List of Abbreviations

BLER	Block Error Rate
BS	Base Station
CSI	Channel State Information
DL	Downlink
EXP/PF	Exponential/Proportional Fair
FNBW	First Null Beam Width
GoB	Grid of Beams
GoP	Group of Pictures
HMD	Head Mounted Display
HOL	Head Of Line
LoS	Line-of-Sight
MI-ESM	Mutual Information Effective SINR Mapping
M-LWDF	Maximum-Largest Weighted Delay First
MR	Maximum Ratio
MRC	Maximum Ratio Combining
OLLA	Outer Loop Link Adaptation
PF	Proportional Fair
PRB	Physical Resource Block
RS	Reference Signal
RX	Receiver
SINR	Signal-to-Interference-plus-Noise Ratio
TB	Transport Block
TBS	Transport Block Size
TDD	Time Division Duplex
TTI	Transmission Time Interval
TX	Transmitter
UE	User Equipment
UL	Uplink
URA	Uniform Rectangular Array

List of Symbols

Latin alphabet

not perfectly alphabetically ordered *yet*

a_ϕ	azimuth lower limit for GoB
a_θ	elevation lower limit for GoB
b_ϕ	azimuth upper limit for GoB
b_θ	elevation lower limit for GoB
B	bandwidth
$BLER_0$	target BLER
C_t	coordinates of centre of the table
C_m	coordinates of centre of mass of the room
d_f	distance to the front of the user for camera placement
d_F	Fraunhofer distance
d_s	distance to the side of the user for camera placement
d_{out}	distance outwards from the head centre for HMD antenna offset
d_{up}	distance upwards from the head centre for HMD antenna offset
D	largest dimension of radiator to estimate effective area
E_b	energy per bit
$\boldsymbol{H}_{bul}$	channel matrix between BS b and UE u in layer l
I	interference
j	imaginary unit $\left(j = \sqrt{-1}\right)$
k_B	Boltzmann constant
L_{max}	maximum latency for the radio link
N_0	noise power spectral density
N_r	number of receive antennas
N_t	number of transmit antennas
N_{users}	number users or participants
N_{phy}	number of physical users
N_{vir}	number of virtual users
N_{CSI}	number of beams with CSI-RS
N_{bs}	number of BSs
N_{ue}	number of UEs
N_{cam}	number of cameras
N_x	number of antennas along the x-dimension
N_y	number of antennas along the y-dimension
N_{slots}^{DL}	number of DL slots in a transmission period
N_{slots}^{UL}	number of UL slots in a transmission period

N_{slots}^{TDD}	number slots in a transmission period
N_{slots}^{CSI}	number slots between CSI updates
N_{slots}^{SCH}	number slots between scheduling updates
N_{ant}^{UE}	number of antennas on the UE side
N_{ant}^{BS}	number of antennas on the BS side
NF_{BS}	BS noise figure
NF_{UE}	UE noise figure
NF_r	noise figure at the receiver
$N_{PRB,bu}$	number of PRBs allocated for link between BS b and UE u
$N_{PRB,bul}$	number of PRBs allocated for link between BS b and UE u in layer l
N_{symb}^{PRB}	number of symbols per PRB
N_{bits}^{symb}	number of bits per symbol
$N_{infobits}^{symb}$	number of information bits per symbol
N_{bits}^{slot}	number of bits per slot
N_{bits}^{PRB}	number of bits per PRB
$N_{bits,bul}$	number of bits to be sent between BS b and UE u in layer l
p	scheduling priority
P_N	noise power
P_u	position of user u
P_r	received power
P_s	signal power
P_t	transmit power
$P_{t,max}^{UE}$	maximum transmit power per UE
$P_{t,max}^{BS}$	maximum transmit power per BS
$P_{r,bu}^{UE}$	received power by the UE for a link between BS b and UE u
$P_{t,bu}^{BS}$	transmit power at the BS for a link between BS b and UE u
$P_{t,bu}^{UE}$	transmit power at the UE for a link between BS b and UE u
$P_{t,bul}$	transmit power for a link between BS b and UE u in layer l
$P_{t,bul}^{UE}$	transmit power at the UE for a link between BS b and UE u in layer l
$P_{t,bul}^{BS}$	transmit power at the BS for a link between BS b and UE u in layer l
P_{IaCI}	interference power from intra-cell interference
P_{IeCI}	interference power from inter-cell interference
P_{ILI}	interference power from inter-layer interference
Q_m	modulation order
r_ϕ	azimuth resolution for GoB
r_θ	elevation resolution for GoB
r_t	radius of table
r_u	radius of user circumference of disposition
$r_{DL/UL}$	ratio between DL and UL application throughputs
$r_{P/I}$	ratio between P-frame and I-frame sizes
R_b	instantaneous bit rate
$\hat{R}_b$	estimated bit rate
R_c	code rate
R_{packet}	packet arrival rate
R	instantaneous throughput
$\overline{R}$	average application throughput
$\overline{R}_{DL}$	average application throughput in the DL
$\overline{R}_{UL}$	average application throughput in the UL

R_F	frame rate
s	speed of user head position change
s_{TDD}	TDD split
$SINR_{eff}$	effective SINR, i.e. aggregated over all scheduled PRBs
$SINR_i$	SINR of the i-th PRB
S_I	size of I-frame
S_P	size of P-frame
S_{packet}	size of a packet
S_{GoP}	size of a GoP
S_{TB}	size of a TB
$S_{TB,max}$	maximum size of a TB
T	noise temperature
T_{sim}	simulation duration
$T_{slot,\mu}$	slot duration for numerology μ
T_{rot}	interval between consecutive orientations when head rotating
$\boldsymbol{w}$	vector of beamforming weights
$\boldsymbol{w}_{\phi,\theta}$	beamforming weights from a GoB with direction (ϕ, θ)
$\boldsymbol{w}_{i,j}$	beamforming weights vector with grid indices (i, j), from a GoB
$\boldsymbol{w}_{bu}^{BS}$	beamforming weights between BS panel b and UE u, at the BS
$\boldsymbol{w}_{bu}^{UE}$	beamforming weights between BS panel b and UE u, at the UE
$\boldsymbol{w}_{t,bul}$	transmit weights vector between BS panel b and UE u, in layer l
$\boldsymbol{w}_{r,bul}$	receive weights vector between BS panel b and UE u, in layer l

Greek alphabet

α_P	power compensation factor for UL power control
α_u	angle from the centre of the table to user u
β_x	upper limit on uniform distribution for rotation around x axis
β_y	upper limit on uniform distribution for rotation around y axis
β_z	upper limit on uniform distribution for rotation around z axis
γ	burstiness parameter for application traffic
γ_{OLLA}	step size for OLLA parameter (Δ_{OLLA}) update
Δ_{OLLA}	outer loop link adaptation step
η_{OH}	efficiency due to overhead
η_{slot}	efficiency in bit rate from slot format
η_{CSI}	efficiency from CSI acquisition
λ	wavelength
o	overlap parameter for application traffic
σ_x	standard deviation of normal distribution for position coordinate x
σ_y	standard deviation of normal distribution for position coordinate y
σ_z	standard deviation of normal distribution for position coordinate z
τ_{CSI}	CSIs delay, in number of TTIs
τ_{ACK}	delay before acknowledgement, in number of TTIs

