Impact of centralized-radio access network architecture on 5G performance

Tania Alhajj

COLLEGE SCIENCES
DOCTORAL POUR L'INGENIEUR
BRETAGNE ET LE NUMERIQUE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

THÈSE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPÉRIEURE
MINES-TÉLÉCOM ATLANTIQUE BRETAGNE
PAYS-DE-LA-LOIRE - IMT ATLANTIQUE

ÉCOLE DOCTORALE N° 648
SPIN : Sciences pour l'ingénieur et le numérique
Spécialité : *Informatique*

Par

Tania ALHAJJ

Impact of centralized-radio access network architecture on 5G performance

Thèse présentée et soutenue sur le campus de Rennes d'IMT Atlantique, le 17 mai 2023
Unité de recherche : IRISA
Thèse N° : 2023IMTA0342

Rapporteurs avant soutenance :

Nadjib AIT SAADI Professeur, UVSQ Université Paris-Saclay
Philippe MARTINS Professeur, Télécom Paris

Composition du Jury :

Présidente :	Thi-Mai Trang NGUYEN	Professeur, Université Sorbonne Paris Nord
Examinateurs :	Nadjib AIT SAADI	Professeur, UVSQ Université Paris-Saclay
	Karine AMIS	Professeur, IMT Atlantique
	Lila BOUKHATEM	Maître de conférences HDR, Université Paris-Sud
	Philippe MARTINS	Professeur, Télécom Paris
	Loutfi NUAYMI	Professeur, IMT Atlantique
Dir. de thèse :	Xavier LAGRANGE	Professeur, IMT Atlantique

Acknowledgments

I would like to start my thesis by thanking some people without whom this work would never have been accomplished.

First of all, I would like to give special thanks to my thesis director, Professor Xavier Lagrange, whose constant support, knowledge, expertise, advice, and patience have been as valuable for my work as for me.

Thanks to Professor Karine Amis from Lab-STICC for the valuable discussions and feedback I received during my thesis work. It was a pleasure to collaborate with her on this thesis.

Thanks to Professors Nadjib Ait Saadi and Philippe Martins who accepted to be the reporters of this thesis, and to Professors Thi-Mai Trang Nguyen, Loutfi Nuaymi, and Dr. Lila Boukhatem who accepted to be part of my thesis defense jury. I am very honored that my work has been evaluated by them.

I would also like to thank the Adopnet team members for the fruitful meetings and discussions. My sincere thanks go to Nicolas Huin whose experience and suggestions made this thesis richer. I would also like to thank Juan Vargas, Julien Saint Martin, and Cesar Vargas for their experiments and/or our discussions.

In IMT Atlantique, I was part of the SRCD department whose members are very kind and make working life better. Thanks for all the good memories. The most fun part of this journey was sponsored by ADER. I want to thank all ADER members and ex-members. I won't mention names to avoid forgetting anyone. I'm really happy to have met people as kind and caring as you!

I would also like to thank my family and friends who have become a family. Whether they are in France, Lebanon, Australia, KSA, or any other corner of the world, their daily encouragement and presence have been invaluable to me.

I am deeply grateful to my siblings Tony, Dayane, and Abdo and my parents Juliette and Michel for their support and unconditional love. They have been present at every stage of my life, including my three years of thesis. I am proud to have such an exceptional family.

Last but certainly not least, I express my appreciation and love to my partner, Manuel. His constant support and unconditional love have been and always will be essential to my success. I am lucky to have him in my life.

[0]Presenting my work in a structured and professional manner was made easier by the template provided in https://olivier.commowick.org/thesis_template.php.

Contents

List of Figures

List of Tables

List of Acronyms

List of Variables and Parameters

List of variables and parameters for Part II

A	Indoor area length
A_{m}	Antenna maximum attenuation
B	Indoor area width
$d_{i,j}$	Distance between RU_i and UE_j
$E^{\mathrm{UL},q}_{\mathrm{CU,PRACH},T}$	Energy consumed by the CU for PRACH occasions during T in mode q
$E^{\mathrm{DL},q}_{\mathrm{CU,SSB},T}$	Energy consumed by the CU for SSB transmissions during T in mode q
$E^{n,q}_{\mathrm{CU,data},T}$	Energy consumed by the CU for data transmission during T in mode q and direction n
E^{n}_{CU}	Energy consumed by the CU in direction n
$E^{\mathrm{UL},q}_{\mathrm{RU,PRACH},T}$	Energy consumed by the RU for PRACH occasions during T in mode q
$E^{\mathrm{DL},q}_{\mathrm{RU,SSB},T}$	Energy consumed by the RU for SSB transmissions during T in mode q
$E^{n,q}_{\mathrm{RU,data},T}$	Energy consumed by the RU for data transmission during T in mode q and direction n
E^{n}_{RU}	Energy consumed by the RU in direction n
$E^{n}_{\mathrm{SP,f}}$	Packet length independent signal processing energy consumption in direction n
$E^{n}_{\mathrm{SP,v}}$	Packet length dependent signal processing energy consumption in direction n
E^{n}_{SP}	Energy consumed for signal processing in direction n
$E^{\mathrm{UL},q}_{\mathrm{total,PRACH},T}$	Total energy consumed for PRACH occasions during T in mode q
$E^{\mathrm{DL},q}_{\mathrm{total,SSB},T}$	Total energy consumed for SSB transmissions during T in mode q
$E^{n,q}_{\mathrm{total,data},T}$	Total energy consumed for data transmission during T in mode q and direction n
f_c	Central frequency
$G(\theta_{i,j})$	Linear antenna gain as function of $\theta_{i,j}$
G_{A}	Antenna gain
$G_{\mathrm{dB}}(\theta_{i,j})$	Antenna gain in dB as function of $\theta_{i,j}$

I	Number of deployed RUs
i	Index of an RU
J	Total number of users
j	Index of a UE
M_{RB}	Available RBs in one time slot
$m(i,j,\nu)$	Number of RBs needed to serve UE_j by RU_i in configuration ν
$m^n_{\mathrm{RB},j,q}$	Number of RBs needed to serve UE_j in mode q and direction n
$m^{\mathrm{PRACH}}_{\mathrm{RB}}$	RBs used for preamble reception
$m^{\mathrm{SSB}}_{\mathrm{RB}}$	RBs used for SSB transmission
$N^{\mathrm{DL}}_{\mathrm{NF}}$	UE noise figure
$N^{\mathrm{UL}}_{\mathrm{NF}}$	BS noise figure
N^n_{Pkt}	Number of transmitted or received packets
N^n_{p}	Noise power in direction n
n	Transmission direction (DL or UL)
P_{ADC}	Analog to digital converter power consumption
P_{BPF}	Band pass filter power consumption
$P^n_{\mathrm{CU,f}}$	Fixed CU power consumption in direction n
P_{DAC}	Digital to analog converter power consumption
$P_{\mathrm{E/0}}$	E/O converter power consumption
P_{LNA}	Low noise amplifier power consumption
P_{LPF}	Low power amplifier power consumption
P_{Mixer}	Mixer power consumption
$P_{\mathrm{O/E}}$	O/E converter power consumption
$P^n_{\mathrm{RU,f}}$	Fixed RU power consumption in direction n
$P_{\mathrm{SP,f}}$	Fixed signal processing power consumption
$P^n_{\mathrm{r},i,j,q}$	Received power between UE_j and RU_i in mode q and direction n
$P^n_{\mathrm{t},q}$	Transmission power in mode q and direction n
$\mathbb{P}^n_{\mathrm{blocking},1,q}$	Blocking probability for 1 users in mode q and direction n
$\mathbb{P}^n_{\mathrm{blocking},J,q}$	Blocking probability for J users in mode q and direction n
$\mathbb{P}^n_{\mathrm{out},q}$	Outage probability in mode q and direction n
q	Mode index (Mode 1 or 2)
R^n_{T}	Target data rate in direction n
$R^{\mathrm{DL}}_{(i,j,\nu)}$	Perceived data rate by UE_j from RU_i in configuration ν

$R^n_{j,q}$	Perceived data rate by/from UE_j in mode q and direction n
r_0	Path-loss reference distance
s_i	Activity state of RU_i
T	Reference period of study
T_K	Receiver temperature
T_s	Time slot duration
T_n	Transmission duration in direction n
W_total	System bandwidth
w_RB	One RB bandwidth
$\boldsymbol{x}$	Order-3 tensor variable that determines the service of all users
$x_{i,j,\nu}$	Variable that associates a UE to an RU in a configuration
x_i	X axis index of RU_i
x_j	X axis index of UE_j
y_ν	Variable that determines the number of RBs occupied on the system level
y_i	Y axis index of RU_i
y_j	Y axis index of UE_j
α	Path-loss exponent
γ^n_min	Minimum SNR in direction n
$\gamma_{i,j,\nu}$	Received SINR by UE_j served by RU_i in configuration ν
$\gamma^n_{j,q}$	Received SINR from/at UE_j in mode q and direction n
δ_BW	Bandwidth correction factor
δ_SNR	SNR correction factor
$\eta_\mathrm{AC/DC}$	AC/DC power supply gain
η_PAE	Power amplifier gain
θ_3dB	Half power beam width
$\theta_{i,j}$	Angle between RU_i and UE_j, measured from the antenna boresight
μ	NR numerology
ν	Index of configuration
$\xi_{i,j}$	Normal random variable representing shadowing between RU_i and UE_j
σ_dB	Shadowing standard deviation
$\tau^n_{j,q}$	Resource utilization rate to serve UE_j in mode q and direction n

