'s research uses a relatively simple energy model and considers that the UE does not consume power during a BWP switch, even though the energy consumption of the UE during the RF retuning and the modem warm-up that is necessary for a BWP switch should be incorperated. Furthermore, the energy consumption is independent of the operating frequency and antenna characteristics and only depends on the bandwidth and not on the numerology.
The research is only performed on switching between default and active BWPs. Another interesting use-case is the switching between active to active BWPs using the DCI-based messaging system, when there is still an active traffic rate, but can be adapted to a minimum requirement for the BWP. The average scheduling delay in Ramaswamy et al.'s model is a function of the switch delay (in a constant number of slots) and the average number of grants in a cycle. The switch delay is normally dependent on the numerology of the BWPs, which is excluded in this research. Different numerologies change the slot duration, whereas this model only assumes a fixed slot duration of 1 ms. It is unknown how a multi-numerology scenario would perform with their model. Also, their suggestion for packet shaping might not always be possible, since e.g. URLLC requires low latencies and must transmit small packages more frequently.
Another paper by Fuad Abinar et al. [4] presents a system-level evaluation of a 5G NR deployment with dynamic BWP adaptation. This paper also focuses on the BWP inactivity timer and its impact on the power savings and traffic latency. Their system-level simulations indicate that there is no effect on the system performance under high loads. This is based on the absence of BWP switches, because the UE BWP inactivity timer does not expire and stays in its active BWP. Since more BWP switches cause more overhead in terms of delay. Similarly to the results of Ramaswamy et al, for a low load and low BWP inactivity timers, the UE will stay in the active BWP for short periods and many BWP switches are instantiated, causing a lot of overhead in latency. They do note that BWP adaptation enables power savings if the bandwidth of the default BWP is smaller than the bandwidth of the active BWP.

3.3 Analysis of Haghshenas et al.'s Coexistence enhancement using the BWP

The LBT procedure that is mandatory in the unlicensed bands causes channel access uncertainty, because nodes must first sense before they can access the channel. A node might sense that the channel is busy, preventing it from direct transmission. This uncertainty is unwanted and causes challenges for the high demanding requirements of the 5G NR services, that will surely degrade in performance in NR-U. Haghshenas et al. [3] introduces an innovative algorithm based on LBT, that leverages the BWP feature to provide gNBs more transmission opportunities in the unlicensed bands.

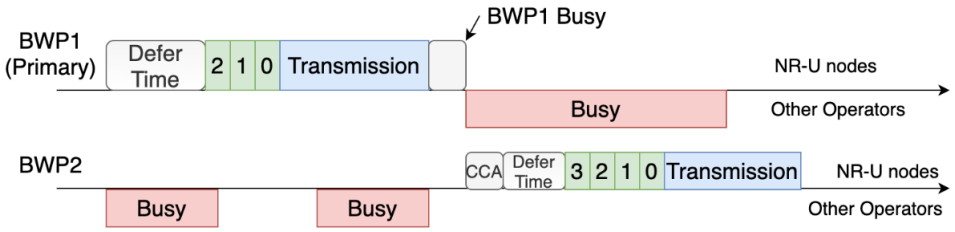


Figure 3.1: The proposed LBT procedure over multiple BWPs from Haghshenas et al. [3]

In their proposed algorithm, the gNB is configured with a number of BWPs with each a bandwidth of 20 MHz, to comply with standardization of the unlicensed bands. One of the BWPs is assigned as the primary BWP and the remaining as secondaries. When there is data available for transmission, one of the standard LBT categories is initiated on the primary BWP, in their simulations they use the LBT category 3 that has a fixed contention window. The values used for the CCA and each backoff slot are commonly used values and are $T_{CCA} = 34\ \mu s$ and $T_{bs} = 9\ \mu s$ respectively. If the channel is determined to be idle, transmission can occur at the primary BWP. If the channel is determined to be busy, instead of going into the backoff procedure, a shorter CCA procedure of unknown duration is executed on the secondary BWPs. It starts with the short CCA on one of the secondary BWPs. If the CCA determines the channel to be busy, then the short CCA is performed on the next secondary BWP. A standard LBT procedure (e.g. category 3), including backoff procedure, will be performed on the first channel of a secondary BWP that is determined to be idle by the shorter CCA. Then, when the backoff completes, the gNB will proceed transmission on this secondary BWP. In the case that the short CCA determines all secondary BWP channels to be busy, then the short CCA will

continue on all BWPs sequentially until one becomes available.

Analyzing the proposed LBT procedure from Haghshenas et al, a different behaviour is observed compared to the standard LBT. Normally, the backoff procedure is only performed when the CCA determines the channel to be busy. According to Figure 3.1, the backoff procedure is performed on BWP2, even though the CCA has determined the channel to be idle. It could be that after the short CCA is performed on the secondary BWP, the standard LBT procedure of the primary BWP is continued. This means that, because the CCA on the primary BWP determined that channel to be busy, a backoff is performed on another channel. Since the sensing procedures switches from channel, the LBT procedure must be independent. For higher spectral efficiency, it will be better to either immediately start transmission when the short CCA determines the channel to be idle, or perform the LBT according to its standard after the short CCA. The latter would mean that if the short CCA succeeds, another CCA follows that will determine whether transmission can start, or that the gNB must go in backoff. Otherwise, the gNB will unnecessary postpone its transmissions on the secondary BWPs.

In their simulations, they only incorporate two BWPs, one primary and one secondary, that are being used by 3 gNBs and 15 UEs of the NR-U network. The channel of the primary BWP is also occupied by 3 Wi-Fi APs and 15 STAs, while the secondary BWP is occupied by 2 AP and 10 STAs. They run a baseline scenario, where only one BWP is activated and the NR-U network can only use the primary BWP. The second scenario uses their proposed LBT procedure and activates the second BWP. All simulation runs are done over an increasing traffic load. The results of their simulations show that their proposed LBT procedure improves both the throughput and latency of both networks, where the total system throughput is even increased by 50%, compared to the baseline scenario where only one BWP is active.

It is unclear if the performance of all Wi-Fi APs are incorporated, or only the three APs on the primary BWP. Since the channel occupancy table excludes the second channel for the Wi-Fi network, it is assumed that only the performance of the three APs is evaluated. Because the channel of the primary BWP becomes less congested due to the gNB's switch to the secondary BWP, the Wi-Fi network operating on channel 1 benefits from this and obtains more airtime, thus improves in throughput and latency. However, the impact on the other Wi-Fi network remains unknown. Because the gNB can now also transmit on a second channel, that channel becomes more congested and could have a negative impact on the Wi-Fi network operating on that channel.

While their research shows promising results for leveraging the 5G NR BWP in the unlicensed bands, there are some key aspects missing. As explained in Section 2.2.3, switching the gNB and UE connection to another BWP requires signalling between the two nodes and time for the UE to tune its radio frequency and allow the modem to warmup. 3GPP specifies the delays that are required for specific UE types, which must be incorporated as necessary time duration in which the UE can not transmit or receive any data, before transmission can restart on the next active BWP. The research of Haghshenas et al. ommits the standardization (BWP switch delay and control signalling) of the BA feature, and does not take into account the impact that this can have on the throughput and latency of the NR-U network. This thesis will further investigate the exploitation of the BWP in the unlicensed bands, including the standardization of the BA feature, to enhance the throughput of the NR-U network.

Chapter 4

System Model

This chapter describes the system model that has been designed to provide quantifiable answers for the research question. Section 4.1 describes the scenario and context of this thesis. Thereafter, the design of the gNB model is explained in Section 4.2 in several components. A number of components form the proposed heuristics for defining a set of BWPs for the UEs of a gNB. The other components solve the BWP time scheduling with another proposed heuristic. In Section 4.3, a brief description of the Wi-Fi AP model is given. Both the gNB and Wi-Fi models are later implemented in the simulator described in Chapter 5.

4.1 Scenario and context

We consider a wideband 5G NR-U gNB serving N_{UE} UEs in Carrier Aggregation (CA) mode, where a licensed carrier is used for the control plane (e.g. control switching of BWP) and a unlicensed carrier for DL transmission. The gNB shares the 5 GHz unlicensed spectrum with a single Wi-Fi AP with a bandwidth defined by BW_{AP}, that serves a number of STAs defined by N_{STA}. We consider this unlicensed coexistence scenario to be in an Urban Micro (UMi) environment, and adopt 3GPP Urban Micro channel models, where the transmitters are mounted below rooftop levels of surrounding buildings [24]. All receivers, both 5G NR-U UEs and Wi-Fi STAs, are distributed uniformly over a given deployment area. We assume that all transmitters use a single antenna and have the same maximum Equivalent Isotropic Radiated Power (EIRP) P_{tx} defined in the ETSI EN 301 893 standard [25] that is used for omnidirectional transmission. For simplification, we consider the Wi-Fi network to only contain Wi-Fi 5 devices (IEEE 802.11ac), that can only serve its users one at a time via OFDM, in comparison to Wi-Fi 6 (IEEE 802.11ax) that has the option to use OFDMA. The gNB's unlicensed carrier has a bandwidth defined by BW_{gNB} on which the UEs' BWPs are scheduled. The maximum value for BW_{gNB} is 100 MHz,

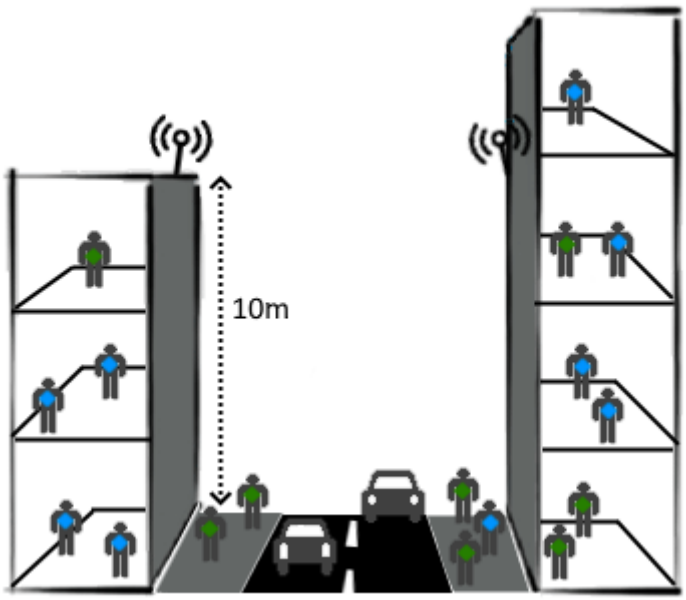


Figure 4.1: Simple illustration of the UMi Street Canyon environment used in the context of this thesis. Basestations are located below rooftop levels at a height of 10m, and users are either outdoor or indoor at different heights.

this is the maximum bandwidth that a gNB can use in FR-1. The total investigated bandwidth is denoted by BW_{total}. Normally, channels are configured with bandwidth sizes of 20 MHz in the UNII standard. However, since we also design this model to be able to coexist with other NR-U networks that can have smaller bandwidth sizes, a smaller channel sensing granularity is required. We define BW_{ch} to be the channel bandwidth and is also the granularity of the gNB's spectrum sensing.

4.2 Models for the gNB

The gNB will define a set of BWPs B that it can use to transmit in downlink to its connected UEs. Let a BWP be defined by $b :=< \mu, BW, f_c > \in B$, where BW is the bandwidth of b in MHz and f the start frequency of b. The start frequency is used instead of carrier frequency for simplification in the mathematics. $\mu \in \{0, 1, 2\}$ is the numerology of b, which directly translates to the subcarrier spacing $\Delta f = 15 * 2^{\mu}$ (kHz), as shown in Table 2.1. Note that we do not consider the extended cyclic prefix in the numerology configuration. The set B is derived from the SLA requirements of the service s of each connected UE $u \in U$ and the current channel state, which will be explained in the sections below. Each service is defined by its SLA requirements:

$s :=< R_{min}, T_{slot}^{max} >$, where R_{min} is the minimum acceptable data rate and T_{slot}^{max} the maximum acceptable slot duration [26]. The R_{min} and T_{slot}^{max} parameters of a service are randomly chosen for each UE. A homogeneous traffic model is used for all UEs in the NR-U network and the UEs are distributed uniformly over the UMi environment with respect to the gNB, from which they need to have a minimum distance of 10m.

There are two problems that we have to solve with the model of the gNB.

- 1. How do we define the set B?
- 2. How do we handle the time scheduling with set B?

To answers those questions, we seperate the gNB's operation into two main phases. The *pre-processing phase*, which focuses on solving question 1, to define an optimal set B and the *scheduling phase*, that handles the time scheduling of the set B. In the pre-processing phase, the gNB scans the medium for its current state and defines so-called *"LBT blocks"* as the set L and afterwards the BWP set B based on this state. The scheduling phase performs separate LBT procedures on each channel with bandwidth BW_{ch} within each LBT block $l \in L$ and schedules the UEs that have payloads to transmit. The pre-processing phase is designed to initialize the set B and provide a synchronization in the channel sensing policy through L. This initialization is repeated on a certain period defined by T_{pp}, since it is assumed that the state of the medium does not change more frequent than T_{pp}. This parameter is not further used in this thesis, due to short simulation times, explained in Chapter 5. The goal of the pre-processing phase is to calculate an optimal B and L, such that the throughput of the gNB is maximized in the scheduling phase. The scheduling phase transmits as many UE payloads on BWPs as possible per LBT block that gains access to its channel bandwidth, per T_{pp}.

4.2.1 Proposed BWP configuration heuristic

This section describes the proposed BWP configuration heuristic to define an optimal B, that happens in the so-called pre-processing phase of the gNB. This phase is divided into several subphases that are executed sequentially in time. At first, the spectrum is sensed with a granularity of BW_{ch}. At the same time, the gNB calculates the BWP requirements for each $u \in U$, that includes the numerology and bandwidth. After the total spectrum with bandwidth BW_{total} is sensed, an optimal set L is calculated. When L is known, the set B can be created by deriving a start frequency f next to the BWP requirements for each b of $u \in U$. Thus, B holds the BWPs of all UEs, where a $b \in B$ can also be shared between multiple $u \in U$. Each u can be assigned up to four BWPs, but only one active BWP must be set. We denote the set

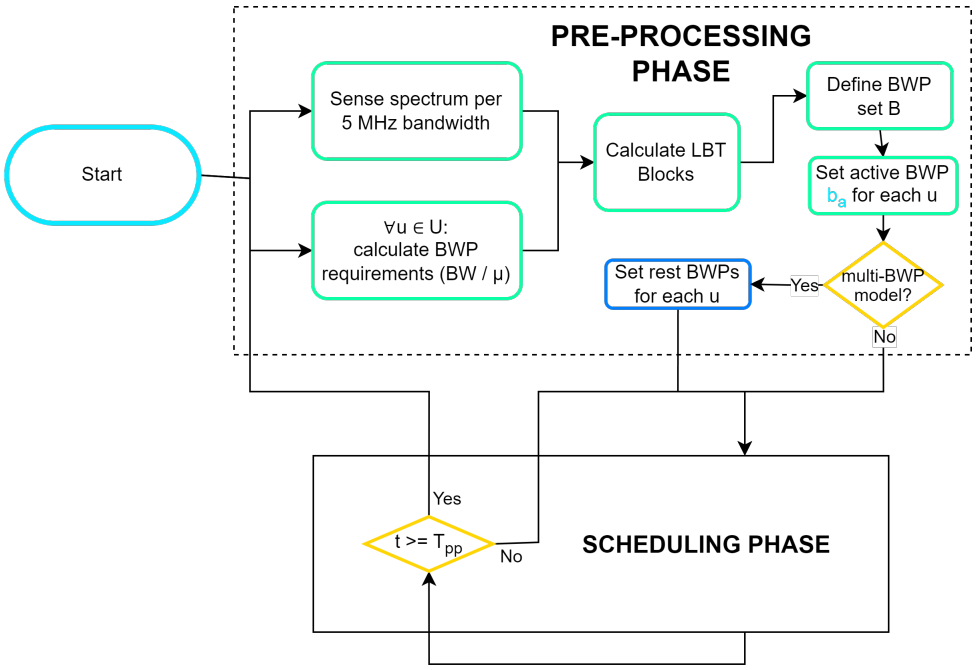


Figure 4.2: Flowchart of the pre-processing phase of the gNB. The goal is to define an optimal set of BWPs B based on the requirements of the UEs and the state of the medium.

BWPs of each u and the active BWP of each u by $B_u \subset B$, and $b_a \in B_u$ respectively. We further elaborate on each subphase in the next sections.

UE BWP requirements

As explained in Section 2.2.2, a BWP b is defined by its numerology, bandwidth and carrier frequency. In deriving the set BWPs B_u for each UE $u \in U$, we must first translate the SLA requirements of the service s of u to numerology and bandwidth (μ, BW), shown in Algorithm 1. The start frequency f is not important at this point, since this will be a frequency allocation problem at a later stage. The algorithm first decides on a starting μ directly from the acceptable maximum slot duration T_{slot}^{max}, and preferably chooses the lowest μ possible, due to the decrease in robustness of OFDM against channel delay spread when using a higher μ having a lower symbol duration [6]. However, it is not always possible to pick the lowest μ, due to the bandwidth limitations for each Δf of μ, shown in Table 2.2. When the as minimum as possible μ is defined, we derive the possible bandwidth values from Table 2.2 for μ in the set B_μ. We then calculate the maximum datarate R_{max} (in Mbps) that it can transmit on each BWP(μ, BW) $\forall BW \in B_\mu$, from lowest to highest BW. By doing this, we define a BWP for the UE with the lowest μ and bandwidth possible. The lowest μ to have a higher robustness against ISI due to the channel delay spread.