Sets:

$\mathbb{N}$	natural numbers
$\mathbb{N}_0$	natural numbers including zero
$\mathcal{B}$	base station panels
$\mathcal{U}_b$	users served by base station b
$\mathcal{L}_{bu}$	layers scheduled between base station b and user equipment u
$\mathcal{F}$	frequencies to simulate

Other nomenclature

$\boldsymbol{A}$	matrix
$\boldsymbol{a}$	column vector
$\lvert\boldsymbol{a}\rvert$	euclidean norm of vector $\boldsymbol{a}$
$\boldsymbol{A}^{\mathsf{T}}$	transpose of $\boldsymbol{A}$
$\boldsymbol{A}^{\mathsf{H}}$	Hermitian of $\boldsymbol{A}$, also know as the transpose conjugate of $\boldsymbol{A}$
$\lceil a \rceil$	ceil a, i.e. round up a to the nearest integer
$\lfloor a \rfloor$	floor a, i.e. round down a to the nearest integer

1

Methodology

Contents

1.1 Radio Access Network

In this section, we detail the functions executed by the network equipment to enable data transmission. The network equipment needs to acquire CSI and manage resources accordingly to cope with the incoming application traffic and fulfil service requirements. Firstly, we go over important considerations and assumptions, namely regarding multi-layer transmissions, concentrating on the Downlink (DL), among other miscellaneous but relevant matters. Then we list the steps required to simulate a Transmission Time Interval (TTI) and all processing associated with making the right choices when transmitting and receiving. We summarise these steps with a flowchart and proceed to detail each one.

Firstly, we opt for a Grid of Beams (GoB)-based beamforming approach. With the growing number of antennas at the receivers, full channel knowledge is practically unobtainable, and we need to resort to more overhead-efficient approaches.

Secondly, we address the considerations regarding multi-layer transmission. To reiterate, the difference between single-layer and multi-layer operation is the number of independent streams transmitted per User Equipment (UE). And to transmit independent streams or layers, there must be some orthogonality mechanism that renders such layers independent. The orthogonality domain we are concern with is orthogonality in space. However, by opting GoB-based beamforming although we save in overhead, we loose considerably in transmission flexibility. Free-format beamforming would allow us to send independent layers in the same direction, only focusing different antennas at the reception. But using a GoB we do not have enough beams to do that, instead we have to resort to completely different propagation paths, with paths beyond the first not being the Line-of-Sight (LoS). This would not yield insignificant improvements.

Another option would be to resort to polarisation orthogonality. Instead of using all antenna elements to perform a transmission, we may use the antennas oriented in a given direction to send one layer and the elements oriented perpendicularly to send another. Note that the same beam in the GoB can be used for the different polarisations when they are to be sent over the same path. However, the moments in time where inter-polarisation interference is small, e.g. less than 20 dB, are rare. In other words, often antennas with a given orientation at the receiver get signal from both polarisations at the transmitter. Thus, it would require considerably more complicated interference estimations algorithms to do multi-layer transmissions polarisation-based. This is why we opt for single-layer transmission using all antennas both in the Transmitter (TX) and in the Receiver (RX). Nevertheless, the vast majority of modelling in this section is agnostic to the number of layers.

Thirdly, although in Section **??** we modelled the location of all UEs in the system and application traffic for UL and DL, for conciseness in this section we describe the model for DL transmission procedure and thus we do not consider cameras. In the DL, he number of UEs N_{ue} equal to the number of physical users N_{phy}.

In essence, we need to list all procedures that can happen in a TTI. Some may not happen every TTI and we need to state in what circumstances they do happen. See in Figure 1.1 a flowchart of the main steps required to simulate a (DL) TTI.

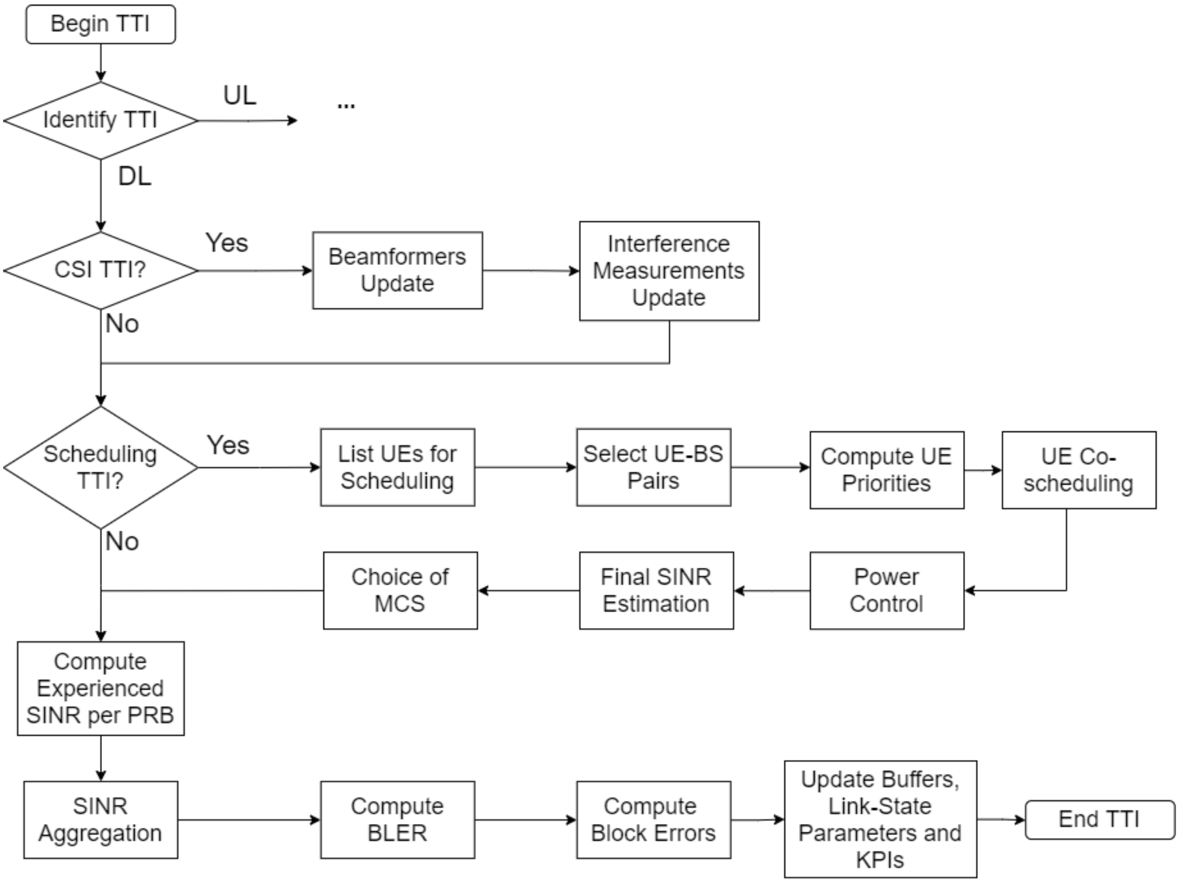

Figure 1.1: Flowchart for of simulation steps for each TTI.