Résumé Français

Introduction

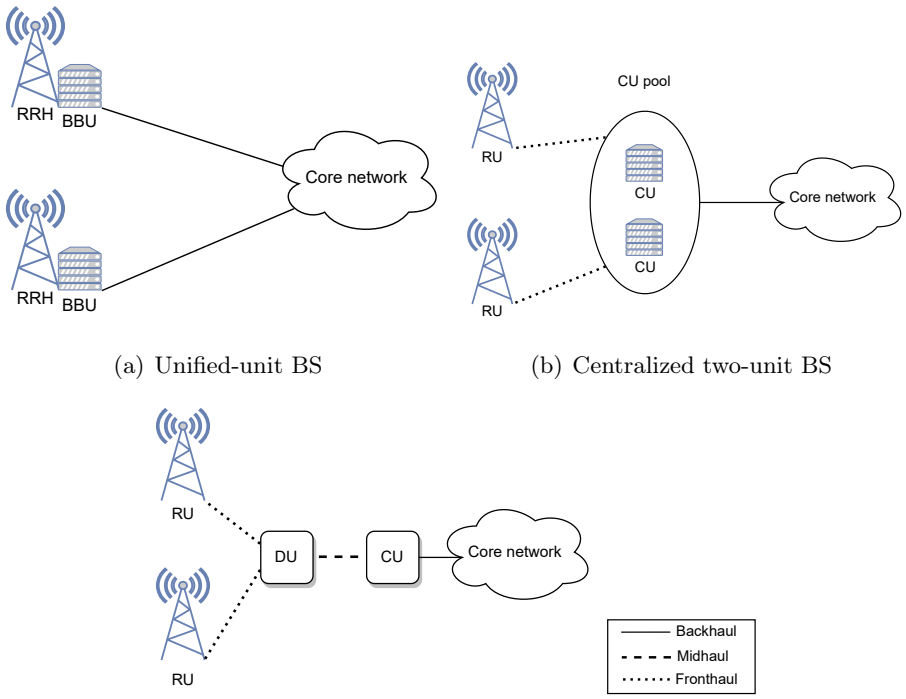
1. **Introduction générale**

   Depuis les années 1980, des générations de réseaux mobiles sont apparues, ont évolué et continuent d'évoluer. Cinq générations ont été développées et la sixième est actuellement en cours de développement. Après quatre générations, des débits de données plus élevés, une meilleure sécurité, une consommation d'énergie optimisée, une capacité accrue et d'autres facteurs sont disponibles. Toutes ces améliorations avaient pour but d'améliorer la vie des gens, de répondre à des demandes croissantes et de les satisfaire. Ensuite, la 5ème génération (Fifth Generation (5G)) est apparue non seulement pour servir les personnes, mais aussi pour servir les choses. De nouveaux types de services sont destinés à être fournis par la 5G par rapport à ses prédécesseurs. Les variantes Enhanced Mobile Broadband (eMBB), Massive Machine-Type Communications (mMTC) et Ultra Reliable Low Latency Communications (URLLC) sont les trois types de services identifiés par l'UIT dans le cadre de la norme International Mobile Telecommunications-2020 (IMT-2020).

   L'eMBB fournit principalement des débits de données très élevés. Le mMTC est conçu pour desservir un grand nombre d'appareils. Ces deux types sont des extensions de services connus des générations précédentes. Néanmoins, les débits de données élevés et le grand nombre d'appareils peuvent augmenter la consommation d'énergie du réseau, ce qui limite le déploiement pour des raisons de coût et d'environnement. C'est l'un des défis à relever pour fournir des services dans les réseaux 5G. URLLC est un nouveau type de service. URLLC consiste en des exigences strictes pour soutenir les applications critiques telles que la chirurgie à distance, la conduite autonome et les réseaux électriques intelligents [5G-Americas 2018]. Les exigences strictes sont une très grande fiabilité et une très faible latence. Néanmoins, la combinaison d'une fiabilité élevée et d'une latence très faible présente de nombreux défis, car il s'agit d'exigences contradictoires [Soret *et al.* 2014]. Pour atteindre l'une, il faut parfois sacrifier l'autre. Par exemple, en général, la fiabilité est obtenue par des retransmissions qui augmentent le temps de latence, et la satisfaction de ces deux exigences peut nécessiter la dégradation d'autres performances du réseau, telles que la capacité.

2. **Le réseau d'accès**

   Le réseau d'accès radio (Radio Access Network (RAN)) est constitué de stations de bases (Base Stations (BSs)) et d'antennes qui couvrent des zones spécifiques. Le RAN a évolué avec les générations de réseaux mobiles, permettant une évolution architecturale où le RAN centralisé (Centralized-RAN



(a) Unified-unit BS

(b) Centralized two-unit BS

(c) Centralized three-unit BS

Figure 1: Évolution du RAN.

(C-RAN)) a été introduit. La BS représentée sur la Figure 1.1(a) était une unité unifiée contenant une entité radio (Remote Radio Head (RRH)) et une unité de bande de base (Baseband Unit (BBU)) placées dans chaque cellule. Les RRH et BBU ont ensuite été séparées en deux unités distinctes, comme le montre la figure 1.1(b). L'RRH communique, via une liaison frontale, avec la BBU connectée au réseau coeur. En 2010, l'architecture C-RAN a été présentée pour la première fois [C.M.R.Institute 2010] [Lin *et al.* 2010]. C-RAN sépare le RRH, également appelée Radio Unit (RU), mise en œuvre sur le site cellulaire, de la BBU, appelée Centralized Unit (CU), mise en œuvre avec d'autres CUs et centralisées dans un groupe de CU ou CU pool, comme le montre la figure 1.1(c). Le C-RAN a fait l'objet d'une attention particulière dans le contexte de la 5G. Dans la 5G, la BS est divisée en trois unités : CU, Distributed Unit (DU) et RU. CU et DU sont reliées par une liaison fédératrice (backbone), comme l'illustre la Figure 1.1(d).

Entre autres avantages, C-RAN permet de réduire les coûts de déploiement du réseau tout en améliorant la couverture du système, la consommation d'énergie et la flexibilité. Quelques exemples d'amélioration des performances du C-RAN sont donnés ci-après. Plusieurs BSs partagent la même infrastructure à coût élevé dans un lieu centralisé. Les têtes radio à faible coût sont réparties sur les sites cellulaires, ce qui améliore la couverture du réseau et

réduit les coûts de déploiement du réseau.

3. **Objectifs et contributions.**

   Pour tirer parti du C-RAN, cette thèse s'articule autour de deux axes principaux. Dans le premier, nous étudions différentes répartitions fonctionnelles du C-RAN afin d'obtenir simultanément une fiabilité élevée et une faible latence, et les principales contributions sont les suivantes :

   - Nous analysons l'impact des différentes répartitions fonctionnelles du C-RAN sur la latence et la fiabilité sur la voie montante. Deux modes de réception sont étudiés : réception unique et réception multiple. Dans le mode de réception unique, le méchanisme de correction dérreur (Hybrid Automatic Repeat reQuest (HARQ)) est mis en œuvre localement dans chaque site cellulaire, ce qui réduit le temps d'aller-retour (Round Trip Time (RTT)). Dans le mode de réception multiple, le HARQ est centralisé, ce qui augmente le RTT, mais offre une coopération centralisée pour une meilleure diversité spatiale. Deux modes de réception unique et deux techniques de combinaison pour la réception multiple sont examinés.
   - Proposition d'une expression mathématique pour évaluer la probabilité d'erreur en utilisant HARQ with Chase Combining (HARQ-CC) avec la technique Maximum Ratio Combining (MRC) pour la macro-diversité. L'expression permet de calculer la probabilité d'erreur lors de chaque transmission HARQ-CC. Elle dépend de la combinaison spatiale des réceptions multiples avec MRC et de la combinaison temporelle avec les transmissions HARQ-CC précédentes.

   Dans la deuxième partie, nous étudions la consommation d'énergie d'un C-RAN en environnement intérieur dans les régimes de charge faible et moyenne. Les principales contributions sont les suivantes :

   - Évaluation de la consommation d'énergie de la BS avec l'architecture C-RAN dans un environnement intérieur en régime de faible charge. Deux modes de transmission sont étudiés : la transmission multiple et la transmission unique. Nous calculons la consommation d'énergie du BS pour chaque mode lors des transmissions des signaux de contrôle et des données sur la liaison montante (UpLink (UL)) et la liaison descendante (DownLink (DL)). Nous analysons également la couverture du réseau et la probabilité de blocage pour chaque mode.
   - Formulation d'un problème d'optimisation visant à minimiser la consommation d'énergie da la BS tout en augmentant la capacité du système dans les systèmes à charge moyenne. On alloue les User Equipments (UEs) aux RUs et on attribue les ressources radiofréquences dans DL. Nous évaluons la consommation d'énergie et la capacité du réseau obtenues en résolvant ce problème d'optimisation.

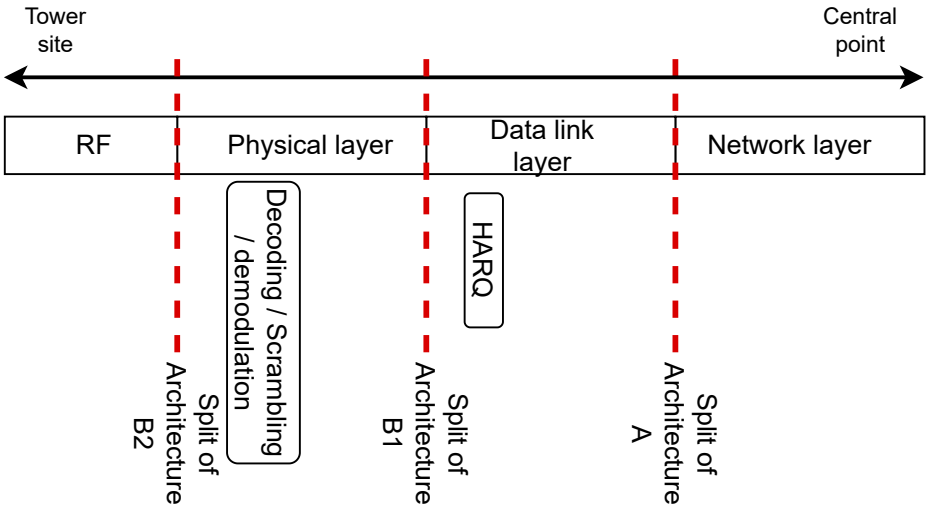


Figure 2: Répartitions des fonctions des architectures A, B1 et B2.