And a BWP with the lowest BW to use as minimum resources as possible to meet the users SLA requirement of the service and provide better spectral efficiency.

Equation 4.1 shows the formula for the number of PRBs that are possible within a certain (μ, BW) of a BWP, from which the values of Table 2.2 are calculated. BW_g is the required guard bandwidth [11] that belongs to the specific (μ, BW) combination. The $15 * 10^3 * 2^\mu$ indicates the subcarrier spacing for the given numerology μ. Since there are 12 subcarriers within a single PRB, the subcarrier spacing is multiplied by 12. The formula for R_{max} is shown in Equation 4.2, adopted from [27] and is dependent on the Modulation Code Scheme (MCS), number of PRBs N_{PRB} and μ.

$$N_{PRB} = \left\lfloor \frac{BW - BW_g * 2}{15 * 10^3 * 2^\mu * 12} \right\rfloor \tag{4.1}$$

$$R_{max} = \left(\frac{V * N_{PRB} * 12}{T_s^\mu} * (1 - OH) \right) * 10^{-6}, \quad \text{where } V = v_{layers} * Q_m * SF * R_c \tag{4.2}$$

v_{layers} is the number of transmission streams and equals only 1 downlink in our case, Q_m is the modulation order, $SF \in \{0.4, 0.7, 0.8, 1\} = 1$ the scaling factor, and R_c the code rate. Q_m and R_c are derived from the MCS, which depents on the Signal-To-Noise Ratio (SNR) of the user. The OFDM symbol period belonging to each μ is denoted as $T_s^\mu = \frac{10^{-3}}{14*2^\mu}$ for the standard cyclic prefix. The overhead is denoted with OH and is defined to be 0.14 for downlink in FR-1 [27].

If R_{max} can hold the minimum data rate R_{min} of s, the (μ, BW) combination is returned. If the R_{min} can not be satisfied on any of the (μ, BW) combinations, μ is increased. The highest data rates are achieved on $\mu = 1$, since they provide the most PRBs/s, which can be seen in Table 2.2. It might be possible that the SNR of a user is too low to provide any BWP that can meet the service requirements. A low modulation coding scheme is obliged for low SNR values, which limits the transmission rate and forces the bandwidth to be higher then the maximum possible out of Table 2.2. In this matter, we return *None* to define the BWP requirement after the creation of the LBT blocks in chapter 4.2.1, where we scale the bandwidth of the BWP down to the bandwidth of the largest LBT block bandwidth. Note that at this subphase, a BWP is not yet defined, since f is not yet defined.

Algorithm 1: UE BWP requirements calculation

Data: Service $s :=< R_{min}, T_{slot}^{max} >$, Three sets of possible BW values $B\mu$ $\forall \mu \in \{0, 1, 2\}$
; /* BW values from Table 2.2 */

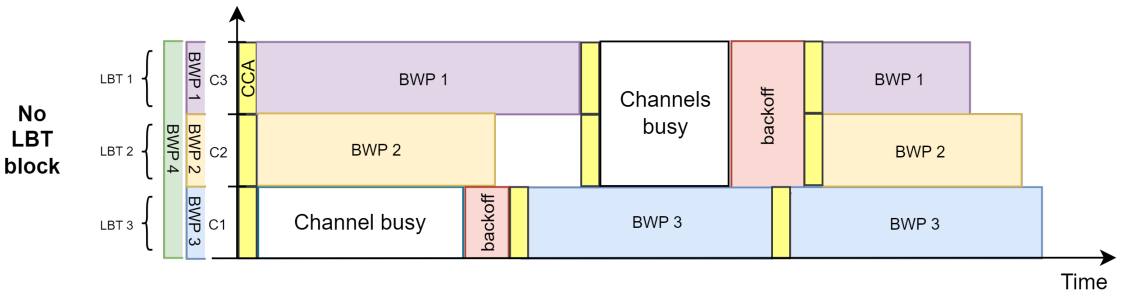
Result: The numerology and BW requirement (μ, BW) for the BWP

if $T_{slot}^{max} \geq 1$ **then**
  $i \leftarrow 0$;
else if $0.5 \leq T_{slot}^{max} < 1$ **then**
  $i \leftarrow 1$;
else
  $i \leftarrow 2$;
end
for $\mu \leftarrow i$ **to** 2 **by** 1 **do**
  for $BW \in B_\mu$ **do**
    $SNR =$ get_UE_SNR(BW);
    $MCS =$ get_MCS_from_SNR(SNR);
    $N_{PRB} =$ get_num_PRBs(μ, BW) ; /* See Table 2.2 */
    $R_{max} =$ get_BWP_max_datarate(μ, N_{PRB}, MCS) ; /* See Equation 4.2 */
    if $R_{max} \geq R_{min}$ **then**
      return (μ, BW)
  end
end
return None ; /* If requirement can not be met. */

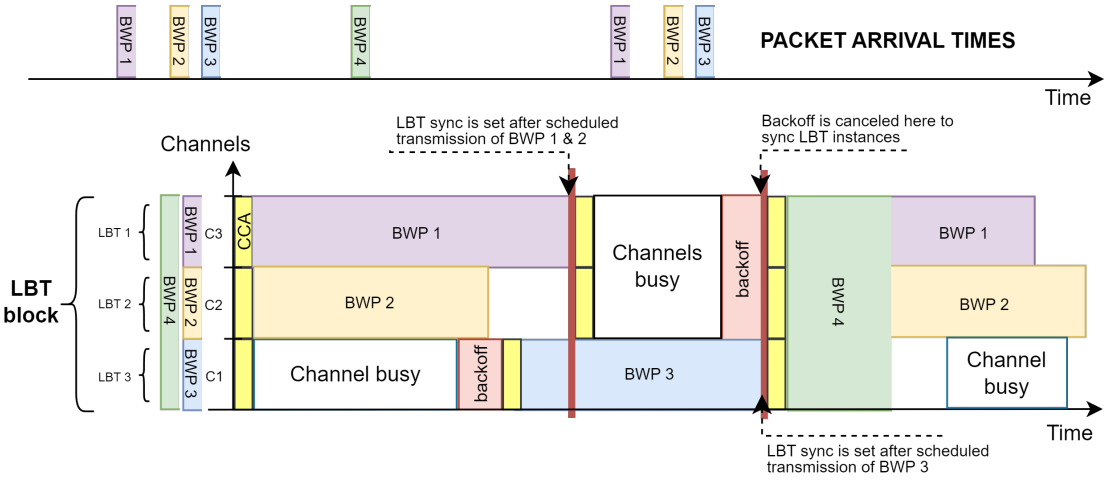
LBT blocks

The gNB will be defined with a number of BWPs divided over and shared by the UEs of the gNBs network. There is no limitation in frequency allocation of these BWPs and can be allocated adjacent, partly overlapped or fully overlapped to each other. Overlapped BWPs adds extra complexities in the spectrum sensing and access policies. Figure 4.3a shows an example of three separate LBT instances, each sensing a smaller part of the spectrum. In this example, there is a larger BWP (Green BWP 4) that overlaps with three smaller ones (BWP 1, 2 and 3). BWP 4 requires all three LBT instances to finish the sensing simultaneously in order to transmit on this BWP, assuming all channels are idle. When there is no synchronization between the three LBT instances, it is highly unlikely that the LBT instances will align at any point in time, thus the larger BWP will not be able to get access to the medium. LBT blocks are introduced in this thesis to provide a synchronization between LBT instances to

provide access to the spectrum for BWPs with larger bandwidth sizes that overlap BWPs with smaller bandwidth sizes. Next to this, an LBT block is always overlapping with a maximum of 1 neighbouring RAT. A set of BWPs can then be allocated within the LBT block, such that this set of BWPs can be placed adjacent, partly or fully overlapped with each other, but they only overlap with at most 1 other RAT.



(a): No LBT block, so there is no synchronization between the LBT instances. As a result, BWP 4 is not able to transmit, because the three separate LBT instances will never sense the spectrum to be idle at the same time.



(b): LBT block to provide synchronization between the LBT instances. The LBT synchronization event is set at the end of the latest transmission end of a BWP. Because BWP 3 is scheduled before the synchronize event, the event is moved to the end of the transmission of BWP 3.

Figure 4.3: Spectrum sensing and access policies with and without LBT block.

We define the set of all users BWP requirements R, where $r_u :=< \mu, BW > \in R \; \forall u \in U$. Let BW_{max} denote the maximum bandwidth required among all UEs in U. Let an LBT block l and an occupied channel c be defined by the same tuple: $< f, BW >$, where $f \in BW_{ch}k \mid k \in \{0, 1, \; ... \; , (BW_{gNB} - BW_{ch})/BW_{ch}\}$ is the start frequency and $BW \in \{BW_{ch}k \mid k \in \{1, 2, ... \; , BW_{gNB}/BW_{ch}\}\}$ the bandwidth size.

We define the set L_A that consists of all possible $l :=< f, BW > \forall f, \; \forall BW$, creating a total size for L_A of $\sum_{f=0}^{(BW_{gNB}/BW_{ch})-1} BW_{gNB}/BW_{ch} - f$. Next to this, we define $L \subset L_A$ to be the set of optimal LBT blocks that must be constructed from the LBT blocks in L_A. The number of possibilities to construct the set L is shown in Equation 4.7. However, not all possibilities of L are valid, and we must form the set L within Constraints 4.4, 4.5 and 4.6. The goal of LBT blocks selection is to construct the valid, optimal set $L \subset L_A$, such that L maximizes the amount of BW requirements of R that can fit within all $l \in L$.

$$max\left(\sum_{l \in L} \sum_{r_u \in R} f(l, r_u)\right) \quad \text{where } f(l, r_u) = \begin{cases} \left\lfloor \frac{BW_l}{BW_{r_u}} \right\rfloor * BW_{r_u} & \text{if } BW_l \geq BW_{r_u} \\ -BW_{r_u} & \text{otherwise} \end{cases} \tag{4.3}$$

subject to:
The constraint that any LBT block $l \in L$, can not overlap in the frequency domain with any other LBT block $l' \in L$. The function $g(l, l')$ returns 1 if l and l' partly or fully overlap in frequency domain. It returns 0 if l and l' do not overlap at all in frequency domain.

$$\sum_{l' \in L | l \neq l'} g(l, l') = 0 \quad \forall l \in L \quad \text{where } g(l, l') = \begin{cases} 1 & \text{if } (f_l = f_{l'}) \vee \\ & (f_{l'} < f_l + BW_l \leq f_{l'} + BW_{l'}) \vee \\ & (f_l < f_{l'} + BW_{l'} \leq f_l + BW_l) \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

And the constraint that we allow any l to only overlap with at most 1 occupied channel in the spectrum. The function $h(l, c)$ returns 1 if the end frequency of the occupied channel c (which is the start frequency f_c of c plus the bandwidth BW_c of c) is greater than the start frequency f_l and smaller or equal than the end frequency of l (which is $f_l + BW_l$). $h(l, c)$ returns 0 if l does not overlap with c.

$$\sum_{c \in C} h(l, c) \leq 1 \quad \forall l \in L \quad \text{where } h(l, c) = \begin{cases} 1 & \text{if } f_l < f_c + BW_c \leq f_l + BW_l \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

And finally the constraint that we do not allow any l to partly overlap with a c, LBT blocks can only fully overlap an occupied channel. Thus neither the start frequency

f_l or end frequency ($f_l + BW_l$) of any l can not lie within the start and end frequency of any c.

$$\sum_{c \in C} k(l,c) = 0 \quad \forall l \in L \quad \text{where } k(l,c) = \begin{cases} 1 & \text{if } (f_c < f_l < f_c + BW_c)\ l \wedge \\ & (f_c < f_l + BW_l < f_c + BW_c) \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

The number of unique L that can be constructed from L_A:

$$\sum_{k=1}^{N} \frac{N!}{k!(N-k)!} \quad \text{where } N = BW_{gNB}/BW_{ch} \tag{4.7}$$

Note that not all number of unique L are within the constraints. These number of L would all be valid if $|C| = 0$. When there are occupied channels in the spectrum, LBT blocks must be defined within the above constraints. An example of defining an LBT block from $f = 0$, the start of the first LBT block to the end frequency of the first LBT block is shown in Figure 4.4 by the green arrows.

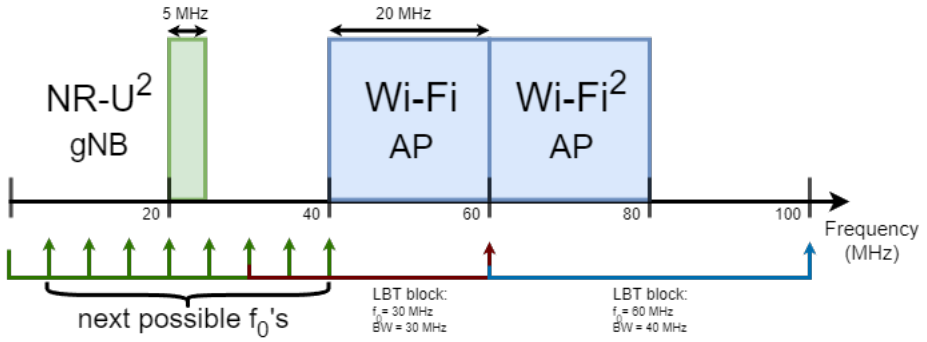


Figure 4.4: An example of a number of options to create the LBT blocks from. The spectrum is occupied by 2 Wi-Fi APs and 1 other NR-U gNB. Each LBT block can only fully overlap the bandwidth of another RAT, not partially and at most one. The distance between the green arrows equals BW_{ch}.

The optimization problem is solved using a heuristic shown in Algorithm 2. The algorithm compares different sets of LBT blocks by using the maximization function of Equation 4.3. It has a time complexity of $\mathcal{O}(R + C)$, where the size of R only depends on the $r \in R$ with a unique bandwidth size, so there are no r with duplicate bandwidths. The first set of LBT blocks is decided based on the current BW_{max} in the set R. The algorithm tries to define as many LBT blocks that have a bandwidth equal to BW_{max}. It does this by trying to create an LBT block with bandwidth size BW_{max} and calculates the amount of occupied channels

$c \in C$ it overlaps. If the LBT block can not be created either due to the end frequency of the LBT block that lies within an occupied channel, or Constraint 4.5 is not met, then the decision is made to place the end frequency of the current LBT block at the end frequency of the first overlapped LBT block. This minimizes the used bandwidth in the case that the bandwidth of the LBT block does not equal BW_{max}, and more bandwidth is left for the creation of the next LBT blocks. The second set of LBT blocks is then decided based on the second largest bandwidth in R and is denoted as the new BW_{max}. This continues until all non-duplicate bandwidth values in R have been used as BW_{max} in the creation of the set L.

Algorithm 2: LBT blocks heuristic

Data: set of occupied channels C and a set of BWP requirements R
Result: set of optimal LBT blocks L_O
$O \leftarrow -\infty$; /* Result of maximization */
$L_O \leftarrow \{\}$; /* Initialize empty set for optimal L */
sort C on increasing f_0;
while *R is not empty* **do**
  $BW_{max} =$ get_max_bw(*R*);
  Remove all $r \in R$ where $BW_r = BW_{max}$;
  $L \leftarrow \{\}$;
  $f_l \leftarrow 0$; /* Variable for start frequency position of each l */
  $f_e \leftarrow BW_{max}$; /* Variable for end frequency position of each l */
  while $f_l \neq BW_{total}$ **do**
    /* Let C_p be a list of all occupied channels that are within f_l and f_e */
    $C_p \leftarrow$ all $c \in C$ where $f_c \geq f_l$ and $f_c < f_e$;
    if $(|C_p| = 0)$ *or* $(|C_p| = 1$ *and* $f_e > f_{C_p(1)} + BW_{C_p(1)})$ **then**
      /* If l from f_l to f_e either overlaps no c or fully overlaps max one c */
      $L.insert(Channel(f_l, f_e - f_l))$;
      $f_l = f_e$;
    else if $(|C_p| = 1$ *and* $f_e \leq f_{C_p(1)} + BW_{C_p(1)})$ *or* $(|C_p| > 1)$ **then**
      /* If l from f_l to f_e either partly overlaps one c or overlaps more than one c */
      $L.insert(Channel(f_l, f_{C_p(1)} + BW_{C_p(1)} - f_l))$;
      $f_l = f_{C_p(1)} + BW_{C_p(1)}$;
    if $f_l + BW_{max} \leq BW_{total}$ **then** $f_e = f_l + BW_{max}$;
    else $f_e = BW_{total}$;
  end
  $v \leftarrow 0$;
  for $l \in L$ **do**
    for $r \in R$ **do**
      if $BW_l \geq BW_r$ **then** $v = v + \lfloor BW_l / BW_r \rfloor * BW_r$;
      else $v = v - BW_r$;
    end
  end
  if $v > O$ **then**
    $O = v$;
    $L_O = L$;
end
return L_O;

Define the BWP set

Given is the LBT blocks set L, derived from Section 4.2.1 and the updated set of BWP requirements R, such that each $r_u \in R$ fits in at least one LBT block $l \in L$. Let BW^l_{max} be the bandwidth of the l with the largest bandwidth. If any r_u has a bandwidth larger than BW^l_{max}, then $BW_{r_u} = BW^l_{max}$. It is now the task of the gNB to assign a frequency position f to each $r_u \in R$, to define a BWP b_u for each user $u \in U$, to form the set of BWPs B. Where each $b \in B$ is placed within a single $l \in L$.