Several verifications are made to decide whether some procedures should take place. The first is to identify the nature of the current TTI - UL and DL TTIs have different steps. The second is checking whether CSI should be updated. Thirdly, it is to verify whether the current user scheduling information for that TTI is to be updated. Only after those verifications and respective procedures, the transmissions scheduled for the present TTI are processed.

The very first step is assessing the nature of the TTI depends on the slot-structure and TDD split.

TDD Split and Slot Format

We recognise two options. The first is to use self-contained slots, at a cost of about $2/14 \approx 14\%$ lower bitrate since 2 out of 14 symbols are used for guard and control, but having the benefit of feedback about block errors in the same TTI, thus allowing triggering retransmissions of the lost information the next TTI. This way the likelihood of packet dropping due to transgressions of time constrains is reduced since latency is reduced, leading to more opportunities to transmit the data on time. The second is simpler and more throughput-efficient, at the cost of latency performance. It consists on using slots that only have DL/Uplink (UL) symbols, respectively, and we ignore the guard time in the transition slot.

Therefore our definition of UL/DL split, or Time Division Duplex (TDD) split depends on the option. Let us define s_{TDD} as the ratio between UL and DL slots. We represent this ratio as $N_{slots}^{DL} : N_{slots}^{UL}$, e.g. 4:1, meaning that for each UL slot there are 4 DL slots. The slot structure is completely defined by the number of slots in a transmission period N_{slots}^{TDD}.

In order to optionally change between both options, we introduce a transport block acknowledgement delay τ_{ACK} (in TTIs) and a slot efficiency η_{slot}. The acknowledgement delay is the number of TTIs before the transmitter receives the acknowledgement, thus $\tau_{ACK} + 1$ is the number of TTIs until the erroneous transport block can be transmitted again. With self-contained slots, $\tau_{ACK} = 0$. Without self-contained slots it depends on the s_{TDD}.

The slot efficiency $\eta_{slot} = 0.86$ in the example of self-contained slots where 14% of symbols are not used for data, and $\eta_{slot} = 0$ in the DL/UL heavy slots. It is applied to the instantaneous throughput R as $R_{modified} = R\eta_{slot}$

As mentioned, we solely present, and posteriorly evaluate, modelling for DL TTIs. Thus after making the distinction between TTIs, the next step is to update the CSI information based on our beamforming strategy. Therefore, let us first state how the GoB is created.

Grid of Beams

To create a GoB we need to know which directions to steer the beam. The beam-steering directions are all possible combinations of values in the azimuthal and elevation angular domains, relative to the antenna boresight (direction perpendicular to the plane the antenna array is inserted). And to create a beam grid in one such domain, one simple way is to use the resolution and the values of the extremes. We define in Equation (1.1) an interpolation function to perform the operation of creating a set of values from a to b, given b strictly greater than a, with intervals of resolution r.

$$F_I(a, b, r) = \{a + i \times r \; \forall \; i \in \mathbb{N}_0 : i \times r \leq b - a\} \tag{1.1}$$

This way, we define in the azimuthal angular domain as $\mathcal{A}_\phi = F_I(a_\phi, b_\phi, r_\phi)$ and the elevation angular domain as $\mathcal{A}_\theta = F_I(a_\theta, b_\theta, r_\theta)$. For instance, if the antenna is positioned in the centre of the room, on the ceiling, pointing downwards, then the most logical approach is a symmetric approach because in that position the coverage of the room would be uniform since we consider our room with equal length and width(for room and user behaviour modelling, see Section **??**). More concretely, the GoB should cover all positions the UEs may potentially be. Thus, given the position and movement of the users in relation to the size of the room described in the example of Section **??**, choosing the lower limits to $a_\phi = a_\theta = -60°$and the upper limits to $b_\phi = b_\theta = 60°$covers all possible UE positions.

The resolutions should depend on the array size. To create a pseudo-non-interfering GoB, where the maximum of the main lobe of one beam points at the a minimum of an adjacent beam, the resolution should be roughly half the First Null Beam Width (FNBW). It is 'pseudo-non-interfering' because the FNBW varies with the direction at which the beam is steered, which causes the maximums to not align perfectly with the nulls. This effect is unnoticeable in adjacent beams, and gets more noticeable the more far apart beams are from each other. So, this method is a simplistic yet effective approach to minimise the interference between beams, but it does not eliminate this interference.

Thus, the possible directions are defined as a cartesian product between the azimuthal and elevation domains, shown in Equation (1.2).

$$\mathcal{D} = \mathcal{A}_\phi \times \mathcal{A}_\theta = \{(\phi, \theta) : \; \phi \in \mathcal{A}_\phi, \; \theta \in \mathcal{A}_\theta\} \tag{1.2}$$

Having the directions, we need the precoder that will construct a beam pointing in that direction. In Equation (1.3) we define the M by N beamforming matrix $\boldsymbol{W}_{\phi,\theta}$ that contains the relative amplitudes and phases that are applied to the signal of each antenna element of an M by N planar array, obtaining as a result a beam directed to ϕ degrees on the horizontal plane and θ degrees on the vertical plane. Note that such planes depend on the orientation of the array and the angles ϕ and θ are null in the interception of both planes, corresponding to the direction orthogonal to the array plane (see Appendix **??** for a complete derivation).

$$\boldsymbol{W}_{\phi,\theta} = \begin{bmatrix} 1 & u_2 & \dots & u_2^{(N-1)} \\ u_1 & u_1 u_1 & \dots & u_1 u_2^{(N-1)} \\ \vdots & \vdots & \ddots & \vdots \\ u_1^{(M-1)} & u_1^{(M-1)} u_2 & \dots & u_1^{(M-1)} u_2^{(N-1)} \end{bmatrix}, \text{ with } \begin{cases} u_1 = e^{-j\pi \sin(\phi)\sin(\theta)} \\ u_2 = e^{-j\pi \cos(\phi)\sin(\theta)} \end{cases} \tag{1.3}$$

Subsequently, to obtain every precoder in the GoB we need to build a precoding matrix for each direction in $\mathcal{D}$. Let us define in Equation (1.4) the set $\mathcal{W}$ containing all precoders $\boldsymbol{W}_{\phi,\theta}$ in the GoB, formed for an M by N Uniform Rectangular Array (URA).

$$\mathcal{W}^{\text{GoB}} = \{\boldsymbol{W}_{\phi,\theta} : (\phi, \theta) \in \mathcal{D}\} \tag{1.4}$$

Figure 1.2 illustrates the result of a cut at zero degrees elevation on beams of two grids. The two grids are built for square antenna arrays, with 16 and 1024 elements, respectively, left and right sides of the figure, hence the noticeably different directivity. Using the 3GPP-defined elements in [1], the maximum directivities are 20 dBi and 38 dBi, respectively, for the 16-element array and for the 1024-element array. Furthermore, since the resolutions were purposely set to match half of the FNBW, the grid on the left spans 120 ° of angular domain, from -60 ° to 60 °, with steps of 30 °, while the grid on the right does so with a resolution of 4 °. In total, this equates to 25 distinct beams of the small array and 961 beams in the larger array.