Impact du C-RAN sur la latence et la fiabilité

1. **Introduction**

   Dans un premier temps, nous avons étudié l'impact du réseau d'accès sur la fiabilité et la latence.

   La diversité temporelle n'est pas adaptée au type de service URLLC. Cependant, HARQ est un mécanisme inévitable pour augmenter la fiabilité dans les réseaux mobiles. Dans notre étude, nous tirons parti du C-RAN pour faciliter la coopération entre les RUs afin de bénéficier de la diversité spatiale dans le but de réduire les délais potentiels produits par le mécanisme HARQ.

2. **Architectures du C-RAN étudiées**

   Notre but est d'obtenir à la fois une haute fiabilité et une faible latence pour des services de type URLLC. Nous comparons trois architectures (A, B1 et B2) du réseau d'accès dont les répartitions des fonctions sont représentées dans la Figure 4.2. Dans toutes les architectures, nous nous concentrons principalement sur l'emplacement de la couche Media Access Control (MAC) et sur le processus de décodage. La couche MAC contient le processus HARQ-CC pour la correction des erreurs.

   Dans l'architecture A, tous les traitements sont effectués sur le site de la cellule. Ici, on considère une réception unique. Il s'agit de la répartition

fonctionnelle A dans [ecp 2019]. L'RU-DU reçoit le paquet, le décode et en stocke une version. Si le décodage est réussi, le paquet est envoyé au CU. En cas d'erreur, une retransmission est déclenchée par l'RU-DU car c'est là que la couche MAC se trouve. Le paquet retransmis est combiné avec ses versions précédemment reçues. Ce processus est répété jusqu'à ce que le paquet soit correctement décodé. Dans l'architecture A, nous considérons deux cas de réception : réception par la RU-DU la plus proche et réception par la meilleure RU-DU (celle qui reçoit le rapport signal sur bruit (Signal to Noise Ratio (SNR)) le plus élevé).

Dans l'architecture B1, on considère des réceptions multiples. Chaque RU sur les sites des cellules qui reçoit le paquet transmis le décode et le transmet au point central où la couche MAC est implémentée dans la DU-CU. C'est la répartition fonctionnelle D dans [ecp 2019]. Une erreur se produit si I RUs ne parviennent pas à décoder correctement le paquet. En cas d'erreur, la couche MAC au niveau du point centralisé demande une nouvelle transmission. Ceci est répété jusquà la bonne réception du paquet.

Dans l'architecture B2, nous considérons aussi des réceptions multiples. Chaque RU recevant le paquet de l'UE le transmet au point central où la couche MAC est implémentée dans la DU-CU. Le décodage et HARQ-CC sont centralisés et réalisés par la DU-CU. Il s'agit de la répartition fonctionnelle traditionnelle décrit dans [Duan *et al.* 2016]. Dans la DU-CU, les signaux sont combinés par la technique MRC. Les SNRs des I signaux sont additionnés. Ensuite, le processus de décodage a lieu. En cas d'erreur, une nouvelle transmission est lancée depuis le point central. Chaque version reçue du signal est stockée pour être combinée avec les réceptions suivantes en cas de retransmission. Ce processus est répété jusqu'à ce que le paquet soit correctement décodé.

3. **Distribution du délai**

   Le temps nécessaire pour la transmission et la propagation du paquet depuis l'UE vers le réseau dépend de la répartition de fonctions au niveau du C-RAN.

   Dans l'architecture A, les retransmissions ont lieu entre l'UE et l'unité radio (emplacement du HARQ) jusqu'à ce que le paquet soit reçu avec succès. À ce stade, le paquet correctement décodé est transmis à l'unité centrale. Le délai d'un envoi erroné est donc calculé en fonction du temps de transmission et de propagation entre l'UE et l'unité radio. Seul le temps nécessaire à la dernière (bonne) réception prend en compte le temps de transmission et de propagation entre l'unité radio et l'unité centrale en plus. Dans les architectures B1 et B2, chaque fois qu'un paquet est transmis, toutes les unités radio le transmettent à l'unité centrale, où HARQ est implémenté. Dans l'unité centrale, la décision de retransmission ou pas est prise. Ainsi, pour ces deux architectures, le temps écoulé pour chaque envoi est calculé par le temps de transmission et de propagation entre l'UE et l'unité radio et entre l'unité radio et l'unité centrale.

   Le temps nécessaire pour recevoir correctement un paquet est fonction du

nombre de transmissions l ($l-1$ mauvaises réceptions et une bonne). Par conséquent, la distribution de l est nécessaire pour connaître le délai et sa distribution.

Pour obtenir la distribution du nombre de transmissions, nous évaluons la probabilité de ne pas décoder correctement le paquet lors de chaque transmission.

Pour l'architecture A, une erreur se produit pendant la transmission k si l'unité qui reçoit l'UE (la plus proche ou la meilleure) ne parvient pas à décoder correctement le paquet.

Pour l'architecture B1, une erreur se produit pendant la transmission k si aucune des unités de réception ne parvient à décoder correctement le paquet provenant de l'UE.

Pour l'architecture B2, une erreur se produit si, après avoir combiné les signaux reçus dans les différentes RU au moyen de la MRC, le décodage du paquet échoue lors de la $k$ème transmission.

4. **Résultats**

   Les approches analytiques et par simulations ont donné les mêmes résultats. Les fonctions de répartitions complémentaires (Complementary Cumulative Distribution Functions (CCDFs)) représentées dans la Figure 3 résument les résultats de cette partie. Elles représentent la probabilité de ne pas recevoir correctement un paquet au cours d'une certaine période. En d'autres termes, cette CCDF représente le taux de perte de paquets.

   Cette figure compare les trois architectures avec tous les cas étudiés. Nous pouvons voir l'amélioration produite en recevant par la meilleure BS par rapport à la plus proche. Une amélioration supplémentaire est observée lorsque l'on reçoit depuis 4 BSs avec l'architecture B1. L'amélioration augmente lorsque l'on utilise MRC avec l'architecture B2. En fait, la sommation des signaux avant le décodage augmente les chances d'un bon décodage, c'est-à-dire une grande fiabilité, et diminue le nombre de retransmissions, c'est-à-dire une faible latence. La différence semble significative pour les faibles taux de perte. Si l'on prend un taux de perte de 10^{-2}, la différence entre le délai de l'architecture B1 et celui de l'architecture B2 est de $0,06$ ms, avec un délai plus court pour l'architecture B1. En revanche, pour une ultra-fiabilité de 10^{-6}, cette différence s'élève à environ 3.59 ms, avec un délai plus court pour l'architecture B2. Ainsi, pour les applications tolérantes aux erreurs, l'architecture B1 est suffisante. Cependant, pour les applications URLLC, l'architecture B2 avec MRC est meilleure.

5. **Conclusion**

   Dans la première partie, nous avons étudié l'impact de l'architecture du RAN et de la macro-diversité sur la fiabilité et la latence avec trois répartitions fonctionnelles différentes.

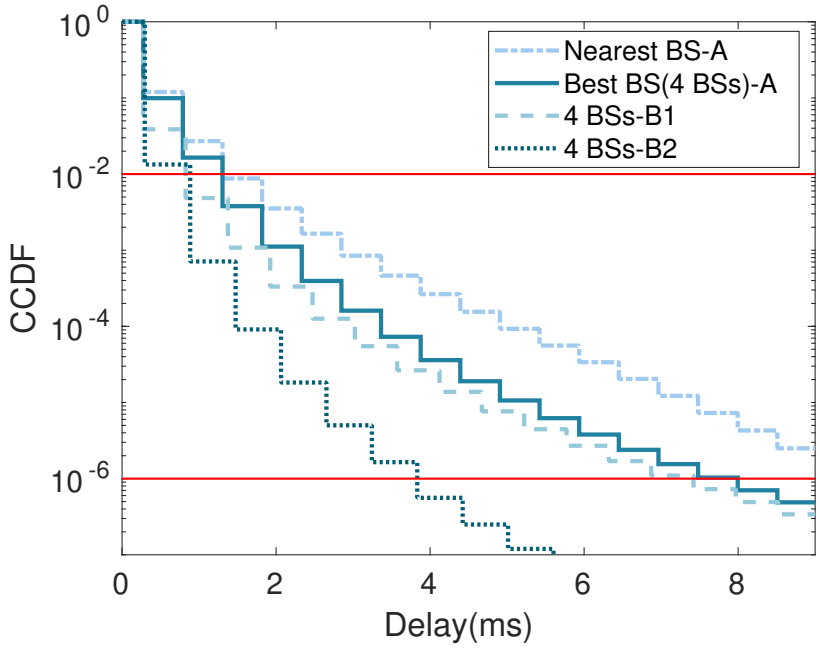


Figure 3: CCDF du délai des architectures A (2 cas), B1 et B2.