We formulate this frequency allocation problem as a bin packing problem and solve it using a modification of the Best Fit Bin Packing algorithm. Bin packing is a classical combinatorial optimization problem with the goal to pack a sequence of items (represented by R in this matter) into the smallest possible number of bins [28], which we initially define as the set $J = L$. Best Fit Bin Packing tries to pack an item into the most full bin (highest load), and opens a new bin if this is not possible. Each bin and item is represented by an $l \in J$ and $r \in R$ respectively and the algorithms tries to pack each r into an l. The capacity of a bin l is the bandwidth of l, and the size of each r is also the bandwidth of r. Whereas in the original algorithm, the bins have a fixed size, the modified algorithm could have different values for the bandwidth of each item in L. When each bin $l \in J$ has been opened and the current r does not fit anywhere, the set J is extended with all $l \in L$. All these new l bins are marked as *zero load*, and *closed*, to provide the same LBT blocks, but new bins that can be opened again for the current r. Thus, in the modified algorithm, bins are added as groups of the content of L. As in the original Best Fit algorithm, the goal is to minimize the amount of opened bins in J. To translate this to the context of placing BWP requirements into LBT blocks, it is tried to minimize the number of times that an LBT block has to be opened, and is formulated as:

$$min\left(\sum_{l \in J} O(l)\right) \quad \text{where } O(l) = \begin{cases} 1 & \text{if } l \text{ is marked as } \textit{opened} \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

subject to:
The lower bound that we must at least use one LBT block: $min(\sum_{l \in J} O(l)) \geq 1$.
The one dimensional capacity (in bandwidth) of each $l \in J$ can not be exceeded by the size (in bandwidth) of each r that has been allocated in that l.

$$\sum_{r \in R} BW_r x_{rl} \leq BW_l, \ \forall l \in J \tag{4.9}$$

where $x_{rl} = 1$ if BWP requirement r is allocated in $l \in J$. $x_{rl} \in \{0, 1\}$.
Each r can only be allocated in a maximum of 1 l.

$$\sum_{l \in J} x_{rl} = 1 \tag{4.10}$$

Best Fit Bin Packing has a approximation guarantee of *BF(R)* $\leq \lfloor 1.7 * OPT) \rfloor$. Thus, if the optimum amount of bins equals *OPT*, the algorithm will produce at all times a set B with at most $\lfloor 1.7 * OPT \rfloor$ bins used [28]. The time complexity for Best Fit is $\mathcal{O}(|R| \log(|R|))$ [29]. Where a set of BWPs B is returned for a given set of LBT blocks L and BWP requirements R, where $r :=< \mu, BW > \in R$. Each r has now been assigned a frequency position and has become a BWP $b \in B$.

Single vs Multi-BWP

The final step is to assign a set of BWPs and an active BWP to each user of the gNB, we denote this by B_u and $b_a \in B_u$ respectively $\forall u \in U$. Each $b \in B$ corresponds to an $r \in R$, thus each b formed by r_u of user u is placed in B_u. This first $b \in B_u$ is also assigned as the users first Active BWP b_a. For the exploitment of BWPs in the multi-BWP model, we want the possibility to copy the BWP that corresponds to r and place it on a different l. This is done by adding a duplicate value for each r in R. The bin packing algorithm assures that a second BWP for the UE is allocated in another LBT block.

4.2.2 Proposed BWP scheduling heuristic

This section describes the proposed BWP scheduling heuristic to optimally schedule the set B in time when the gNB has data to transmit to its UEs. This happens during the so-called scheduling phase of the gNB and is executed for a longer duration than the pre-processing phase. In the scheduling phase, the channel access procedures and transmissions take place. The synchronized LBT instances of Figure 4.3b are initialized per LBT block. Each LBT block will run in parallel on the gNB, such that the all events in the scheduling phase are within an LBT block and are independent of each other. You could see this as multiple tiny gNBs that cooperate together. We therefore focus on the behaviour of a single LBT block, which can be divided into several sub-phases. Figure 4.5 shows the flowchart of the entire scheduling phase.

Spectrum sensing and access policies

The LBT Blocks are running parallel from each other, thus perform their spectrum sensing and access procedures independent from each other. We use CAT4-LBT

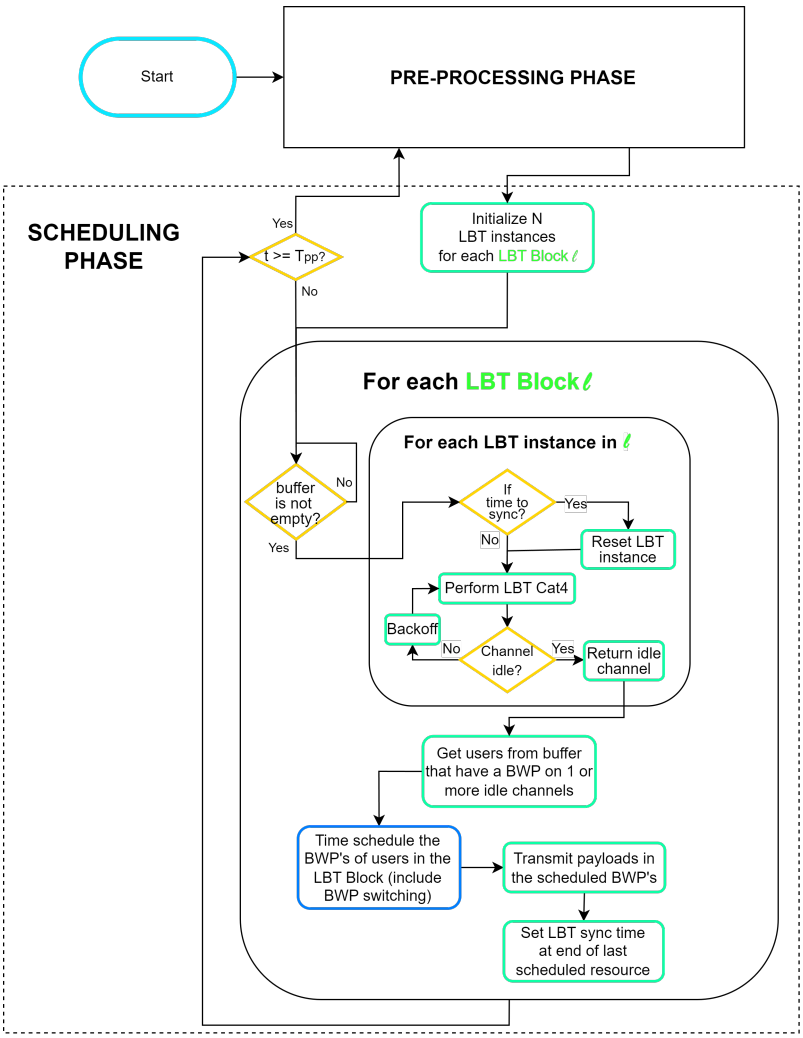


Figure 4.5: Flowchart of the scheduling phase of gNB. When there is data to transmit, LBT Category 4 is performed on each LBT block and BWPs are scheduled based on the channels that are determined idle.

for all performed sensing and have a similar behaviour of an Load Based Equipment (LBE) system. Even though the LBT blocks are independent of each other, the gNB's buffer is shared between them. As an LBE system, it only acts as soon as there is data in the buffer to transmit. Thus, all LBT blocks are simultaneously sensing their respective channels with a bandwidth of BW_{ch} with their synchronized LBT instances. The LBT instances of an LBT block are synchronized at the end of scheduled transmission, shown in Figure 4.3b. When a new transmission is scheduled before the synchronization event, the event is updated to the end of that transmission time, as shown in Figure 4.3b. So, any LBT instance that does not schedule a new transmission, is synchronized with the last transmission of another LBT instance. The LBT instances that are in a backoff procedure also reset and start with the initial CCA check. This allows the LBT block to schedule larger BWPs.

BWP scheduling

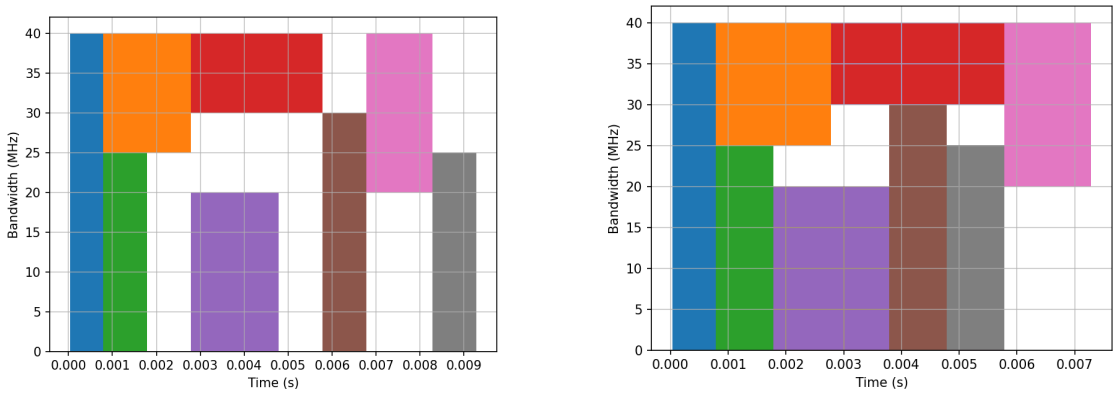
When an LBT instance, or multiple LBT instances have determined their channel to be idle, the gNB must obtain the set of users that have a BWP available on these idle channels. For the single-BWP model, this can only be the users first-active BWP. For the Multi-BWP model, this can be any BWP of the user. Since the LBT blocks are asynchronous from each other, the first LBT block that has a BWP on an idle channel can remove the user corresponding to that BWP from the queue. It is therefore difficult to preserve a First-In First-Out (FIFO) way of ordering the queue, since users that do not have an idle BWP, can not be handled at that point in time. So, the users that have data in the queue and an idle BWP within the respective LBT block are obtained and are still handled in a FIFO way.

At this point, from the perspective of within any LBT block, we have obtained the set of users that have an idle BWP in this LBT block. Based on the payload size and MCS, the time duration that is necessary to transmit the payload size on the BWP can be calculated, defining the time-frequency resource block. The minimum time granularity for a data transmission on the NR-U network is in the order of symbols, while the granularity of the airtime is in the number of slots. This means that the time duration of the payload transmission on a BWP is at minimum one slot. Dependent on the numerology of the BWP, the symbol duration varies. For the multi-BWP model, if this BWP is not the current active BWP, the BWP switch delay of the UE must be added to the scheduling time of the defined resource block, which can be obtained from Table 2.3. If not all the UEs obtained from the queue can be scheduled within this TXOP frame, the UEs that have arrived later in the queue will have a lower priority and be unscheduled. All unscheduled UEs will be returned to the queue in the same order that it was taken from it.

Since BWPs are already allocated in the frequency domain, when a number of users have a payload to transmit, the gNB only has to schedule the respective BWPs of these users in time. Since we consider a FIFO fashion of prioritizing users in the queue, we can simply schedule the BWPs on the priority of the queue. Starting from a relative scheduling time $t = 0$, the scheduling algorithm loops over the users in FIFO priority and tries to schedule as many BWPs possible at $t = 0$ that can fit within the bandwidth limits, it is therefore possible that a user with lower priority is placed earlier in time. Note that the starting time of a BWP must be shifted in the case that the BWP is not the currently active one of the user in the multi-BWP model. When all users BWPs have been checked and no more BWPs fit within the bandwidth limits, the relative scheduling time is updated to the time at which the largest transmission duration of any currently scheduled BWP ends. An example

result of this scheduler is shown in Figure 4.6a. The figure shows quite some gaps in between the time-frequency resources and it can be clearly seen that the result is far from optimal.

We update the scheduler with the goal to obtain a more optimal time-frequency scheduled set of BWPs. To quantify this, the updated scheduler should in general transmit each payload of all (to be scheduled) users faster in time than the baseline scheduler. The updated scheduler still handles the users in the order of the queue at first, but for each current users BWP that is ought to be scheduled, the other BWPs of users with lower priority are compared in time with the current users BWP. The BWP that can be placed at the earliest point in time is given priority over the standard queue priority. This creates in general less gaps between the time-frequency resources, see Figure 4.6b. A quantifiable comparison between the two schedulers



(a): The baseline (old) scheduler shows time gaps in between some BWPs.

(b): The updated (new) scheduler is able to place the BWPs more efficient in the spectrum.

Figure 4.6: Comparison of the new and old time scheduler of 8 BWPs within an LBT block of 40 MHz. It can be seen that the updated scheduler can schedule the same set of BWPs in a shorter time window.

is obtained by running a script that executes both schedulers a number of times with a set of BWPs for each iteration. This shows a fair comparison in which the better scheduler should schedule the same set of BWPs and have a smaller total transmission time. In each iteration, the number of users to schedule, the BWP of each user, the number of LBT blocks and the LBT block bandwidth is random. This creates a different number of samples for a different number of users and BWPs on each LBT block, where in this comparison only the first LBT block is used. The gNB in the script always has 15 UEs, we see a number of users/BWPs varying from 2 to 10 on the first LBT block of the gNB. To obtain a valid comparison, we only om-

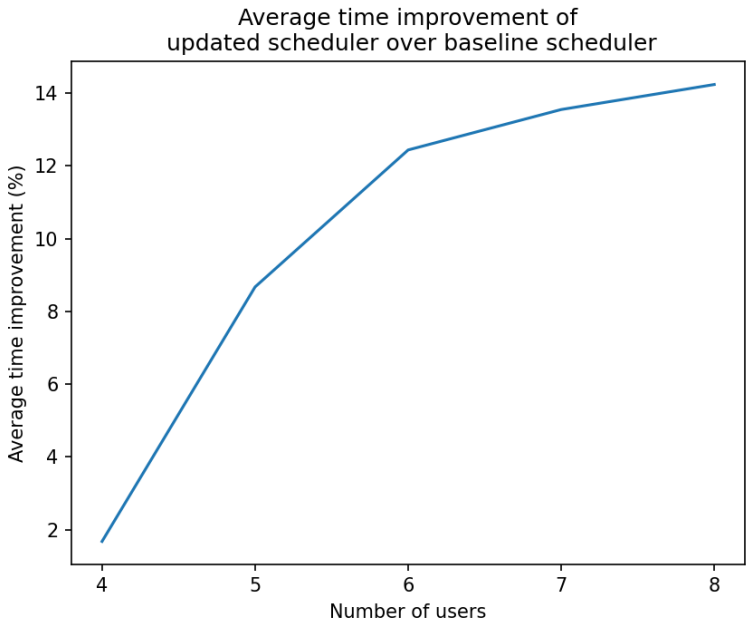


Figure 4.7: The average time improvement (in %) of the updated BWP scheduler compared to the baseline BWP scheduler for the number of users.

mit the comparison with the number of users/BWPs that have enough samples. At increasing number of samples, the time improvement of the updated scheduler per number of users stagnated to the results shown in Figure 4.7.

4.3 Model for the Wi-Fi network

4.3.1 Initialization

The focus of this thesis lies at the gNB, it is therefore decided to keep the model of the Wi-Fi network rather simple. Similarly to NR-U, the traffic model is homogeneous for the entire Wi-Fi network and also homogeneous for both NR-U and Wi-Fi, to provide a more fair setup. The STAs are distributed uniformly over the UMi environment with respect to the AP, with a minimum distance of 10m from the AP. The start frequency f_{AP} of the channel of the AP is a randomly chosen over the investigated piece of unlicensed spectrum with a bandwidth of BW_{total}, and can have any value of $f_{AP} \in BW_{ch}k \mid k \in \{0, 1, \ ... \ , (BW_{gNB} - BW_{AP})/BW_{ch}\}$.

4.3.2 Spectrum access policies and user scheduling

The Wi-Fi AP uses the CSMA/CA procedure to obtain access to its channel. This procedure is exactly similar to the CAT4-LBT used by the gNB in this thesis, because the RTS/CTS Exchange of CSMA/CA is not used and all the same parameters of

Table 2.4 are used for both CSMA/CA and CAT4-LBT for a fair comparison. The main difference, is that the AP senses with a bandwidth granularity of BW_{AP}, since it uses OFDM, where the entire channel bandwidth is allocated to a single user at a time. The gNB senses with a bandwidth granularity of BW_{ch}. Therefore, when only a small portion of the AP's channel is occupied by another RAT, the entire channel is sensed as busy and the AP has to backoff. Similarly to NR-U, for simplification, we always assume a perfect channel condition without any packet collisions. The CW is therefore never updated by the binary exponential function.
The user data transmissions are also similar to NR-U, handled in a FIFO manner, where the users are simply prioritized by the time order in which they have data ready to transmit. In contrast to the gNB, the time duration of a payload of a user is simply defined by the payload size and data rate, which in terms is defined by the MCS, which is derived from the SNR of the receiving STA. A Wi-Fi transmission does not have to use an entire symbol, and can partially use symbols to transmit a payload. The granularity of the transmission time is therefore in the order of microseconds. The AP tries to schedule as many STAs that have data to transmit, in the FIFO order of the buffer within the given TXOP time.

Chapter 5

Simulator Framework

A simulator is required to give quantifiable answers to the research questions. Existing simulator frameworks such as NS-3, do not support the specific requirements that are necessary for this thesis. This includes BWP configuration (including numerology) and the BA feature, all in an unlicensed spectrum scenario. Therefore, a new simulator is developed in python, which we refer to as the Unlicensed Spectrum Simulator (USS) from now on. The USS is a continuous time simulator that can simulate a number of basestations contending for a number of channels in the unlicensed spectrum. Most important, with this design freedom, it is possible to implement the gNB design from Chapter 4. This comes at the cost of some aspects, explained later in this chapter, that makes the radio simulator less realistic compared to other simulator frameworks.