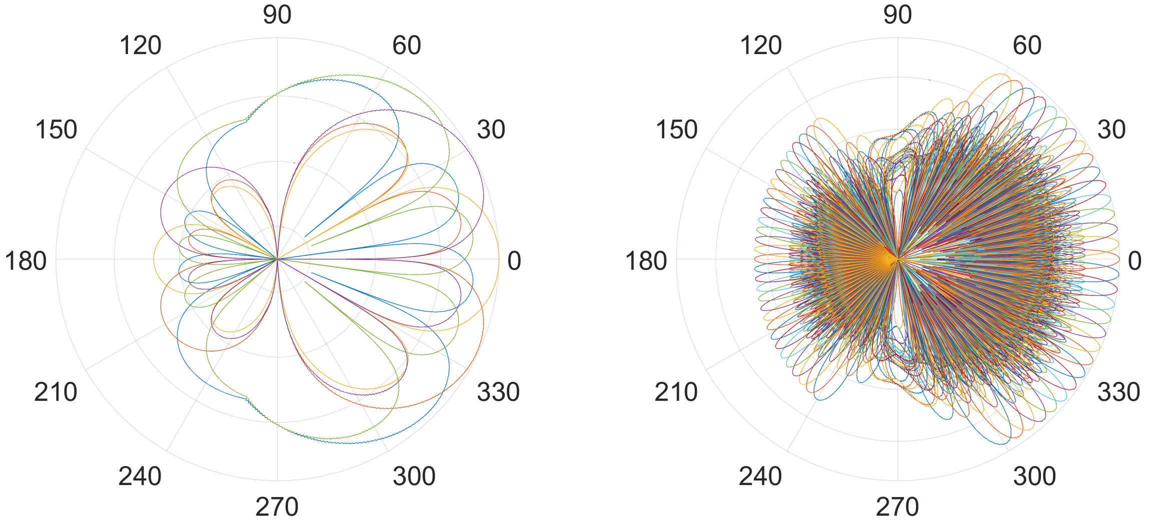

Figure 1.2: GoB azimuth cuts for 4 by 4 (left) and a 32 by 32 (right) antenna element array.

1.1.1 Channel State Information

CSI updates happen every N_{slots}^{CSI} slots or TTIs. Not all TTIs have CSI updates because the channel does not change enough to be worth updating that frequently, and due to prohibitive overheads since reference signals are sent in place of data.

The overheads associated with different Channel State Information (CSI) feedback schemes is not modelled. The overhead would depend on the type and quality of said measurements, thus we simply define a CSI-slot efficiency η_{CSI} meant to reduce the bit rate of CSI slots.

CSI is required for operation of two important mechanisms. First, to direct and receive signals optimally, in accordance with the paths where attenuation is lower. As such, it is used to update matching beamformers at the transmitter and receiver, or beam pairs. Second, to assess received power and interference, which are crucial to estimate channel quality, which is then used to, e.g. determine which MCS to use.

Beam pairs Update

To update the best beam pairs between UEs and BS panels, the BS should transmits N_{CSI} beamformed CSI-RSs and the UE reports how well it received each RS. However, this would require a mechanism for the BS to identify, based on previous channel measurements, which beams are more likely to best serve the UE. As such, instead we check all beams in the GoB to assess which best suit the channel. Then we keep received power information about N_{CSI} of them.

The best beam pairs are chosen to maximise the channel gain achieved from performing a transmission with a given GoB beam, with a best effort reception using MRC. Therefore, for a link between UE u and BS panel b, the beamformer on the BS side $\boldsymbol{w}_{bu}^{BS}$ is always a beam-steering vector from the GoB, i.e. $\boldsymbol{w}_{bu}^{BS} \in \mathcal{W}_{b}^{GoB}$. And the UE-side beamformer $\boldsymbol{w}_{bu}^{UE}$ is always the Maximum Ratio (MR) beamformer that fits perfectly the use of the BS beamformer and the channel $\boldsymbol{h}_{bu}$, given in Equation 1.5. Thus, $\boldsymbol{w}_{bu}^{BS}$ is $N_{ant}^{BS} \times 1$, $\boldsymbol{w}_{bu}^{UE}$ is $1 \times N_{ant}^{UE}$, and $\boldsymbol{H}_{bu}$ is $N_{ant}^{UE} \times N_{ant}^{BS}$ where N_{ant}^{UE} and N_{ant}^{BS} are the number of single-polarised antenna elements at the UE and BS panel antenna arrays, respectively.

$$\boldsymbol{w}_{bu}^{UE} = \frac{\left(\boldsymbol{H}_{bu} \cdot \boldsymbol{w}_{bu}^{BS}\right)^{H}}{|\boldsymbol{H}_{bu} \cdot \boldsymbol{w}_{bu}^{BS}|} \tag{1.5}$$

When $\boldsymbol{w}^{UE}$ is a MR beamformer, the channel gain under transmit and receive beamforming is a real number. Therefore, to choose the $\boldsymbol{w}$ that achieves the highest gain, we simply have to choose $\boldsymbol{w}$ that results in highest norm of its internal product with the channel. This shortcut is represented in (1.6). In essence, this means that because we are computing the UE-side beamformer already taking into account the transmit-side beamformer, to maximise the norm of the received signal it is sufficient to choose the appropriately the transmit-side beamformer.

$$\boldsymbol{w}_{bu}^{BS} = \underset{\boldsymbol{w} \,\in\, \mathcal{W}_b^{GoB}}{\operatorname{argmax}} \left|\boldsymbol{w}_{bu}^{UE} \cdot \boldsymbol{H}_{bu} \cdot \boldsymbol{w}\right| = \underset{\boldsymbol{w} \,\in\, \mathcal{W}_b^{GoB}}{\operatorname{argmax}} \left|\boldsymbol{H}_{bu} \cdot \boldsymbol{w}\right| \tag{1.6}$$

When $N_{CSI} > 1$, instead of the best beamformer, we save the N_{CSI} best GoB beamformers. For sake of practicality, let us assume $N_{CSI} = 1$ for now on. Furthermore, beam pairs computed in this way profit from beam-reciprocity, i.e. the beams used for receiving can be used for transmitting as well. And doing this way, the received power is already present (see Equation (1.7)), thus we only need to update the interference now.

$$P_{r,bu}^{UE} = P_{t,bu}^{BS} \left|\boldsymbol{w}_{bu}^{UE} \cdot \boldsymbol{H}_{bu} \cdot \boldsymbol{w}_{bu}^{BS}\right|^2 \tag{1.7}$$

Interference Measurements Update

To measure interference, the Base Station (BS) should schedule an empty UE-specific RS for interference measurements. It should result in measuring the power received by the interfering sources. The main drawback is the outdatedness of the measurement since in takes around 4 TTIs until the information is available. Therefore, when the interference measurement is available, it refers to e.g. 4 TTIs ago.