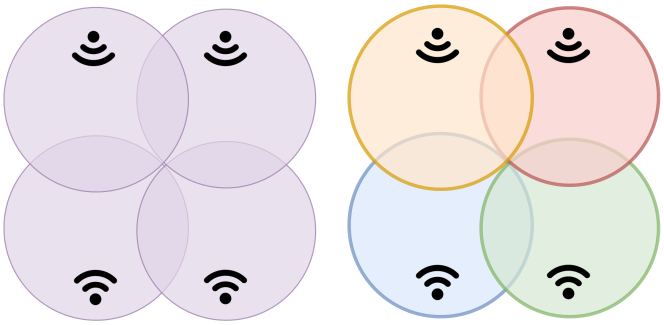
La macro-diversité avec les architectures B1 et B2 a eu un meilleur impact sur la fiabilité et la latence qu'une réception unique avec l'architecture A. En raison du gain de diversité, moins de transmissions sont nécessaires. Par conséquent, nous pouvons constater que nous obtenons à la fois une haute fiabilité et des délais plus courts avec les architectures B1 et B2, même avec un RTT plus long.

Impact du C-RAN sur la consommation d'énergie de la station de base

1. **Introduction**

   Dans un deuxième temps, nous avons étudié l'impact du réseau d'accès sur la consommation d'énergie de la BS dans des régime à faible et moyenne charges.

   La consommation d'énergie augmente en raison de l'évolution de la demande dans les réseaux 5G sans fil. En raison des préoccupations environnementales, l'amélioration de l'efficacité énergétique des systèmes est récemment devenue une question majeure. Certains cas d'utilisation présentent de fortes variations de charge (par exemple, les terminaux aériens), et le réseau est souvent construit pour desservir simultanément un grand nombre d'utilisateurs actifs. Cependant, périodiquement (par exemple, la nuit ou le week-end dans les centres d'affaires), seuls quelques appareils sont actifs. L'un des moyens d'améliorer l'efficacité énergétique des systèmes sans fil consiste à rendre le



(a) Mode 1 (b) Mode 2/3

Figure 4: Modes de fonctionnement.

réseau réactif aux variations de charge.

2. **Modes de fonctionnement**

   Dans cette parite, nous considérons des systèmes micro-cellulaires avec des régimes de charge faibles à moyens, afin d'évaluer l'impact du réseau d'accès sur la consommation d'énergie de la station de base.

   Nous comparons 3 modes de fonctionnements :

   Dans les systèmes à faible charge, il suffit de servir un UE à la fois. Ainsi, en mode 1, tous les RUs servent simultanément un UE à la fois. Dans ce mode, le système est une grande cellule et les UEs ignorent la présence de plusieurs RUs (voir Figure 4(a)). Une RU virtuelle, désignée RUtextesubscript$I + 1$, fonctionne dans ce mode comme si toutes les RUs étaient actives pour desservir un utilisateur simultanément. Dans le mode 2, une RU peut être active à la fois et peut servir un UE sur un bloc de ressources (Resource Block (RB)) spécifique. Chaque RU est située dans une cellule indépendante (voir Figure 4(b)). En mode 2, l'UE est attribué à l'RU qui fournit le rapport signal sur bruit et interférence (Signal to Interference and Noise Ratio (SINR)) le plus élevé. Dans les modes 1 et 2, il n'y a pas de réutilisation des ressources. Dans un système à charge moyenne, la desserte de plusieurs UEs en parallèle est importante pour écouler la charge requise (jusqu'au nombre d'RUs utilisateurs desservis sur le même RB), mais aucune technique spécifique n'est nécessaire, comme la formation de faisceaux (jusqu'au nombre de faisceaux d'utilisateurs desservis sur le même RB). Si nécessaire, plusieurs RUs peuvent émettre sur le même RB en l'absence de contraintes d'interférence. À cette fin, le mode 3 est un mélange des modes 1 et 2, avec la possibilité de réutiliser les ressources.

3. **Allocation de ressources radio**

   Nous considérons I RUs implémentées dans une zone rectangulaire et liées à une CU. Les I RUs partagenr les ressources radio disponibles dans le système.

   En mode 1, pour servir un utilisateur, toutes les RUs utilisent simultanément

le même RBs. Lorsqu'une ressource est occupée, les I RUs émettent en même temps le même signal vers l'utilisateur. Dans le mode 2 une RU (la meilleure) utilise le nombre d'RBs nécessaire pour servir un utilisateur. Lorsqu'un RB est occupé, une RU émet et les $I - 1$ autres RUs sont inactives. Dans les modes 1 et 2, un RB est entièrement dédiée à un utilisateur, qu'il soit desservi par toutes les RUs (mode 1) ou par la meilleure (mode 2). Néanmoins, le fonctionnement indépendant de chaque unité (comme dans le mode 2) signifie que les ressources peuvent être réutilisées entre différentes RUs. Cela permet d'augmenter le nombre d'utilisateurs pouvant être servis. Dans le mode 3, plusieurs configurations sont possibles : une configuration similaire au mode 1, des configurations similaires au mode 2 et des configurations avec réutilisation des ressources. Dans ce cas, l'allocation des ressources n'est pas aussi simple que dans les modes 1 et 2. Nous posons donc un problème d'optimisation pour gérer l'allocation des ressources radio en mode 3 tout en minimisant la consommation d'énergie.

Nous imposons le même débit cible en DL pour chaque UE, la même puissance d'émission pour les RUs et un nombre maximal d'RBs, qui définit les contraintes. Le problème d'optimisation qui en résulte est formulé sous la forme d'une programmation linéaire en nombres entiers mixtes (Mixed Integer Linear Programming (MILP)). Des simulations sont effectuées à l'aide de Python CPLEX API.

4. **Résultats**

   La Figure 5 montre la consommation d'énergie de la BS pendant la transmission de données en DL dans les trois modes. Les courbes se terminent à $J = 32$, 17 et 50 pour les modes 1, 2 et 3, respectivement (valeurs qui correspondent à une probabilité de blocage de 10^{-2}).

   Bien que toutes les RUs transmettent en mode 1, la consommation d'énergie est inférieure à celle du mode 2, dans lequel une RU dessert un UE. Les transmissions simultanées améliorent le SINR perçu par l'UE et réduisent le nombre de RBs où la BS est active. Cela montre la prédominance de la consommation d'énergie de CU à une faible puissance de transmission dans un contexte micro-cellulaire.

   La consommation d'énergie du mode 1 est similaire à celle du mode 3 pour $J \leq 32$. La consommation d'énergie augmente linéairement avec le nombre d'UEs jusqu'à 45 UEs en mode 3. Pour cette charge et au-delà, la réutilisation des ressources devient importante afin de pouvoir servir ce nombre d'utilisateurs, et l'augmentation de la consommation d'énergie n'est plus linéaire et s'accélère. En effet, la réutilisation des ressources entraîne une augmentation des interférences, ce qui nécessite davantage de RBs pour desservir les UEs et entraîne une consommation d'énergie importante.

5. **Conclusion**

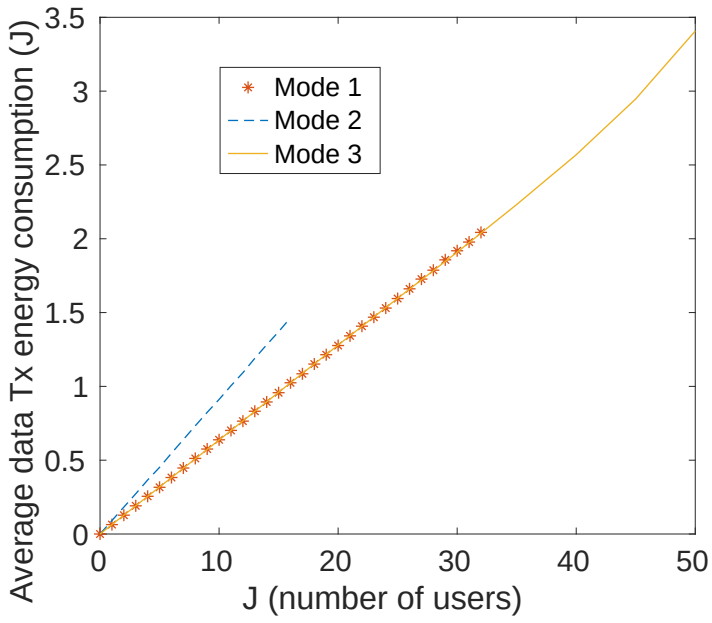


Figure 5: Energie consommée lors de la transmission des données en DL.

Dans cette partie, nous avons montré que le mode 1 est efficace pour réduire la consommation d'énergie dans les régimes de faible charge, tandis que la combinaison des modes 1 et 2 dans le mode 3 permet d'obtenir à la fois une faible consommation d'énergie et une capacité raisonnable pour un nombre d'utilisateurs allant jusqu'à 50 UEs.

Conclusion

Cette thèse a évalué l'impact de l'architecture RAN sur la fiabilité, la latence et la consommation d'énergie. L'évolution de l'architecture RAN a conduit à une architecture centralisée dans laquelle certaines fonctions de la BS peuvent être centralisées en amont dans le réseau et d'autres fonctions sont distribuées sur les sites cellulaires.

Dans la première partie, nous avons montré que l'HARQ avec la macro-diversité dans un C-RAN a un effet bénéfique sur la fiabilité et la latence par rapport à une réception unique avec un RAN traditionnel. En raison du gain de diversité, moins de retransmissions sont nécessaires. Par conséquent, nous obtenons à la fois une haute fiabilité et une faible latence même avec un RTT plus long dû à la centralisation du mécanisme HARQ.

De plus, avec l'utilisation accrue des réseaux sans fil et l'augmentation des émissions de CO2, il devient obligatoire de trouver des solutions pour réduire la consommation d'énergie des réseaux.