5.0.1 Environment and pathloss models

In the USS, it is possible to build your own simple environment, such that you can decide the square meters of area, create indoor/outdoor spots and distribute UEs and STAs corresponding to a gNB and AP respectively. The models implemented in the USS are adopted from 3GPP and consist of a pathloss model with log-normal shadow fading, a Line-of-sight (LOS) probability and an Outdoor-to-Indoor (O2I) building penetration loss model [24]. These models are different for different scenarios, we adopted the ones corresponding to the UMi Street Canyon, for which the environment is build in Chapter 6 accordingly for the simulation runs. Figure 5.1 shows an example environment of $50x50m^2$ that has one building (the grey area) and an outside area (white area). Using the 3 dimension coordinate system (x,y,z), six UEs are distributed over the total area, from which three are outdoors and three are indoors, all at the height of $h_{UT} = z = 2.5m$. The entity in the left top is the gNB at the height of $h_{BS} = z = 10m$ and the lines between each UE and the gNB show the path attenuation in 3D space (dB). The distance between the gNB and a

UE can be derived with Equation 5.3, where d_{2D-out} is the two dimensional distance on the (x, y) plane, that is outside the building and d_{2D-in} the distance inside the building, see Figure 5.2. Thus $d_{3D-in} = 0$ and $d_{2D-in} = 0$ for outside users. It can be noted that users within the building are experiencing a higher pathloss, thus a lower receive power due to penetration loss of the building. In the pathloss calcu-

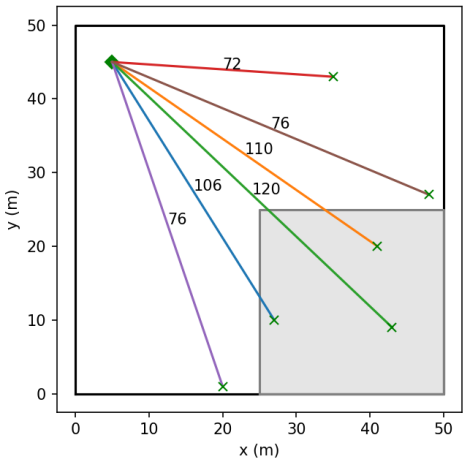
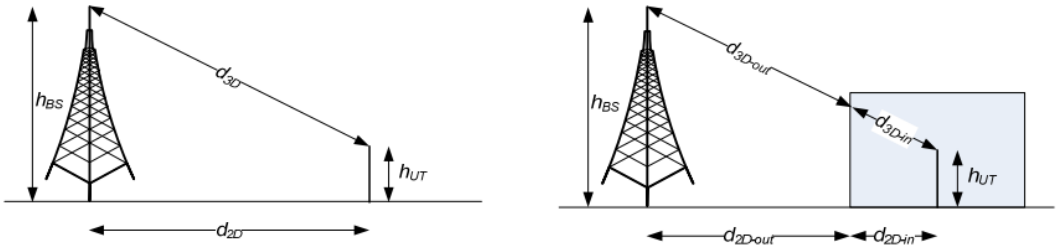


Figure 5.1: Top view of the pathloss (dB) between multiple UEs and their gNB in the USS. A number of UEs are located inside a building (grey area) and the others are outside (white area).

lation between the gNB and the UE, it is first derived whether the UE is inside or outside. If outside, it is decided whether the UE is in LOS with the gNB using the LOS probability Equation 5.4. The formulas for the pathloss for both LOS and Non-line-of-sight (NLOS) from [24] are defined as in Equation 5.1 and 5.2 respectively. The shadow fading is included as a sample of a normal distribution with the pathloss as mean value and standard deviation σ_{SF}, where $\sigma_{SF} = 4$ and $\sigma_{SF} = 7.82$ for LOS and NLOS respectively.

$$
\begin{aligned}
PL_{UMi-LOS} &= \begin{cases} PL_1 & \text{if } 10m \leq d_{2D} \leq d'_{BP} \\ PL_2 & \text{if } d'_{BP} \leq d_{2D} \leq 5km \end{cases} \\
PL_1 &= 32.4 + 21\log_{10}(d_{3D}) + 20\log_{10}(f_C) + N(0, \sigma^2_{SF}) \\
PL_2 &= 32.4 + 40\log_{10}(d_{3D}) + 20\log_{10}(f_C) - 9.5\log_{10}(d'_{BP})^2 \\
&\quad +(h_{BS} - h_{UT})^2 + N(0, \sigma^2_{SF})
\end{aligned} \tag{5.1}
$$

$$
\begin{aligned}
PL_{UMi-NLOS} &= \text{max}(PL_{UMi-LOS}, PL^{'}_{UMi-NLOS}) \\
&\quad \text{for } 10m \leq d_{2D} \leq 5km \\
PL^{'}_{UMi-NLOS} &= 35.3\text{log}_{10}(d_{3D}) + 22.4 \\
&\quad 21.3\text{log}_{10}(f_C) - 0.3(h_{UT} - 1.5)
\end{aligned} \tag{5.2}
$$



(a): All distance definitions used for users that are outside. The propagation path between the basestation and user is only outside.

(b): All distance definitions used for users that are inside. Now the path consist of a distance outside plus a distance inside.

Figure 5.2: The distance definitions adopted from 3GPP TR 138 901 [24].

$$
d_{3D} = d_{3D-out} + d_{3D-in} = \sqrt{(d_{2D-out} + d_{2D-in})^2 + (h_{BS} - h_{UT})^2} \tag{5.3}
$$

$$
Pr_{LOS} = \begin{cases} 1 & \text{if } d_{2D-out} \leq 18m \\ \frac{18}{d_{2D-out}} + \text{exp}(-\frac{d_{2D-out}}{36})(1 - \frac{18}{d_{2D-out}}) & \text{if } 18m < d_{2D-out} \end{cases} \tag{5.4}
$$

When the user is inside, the penetration loss model is included to add the loss of the signal through the walls of the building, shown in Equation 5.5.

$$
PL = PL_b + PL_{tw} + PL_{in} + N(0, \sigma_P^2) \tag{5.5}
$$

Where PL_b is the basic outdoor pathloss from Equation 5.1 or 5.2. PL_{tw} the building penetration loss through the external wall, PL_{in} the inside loss and σ_P^2 the variance for the penetration loss, added by a normal distribution sample. For simplifications of the penetration loss model and suggested by [24] for backwards compatibility with TR 36.873 [30] for sub 6 GHz carrier frequencies, we use the parameter values shown in Table 5.1.

Parameter	Value
PL_{tw}	20 dB
PL_{in}	0.5 d_{2D-in}
σ_P	0 dB
σ_{SF}	7 dB (replacing the value in Equation 5.1 or 5.2)

Table 5.1: Parameters for the penetration loss model from [30].

5.0.2 BWP Generation

As explain in Chapter 4, during the pre-processing phase of the gNB, an optimal set of LBT blocks is defined, together with a set of BWPs that are placed within these LBT blocks using the Bin Fit Packing algorithm. For each simulation run, the Wi-Fi AP is first randomly given a channel in the investigated spectrum, before the gNBs pre-processing phase starts. This indicates that the AP should already have been active in the spectrum, before the gNB performs its pre-processing phase. Based on the position of the Wi-Fi AP, the LBT blocks and BWPs are defined. Figure 5.3 shows a few examples of generated BWPs, based on different frequency locations of the Wi-Fi AP. In some cases, three LBT blocks are generated and in others only two. This can be observed by looking at clear boundaries where BWPs do not overlap. The figure in the bottom right shows an example of BWPs generated in the multi-BWP model, where each BWP requirement is duplicated as input, such that two similar BWPs on different channels and different LBT blocks are outputted.

5.0.3 Selective UE BWP switching

As an addition to the system model described in Chapter 4, we provide the option to either let all UEs exploit the BA feature or only a selective number of UEs. More specifically, either all UEs can switch between their set of BWPs or only the UEs that have their first activated BWP overlapping with the Wi-Fi AP can switch between their set of BWPs. The downside of the first option as it was originally designed in Chapter 4, is that it can also schedule more UEs on the channel of the Wi-Fi AP. By the Bin Fit Packing algorithm of Section 4.2.1, the UEs are distributed over the entire BW_{gNB}, and only a number of UEs have a first active BWP overlapping with the AP channel. While providing the BWP switch option for all UEs, this can positively impact their latency. However, scheduling more UE on the crowded parts of the spectrum where Wi-Fi operates, negatively impacts the fairness and throughput of the Wi-Fi network. As shown in Figure 5.4, it can be noted that for a high load of 45 UEs on the gNB, the AP is not able to obtain the same average throughput for 45 STAs or more. It is therefore decided to further investigate the multi-BWP model of the gNB where only

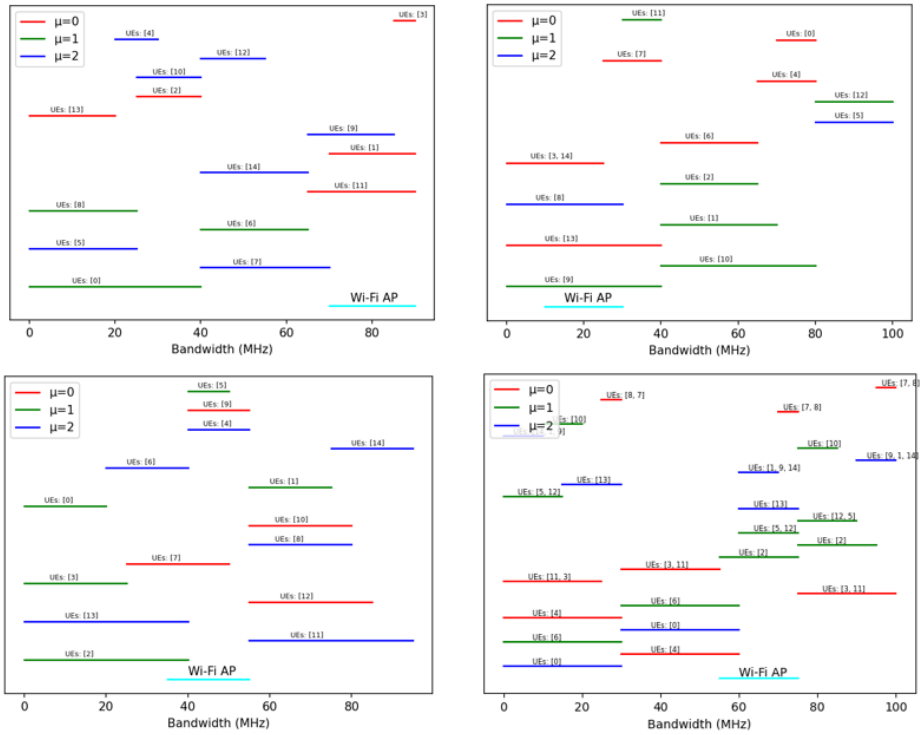


Figure 5.3: Four examples of generated BWPs, as a result of the pre-processing phase of the gNB. The figure in the bottom right is a result of the multi-BWP gNB model, where each BWP requirement from the input is duplicated. This results in the same BWP configuration (numerology and BW) being at least twice in the total BWP set, on different channels and different LBT blocks. The LBT blocks might be difficult to distinguish, but in the bottom right figure, the LBT blocks range from: 0-30 MHz, 30-75 MHz and 75-100 MHz.

a selective number of UEs that have their first active BWP overlapping with the AP can use the BA feature.

5.0.4 User data rates and data transmission

The rate at which a user can receive its data depends on the link quality between the receiving user and its corresponding transmitting basestation. For simplification, we only consider the pathloss PL as explained in Section 5.0.1 and the thermal noise N_T to impact the SNR of the user. Interference by other nodes are neglected and we always assume a perfect channel condition. The SNR of a user is calculated as follows: $SNR = P_{tx} - PL - N_T$. For both the gNB and the AP, we adopted the same SNR to MCS table from [14], which translates the obtained users SNR value in the range of [-23, 40] dB to the MCS that it can use.

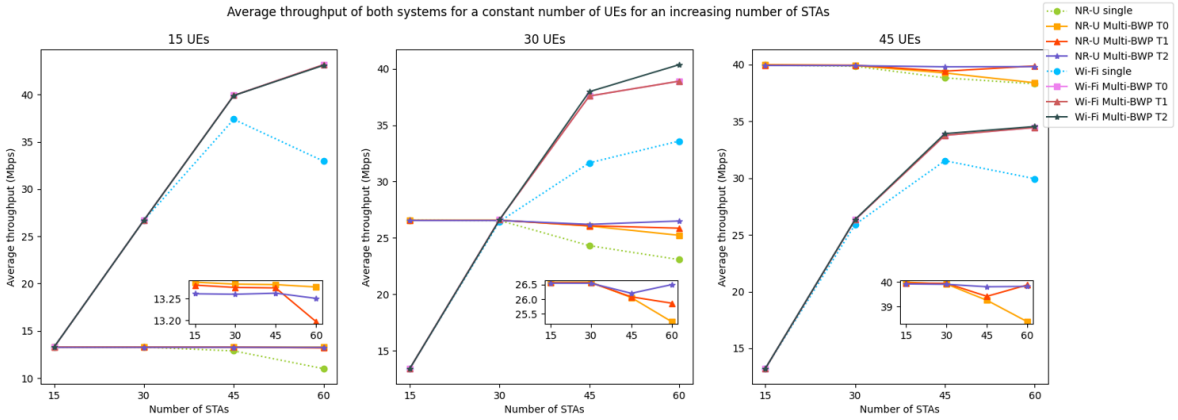


Figure 5.4: Average throughput for both the gNB and AP where all UEs can use the BA feature. The datapoints of each figure are calculated over a number of simulation runs.

The data rate R_w for a Wi-Fi STA is calculated as shown in Equation 5.6, almost similar to maximum data rate Equation 4.2 of a BWP.

$$R_w = \frac{N_t * V}{T_{dft} + T_{gi}} 10^{-6} \quad \text{where } V = v_{layers} * Q_m * SF * R_c \tag{5.6}$$

Where N_t are the number of tones, which are the number of subcarriers in a Wi-Fi Resource Unit (RU). So, the number of subcarriers is calculated as $N_t = BW_{AP}/\Delta f_{AP} - 12$, where 12 subcarriers are subtracted since they are used as null and guard subcarriers in the RU. T_{dft} is the OFDM symbol duration and T_{gi} the guard interval in time. Since Wi-Fi does not have to use an entire OFDM symbol for data transmission, the time duration of users payload is then simply derived from the payload size and the data rate of Equation 5.6.

The data transmission between the gNB and the UE in the NR-U network is implemented a little different. Since we consider a granularity of slots for the gNB, a number of symbols is derived from the MCS. Each slot consist of 14 OFDM symbols, so the number of symbols are divided by 14 and rounded to the higher integer to obtain the number of slots required to transmit the payload.

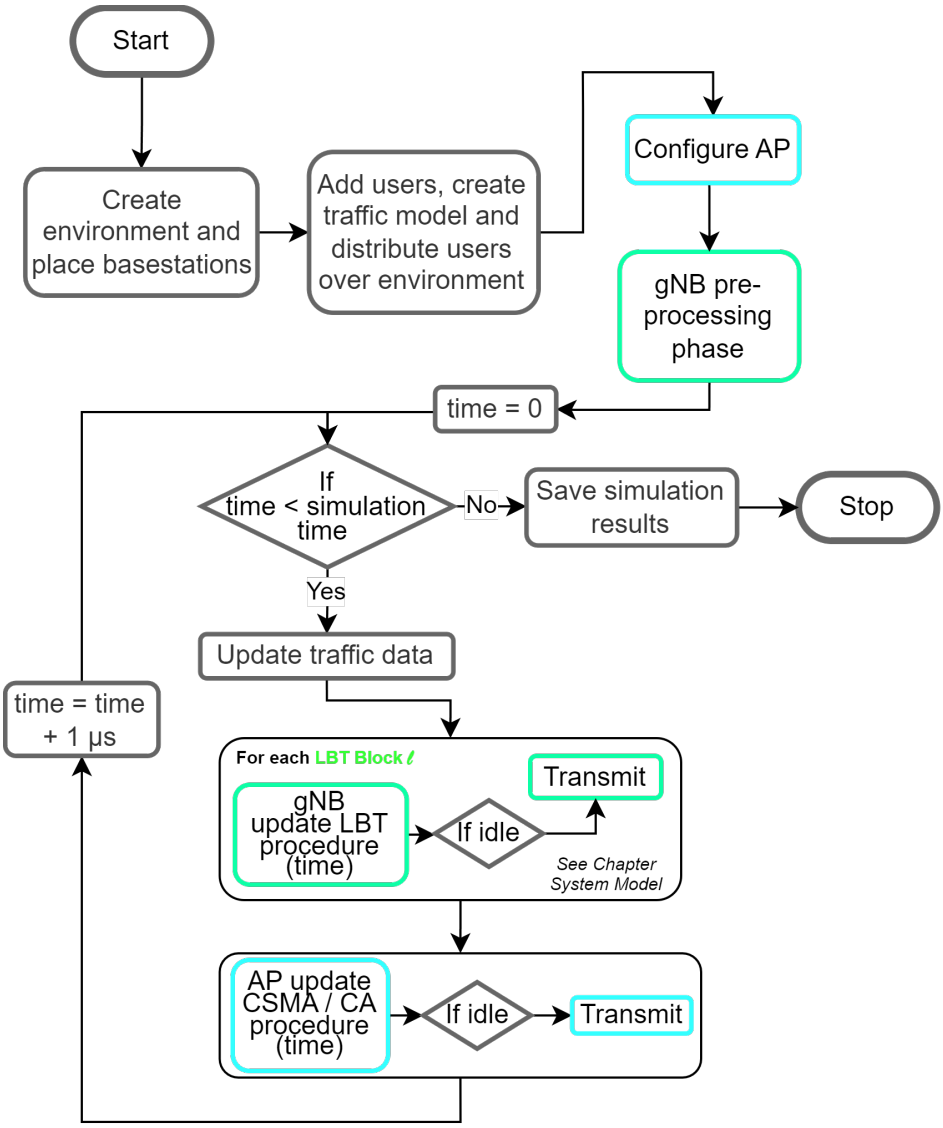


Figure 5.5: Flowchart of the continuous time USS. First the scenario and context is setup, including the configuration of the AP and the pre-processing phase of the gNB. Afterwards, the gNB and AP each sense and transmit into the medium for a given simulation time.