To simulate this process, we report the experienced interference from transmission that occurred τ_{TTI} TTIs back.

A major disadvantage of estimating the interference in this manner comes from the fact that the experienced interference is extremely dependent on current scheduling. If the scheduled UEs or beamformers in use change, then it is expected a major change in the experienced interference to take place, thus possibly rendering the measurement completely invalid. We foresee precise interference estimation algorithms, perhaps driven by learning mechanisms, to be a future direction of work. We discuss this matter further in Section **??**.

Before continuing to the update of the scheduling information, it is worth introducing two steps used more than once by our model of the network equipment, the link adaptation mechanism and the instantaneous bit rate computation from an SINR.

Link Adaptation

An intuitively important step is to choose which MCS to use for transmission. Since choosing one too high leads to only errors and one too low wastes resources, this choice must be on par with the channel as often as possible. One smart mechanism to achieve this adapts the MCS choice according with block errors.

We called it the Outer Loop Link Adaptation (OLLA) mechanism and it is UE-specific. Every time a MCS is estimated, it is adjusted with the OLLA parameter as a last step. Each OLLA parameter Δ_{OLLA} is initialised at zero and is updated in every TTI the given UE is scheduled. When a Transport Block (TB) scheduled to/from is successfully transmitted, the OLLA parameter is updated according with Equation (1.8), but if the block has errors, Equation (1.9) is used instead.

$$\Delta_{OLLA} = \Delta_{OLLA} + BLER_0 \times \gamma_{OLLA} \tag{1.8}$$

$$\Delta_{OLLA} = \Delta_{OLLA} - (1 - BLER_0) \times \gamma_{OLLA} \tag{1.9}$$

Observe the subtlety of the asymmetry in update. The term that multiplies the step size γ_{OLLA} is much bigger in (1.9) than in (1.8), since $BLER_0$ is usually 0.1 or smaller. It is a defensive approach, to take bigger steps towards more conservative MCSs when there are errors because it is always better to have some bitrate than no bitrate. Contrarily, the progression to increasing the MCS is slower.

The OLLA parameter adjusts the MCS choice by flipping an appropriately biased coin and adding either $\lfloor \Delta_{OLLA} \rfloor$ or $\lceil \Delta_{OLLA} \rceil$ to the CQI index estimated in the previous step. An appropriately biased coin in this situation is a coin that selects to round down the OLLA parameter with a probability of $\lceil \Delta_{OLLA} \rceil - \Delta_{OLLA}$. This makes sense because Δ_{OLLA} is decreased when a block has errors, thus making more likely that the MCS is reduced when the link has worse quality than expected. When the block does not have errors, it makes it more likely to increase the MCS estimate, such that a good link condition can be taken advantage of to increase the bitrate. Note that this formulation still works as supposed for negative values, i.e. the OLLA mechanism works for increasing and decreasing the MCS.

Instantaneous Bit rate

To compute the instantaneous bit rate allows us to quantify the value of serving each user. Then we can weigh options against each other regarding fairness, maximum aggregated throughput, or likelihood of fulfilling latency constraints. Therefore, computing estimated and realised bit rate is fundamental to the operation of this wireless system.

The SINR is used to choose the MCS from the MCS curves represented in Figure 1.3, and equations in Appendix **??**. The point at which each MCS curve intercepts the BLER probability of 10% is marked, and the MCS that corresponds to each CQI index is in Table **??**. The MCS choice consists on selecting the first MCS that achieves a lower percentage of block errors than the Block Error Rate (BLER) target $BLER_0$ is chosen. Usual values for $BLER_0$ are 10% or lower.

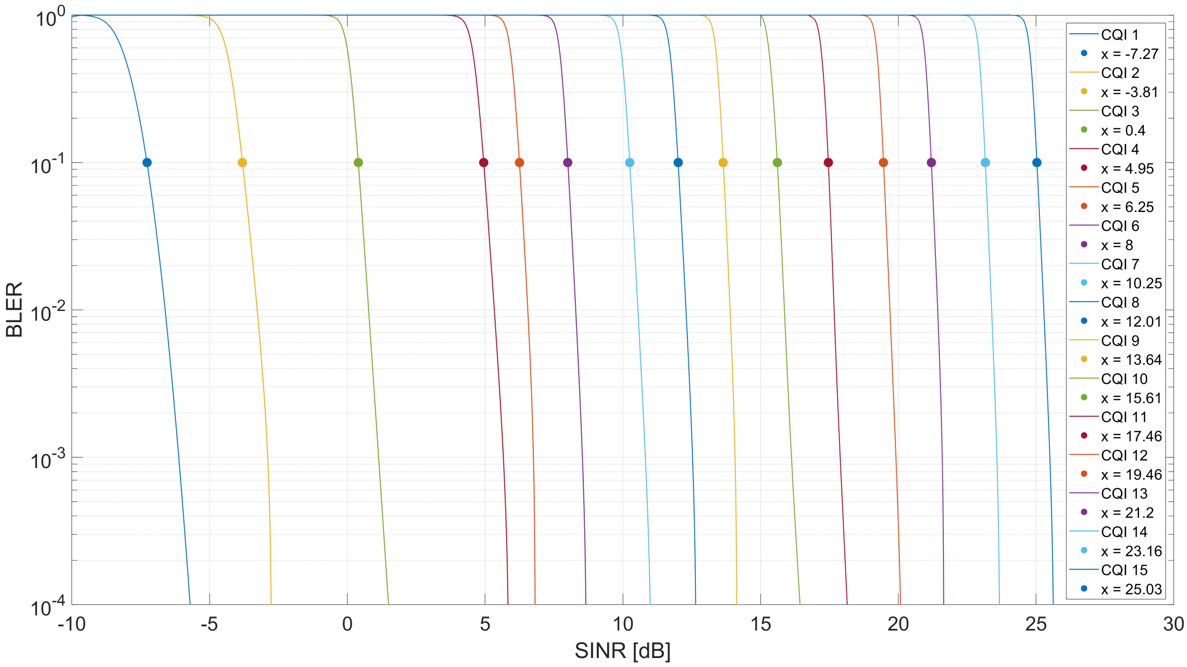

Figure 1.3: BLER curves for all MCSs. Simulated with Vienna Link-Level Simulator [**?**].