Dans la deuxième partie, nous montrons que la transmission simultanée par tous les RUs est efficace pour réduire la consommation d'énergie en contexte micro-

cellulaire dans les régimes de faible charge, tandis que la combinaison de ce mode avec la transmission indépendante et la réutilisation des ressources permet d'obtenir à la fois une faible consommation d'énergie et une capacité à écouler une charge moyenne.

Chapter 1

Introduction

Contents

1.1 General introduction

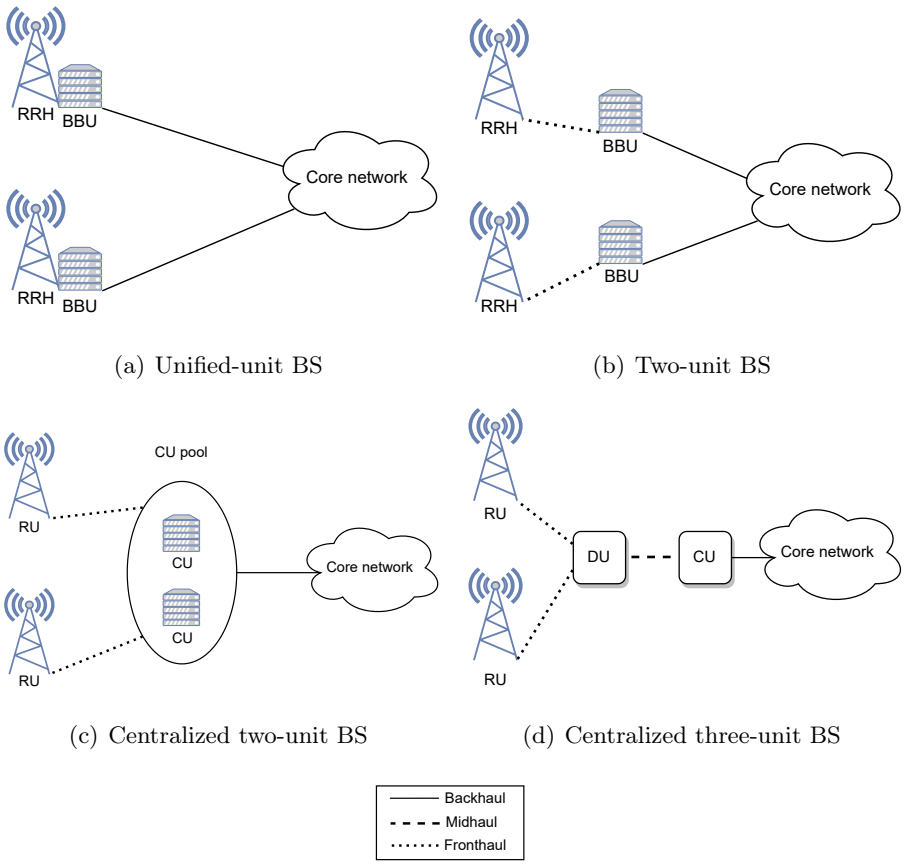
Mobile network generations have appeared, evolved, and are still evolving since the 1980s. Five generations have been developed, and the sixth is under development. It all started with giant, expensive cell phones and a network designed and deployed only for one task: making voice calls. Phone call quality, coverage, and security were poor at the time. The data rate was so low that it did not allow advanced services to work. Later, all these problems have been improved until the 4th generation. After four generations, higher data rates, improved security, optimized energy consumption, increased capacity, and more factors are available. All the updates were intended to improve people's lives, respond to increased demands, and satisfy them. This evolution made it possible to access faster high-quality services like gaming, video conferencing, and much more. Briefly, it offered high-speed internet access from anywhere at any time. Then, the Fifth Generation (5G) emerged not only to enhance all the previous improvements and to serve people but also to serve things. New types of services are destined to be performed by 5G compared to its predecessors. Enhanced Mobile Broadband (eMBB), Massive Machine-Type Communications (mMTC), and Ultra Reliable Low Latency Communications (URLLC) are the three types of services identified by the International Telecommunication Union (ITU) within the framework of the International Mobile Telecommunications-2020 (IMT-2020) standard.

The eMBB mainly provides very high data rates. The mMTC is meant to serve a massive number of devices. Both types are a kind of extension of the services known from previous generations. Nevertheless, the high data rates and large number of devices can increase network power consumption, limiting deployment for cost and environmental reasons. This is one of the challenges for service delivery in 5G networks. URLLC is the type of service of a new kind. URLLC consists of a range of stringent requirements to support mission-critical applications such as

remote surgery, autonomous driving, and smart energy [5G-Americas 2018]. The strict requirements are ultra-high reliability and very low latency. Nonetheless, the combination of high reliability and very low latency presents many challenges because they are conflicting requirements [Soret *et al.* 2014]. Achieving one may mean sacrificing the other. For example, in general, reliability is achieved through re-transmissions that increase latency, and meeting both requirements may require degrading other network performance, such as capacity.

Since the first mobile network generation, a part of the mobile system has been built between any User Equipment (UE) and the core network. This is the Radio Access Network (RAN) that provides resource access and coordinated management across radio sites. RAN is constituted of Base Stations (BSs) and antennas that cover specific areas. A UE communicates with a BS linked to the core network via the backhaul. RAN has evolved with the mobile network generations reaching an architectural evolution where the Centralized-RAN (C-RAN) was introduced. The BS represented in Figure 1.1(a) was a unified unit containing a Remote Radio Head (RRH) and a Baseband Unit (BBU) distributed in each cell. The RRH and BBU were then separated into two distinct units, as illustrated in Figure 1.1(b). The RRH communicates, via a fronthaul, with the BBU connected to the core network through the backhaul. In 2010, the C-RAN architecture was firstly introduced [C.M.R.Institute 2010] [Lin *et al.* 2010]. The C-RAN separates the RRH, also referred to as Radio Unit (RU), implemented on the cell site, from the BBU, referred to as Centralized Unit (CU), implemented with other CUs and centralized in a CU pool, as shown in Figure 1.1(c). C-RAN has received particular focus in the context of 5G. In 5G, the BS is further split and has three units: CU, Distributed Unit (DU), and RU. The CU and DU are linked through a midhaul, as illustrated in Figure 1.1(d).

Among other benefits, C-RAN offers scope for reducing network deployment costs while improving system coverage, power consumption, and flexibility. Some C-RAN performance enhancement examples are given hereafter. Multiple BSs share the same high-cost infrastructure in a centralized location. Low-cost radio heads are distributed across cell sites, improving network coverage and reducing network deployment costs. Multiple services with different requirements are supported by 5G networks. The network must cooperate with all services by ensuring a different Quality of Service (QoS) for each user requesting any type of service. Due to its flexibility, C-RAN can handle users asking for different types of services like eMBB and URLLC concurrently [Tang *et al.* 2019]. With the centralization and virtualization of BBUs, fewer BBUs can be used, reducing BS power consumption [Bassoli *et al.* 2017]. While C-RAN can reduce BS power consumption, it can be further reduced by taking into account the network load. Data transmission and network deployment must be efficient to reduce power consumption at low and high network loads.



(a) Unified-unit BS

(b) Two-unit BS

(c) Centralized two-unit BS

(d) Centralized three-unit BS

Figure 1.1: RAN BS evolution.

1.2 Thesis objectives and contributions

To take advantage of the C-RAN, this thesis is conducted along two main axes. In the first one, we study different C-RAN functional splits in order to achieve high reliability and low latency simultaneously and the main contributions are:

- Analysis of the impact of multiple C-RAN functional splits on latency and reliability in the UpLink (UL) direction. Two reception modes are examined: single reception and multiple reception. In the single reception mode, the Hybrid Automatic Repeat reQuest (HARQ) mechanism is implemented in each cell site, resulting in a short Round Trip Time (RTT). In the multiple reception mode, this mechanism is centralized, resulting in a long RTT, but with centralized cooperation providing spatial diversity. Two propagation models are considered: one based on distance and the other on distance and shadowing. In the distance-based propagation model, we also propose a flexible functional division based on thresholding according to the user's location.

- Proposition of a mathematical expression for evaluating the probability of error using HARQ with Chase Combining (HARQ-CC) with the Maximum Ratio Combining (MRC) technique for the macro-diversity. The expression allows the computation of having an error during each HARQ-CC transmission. The evaluation of this error probability depends on two main combinations. First, combining in the space domain the multiple receptions using MRC. Then, combining in the time domain each spatially-combined reception with all the previously received HARQ-CC transmissions.

In the second one, we study an indoor C-RAN BS energy consumption and capacity in low and medium load regimes and the main contributions are:

- Evaluation of BS energy consumption with C-RAN architecture in an indoor environment in a low-load regime where unused resources can be exploited. We adopt two transmission modes: multiple and single transmission. In the first one, all RUs serve the UE simultaneously. In the second one, a UE is served by one RU. We provide the BS energy consumption computation, that integrates transmission power and energy consumed for processing, for each mode during control and data transmissions in the UL and DownLink (DL). In addition to the energy consumption, we analyze the network coverage and the blocking probability of each mode, i.e. the probability of not being able to accommodate a certain number of users with the available resources.

- Formulation of an optimization problem that minimizes the BS energy consumption while increasing the system capacity to provide service in medium-load systems. We jointly assign UEs to RUs and allocate time-frequency radio resources in the DL. We provide the BS energy consumption computation with its different components as a function of the decision variables. We evaluate the energy consumption and the network capacity achieved by this optimization problem resolution.

1.3 Thesis outline

This thesis is divided into eight chapters including the introduction (this chapter). Below is a brief description of each chapter.

In Chapter 2, we overview some technical notions that are covered in this thesis. We give a brief 5G background related to its use cases and some protocol stack functions. We then present the evolution of the C-RAN architecture, whose impact on the performance of 5G networks is studied in this thesis. Then, the thesis is divided into two parts.