Chapter 6

Performance Evaluation

This simulator framework of Chapter 5 is used to perform a number of simulations. The data generated from these simulations are graphed for several metrics, including throughput, medium access latency, airtime utilization, number of BWP switches and switch overhead. Which are used to perform an extensive analysis in order to give answer to the research questions of this thesis. First, the simulation configuration is described. Second, the results are categorized and evaluated as the metrics in the same order as above.

6.1 Simulation configuration

6.1.1 Environment

The environment of the simulator is build to represent a street canyon and for the greater parts uses the parameters recommended by the 3GPP for the UMi Street Canyon model [24]. The main difference is the suggested dimensions and shape of the environment. Since we only consider simulation of one gNB, we focus on a single cell environment and give this, for simplification, a square area of 100 by 100 meters. 3GPP suggests a hexogonal grid as the cell layout with an intersite distance (ISD) of 200m. However, they do note that the typical open area is in the order of 50 to 100m. Therefore, a street canyon was built shown in Figure 6.1.

6.1.2 Traffic model

We consider a homogeneous traffic model over all users in the environment. Thus, both UEs and STAs have the same traffic model. At first, a gaming traffic model and video streaming model, both suggested by 3GPP in a paper by Navarro-Ortiz et al. [31] were adopted in the simulator. However, due to relatively low byte sizes of the packets in both traffic models, the airtime differences between NR-U and Wi-Fi

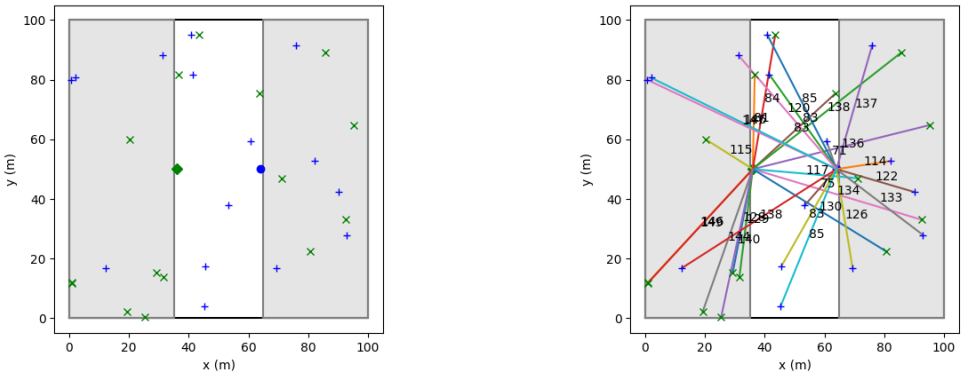
Parameters		UMi - street canyon
Environment dimensions	Total area	100 x 100 m^2
	Buildings	35 x 100 m^2
	Street	30 x 100 m^2
BS antenna height h_{BS}		10m
UT location	Outdoor/indoor	Outdoor and indoor
	LOS/NLOS	LOS and NLOS
	Height h_{UT}	$1.5m \leq h_{UT} \leq 22.5m$
Indoor UT ratio		80%
Min. BS - UT distance (2D)		10m
UT distribution		uniform

Table 6.1: Parameter values for the modified UMi - street canyon environment adopted from [24] used in the simulation.

were significant. In the paper of Patriciello et al. [17], they mention that the airtime of an NR-U device is occupying the channel almost three times more than a WiGig device, because of the mimimum resource allocation granularity. In their research, this is an entire OFDM symbol for NR-U, while for WiGig, this OFDM symbol duration does not have to be finished. In this thesis, the granularity difference is even larger, since we consider a slot-based granularity for the NR-U devices, where the number of symbols are divided by 14, since there are 14 symbols in a slot and rounded to the nearest higher integer of slots. On top of this, the subcarrier spacing of the Wi-Fi network is much higher (312.5 kHz), compared to a maximum subcarrier spacing of 60 kHz for NR-U. This results in a much lower symbol duration of the Wi-Fi network, but gives you less PRBs on the same bandwidth. Thus, if the packet sizes stay small enough to only occupy a few symbols, the airtime for Wi-Fi comes in the order of microseconds, while the NR-U transmission for that same packet would be fit into a slot in the order of milliseconds.
Therefore, we consider a File Transfer Protocol (FTP) traffic model of a file with unlimited size. For the packet size, we use the Largest Extreme Value (LEV) distribution adopted from the 3GPP gaming traffic model with the location parameter $P_a = 1500$ bytes and the scale parameter $P_b = 36$ bytes. The arrival time of a packet is also a LEV distribution with $t_a = 12.5$ ms and $t_b = 2$ ms. This creates an average data rate of $P_a * (1/(t_a * 10^{-3})) * 8 = 0.96$ Mbps.

6.1.3 Different gNB models

In the simulation runs, we compare a number of different gNB models that can give some insights on the BWP switching performance in the unlicensed spectrum. The



(a): Street canyon environment with a gNB (green) and AP (blue) with both 15 users.

(b): Street canyon including pathloss (dB) shown between the gNB and its UEs and the AP and its STAs.

Figure 6.1: Topview of the street canyon environment used in the simulation. This shows one gNB (green diamond) and one AP (blue dot), both mounted below rooftop levels at 10m height. Both basestations have each 15 users distributed over the environment.

models are as follows:

1. Single-BWP model: where each UE is only assigned a single BWP, and can only use this BWP for the rest of the simulation.
2. Multi-BWP model: where each UE is assigned two or more BWPs, and can switch between them based on which BWP has earlier access to its channel according to the CAT4-LBT procedure. There are three variants for this model, based on the delay types of Table 2.3:
   (a) Multi-BWP Type 0 model: there is no delay required at the UE side between a BWP switch. This is an unrealistic model;
   (b) Multi-BWP Type 1 model: this introduces a delay of type 1 as shown in Table 2.3 in the granularity of slots, dependent on the smaller numerology of the two BWPs associated with the switch;
   (c) Multi-BWP Type 2 model: similar to the Type 1 model, but with an increased switch delay time, shown as type 2 in Table 2.3. These delay times are the minimum requirement for a UE to have. This models forms the most extreme case, with the longest delays.

The single-BWP is compared against all three variants of the multi-BWP model in each simulation run. There is never a combination of multi-BWP variants deployed, all UEs either have a switch delay based on the Type 0, 1 or 2 model.

Parameter	NR-U	Wi-Fi
Average data rate / user	0.96 Mbps	0.96 Mbps
BW_{gNB} / BW_{AP}	100 MHz	20 MHz
Simulation time	1 s	
Total observed bandwidth BW_{total}	100 MHz	
Minimum channel bandwidth BW_{ch}	5 MHz	
Minimum time units	1 μs	
gNB multi-BWP	[true, false]	-
Multi-BWP variants	[T0,T1,T2]	-
BA feature for all UEs	false	-
N_{UE} / N_{STA}	[15,30,45]	[15,30,45,60]
Subcarrier spacing Δf	15/30/60 kHz	312.5 kHz
CW	15	15
TXOP	4 ms	4 ms
CCA time duration T_{CCA}	34 μs	34 μs
Backoff slot duration T_{bs}	9 μs	9 μs
Carrier frequency f_C	5 GHz	5 GHz
Height of users h_{UT}	1.5 - 22.5 m	1.5 - 22.5 m
Height of basestation h_{BS}	10 m	10 m
Tx power of basestation P_{tx}	30 dBm	30 dBm

Table 6.2: The parameters used in simulation runs of the USS. Parameter values in square brackets show the values that were changed over different simulation runs.

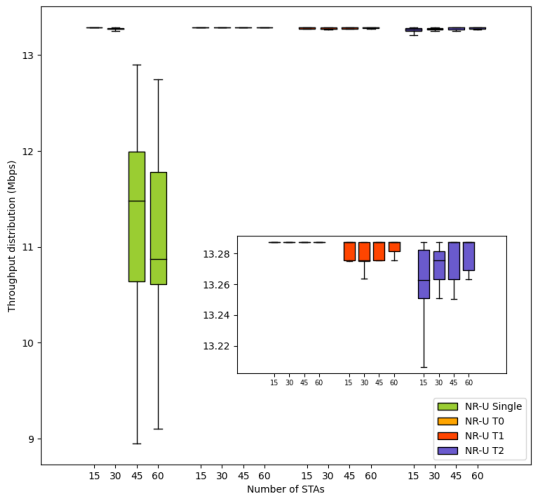
6.2 Throughput performance

6.2.1 Throughput improvement of the multi-BWP gNB

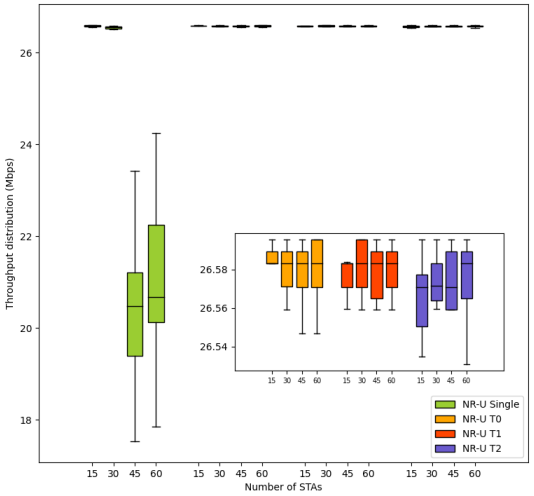
The main goal of leveraging BWPs in the unlicensed spectrum is to increase the throughput of the gNB. Figure 6.2 shows the distribution of the throughput for each gNB model over multiple simulation runs. In each simulation run, the generated BWPs for the gNB are different, because the Wi-Fi network chooses another channel. If we focus on the NR-U single-BWP model (green data set), it can be observed that overall, the median has a decreasing trend for an increasing number of STAs for all number of UEs. Thus, for an increasing load on the AP, the throughput of the gNB will start to decrease. The largest drop in throughput can be noted between 30 and 45 STAs. This can be explained because the load on the AP becomes large enough to start starving the UEs that have a BWP on that same channel. As shown in Figure 6.11, the airtime utilization for the AP on its 20 MHz channel in coexistence with

the multi-BWP gNB model can become 95% at 45 STAs. For the single BWP gNB model, the average airtime utilization of the AP stagnates after 30 STAs to around 60 to 65%. Thus, starting from 45 STAs, both the gNB and AP have a negative impact in throughput for any load on the gNB. For a relatively low load on the AP (15 to 30 STAs), the distribution of the throughput is condensed. For a relatively higher load (45 to 60 STAs), the distribution of the throughput is more spread out and can go up to a difference of approximately 45% (For the NR-U Single boxplot of 45 UEs and 45 STAs). This more spread out gNB throughput values for 45 and 60 STAs in the NR-U Single dataset set can also be explained by the high load and airtime utilization of the AP. Since the AP tries to access the medium more often, there is no guarantee that the gNB can transmit all its packets on that channel, but there might be cases where the gNB wins the contention of the channel more than the AP and as a result has a higher throughput.

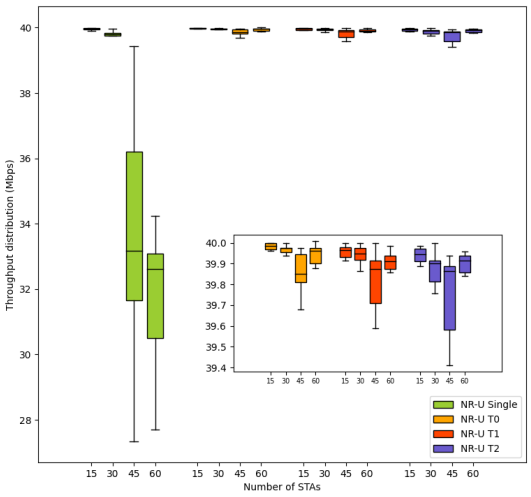
The multi-BWP gNB models are shown in Figure 6.2 as an orange, red and purple data set for delay type 0 (T0), delay type 1 (T1) and delay type 2 (T2) respectively. The throughput of the gNB between these multi-BWP models and the single-BWP model increases for 45 and 60 STAs, for all load on the gNB. Since there is no overlap in the minimum values of the T0, T1 and T2 boxplots and the maximum values of the single-BWP boxplots for 45 and 60 STAs, it can be stated that throughput for the multi-BWP gNB is always higher for these AP load values than the single-BWP gNB. For 15 and 30 STAs on the AP, the throughput between the single-BWP and the multi-BWP gNB can vary from -0.5% to 0.5%, which is a difference that can be neglected. Figure 6.3 shows the average throughput of both the gNB and AP over the multiple simulation runs. AP Single means the performance of the AP in coexistence with the single-BWP gNB model, and AP T0/T1/T2 in coexistence with the multi-BWP gNB models. The percentage labels show the improvement of the AP and gNB when the gNB uses the multi-BWP T2 model compared to when the gNB uses the single-BWP model. The other multi-BWP models are not labeled, since the average throughput differences between the multi-BWP models is negligibly small. We therefore use the T2 model as a representation for the overall average throughput improvement. The average gNB throughput increase for the multi-BWP models ranges from 16% up to 30% for 45 and 60 STAs. The largest increase can be observed when the gNB has 30 UEs, this is where the gNB with the single BWP model performs worst, due to the number of UEs that are on the same channel as the Wi-Fi AP being too little relative to the number of STAs. This means that these UEs are being dominated by the high load of the Wi-Fi AP. It is striking that for 15 and 45 UEs, the gNB throughput increases for the multi-BWP gNB model are about the same. Overall, the multi-BWP model of the gNB is able to transmit all of the data



(a): The throughput distribution for a load of 15 UEs on the NR-U network.

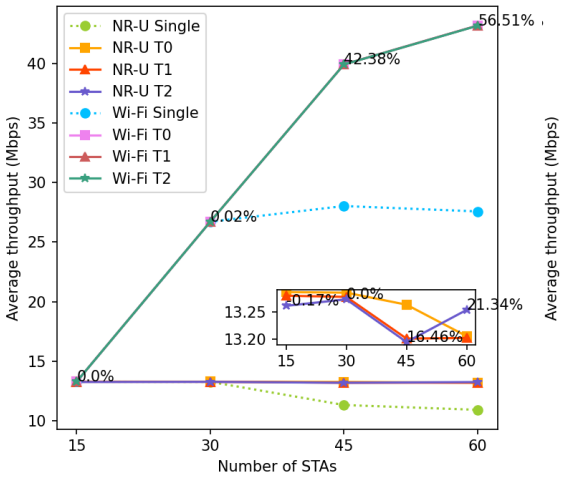


(b): The throughput distribution for a load of 30 UEs on the NR-U network.

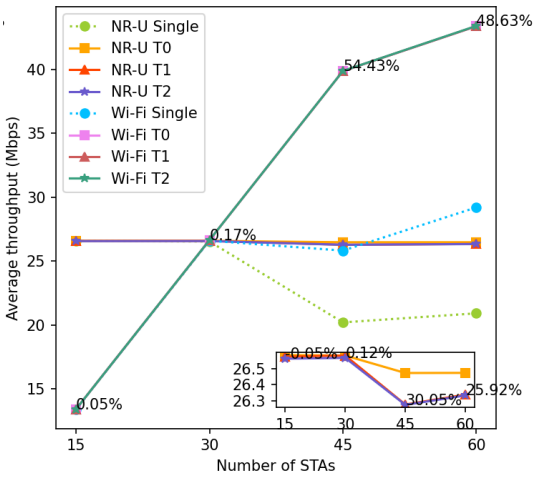


(c): The throughput distribution for a load of 45 UEs on the NR-U network.

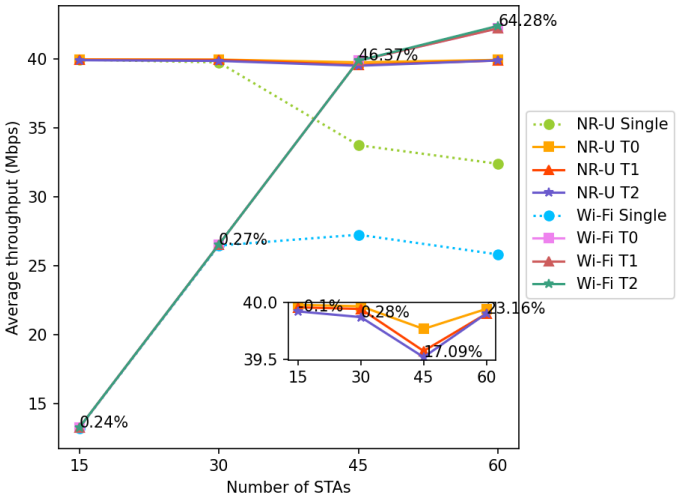
Figure 6.2: The throughput distribution of multiple simulation runs for constant number of 15, 30 and 45 UEs and an increasing number of STAs.