The selected MCS is then adjusted with the link adaptation parameter as described in Subsection 1.1.1. The resultant MCS tells us the number of bits in a symbol N_{bits}^{symb}, which can be computed from the modulation order M, through $\log_2(M)$. To compute the bits per PRB we assume all symbols/REs in a PBR are used for data, i.e. $N_{symb}^{PRB} = 168$, we multiply by N_{bits}^{symb} and account for the code rate that makes further reduces the quantity of data transmitted. In essence, Equation (1.10) computes the bit rate by dividing the number of bits transmitted in a PRB by the duration of that PRB, which corresponds to a slot duration $T_{slot,\mu}$, that depends on the numerology μ.

$$R_b = \frac{N_{bits}^{symb} \times N_{symb}^{PRB} \times R_c}{T_{slot,\mu}} \tag{1.10}$$

To counterbalance the excessive assumptions, like assuming all symbols are used for data, we account for all bit rate reductions, namely due to overheads and self-contained slots and channel state information, respectively, by multiplying the efficiencies η_{OH}, η_{slot} and η_{CSI}. Equation (1.11) has the final bit rate efficiency η. An estimation for the instantaneous throughput R is $R = R_b \times \eta$.

$$\eta = \eta_{OH} \times \eta_{slot} \times \eta_{CSI} \tag{1.11}$$

Scheduler

A scheduler task is to attribute priorities to UEs according to a trade-off of resource sharing fairness, achieving the maximum instantaneous aggregated throughput, or attain the lower average latencies, to name a few. These priorities allow us to select UEs by order of importance according to the weighted trade-off relation we choose.

The most common and widely used scheduler is the Proportional Fair (PF), presented in (1.12). PF weights the instantaneous R and average $\overline{R}$ throughputs to balance immediate reward and fairness across users. The average $\overline{R}$ is computed across a t_w TTI window - Equation (1.13) shows how it is updated as long as $t \geq t_w$, otherwise, $t_w = t$ is used with the same expression.

$$p = \frac{R}{\overline{R}} \tag{1.12}$$

$$\overline{R}(t) = \left(1 - \frac{1}{t_w}\right)\overline{R}(t-1) + \frac{1}{t_w}R(t-1) \tag{1.13}$$

As seen, PF does not consider latencies. Yet, for our case where each user has the same amount of transmitted and received data, the PF acts as a latency-aware scheduler by weighting fairness heavily, not leaving any user waiting for long. However, it may not perform as well as latency-aware alternatives.

Two latency-aware alternatives are Maximum-Largest Weighted Delay First (M-LWDF) [2] and Exponential/Proportional Fair (EXP/PF) [3]. The first is almost as simple as the PF, only weighting the Head Of Line (HOL) latency as well. EXP/PF is more complex and considers a maximum delay and increases priorities exponentially as latencies approach the limit. Both use the PF ratio described in (1.12). M-LWDF outperforms EXP/PF in practically every scenario [2], unless when there is serious dispute for resources. Therefore, both M-LWDF and EXP/PF may prove to be good options in the future.

1.1.2 User Scheduling

Analogous to the CSI update procedure, the scheduling information is only updated every N_{slots}^{SCH} TTIs. An update on the scheduling information consists on refreshing which UEs are considered for scheduling and which BS panels are used. This choice is simple: only UEs with non-empty buffers are examined to be part of the scheduled UEs list; and each UE is served by the BS panel with the best beam pair to that UE. Then a more complex procedure takes place, the derivation of the MCS for transmission to certain UEs and in specific time-frequency resources.

MCS derivation, along with UE scheduling and resource allocation can be summarised in some steps.

SINR Estimation

The received powers for the best N_{CSI} beams have been reported in the CSI acquisition step, as well as the experienced interference, although both are outdated. Also, the channel gain can be derived directly knowing the transmit power that was used. Thus, we assume an equal distribution of the maximum transmit power at the BS $P_{t,max}^{BS}$ over the number of scheduled UEs with non-empty buffers. And the only missing piece in the SINR expression is the noise.

We use wideband scheduling, i.e. allocating all available spectrum to every transmission, relying in spatial separation to prevent excessive interference. Therefore, assuming B to be the system bandwidth, using thermal noise we would get a noise power P_N given by Equation (1.14), with the Boltzmann constant $k_B = 1.380649 \times 10^{-23}$ J/K, the noise temperature T and an upscaling with the receivers' noise figure NF_r. All hardware imperfections are abstracted by considering noise figures in the BS and in the UEs, respectively, NF_{BS} and NF_{UE}, in dB.

$$P_N = k_B T B \times 10^{\frac{NF_r}{10}} \tag{1.14}$$

To summarise, the expression used for SINR estimation uses information from τ_{CSI} TTIs ago on the received power P_s and total interference I. See Equation 1.15

$$SI\hat{N}R_{eff} = \frac{\hat{P}_s}{\hat{I} + P_N} \tag{1.15}$$

Compute UE Priorities with Scheduler

UE priorities are computed based on the estimated instantaneous bitrate, the average bitrate across time, and, depending on the scheduler, head of queue delay and other metrics. Based on latency requirements, in an attempt to have a fair and optimal system, a positive scalar priority is given to each user. See Section 1.1.1 for details on how each scheduler works. Having options between schedulers allows us to know which makes the best decisions based on the achieved performance.

Co-schedule users

This step lists the users to be scheduled together until the next update to the schedule. The co-scheduling rule for a single-BS-panel operation is to add one UE layer at a time to the list, by order of UE priority (computed in the previous step), if the best beams used for those layers are compatible with the previously added UE layers. And we define as compatible beams when the BS-side beam, belonging to the GoB is at least than κ beams apart, with $\kappa \in \mathbb{N}_0$. If κ is 0, then the all layers are accepted. If $\kappa = 1$, then the beams must be different - adjacent beams have a distance of 1, so are still used together. Beams located diagonally adjacent of the GoB are considered to have a distance of 2, hence they may be co-scheduled when $\kappa \leq 2$. Figure 1.4 illustrates the beams that cannot be co-scheduled with certain values of κ, representing in filled circles as incompatible beams, and empty circles as compatible beams with the central orange beam. More generally, the beam distance is defined by the sum of absolute differences of the beam indices in the grid. Mathematically, the beamformers $\boldsymbol{w}_{i,j}$ and $\boldsymbol{w}'_{i',j'}$, having (i, j) and (i', j') as the GoB indices, respectively, are compatible if $|i - i'| + |j - j'| \geq \kappa$.