In the first part, we evaluate the impact of RAN architecture on reliability and latency targeting URLLC use cases. With a C-RAN architecture, in Chapter 3 we make a pilot study to evaluate the impact of the RAN architecture on reliability and latency. This chapter evaluates the latency produced on the radio link with HARQ mechanism to reach the desired reliability. Two aspects are compared: performing re-transmissions locally (short RTT) and from a central point (longer RTT). Then, a flexible switch between both aspects is studied. The promising results of this chapter led to an extension of the performed study. Chapter 4 considers the same aspects as in Chapter 3 but with a more realistic channel model. More combining techniques and a more detailed latency computation are also evaluated in this chapter.

In the second part, we evaluate the impact of RAN architecture on the BS energy consumption in an indoor environment during low and medium loads. Chapter 5 provides the considered system model and operating modes. A C-RAN architecture is considered where RU are distributed to provide the desired coverage and are linked to a CU. The operating modes consider either all RUs as one big cell or each RU as an independent cell. The network coverage, resource usage, and network capacity are evaluated in both modes by Monte-Carlo simulations. Chapter 6 evaluates the energy consumption of the BS in DL and UL for the same model and operating modes considered in Chapter 5. In this chapter, the considered energy consumption model of the BS with C-RAN is given and the computation of the BS energy consumption during control and user data transmissions/receptions is provided. Chapter 7 proposes an optimization problem to increase the system capacity when each RU performs independently. The radio resource is allocated with the objective of minimizing the BS energy consumption.

A final chapter is provided to summarize this thesis and suggest future work.

Chapter 2

5G and RAN evolution technical context

Contents

2.1 Introduction

The objective of this chapter is to provide a general overview of various topics that make up this thesis. We begin with a brief overview of Fifth Generation (5G), tackling the main points of interest in user and control planes. Centralized-RAN (C-RAN)'s architecture is next described with its components and advantages.

2.2 5G use cases

With speedier connection, low latency, ultra-high reliability, and increased bandwidth, 5G is designed to drive societies forward, streamline industries, and significantly improve everyday lives. 5G main requirements are mapped to three main types of services represented below and in Figure 2.1.

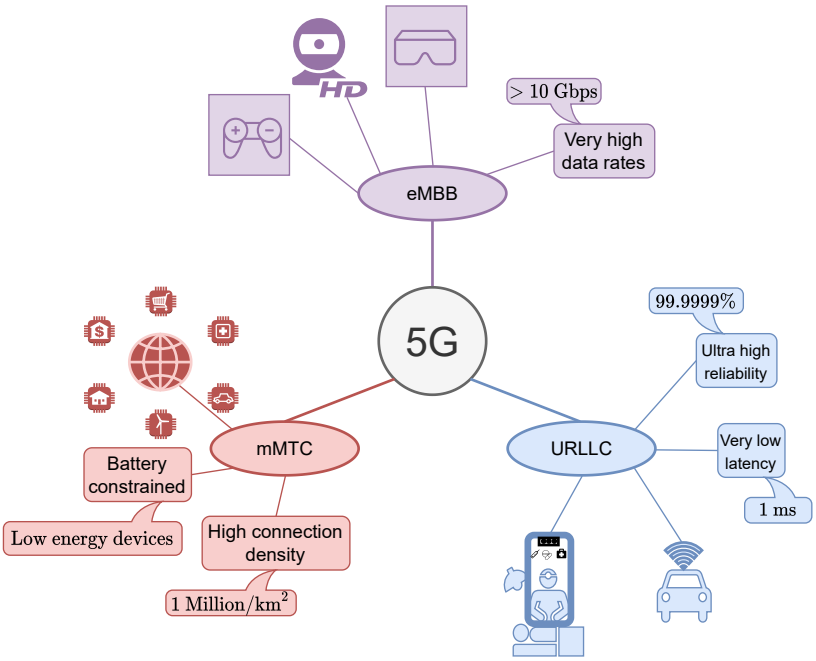


Figure 2.1: 5G types of services: eMBB, mMTC, and URLLC.

1. **Enhanced Mobile Broadband (eMBB)** represents an evolution of Long Term Evolution (LTE) mobile broadband services. It further improves the user experience by speeding up the connection, increasing throughput, and enlarging capacity. The eMBB groups services that require very high data rates: 10 Gbps in UpLink (UL) and 20 Gbps in DownLink (DL). A secondary constraint is low latency: for example, the user plane latency has to be as low as 4 ms [ITU-R 2015][Ateya *et al.* 2018]. Some eMBB applications examples are very high definition videos, virtual reality, and augmented reality.

2. **Massive Machine-Type Communications (mMTC)** is the type of communication where a massive number of devices are connected. For instance, buildings, public infrastructures, and personal devices are connected in a smart city. An mMTC network has to support up to 1 million devices per km^2. It is delay tolerant and manages small data packets. Usually, mMTC handles small and low-energy devices. This communication has thus to ensure a long battery life for these devices.

3. **Ultra Reliable Low Latency Communications (URLLC)** is a new type of service that handles critical applications like telemedicine and autonomous driving. Critical requirements are demanded for these applications, like a very low latency that can reach 1 ms and very high reliability that can be more than 99.999% depending on the application. In this thesis, we focus on this new type of communication in the first part. More details about URLLC are

given in Section 2.2.1.

2.2.1 URLLC in 5G: a new type of service

Healthcare, transportation, manufacturing, haptic/tactile internet, Augmented Reality (AR), and Virtual Reality (VR) are areas where URLLC has a tremendous impact. Some examples of these applications and their targeted latency and reliability are given in Table 2.1. There are good reasons for URLLC's importance in each area. For example, an intervention by a doctor not physically present at the patient's location is necessary for a healthcare emergency. The intervention could be done using a robot. The robot must receive the doctor's commands very quickly and reliably to avoid dangerous actions toward the patient [Siddiqi *et al.* 2019]. Another example is found in tactile internet: cybersickness symptoms might appear if the virtual world and the real action are separated by more than a millisecond [ITU-T 2014].

Now, let us clearly define the latency and reliability required by URLLC applications.

Latency is defined by two main metrics: the control-plane latency and the user-plane latency. The **control-plane latency** is defined as the time it takes a User Equipment (UE) to move from idle state to active state [TR 38.913 2017]. The **user-plane latency** refers to the time it takes to successfully deliver a packet from the entry point of the radio protocol layer 2/3 Service Data Unit (SDU) to the radio protocol layer 2/3 SDU exit point [TR 38.913 2017]. We focus in this thesis on the latter.

Several components contribute to latency in a network architecture. These components are represented in Figure 2.2 and explained hereafter [Parvez *et al.* 2018]. The Radio Access Network (RAN) delay T_{R} is the delay produced on the access network. The backhaul delay T_{B} is the time needed for the connections between the RAN and the core network to be established and the time to transmit on the backhaul network. The core network delay T_{CN} is the time needed for processing in the core network. Transport delay T_{T} is the data transfer duration between the core network and external networks. Counting these components twice (forward and backward) between the source and destination is used to calculate end-to-end latency. In [TS 22.261 2020], the **end-to-end latency** is defined as the period between the source transmitting the message and the destination successfully receiving it.

Our main focus in this thesis is on the delay produced in the RAN (i.e. T_{R}). This is calculated by adding up the transmission duration, the propagation delay, the processing delay, the queuing delay, and the time needed for re-transmissions. Transmission duration is the time needed to transmit the bits on the radio link between the UE and the RAN and on the fronthaul between the Remote Radio Head (RRH) and the Baseband Unit (BBU). Transmission duration depends on the transmission bit rate. Propagation delay measures the time it takes for a signal to travel from its source to its destination (i.e., from the UE to the RRH and from the

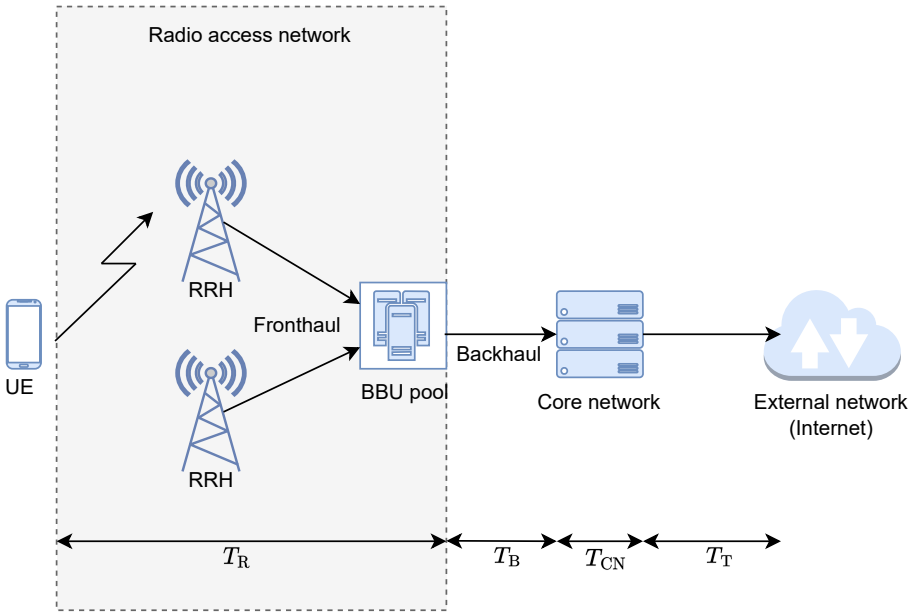


Figure 2.2: End-to-end latency components.