(a): The average throughput for a load of 15 UEs on the NR-U network.

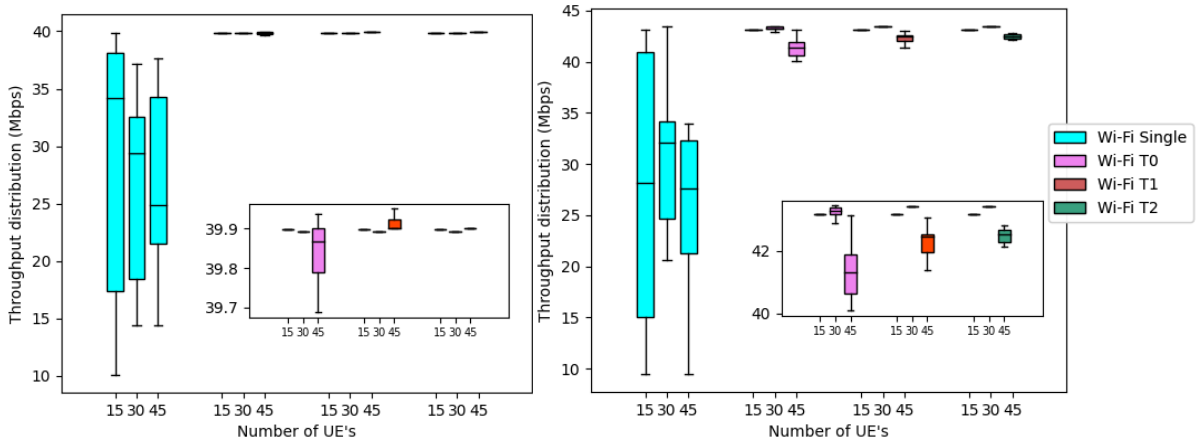


(b): The average throughput for a load of 30 UEs on the NR-U network.



(c): The average throughput for a load of 45 UEs on the NR-U network.

Figure 6.3: The average throughput of both the gNB and the AP over multiple simulation runs for a constant number of 15, 30 and 45 UEs and an increasing number of STAs. The percentage labels are added to the Wi-Fi T2 and NR-U T2 data sets and represent the improvement between their respective ”Single” dataset.



(a): Throughput distribution of the AP with a constant load of 45 STAs.

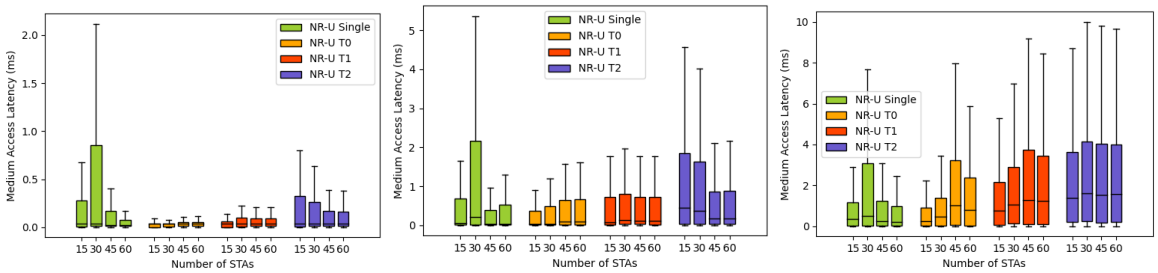
(b): Throughput distribution of the AP with a constant load of 60 STAs

Figure 6.4: Throughput distribution of the AP over multiple simulation runs with a constant number of 45 and 60 STAs and an increasing number of UEs.

from the traffic model, which was around 1 Mbps/user, while the single-BWP model of the gNB starts to decrease in throughput when the load on the AP becomes too high.

6.2.2 Impact on the Wi-Fi throughput

Even though the system model of Chapter 4 is designed for an improved throughput of the gNB, a larger increase can be observed in the Wi-Fi network. Figure 6.3 shows a maximum average increase of 64% for the AP when the gNB with the T2 delay model is implemented, compared to the implementation of the single-BWP gNB model. The same trend as the NR-U Single dataset can be noted for the throughput of the Wi-Fi Single dataset. When the number of STAs is at 45, the medium becomes too crowded and also the AP has to refrain. The main difference is that the throughput of the AP (sky blue dataset) stagnates starting from 30 STAs for an increasing number of its users, while the gNB (yellowgreen dataset) throughput still increases over an increasing number of its users. This is because only a number of UEs are starved by the contention of the medium with Wi-Fi, while all STAs of Wi-Fi have to content with the gNB. For this reason specifically, the AP reaches even higher throughput improvements when the multi-BWP gNB is implemented. The multi-BWP gNB now has more resource options to schedule the UEs that have a BWP overlapping with the Wi-Fi channel, allowing the Wi-Fi AP to use more resources on its channel, releasing the restrained throughput of all its users. However, if we look at the throughput distribution of the AP in Figure 6.4, we can see that the



(a): The MAL distribution with a constant load of 15 UEs on the NR-U network.

(b): The MAL distribution with a constant load of 30 UEs on the NR-U network.

(c): The MAL distribution with a constant load of 45 UEs on the NR-U network.

Figure 6.5: The MAL distribution of the gNB over multiple simulation runs with a constant number of 15, 30 and 45 UEs and an increasing number of STAs.

throughput of the AP under the single-BWP gNB model is not always limited. More specific, there were simulation runs in which the AP did not have a large increase in throughput comparing the multi-BWP and single-BWP gNB model. For 45 UEs, there is always an increase. The more UEs, the chances are that there are more UEs that have a BWP overlapping with the Wi-Fi channel. Therefore, you can see that the median of the throughput decreases for almost each dataset.

6.3 Impact on the latency

6.3.1 Impact on the NR-U Medium Access Latency

As explained in Chapter 5, the assumption is made that the channel quality is only influenced by the pathloss between each user and its basestation. No other status of the channel is tracked. Next to this, there is no receiver side implemented in the simulator, only data packets transmitted on the downlink from the perspective of the basestations. Therefore, the impact on the latency is derived from the MAL metric, which is the time between the packet arrival time (from the generated traffic model) and the start time of the packet transmission. The first thing that can be noticed is that the range of the latency increases from Figure 6.5a to Figure 6.5c, with an increasing load on the gNB. Meaning that for a higher load on the gNB, the average MAL increases for all gNB models. However, the distribution of the gNBs MAL has many extreme outliers, influencing the average for all cases. These outliers show that for a relatively low amount of packets, the MAL is high. See Appendix B for the boxplot with the average values included. These outliers are a relatively low amount

of packets, because only a limited number of UEs are being starved by the Wi-Fi network. Thus, the outliers are the packets that belong to the UEs that have a BWP overlapping with the Wi-Fi network. If we discard these extreme outliers and focus on the trend of the median values, which can be seen in Figure 6.6, it can be seen that the MAL, for all loads on the gNB, increases for the multi-BWP models with increasing delay times. The multi-BWP models perform worse than the single-BWP model in terms of latency and the NR-U T0 has a lower MAL than NR-U T1, and NR-U T1 a lower MAL than NR-U T2. However, the median of the MAL of the gNB focuses on its UEs that do not overlap with the Wi-Fi network. We now take a look at the average MAL of the gNB, shown in Figure 6.7, which takes into account the outliers of the dataset and represent the packets of the UEs that overlap with the Wi-Fi AP. It can now be instantly noted that the MAL has improved for the multi-BWP model compared to the single-BWP model. This is mainly because the MAL of the single-BWP model starts to significantly increase when the Wi-Fi network starts to drain the throughput of the gNB at a load of 45 STAs. However, even at a load of 30 STAs on the AP, the throughput was the same, but the latency seems to deteriorate. This shows us that even though the throughput does not increase at the AP load of 30 STAs, the multi-BWP model does have its latency improvements. The latency differences between the different multi-BWP are still the same between the median and average linechart of the MAL, where the T2 model has the highest MAL and T0 the lowest.

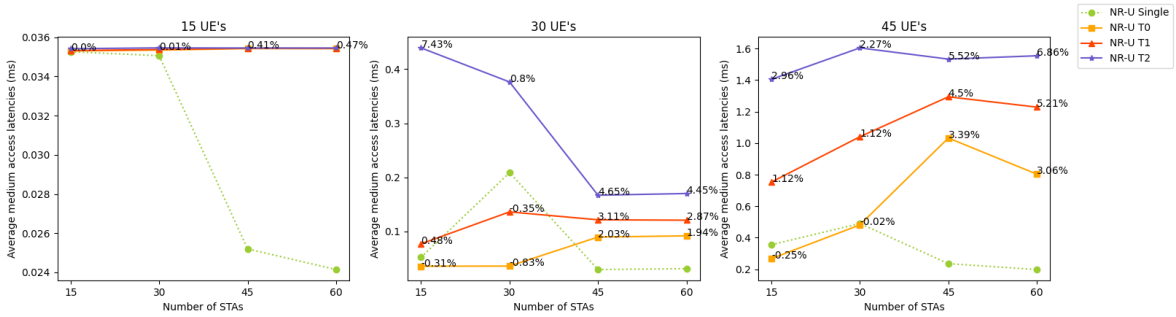


Figure 6.6: The median values of the MAL of the gNB over multiple simulation runs with a constant number of 15, 30 and 45 UEs and an increasing number of STAs.

6.3.2 Impact on the Wi-Fi Medium Access Latency

The distribution of the MAL of the Wi-Fi network shown in Figure 6.9 does not have any outliers, because all of the users packets are influenced when the load increases. Therefore, there is a larger number of Wi-Fi packets from which the latency deteriorates than the gNB. If we look at the Wi-Fi Single dataset, at a load

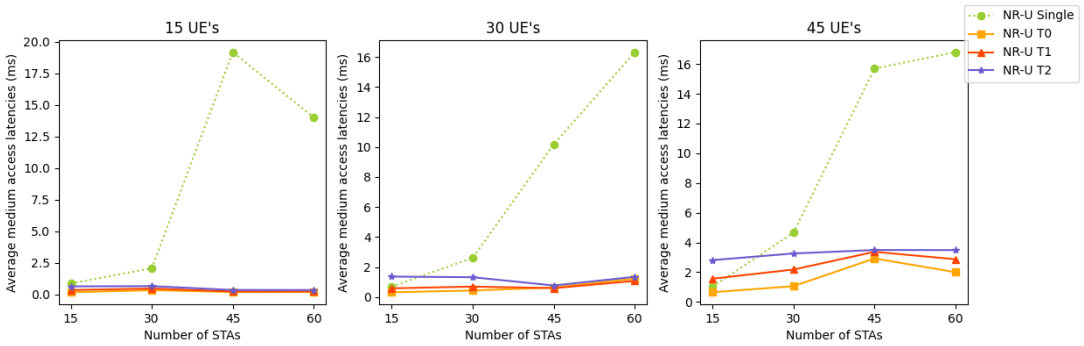


Figure 6.7: The average values of the MAL of the gNB over multiple simulation runs with a constant number of 15, 30 and 45 UEs and an increasing number of STAs.

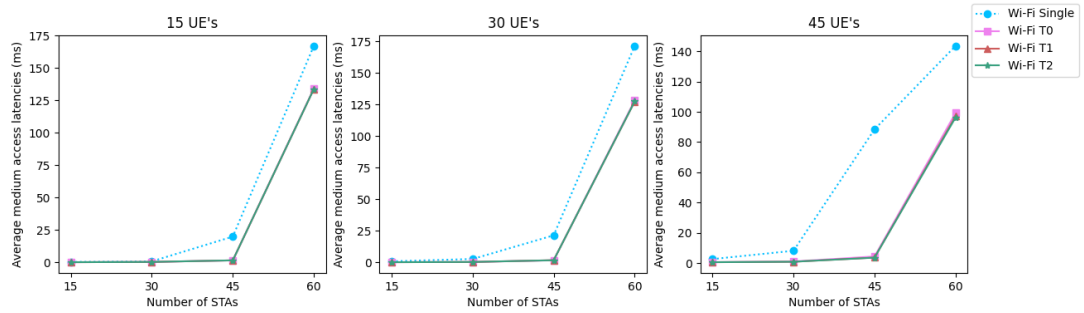


Figure 6.8: The average MAL of the Wi-Fi AP over multiple simulation runs with a constant number of 15, 30 and 45 UEs and an increasing number of STAs.

of 60 STAs, the network is clearly overloaded and the MAL reaches almost 175 ms for a load of 15 and 30 UEs on the gNB. At a load of 45 STAs, the latency also deteriorates, but not that significant. The Wi-Fi AP is not much influenced by the different delay types of the multi-BWP models, as can be seen that the Wi-Fi T0, Wi-Fi T1 and Wi-Fi T2 datasets are closely together. The Wi-Fi network does show some improvement in latency when the gNB implements any of the multi-BWP models, compared to the single-BWP gNB model. The largest improvement is observed at 45 UEs against 45 STAs. At 45 UEs, there are enough UEs that have a BWP overlapping with the Wi-Fi network, such that the Wi-Fi network also starts to show a deterioration in latency at a load of 30 STAs. Thus, the Wi-Fi AP does show an improved latency for a load of 30, 45 and 60 STAs, this improvement increases when the load of the gNB increases.

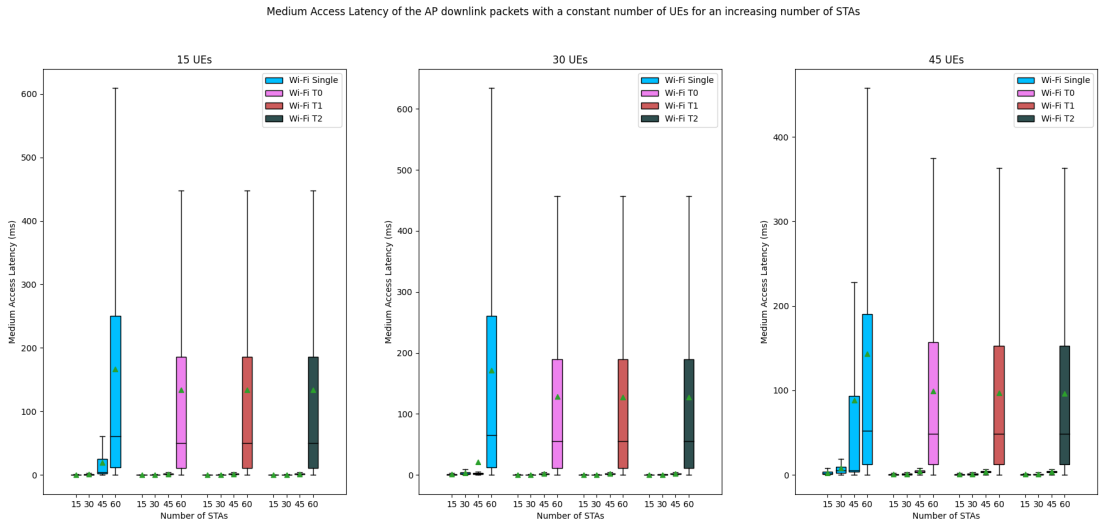


Figure 6.9: The distribution of the the MAL of the Wi-Fi AP over multiple simulation runs with a constant number of 15, 30 and 45 UEs and an increasing number of STAs.

6.4 Airtime utilization

The airtime utilization of a basestation gives us a percentage of the spectrum resources that have been used by that basestation in the simulation time. It gives us a better understanding of the load that is being handled by the basestation. The number of users and their traffic data is hard to translate to the actual load that the basestation has. Therefore, the airtime utilization tells us exactly how much of the spectrum is used by the basestation and how busy its downlink traffic is. Figure 6.11 shows the average airtime utilization of the Wi-Fi AP. This airtime utilization percentage is taken over 20 MHz of bandwidth (the AP bandwidth) and the total simulation time, which gives a clear understanding of the load on the bandwidth of the AP. All measurements are done with both the gNB and AP active, so the AP always has to content with the gNB for the spectrum. Figures 6.11a, 6.11b and 6.11c do not show much difference between each other, meaning that the load of the gNB does not impact the airtime utilization of the Wi-Fi AP. The airtime utilization of the AP linearly increases from 35% under 15 STAs to 65% under 30 STAs. Then, at a load of 45 STAs, the airtime utilization of the Wi-Fi AP under the single-BWP gNB model stops increasing and stagnates for an increasing number of STAs. The multi-BWP gNB models shows that the AP can obtain more spectrum resources for a higher load than 30 STAs on the AP. For 45 STAs, the airtime utilization is at 95%, showing that the same linear increase continues. At this point, the AP uses almost the entire spectrum. Note that a small percentage of airtime is required for the LBT procedure,

meaning that the load at 45 STAs is almost at its maximum. A small increase can be seen when increasing the load to 60 STAs, but this converges under 100%.
The average airtime utilization of the gNB can be seen in Figure 6.10. This airtime utilization percentage is taken over 100 MHz of bandwidth and the total simulation time. Since the bandwidth of the gNB is 5 times as large as the AP bandwidth, the airtime utilization on its own bandwidth is a lot less, and only reaches a maximum of 40% at 45 UEs. For an increasing number of load on the AP, the single-BWP gNB (green dataset) starts to slowly show a decrease in airtime utilization. The airtime of the gNB does not decrease rapidly, because most UEs are not impacted by the increasing airtime utilization of the AP, only the ones that have a BWP overlapping with the AP. For the UEs that do have to share the channel with the AP, the airtime utilization decreases on the single-BWP gNB. The multi-BWP gNB models show that they are able to keep up the airtime utilization at an increasing number of UEs, but show a small decrease for an increasing number of STAs. Comparing the different multi-BWP models, the T0 model (no delay) shows the highest airtime utilization on average for all load values, closely followed by the T1 model, which introduces small delays and the T2 (the highest delays) has an airtime utilization a few percentages under the T0 model.
Figure 6.12 shows the average airtime utilization of both the AP and gNB over a total bandwidth of 100 MHz (the bandwidth of the gNB). This gives a clear view of the relative airtime utilization between the two basestations. Compared to when the airtime utilization was taken over the basestations own bandwidth, the gNB now has a much higher airtime utilization than the AP, since the gNB uses many more resources of the measured time frequency spectrum (5 times the bandwidth of the AP). Only for a small load of 15 UEs, the AP overtakes the gNB in airtime utilization when it has double the load (30 STAs).