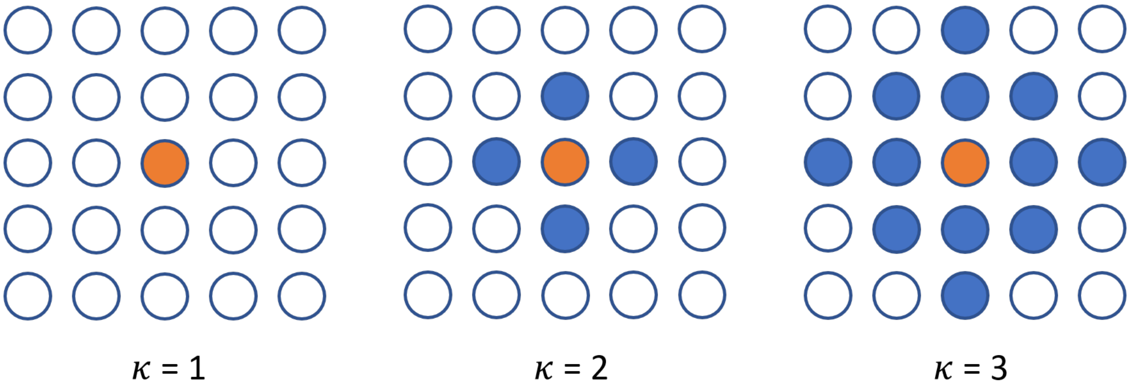

Figure 1.4: Beam co-scheduling incompatibility distance.

In this step there is space for more elaborate algorithms that attempt to choose different combinations of the best N_{CSI} beams of each user, in an attempt to maximise the metrics we care about. Of course, if scheduling one more user considerably reduces the quality of the channel to many others, it is likely not worth doing.

Power Control

Depending on the beamforming strategy, it may be necessary to scale down all precoders due to excessive power per antenna constraints. This, however, does not apply to our case because beam-steering beamformers always have uniform amplitude. The power control process in the downlink is as simple as distributing the maximum total transmit power equally amongst the scheduled UEs.

Final MCS

By now the transmit powers to each user is more accurately defined and the beam pairs are also fixed. Therefore, we obtain an estimation of the SINR for each user, and the MCS to be used for each transmission.

This step concludes the required scheduling procedure. Next follows all computations to simulate a transmission. But, much like previously, we need to introduce two mechanisms in detail before approaching the transmission section. These are an SINR framework to compute the realised SINR in each PRB and an SINR aggregation algorithm to convert vectors of SINRs per PRBs to an effective SINR that characterises the quality of the transmission.

Multi-layer SINR Framework with Beamforming

Although the rest of the network equipment modelling has minor simplifications for single-layer transmission, for future purposes it is derived a multi-layer framework.

To accurately compute the SINR experienced during a transmission, for each scheduled UE and accounting for different channel responses in each PRB of the assigned bandwidth, it is crucial to know what is the power received from any transmitter.

Let l be a layer connecting a set of antennas in a BS b to a set of antennas in a UE u, with $P_{t,l}$ the transmit power and $P_{r,l}$ is the received power in that layer. Then, let $P_{r,ll'}$ be the power received by the layer l receiver using the combiner $\boldsymbol{w}_{r,l}$, transmitted by the layer l' transmitter using the precoder $\boldsymbol{w}_{t,l'}$, with $\boldsymbol{H}_{ll'}$ the matrix that connects the receiver and the transmitter. Equation (1.16) shows these quantities are related.

$$P_{r,ll'} = P_{t,l'} \left|\boldsymbol{w}_{r,l} \cdot \boldsymbol{H}_{ll'} \cdot \boldsymbol{w}_{t,l'}\right|^2 \tag{1.16}$$

The powers are scalars, $\boldsymbol{w}_{r,l}$ is a $1 \times N_r$ vector, $\boldsymbol{w}_{t,l'}$ is $N_t \times 1$ vector and $\boldsymbol{H}_{ll'}$ is a $N_r \times N_t$ matrix, where N_t and N_r are the number of antenna elements at the transmitter and receiver antenna arrays, respectively.

Knowing how to calculate this quantity we can compute the powers of all parts of the SINR expression on a PRB-basis: the signal P_s, the intra-cell interference P_{IaCI}, the inter-cell interference P_{IeCI}, the inter-layer interference P_{ILI} and the noise P_N. Equation (1.17) the expression for the SINR in a given PRB. Subsequently we present equations for each quantity in the SINR expression, along with the rationale behind them.

$$SINR = \frac{P_s}{P_{ILI} + P_{IaCI} + P_{IeCI} + P_N} \tag{1.17}$$

Moreover, and to reiterate, all quantities mentioned in this section are time (TTI) and frequency (PRB) specific. These SINRs need to be posteriorly aggregated in an effective SINR for each transmission in the given TTI. We choose drop the i index to simplify notation, as we did with the TTI since this chapter is TTI-specific.

The received signal power P_s is in Equation (1.18).

$$P_s = P_{r,ll} \tag{1.18}$$

In case of multi-layer transmission, other layers scheduled to/from the same UE may interfere among themselves and the power of inter-layer interference P_{ILI} accounts for this interference by summing the interferences caused in layer l by every other layer l' scheduled between BS b and UE u. See Equation (1.19), where $\mathcal{L}_{bu}$ is the set of layers scheduled between BS b and UE u.

$$P_{ILI} = \sum_{\substack{l' \in \mathcal{L}_{bu} \\ l' \neq l}} P_{r,ll'} \tag{1.19}$$

Interference power contributions from the same cell/BS come from every transmission that takes place to other UEs in the same cell/served by the same BS. See Equation (1.20), where $\mathcal{U}_b$ is the set of users served by BS b.

$$P_{IaLI} = \sum_{\substack{u' \in \mathcal{U}_b \\ u' \neq u}} \sum_{l' \in \mathcal{L}_{bu'}} P_{r,ll'} \tag{1.20}$$

Interference contributions from outside the cell come from all non-serving BSs, all UEs and in all layers. We see this Equation (1.21), where $\mathcal{B}$ is the set of all BS (or BS panels) in the system.

$$P_{IeCI} = \sum_{\substack{b' \in \mathcal{B} \\ b' \neq b}} \sum_{u' \in \mathcal{U}_{b'}} \sum_{l' \in \mathcal{L}_{b'u'}} P_{r,ll'} \tag{1.21}$$

The noise power P_N is computed according with thermal noise expression in 1.14, using the bandwidth of a single PRB, which depends on the numerology as evidenced in Table **??**.

This framework is also applicable when several BS are jointly serving one user, or when one user is transmitting to several BS simultaneously, i.e. Distributed-MIMO (D-

MIMO). This is true because we simply account for power contributions, abstracting from the content of the spatial streams.

Mutual Information Effective SINR Mapping

MI-ESM is an SINR aggregation technique that allows us to attribute one SINR to a transmission where the quality of the channel varies across the transmission band, namely across PRBs. We choose this SINR mapping strategy because [**?**, **?**, **?**, **?**, **?**] show that it unquestionably achieves the very good results without the need of calibration for different MCSs. Equation (1.22) sums how it works.

$$SINR_{eff} = I_k^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} I_k \left(SINR_i \right) \right) \tag{1.22}$$

Above, I_k is the information function that for a given SINR and MCS (with k bits per symbol) gives the bits of information that are conceivably extracted for a transmission with that SINR. For low SINRs, the mutual information is practically zero. As the SINR grows, the quantity of information bits extracted approaches k. Appendix **??** goes into further detail on the mutual information function works.