RRH to the BBU in the UL, and vice versa in the DL). Propagation delay varies with the physical distance between the source and the destination. The processing delay is produced by physical layer processing at the UE, the RRH, and/or the BBU, and it depends on the processing speed. Queuing latency refers to how long packets wait in a queue before they are transmitted, and it increases with network congestion. Finally, the re-transmissions duration is a function of the number of re-transmissions needed for a good reception at the receiver. It increases as channel conditions decrease.

Reliability is defined by 3rd Generation Partnership Project (3GPP) as the probability of successfully transmitting a certain number of bits in a specified amount of time. This time is calculated by the time a small data packet takes to leave the radio protocol layer 2/3 SDU of the transmitter and reach the receiver's radio protocol layer 2/3 SDU of the radio interface [TR 38.802 2017]. The commonly used mechanism to enhance mobile network communications is the Hybrid Automatic Repeat reQuest (HARQ). However, as re-transmissions are sent, HARQ increases reliability at the cost of more significant latency. The first part of this thesis addresses the problem of achieving low latency while achieving high reliability using HARQ. Section 2.3.3.1 explains the HARQ mechanism.

2.3 5G NR protocol stack

5G New Radio (NR) consists of seven protocol entities. Two among these protocol units are specific to the control plane: Non-Access Stratum (NAS) and Radio Resource Control (RRC), one is specific to the user plane: Service Data Application

Table 2.1: URLLC applications requirements.

URLLC application	End-to-end latency (ms)	Reliability (%)	Reference
Health surveillance	100	99.99	[TS 22.261 2020]
Telesurgery	1	99.9999999	[Chen *et al.* 2018]
Smart transportation	5 – 10	99.999	[Chen *et al.* 2018]
Industry automation	10	99.99	[TS 22.261 2020]
Tactile internet	1	99.9999999	[Chen *et al.* 2018]
AR	0.4 – 2	99.999	[Zheng *et al.* 2014] [Elbamby *et al.* 2018]
VR	5 – 10	99.99	[TS 22.261 2020]

Protocol (SDAP), and four are shared between the control and user planes: Packet Data Convergence Protocol (PDCP), Radio-link control (RLC), Media Access Control (MAC), and Physical layer.

The NAS functional layer is between the UE and the Core Network (CN). Some of its functions are security, authentication, paging, and Internet Protocol (IP) address allocation. We don't detail these functions because we will not address them in this thesis. We start by detailing the RRC UE states in the upcoming section.

2.3.1 RRC layer: connection states

RRC entity is responsible of control-plane operations like system information broadcast and connection and mobility management.

In 5G NR a UE can mainly be in three different RRC states represented in Figure 2.3 [Dahlman *et al.* 2018] [TS 38.401 2018]. RRC protocol exchanges signaling between the Base Station (BS) and the user. First, when turned on, the UE is in a disconnected mode, referred to as idle state (RRC-idle). The user is unknown to the RAN. For any communication, the UE proceeds a random access request to initiate a random access process. When successful, this process transforms the UE's state into connected (RRC-connected). In this state, the user is known to the network. These two states are the states defined for the LTE networks. A third state is introduced by 5G NR. In this state, there is no radio connection but the tunnels and connections on the fronthaul and backhaul are maintained and the UE capabilities are stored in the BS. This state is known as the inactive state (RRC-inactive). Like in the idle state, the user's energy consumption is reduced in this state. However, returning to active status is simplified: changing from inactive to active state takes less time than from idle to active state.

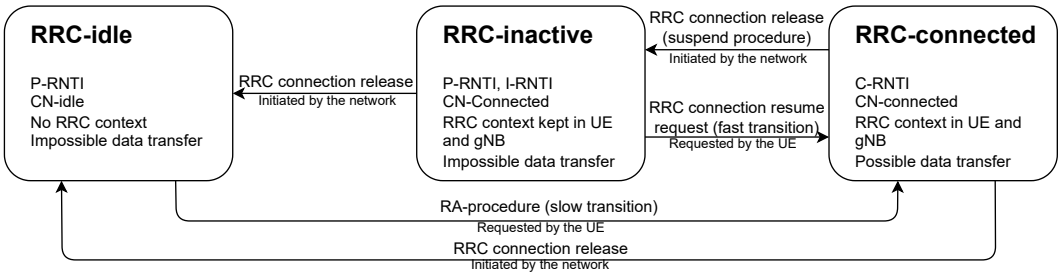


Figure 2.3: RRC states transitions.

2.3.1.1 Idle state and cell search

A UE that is not connected to the network is considered in an idle state (RRC-idle). The UE is also disconnected from the CN and is in CN-disconnected mode. In this state, the UE is in sleep mode and periodically checks the paging messages. The network addresses these messages using a Paging Radio Network Temporary Identifier (P-RNTI). In the idle state, the cell-change decision is taken by the user. This is detailed in Section 2.3.1.4.

A UE in RRC-idle state that needs to use the network searches for a cell to be served by it. BSs broadcast periodically Synchronization Signal Blocks (SSBs) over the beacon channel. The SSB consists of the Primary Synchronization Signal (PSS) and Secondary Synchronization Signal (SSS) along with the Physical Broadcast Channel (PBCH). It provides information needed to the UE in order to know the identity (ID) of its serving BS and synchronize with it. They are sent periodically, without prior knowledge of the cell load, even if the cell is empty. This means there is sometimes a waste of resources, such as unnecessary power consumption. That is why, in NR technology, a minimum amount of information is sent in these messages. Additional information is sent upon user request. Thus, the energy consumed to send this additional information is not wasted. Master Information Block (MIB) and System Information Block 1 (SIB1) are sufficient for the UE to recognize the ID of the potential serving BS.

When operating on high frequency, beam-forming is useful to increase the range of the BS. For beam-forming, the SSBs are sent over sweeping beams to allow the BS to be heard by users in different locations over the cell area. This is called a sweeping beacon channel.

After receiving the necessary information to recognize the cell or the beam, a UE requests random access to begin data transfer, which was impossible during the idle state.

2.3.1.2 Access process

The access process in 5G NR is quite similar to the one in LTE. It consists of four steps in a contention-based scenario and two in a non-contention-based scenario

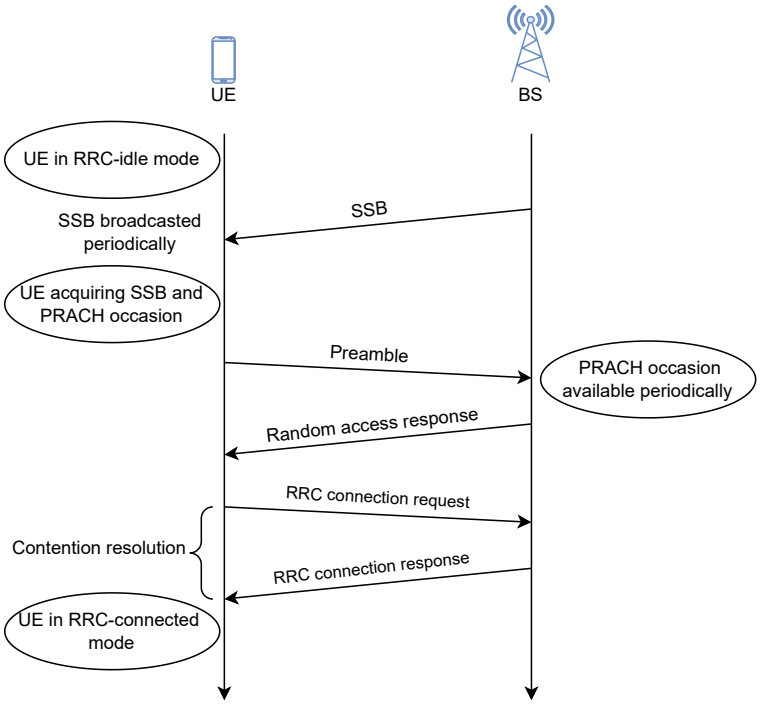


Figure 2.4: Access process in 5G NR.

[Dahlman *et al.* 2018]. After receiving an SSB, a user waits for a transmission opportunity. The transmission occasion or Physical Random Access CHannel (PRACH) occasion is available periodically for the user to request access. Once the PRACH occasion is available, the UE sends a request of attachment to the network by sending a preamble. The network responds with an attachment response message containing an identity related to the preamble previously used by the user. Multiple users may transmit the same preamble in the first step. The possible collision is detected in the third step and resolved in the fourth. The UE is considered connected at the end of these steps (RRC-connected state). An illustration of the above process is in Figure 2.4.

2.3.1.3 Connected state

When connected, the UE is known to the network and is identified by an Cell-RNTI (C-RNTI). The UE is in RRC-connected state, and the RRC context is present in the UE and the radio access network as well. The UE can transmit and receive data and is connected to the core network: this is the CN-connected state. The network provides the UE with neighboring cell frequencies. The UE continuously measures the signaling received from these cells. When the handover is necessary, the network takes care of it according to the measurement report provided by the UE.

When in the RRC-connected state, the network can proceed with an RRC re-

lease request in order to release this connection and put the user in RRC-idle state [TS 38.401 2018]. It can also proceed with an RRC release request using a suspend procedure to switch to the RRC-inactive state and avoid losing contact with the network.

2.3.1.4 Inactive state

In the inactive state, the connection with the core network and the RRC context are maintained. However, transmitting and receiving data is not possible. This state is thus a mix of the connected and idle states. Similarly to the idle state, the UE manages the handover. The network sends periodic paging messages with a P-RNTI. Paging messages are monitored by the UE using the Inactive-RNTI (I-RNTI). This identifier is assigned to the UE by the network during the suspended release to identify it. The UE measures the received powers of different paging messages from different cells. In case of an access need, it selects the cell with high received power. The UE proceeds with an RRC resume request to go back to the connected state. This is faster than the random access when in the idle state because the RRC context and the core network connection are also present. To switch to the idle mode, the network initiates an RRC connection release procedure towards a user in the inactive state [TS 38.401 2018].