6.5 Number of BWP switches and caused overhead

In this section, we investigate the number of BWP switches that are performed by the UEs of the gNB. As a design choice explained in Section 5.0.3, only a selective number of UEs are allowed to switch between their BWPs, namely, the UEs that have their first active BWP overlapping with the Wi-Fi network. This is mainly because a significant decrease in Wi-Fi throughput was noticed on higher loads on the gNB when all UEs are able to switch from BWP. Since the multi-BWP model could also generate secondary BWPs for a number of UEs that overlap with the AP, these UEs are then allowed to switch to that channel whenever it is possible. This only increases the airtime of the gNB on that specific bandwidth, and decreases the airtime of Wi-Fi. Therefore, in the pre-processing phase, only the UE that have their

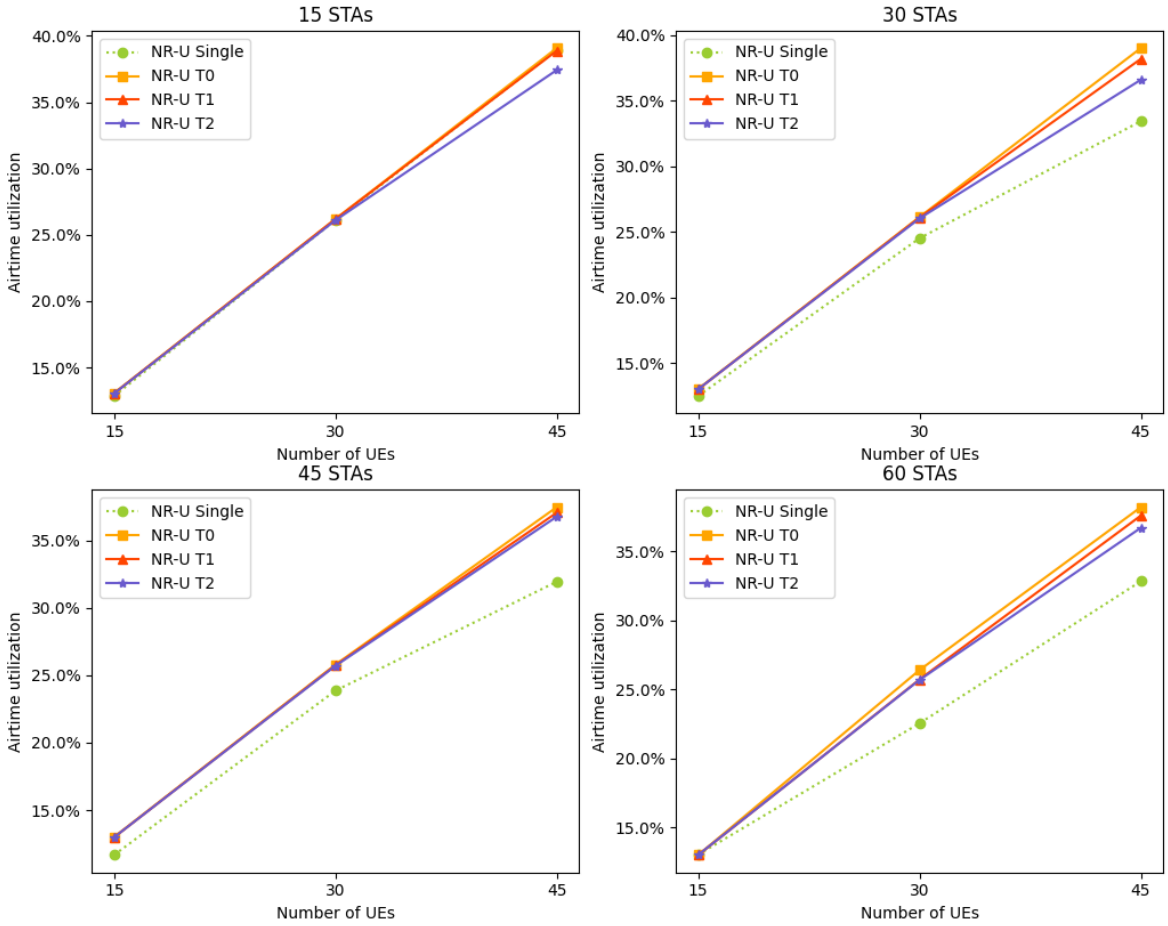
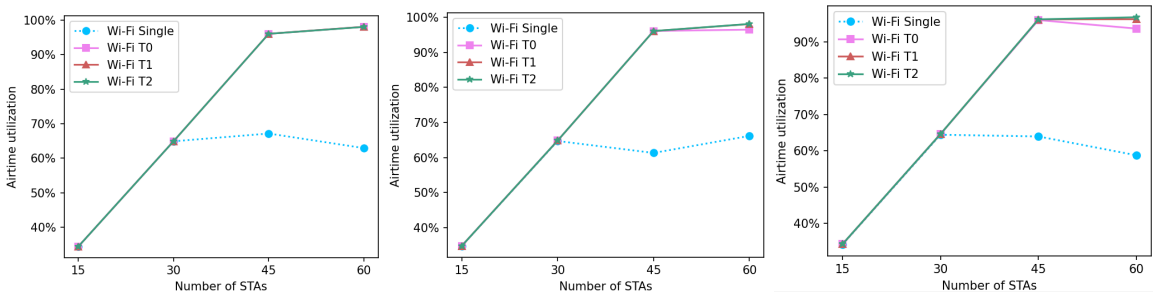


Figure 6.10: The average airtime utilization of the gNB for a constant number of 15, 30, 45 and 60 STAs and an increasing number of UEs.

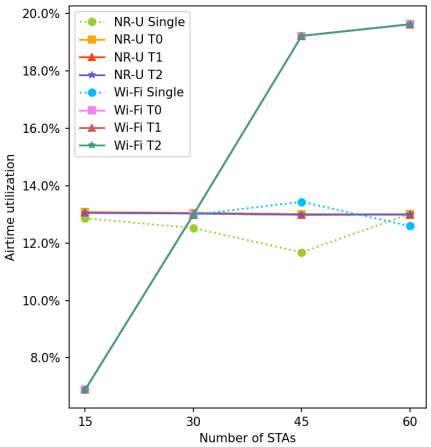


(a): The average airtime utilization for a constant load of 15 UEs on the NR-U network.

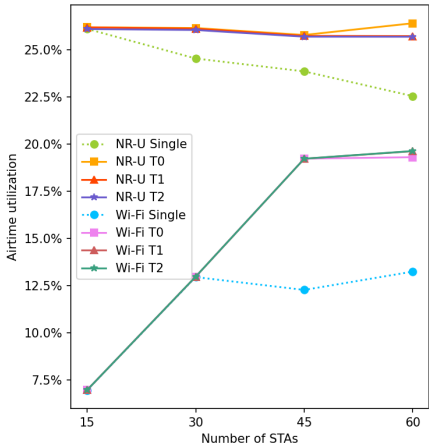
(b): The average airtime utilization for a constant load of 30 UEs on the NR-U network.

(c): The average airtime utilization for a constant load of 45 UEs on the NR-U network.

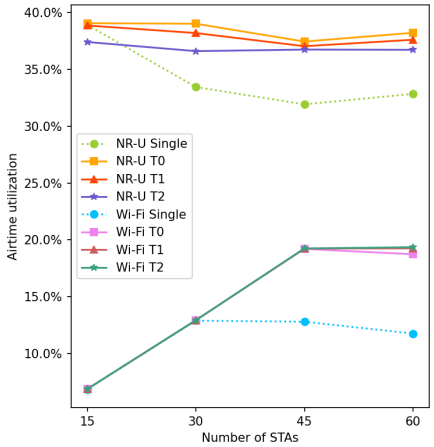
Figure 6.11: The average airtime utilization of the AP on 20 MHz bandwidth for a constant number of 15, 30 and 45 UEs and an increasing number of STAs.



(a): The average airtime utilization for a constant load of 15 UEs on the NR-U network.



(b): The average airtime utilization for a constant load of 30 UEs on the NR-U network.



(c): The average airtime utilization for a constant load of 45 UEs on the NR-U network.

Figure 6.12: The average airtime utilization of both the gNB and AP on 100 MHz of bandwidth over multiple simulation runs with a constant number of 15, 30 and 45 UEs and an increasing number of STAs.

first-active BWP overlapping with the Wi-Fi AP are allowed to switch to a secondary BWP that lies out of this frequency range. Figure 6.13 shows the distribution of the total number of BWP switches by the gNB of the different multi-BWP models. The general trend is that the number of switches is higher for the models with higher delay times. The range of the subfigures of Figure 6.13 also increases, since for a higher load on the gNB, it is obvious that the number of switches increases for all models since there are possibly more UEs that have a first-active BWP placed that overlaps with the Wi-Fi AP. The number of switches also seems to decrease for an increasing number of STAs in the AP for all models. This happens because the load on the AP increases, thus the airtime of the AP increases, which allows for less resources in that part of the spectrum for the gNB. Therefore, the UEs that have a BWP there, have less opportunity to use that BWP and do not switch to it. They will keep using their BWP that does not overlap with Wi-Fi for most of the time.
Figure 6.14 shows the percentage of overhead of the BWP switch time over the total airtime. This time, only the T1 and T2 model are shown, since the T0 model does not have any delay, thus no overhead. Overall, the T2 model has a much higher overhead than the T1 model. Especially for lower loads on the AP, the differences between the median of the T2 compared to the median of the T1 model can go up to 45%. For higher loads on the AP, these differences are closer together. Since in Figure 6.13, the range of the number of BWP switches increases for an increasing load on the gNB, the overhead also increases for an increasing load on the gNB. A maximum of 85% has been reached in a simulation run for a load of 45 UEs on the gNB and 15 STAs on the AP. Meaning that in total, the gNB spend 85% of the total airtime on BWP switching. This does not necessarily mean that this is 85% of unused transmission time, since UEs that need to switch from BWP can be scheduled after other BWPs at the end of a TXOP, as shown in Figure 4.6b.

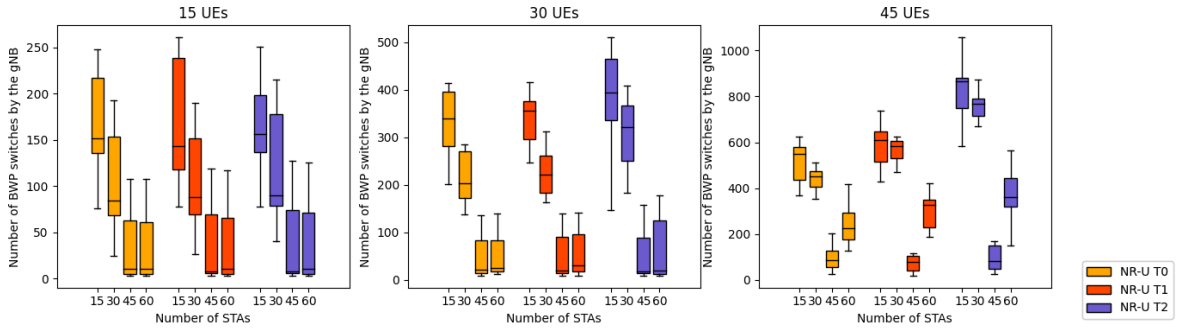


Figure 6.13: The distribution of the total number of BWP switches performed by the gNB for a constant number of 15, 30 and 45 UEs and an increasing number of STAs.

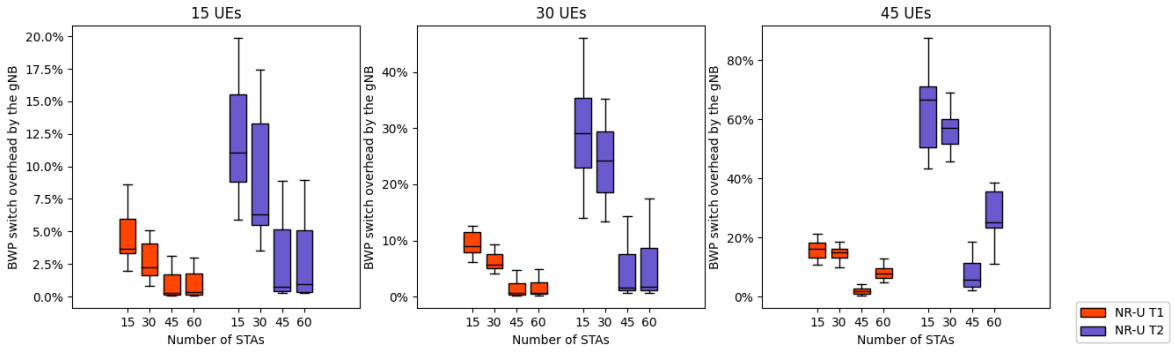


Figure 6.14: The distribution of the BWP switch overhead by the gNB for a constant number of 15, 30 and 45 UEs and an increasing number of STAs. A percentage of the total switch time of the gNB over the total airtime of the gNB.

Chapter 7

Discussion and Future Directions

State-of-the-art research does not discuss the use of BWPs in the unlicensed spectrum a lot. This made it very difficult to research the 5G NR feature as implementation in the unlicensed spectrum. On top of that, to the best of my knowledge, there are no simulators available that implement the BWP and BA features in a unlicensed coexistence scenario. For this specific requirement, a new simulator was designed and developed from scratch. With this design freedom, it was possible to choose the environment, the setup and configuration for both the gNB and AP and even the behaviour of them both.

The decision of making a simulator from scratch has many downsides. The data transmissions are very simplified, channel conditions are assumed to be perfect and transmissions always succeed. Next to this, the simulator is not very optimized, leading to relatively short simulation times. The gNB model of Chapter 4 was designed to be executed in a loop, where the pre-processing phase would be executed periodically, due to a possibly changing environment, including the Wi-Fi AP switching from channel. However, due to the high time complexity of the simulator code, only 1 second of simulation time was doable to perform with the many simulation runs that were required. For each simulation run, this allows us only to perform the pre-processing phase once with a 1 second of the scheduling phase.

The spectrum access method for the Wi-Fi network in the unlicensed bands are described in detail in literature. Wi-Fi is designed to work in the unlicensed spectrum and has an asynchronous way of transmitting data. The radio frame of the NR-U is designed to be synchronous and data is transmitted in structured time slots. There are several deployment modes and multiple forms of LBT procedures are described, however these asynchronous channel access methods require to be synchronized with these slots. For this thesis, we simplified the NR-U deployment to be standalone in the unlicensed spectrum and completely asynchronous without any requirement

of aligning the asynchronous transmissions with the start of a slot using the gap-based or reservation signal, as explained in [22]. As a solution, we round the number of symbols to the nearest slot duration to make up for this in airtime, which could be quite similar to the use of a reservation signal. Since Release 16 of 5G NR, NR-U has the option to decrease this slot-based scheduling all the way down to the granularity of a single OFDM symbol, which can start data transmissions almost immediately after the finish of LBT. As an alternative to the asynchronous and immediate start after LBT with a slot based airtime granularity used for the gNB in this thesis, the mini-slot can be utilized to provide a more realistic deployment and smaller airtime for the gNB data transmissions.

For BWP switching, a control plane is required for the DCI signalling between the gNB and a UE. For a standalone deployment, it means that this control plane is also implemented in the active BWP of a UE. Thus, for every BWP switch that a UE has to make, the gNB has to signal that UE on its current BWP to perform the switch, meaning the BWP first has to get access to the channel, creating possibly large delays. Therefore in this thesis, we considered the control plane between the gNB and each UE to be on a continuous and uninterrupted channel. However, the standalone and asynchronous behaviour in addition with the uninterrupted synchronous control plane could also be difficult to realise in a practical environment due to timing issues and should be investigated.

Both the gNB and AP are very oversimplified in terms of data transmission. The physical layer has been completely omitted, and the MAC layer is only partially and simply implemented, the focus is on the application layer only. For the Wi-Fi AP, 46 bytes have been added as the frame overhead for the MAC layer. For the NR-U gNB, no extra overheads are added due to the relatively larger airtime of the gNB compared to the AP. This is because the data scheduling granularity of the gNB is in terms of symbols, with varying lengths defined by the numerology, and rounded to the nearest slot granularity for the airtime of a transmission. While on the other hand, the AP does not have to use an entire symbol. Also, we used the same parameters for the channel access procedures. Normally, the two different RATs coexist under different installed parameters. Even though some papers describe the larger airtime utilization of NR-U in comparison with Wi-Fi due to this, it could still be a misrepresentation of a realistic implementation. Additional research must be performed to investigate the actual differences of airtime utilization between the gNB and AP when both the physical and MAC layers are implemented at both sides and when their corresponding channel access parameters are used.