Therefore, Equation (1.22) obtains the mutual information achievable in each PRB, averages it and computes the SINR that would achieve that average information. Thus, the effective SINR is determined as the SINR that would yield this average mutual information if it were applied on all PRBs.

1.1.3 Transmission Realisation

Here we obtain the outcomes of the realised transmissions. Firstly, we calculate the number of TBs in which data is separated for transmission, based on the estimated bits to transmit. Secondly, the SINRs each UE experienced in each PRB are computed and then we present how these SINRs can be aggregated in something more easily useable to conclude on overall channel quality, an effective SINR. Finally, effective SINRs result in the success or failure of the transmitted transport blocks according with the MCS used for transmission and then the link quality adaptation mechanism is updated appropriately, as well as buffers and other indicators, such as schedulers fairness indicators, to be used in upcoming transmissions to assure a balanced operation of the system in line with the result of the transmission in the present TTI. As per usual, the steps follow.

Transport Block Size Calculation

To obtain the Transport Block Size (TBS), essentially three ways have been modelled and need to be assessed in simulations. The first, is to consider the same number of TBs in every transmission, N_{TB}. Therefore the numbers of bits to be transmitted $N_{bits,bul}$ is divided equally over TBs and the size of each TB is the same, see Equation (1.23).

$$S_{TB} = \lceil N_{bits,bul}/N_{TB} \rceil \tag{1.23}$$

The second is to consider a maximum TBS $S_{TB,max}$, obtain N_{TB} from Equation (1.24) and then use Equation (1.23). And the third method is to follow the list of steps described in [**?**], making the Transport Block Size depend on the number of layers $\#\mathcal{L}_{bu}$, modulations order M, code rate R_c, number of allocated PRBs $N_{PRB,bul}$ and transmission duration, which we assume to be always T_{slot}.

$$N_{TB} = \lceil N_{bits,bul}/S_{TB,max} \rceil \tag{1.24}$$

This such manner, N_{TB} TBs are sent and the experienced bit rates depend on how many of them are delivered with no errors. If there are no errors, the bit rate computed in Equation (1.10) are achieved, otherwise only a fraction of that bit rate is achieves, corresponding to the successfully transmitted TBs over total TBs.

Compute and Aggregate Realised SINR

Making use of the channel coefficients, we use the transmit powers and scheduled transmissions using certain beam pairs to compute the realised SINR per PRB, as described in the SINR framework described in 1.1.2. Then we aggregate it over the all PRBs into an effective SINR, by applying the procedure described in in 1.1.2.

Compute Block Errors

Subsequently, with the effective SINR $SINR_{eff}$ and the MCS used for the transmission, we look at the correspondent MCS curve in Figure 1.3 and get the resultant $BLER$.

Then we flip a BLER-biased coin to determine whether each block was well received or not.

Update Link Adaptation, Buffers and Performance Indicators

Firstly, the link adaptation mechanism is updated based on the block errors in accordance with subsection 1.1.1.

Then, the information that was successfully transmitted needs to be removed from the buffers. We model an ordered buffer where the information in one transport block has a direct mapping to certain packets. Therefore if that TB gets lost, that packet stays in the buffer.

This means the BLER may cause packets to be arrive out of order. This phenomenon is represented in Figure 1.5 where the size of a TB is set to the same size as a packet for illustration purposes. We see the bits in the transport blocks that didn't arrive successfully are kept in the transmission buffer. Therefore, if those bits are successfully sent in the future, they would be out of order. Note that this is something common in packet networks. Successfully transmitted TBs get their share of packets removed from the buffers.

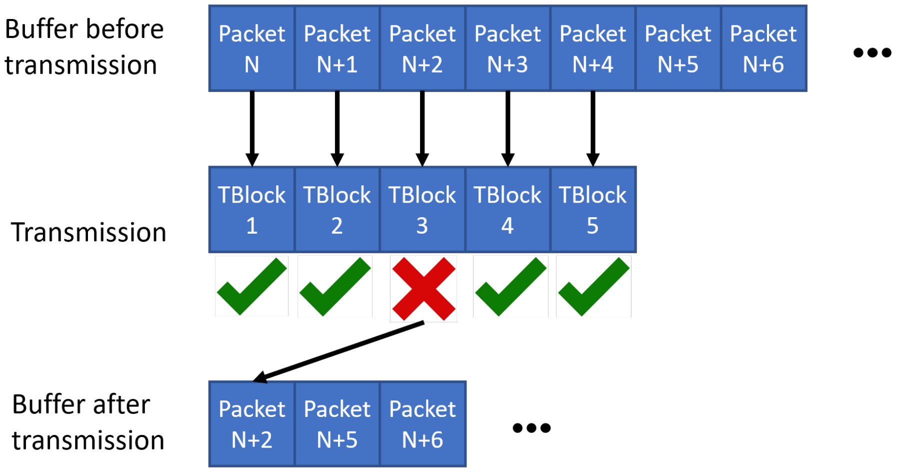

Figure 1.5: Buffer state before and after a transmission, with $S_{TB} = S_{packet}$.

These modelling considerations make the system considerably more realistic. It increases the likelihood that a packet get discarded due to excessive delay, so the latencies supported in a given scenario are higher.

Finally, the experienced bit rates are used to update the average bit rate over time used by the scheduler to influence what UEs are given more importance to serve in light not only of obtaining a fair resource allocation, but also to weigh and decide priorities e.g. regarding latencies versus throughputs.

Conclusion

In this section we followed the required steps to perform data transmission. We started with TTI identification and soon followed Slot format. Then, we modelled

CSI acquisition and presented an intuitive way of creating a GoB, which we made use throughout the section. Afterwards, we presented Link Adaptation and Instantaneous Bit Rate procedures given the important role they play in user scheduling. Subsequently we went through the user scheduling steps. Finally, we introduced a flexible SINR framework and a well known SINR aggregation algorithm, and used them to simulate a data transmission.

Next section we simulate single and multi-user scenarios and assess the relations between several parameters described in this section.

Bibliography

[1] 3GPP, *TS 38.901 - 5G; Study on channel model for frequencies from 0.5 to 100 GHz.*, v16.1.0, Rel. 16, 2020.

[2] F. Afroz, K. Sandrasegaran, and P. Ghosal, “Performance analysis of PF, M-LWDF and EXP/PF packet scheduling algorithms in 3GPP LTE downlink,” *2014 Australasian Telecommunication Networks and Applications Conference (ATNAC)*, 2014.

[3] Jong-Hun Rhee, J. M. Holtzman, and Dong-Ku Kim, “Scheduling of real/non-real time services: adaptive EXP/PF algorithm,” *The 57th IEEE Semiannual Vehicular Technology Conference, Spring*, 2003.