Next to this, the relation between the required bandwidth and the users SLA requirements are also simplified. The BWP (numerology and bandwidth) of a user are defined based on its random generated SLA requirement, which create a rich combination of different BWPs over the UEs. However, the users are all given the same traffic model with a data rate of 1 Mbps. This results in all BWPs with a larger bandwidth to transmit data relatively fast and all BWPs with a lower bandwidth to transmit relatively slow. The traffic models of every UE should align with the UE's SLA requirement, to create a more realistic network scenario.

The number of switches performed by a UE can become quite large, considering that the numbers shown in Figure 6.13 are only on a total simulation time of 1 second. The BA feature is officially designed to save energy for a UE in the 5G NR network, by implementing timers that switch the UE to a BWP that uses less resources when low data rates are detected. However, by exploiting this feature in the unlicensed spectrum, does the feature still result in the savings of energy for a UE?

We take advantage of the flexible OFDMA technology used in the 5G NR physical layer to schedule numerous BWPs with different aspects for users with different requirements simultaneously in time and frequency. However, subcarriers with different numerologies are non-orthogonal to each other, and may interfere with each other, especially for those adjacent ones [32]. This is referred to as inter-numerology interference (INI) and can cause a performance degradation. Additional guard bands could limit the INI, but at the cost of spectral efficiency. What would the performance degradation impact be in terms of throughput and latency in the unlicensed spectrum for the NR-U network? And what techniques can be used to solve this?

The system model designed in Chapter 4 requires the gNB to perform LBT on the entire investigated spectrum when the gNB has data to transmit. In contrast to the research of Haghshenas et al. [3], where the gNB performs a extensive CAT3-LBT on a primary BWP, and a shorter CAT2-LBT on a secondary BWP when the primary BWP disallows access to the channel. This allows for a minimum energy consumption during the sensing procedure. The LBT procedure in this thesis is performed over a large bandwidth and can have serious negative impact on the energy consumption of the gNB and must be further investigated.

Chapter 8

Conclusions and recommendations

To answer the research question formulated in Chapter 1: ”Does the multi-BWP model enhance the coexistence of NR-U and Wi-Fi in the unlicensed bands?”, the three multi-BWP variants of the gNB from the system model of Chapter 4 have been compared against the single-BWP model. We will give direct answers per subquestion that we formulated in Chapter 1.

- To what extent is the throughput improved of the gNB in the multi-BWP model compared to the baseline model?

For an increasing load on the AP, the baseline model of the gNB will start to show degradation in throughput. At 30 STAs on the AP, the load, measured in airtime utilization, of the AP is equal to 65% and the gNB throughput is maintained. At 45 STAs on the AP, the airtime utilization of the AP is at 95% and the throughput of the gNB starts to deteriorate. For an increasing number of UEs on the gNB, there are on average more first-active BWPs that overlap with the Wi-Fi network that are being limited in throughput, which shows the deterioration in throughput for the single-BWP gNB. However, when we look at the throughput distribution of the single-BWP gNB for the highest tested load in Figure 6.2c, we can see that it is sometimes able to maintain its throughput and was able to grasp more resources than the Wi-Fi network on that channel, meaning it dominated the channel contention over Wi-Fi for a number of cases. Then when the number of STAs increases to 60 STAs, the distribution drops again. We can conclude from this, that the throughput in the multi-BWP gNB models is mostly improved when the Wi-Fi AP dominates over the UEs in the single-BWP gNB models that have a first-active BWP that overlaps the AP channel. This improvements is on average at its maximum of 30% when the AP and gNB have an average airtime utilization of around 60% and 24%, at 45 STAs and 30 UEs respectively. The throughput differences between the different variants of the multi-BWP models are negligible, meaning that the introduced delays do not impact the throughput.

- What is the impact on the latency of the gNB in the multi-BWP model compared to the baseline model?

If we consider the median values of the MAL distribution shown in Figure 6.5 as results, we focus on the impact of the MAL for the UEs that do not have a BWP overlapping with the Wi-Fi AP. Figure 6.7 shows us that for those users, the latency deteriorates for all of the multi-BWP gNB models. The largest deterioration is observed in the T2 model and has a median difference of maximum 7.43% in comparison with the single-BWP model. Especially for higher loads on the gNB, the latency values increases for the UEs that do not have to share the channel with the AP for higher delay multi-BWP models.
However, if we consider the average values of the MAL distribution shown in Figure 6.7, we also consider the differences of the MALs of the packets of the UEs that do have to share their channel with the Wi-Fi AP. Because of their channel access uncertainty, they have huge delays, which is only shown lightly by the average MAL of all UEs of the gNB. Because of this heavy impact, the overall latency of the gNB is much improved when comparing the multi-BWP model against the single-BWP model. As expected, the T0 model without any BWP switch delay performs best, followed by the T1 model and at last the T2 model.

- What is the impact on the throughput of the AP in the multi-BWP model compared to the baseline model?

It is very striking that the improvement of the average throughput of the Wi-Fi AP exceeds that of the gNB. As shown in Figure 6.3, the AP has an average throughput improvement of up to 64% when the multi-BWP gNB is deployed in comparison with the single-BWP gNB. This significant improvement is due to all STAs being fit in the small 20 MHz bandwidth of the AP which are all being refrained in throughput in coexistence with NR-U. The multi-BWP gNB model is able to offload the UEs with a BWP overlapping on the APs channel to BWPs that are not overlapping with the AP, giving more resources for the AP and all of its refrained STAs. At 45 STAs on the AP and almost uninterrupted by NR-U, the average airtime utilization of the AP is at 95%. At 60 STAs, the airtime utilization is at 98%, nearing its maximum channel capacity, but never reaches it due to time lost in sensing. Thus, the multi-BWP gNB models allow the Wi-Fi network to maintain its throughput at these high load values.

- What is the impact on the latency of the AP in the multi-BWP model compared to the baseline model?

The MAL values on the AP can become quite large when there is a high load on the Wi-Fi network. The AP under the single-BWP gNB model has normal latency

values until the 45 STAs load threshold, where the airtime utilization uninterrupted is at 95%. At this point, the MAL is starting to get worse. On the other hand, the multi-BWP models of the gNB are able to maintain a normal MAL at this load level. At 60 STAs and 98% uninterrupted airtime utilization for the AP, both the single-BWP and multi-BWP deployment of the gNB can not prevent the huge deterioration of the MAL of the AP. For a larger load on the gNB, the increase in latency starts at lower loads on the AP, but tends to not reach the highest observed MAL. This is because the higher number of UEs on the APs channel causes the latency deterioration at lower AP loads. This is also where the biggest improvement is for the AP under the multi-BWP gNB model. Even though the AP under both the single-BWP and multi-BWP models underwent large deterioration, the multi-BWP does always show improvements compared to the single-BWP model after the 45 STAs threshold.

To conclude the research and give answer to our main research question, we can state that the multi-BWP gNB model does enhance the coexistence of the NR-U gNB and the Wi-Fi AP for this specific scenario, but mostly for higher loads on the AP. Since the load of the gNB is distributed over a higher bandwidth, the problems arise when the load on the AP becomes too large. In terms of throughput, the Wi-Fi network benefits most of the multi-BWP model, since the entire network is impacted by the coexistence due to its smaller bandwidth size and unavailability to offload data on other channels. The UEs that have a first-active BWP on the channel of the AP also benefit from the possibility to choose between multiple BWPs, this improvement can be seen back in the overall throughput results of the gNB and can go in average up to 30%. The latency in terms of the MAL is getting slightly worse for the UEs that do not share the unlicensed channel with Wi-Fi and can be up to 7% higher for the multi-BWP model. At most, this adds up 1.6 ms to the latency. However, the improvement caused by the multi-BWP model of the even larger delays of the UEs that share their channel with Wi-Fi balances the latency deterioration of the other UEs, showing an average improvement in latency on the gNB.

Bibliography

[1] M. Hirzallah, M. Krunz, B. Kecicioglu, and B. Hamzeh, "5g new radio unlicensed: Challenges and evaluation," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 3, pp. 689–701, 2020.

[2] F. Luo, X. Sun, Y. Gao, W. Zhan, P. Liu, and Z. Guo, "Optimal coexistence of nr-u with wi-fi under 3gpp fairness constraint," in *ICC 2022-IEEE International Conference on Communications*. IEEE, 2022, pp. 4890–4895.

[3] M. Haghshenas and M. Magarini, "Nr-u and wi-fi coexistence enhancement exploiting multiple bandwidth parts assignment," in *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2022, pp. 260–263.

[4] F. Abinader, A. Marcano, K. Schober, R. Nurminen, T. Henttonen, H. Onozawa, and E. Virtej, "Impact of bandwidth part (bwp) switching on 5g nr system performance," in *2019 IEEE 2nd 5G World Forum (5GWF)*. IEEE, 2019, pp. 161–166.

[5] " 5G; NR; Requirements for support of radio resource management ," ETSI, Tech. Rep. Version 16.4.0 Release 16, 08 2020, 3GPP TS 38.133 version 16.4.0 Release 16.

[6] H. Yin and S. Alamouti, "Ofdma: A broadband wireless access technology," in *2006 IEEE sarnoff symposium*. IEEE, 2006, pp. 1–4.

[7] V. Ramaswamy, J. T. Correia, and D. Swain-Walsh, "Analytical evaluation of bandwidth part adaptation in 5g new radio," in *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2021, pp. 985–990.

[8] A. Khlass, D. Laselva, and R. Jarvela, "On the flexible and performance-enhanced radio resource control for 5g nr networks," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, 2019, pp. 1–6.

[9] " 5G; NR; Radio Resource Control (RRC); Protocol specification ," ETSI, Tech. Rep. Version 16.1.0 Release 16, 07 2020, 3GPP TS 38.331 version 16.1.0 Release 16.

[10] H. Arslan *et al.*, "Flexible multi-numerology systems for 5g new radio," *Journal of Mobile Multimedia*, 2018.

[11] " 5G; NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone ," ETSI, Tech. Rep. Version 16.5.0 Release 16, 11 2020, 3GPP TS 38.101-1.

[12] " 5G; NR; Multiplexing and channel coding ," ETSI, Tech. Rep. Version 15.3.0 Release 15, 10 2018, 3GPP TS 38.212 version 15.3.0 Release 15.

[13] X. Lin, D. Yu, and H. Wiemann, "A primer on bandwidth parts in 5g new radio," *5G and Beyond: Fundamentals and Standards*, pp. 357–370, 2021.

[14] " 5G; NR; Physical layer procedures for data ," ETSI, Tech. Rep. Version 15.3.0 Release 15, 07 2020, 3GPP TS 38.214 version 16.2.0 Release 16.

[15] B. Chen, J. Chen, Y. Gao, and J. Zhang, "Coexistence of lte-laa and wi-fi on 5 ghz with corresponding deployment scenarios: A survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 7–32, 2016.

[16] "Ieee standard for information technology–telecommunications and information exchange between systems - local and metropolitan area networks–specific requirements - part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications," *IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016)*, pp. 1–4379, 2021.

[17] N. Patriciello, S. Lagen, B. Bojović, and L. Giupponi, "Nr-u and ieee 802.11 technologies coexistence in unlicensed mmwave spectrum: Models and evaluation," *IEEE access*, vol. 8, pp. 71 254–71 271, 2020.

[18] B. Bojović, L. Giupponi, Z. Ali, and M. Miozzo, "Evaluating unlicensed lte technologies: Laa vs lte-u," *IEEE access*, vol. 7, pp. 89 714–89 751, 2019.

[19] " LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures ," ETSI, Tech. Rep. Version 14.2.0 Release 14, 04 2017, 3GPP TS 36.213 version 14.2.0 Release 14.

[20] Y. Huang, Y. Chen, Y. T. Hou, W. Lou, and J. H. Reed, "Recent advances of lte/wifi coexistence in unlicensed spectrum," *IEEE Network*, vol. 32, no. 2, pp. 107–113, 2017.

[21] M. Zajac and S. Szott, "Resolving 5g nr-u contention for gap-based channel access in shared sub-7 ghz bands," *IEEE Access*, vol. 10, pp. 4031–4047, 2022.

[22] K. Kosek-Szott, A. L. Valvo, S. Szott, P. Gallo, and I. Tinnirello, "Downlink channel access performance of nr-u: Impact of numerology and mini-slots on coexistence with wi-fi in the 5 ghz band," *Computer Networks*, vol. 195, p. 108188, 2021.

[23] Y. Kakkad, D. K. Patel, S. Kavaiya, S. Sun, and M. López-Benítez, "Optimal 3gpp fairness parameters in 5g nr unlicensed (nr-u) and wifi coexistence," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 4, pp. 5373–5377, 2022.

[24] "ETSI TR 138 901: 5G; Study on channel model for frequencies from 0.5 to 100 GHz," 3GPP, Tech. Rep. Version 16.1.0 Release 16 TR 38.901, 11 2020.

[25] "ETSI EN 301 893: 5 GHz RLAN; Harmonised Standard covering the essential requirements of article 3.2 of Directive 2014/53/EU," European Telecommunications Standards Institute, white paper V2.1.1, 05 2017.

[26] M. U. Khan, A. García-Armada, and J. Escudero-Garzás, "Service-based network dimensioning for 5g networks assisted by real data," *IEEE Access*, vol. 8, pp. 129 193–129 212, 2020.

[27] " 5G; NR; User Equipment (UE) radio access capabilities ," ETSI, Tech. Rep. Version 17.0.0 Release 17, 05 2022, 3GPP TS 38.306 version 17.0.0 Release 17.

[28] G. Dósa and J. Sgall, "Optimal analysis of best fit bin packing," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2014, pp. 429–441.

[29] D. S. Johnson, "Near-optimal bin packing algorithms," Ph.D. dissertation, Massachusetts Institute of Technology, 1973.

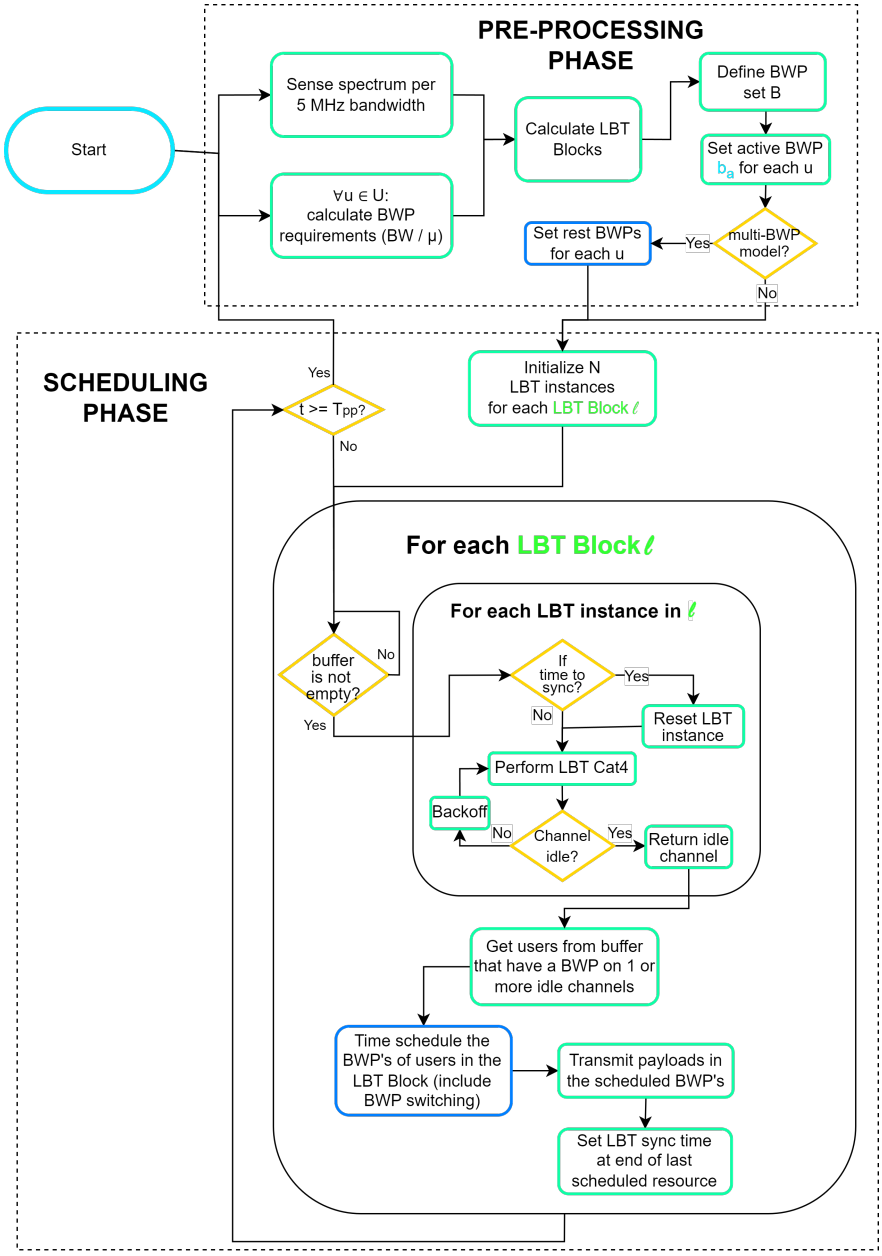
[30] "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on 3D channel model for LTE (Release 12)," 3GPP, Tech. Rep. V12.7.0, 12 2017.

[31] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5g usage scenarios and traffic models," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 905–929, 2020.

[32] X. Zhang, L. Zhang, P. Xiao, D. Ma, J. Wei, and Y. Xin, “Mixed numerologies interference analysis and inter-numerology interference cancellation for windowed ofdm systems,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 8, pp. 7047–7061, 2018.

Appendix A

5G NR-U gNB model



Appendix B

Medium Access Latency distribution of the gNB with average values included