2.3.2 SDAP, PDCP, RLC

SDAP, PDCP, RLC sub-layers and their functions are beyond the scope of this thesis. Nevertheless, we briefly present some functions of each of them. The PDCP is responsible for managing Quality of Service (QoS) flows on the 5G air interface. Among its functions, the PDCP is responsible for IP header compression and decompression, duplicate packet detection and scheduling, and integrity. The functions of the RLC sublayer include error correction and data segmentation and re-segmentation.

2.3.3 MAC layer

MAC layer has multiple functions among which we cite scheduling information reporting and error correction using HARQ. HARQ is of interest to us, so we describe it in detail.

2.3.3.1 HARQ mechanism

Wireless transmissions might be interrupted or distorted due to noise or interference, resulting in reception errors. In order to maintain a reliable communication system, it is necessary to introduce mechanisms for potentially correcting these errors. Automatic Repeat reQuest (ARQ) is an error detection mechanism that, when coupled with Forward Error Correction (FEC), becomes an error correction mechanism called HARQ. In ARQ, during a transmission, the transmitter waits

for feedback from the receiver. The feedback is a short acknowledgment message and can be either an ACKnowledgement (ACK) if the packet is well received or a Negative ACKnowledgement (NACK) otherwise. The received packet is discarded if the receiver receives an erroneous transmission, and a NACK is transmitted back to the transmitter. The transmitter proceeds with a re-transmission upon the NACK reception. The sender also triggers a re-transmission if it receives no feedback messages within a specific duration. This process is repeated infinitely if it is a persistent ARQ until the reception of an ACK by the sender. In the case of truncated ARQ, the process is repeated until a predetermined maximum number of re-transmissions. In FEC, messages are encoded redundantly by the sender. Receivers benefit from redundancy since it allows them to detect and often correct errors throughout a message. This prevents the need to ask for a re-transmission.

In HARQ, a receiver capable of correctly decoding the received packet replies to the sender with an ACK. When the received packet is erroneous, it is not discarded. A NACK is transmitted to the sender to trigger a re-transmission. Packets received during each re-transmission are combined with previous receptions. This is repeated until the sender receives an ACK or the maximum number of re-transmissions is reached. Two types of HARQ define what is re-transmitted by the sender during each re-transmission: HARQ with Chase Combining (HARQ-CC) where the same packet is re-transmitted, and HARQ with Incremented Redundancy (HARQ-IR) where additional redundancy bits are transmitted during each re-transmission.

- HARQ-CC [Benelli 1985] [Chase 1985] involves transmitting a packet and then re-transmitting the same one when needed. A packet is concatenated with Cyclic Redundancy Check (CRC) and parity bits, modulated at the sender, and transmitted in the first place. In response to a NACK received from the receiver, the sender re-transmits the same coded and modulated packet sent in the first place. Upon each re-transmission, the receiver combines the received versions of the same packet using Maximum Ratio Combining (MRC) (described later in 2.4.1.2). This combination increases the Signal to Interference and Noise Ratio (SINR) and thus the chances of good decoding. HARQ-CC is illustrated in Figure 2.5.

- HARQ-IR [Mandelbaum 1974] [Kallel 1990] [Shiozaki 1996] consists of transmitting additional redundancy bits each time a re-transmission is required. After concatenating a packet with CRC and parity bits, it is modulated and sent by the sender. In case of erroneous reception, the receiver replies with a NACK and keeps the packet with the first set of redundancy bits stored. The sender transmits a new set of redundancy bits. At the receiver, a coding gain is achieved. When extra redundancy is sent during each re-transmission, the likelihood of correctly decoding the packet increases. HARQ-IR is illustrated in Figure 2.6.

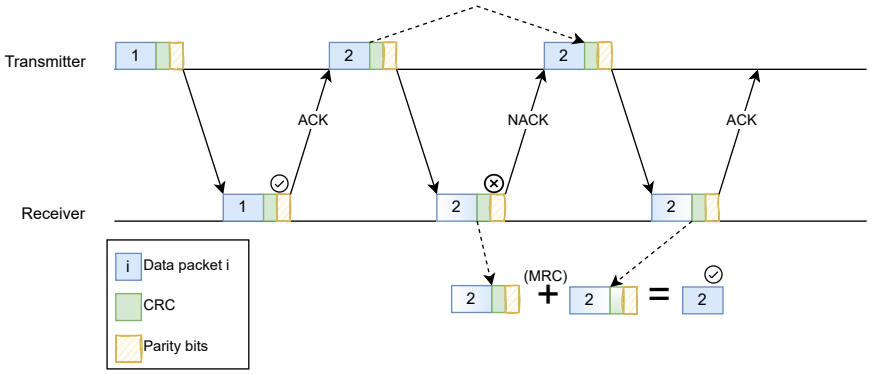


Figure 2.5: HARQ-CC diagram.

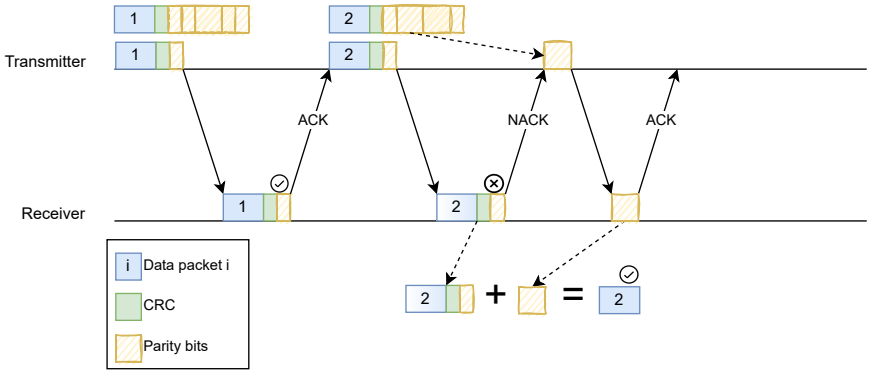


Figure 2.6: HARQ-IR diagram.

Table 2.2: 5G NR numerology variation.

Numerology μ	SCS (KHz)	T_s (ms)	Number of slots per subframe
0	15	1	1
1	30	0.5	2
2	60	0.25	4
3	120	0.125	8
4 (for SSB only)	240	0.0625	16
5 (not defined in Rel 15-16)	480	0.03125	32
6 (not defined in Rel 15-16)	960	0.015625	64

2.3.4 Physical layer

In the physical layer, coding/decoding, modulation/demodulation, multi-antenna mapping, and other functions are handled.

Like LTE, the 5G NR air interface uses Orthogonal Frequency Division Multiplexing (OFDM) modulation since it is the most efficient in the case of multi-path transmissions.

2.3.4.1 5G NR numerology

In 5G NR, a radio frame duration is 10 ms and a subframe duration is 1 ms. Each subframe consists of a set of slots with 14 OFDM symbols per slot. A slot duration is determined by OFDM symbols length, while Sub-Carrier Spacing (SCS) determines each symbol length.

The SCS is variable in 5G NR, and depends on the numerology μ. The SCS can be written as $\Delta f = 15 \times 2^{\mu}$ kHz. As for the time slot duration, it is given by $T_s = \frac{1}{2^{\mu}}$ ms. When $\mu = 0$, we get $\Delta f = 15$ kHz and $T_s = 1$ ms. These values are the same as in LTE, where they were previously unique. Table 2.2 summarizes different numerologies values [TS 38.211 2022].

2.3.4.2 Physical layer technologies

In this section, we introduce some physical layer technologies useful for meeting network requirements in 5G. We briefly discuss Multiple Input Multiple Output (MIMO), beamforming, and macro-diversity.

The term MIMO refers to an antenna technology for wireless communications that uses multiple antennas at both the source (transmitter) and destination (receiver). In 5G NR, MIMO is extended to massive MIMO where an increased number of transmit and receive antennas is used to enhance the network's capacity and coverage.

The deployment of multiple antennas at the BS makes it possible to serve multiple users in parallel: this is the Multi-User MIMO (MU-MIMO) achieved through

beamforming. When a large number of devices are connected, MU-MIMO improves network performance. The transmitter focuses different sets of antenna signals (beams) toward different users to serve them simultaneously on the same resources. Beamforming can be used to increase one user's experience: this is the Single User MIMO (SU-MIMO).

Another type of MIMO is the macro-diversity MIMO, where distant BSs transmit or receive signals in the same resource simultaneously to co-ordinate communications with users. When multiple versions of the same signal are received by a receiver, a combining technique takes place to merge the received signals. Two combining techniques are listed hereafter:

- Selection Combining

  Multiple received signals are sent to a combiner. This combiner in Selection Combining (SC) method chooses the best received signal (Figure 2.7 (a)). That means that the signal with the best Signal to Noise Ratio (SNR) is chosen [Goldsmith 2005]. When all the branches have the same noise power, the highest SNR is equivalent to the highest received power. So to gain time, the signal with the highest power at the receiver will be selected (no SNR determination needed). The SNR at the combiner's output is the highest SNR between all SNRs of the received signals. The outage probability is measured by not receiving from any branch a SNR greater than a threshold SNR:

  $$\mathbb{P}_{\text{out}} = \mathbb{P}(\gamma_1, \gamma_2 ..., \gamma_M < \gamma_{\text{T}}), \tag{2.1}$$

  where $\mathbb{P}_{\text{out}}$ is the outage probability, M the number of hearing BSs, γ_i the SNR corresponding to the signal heard by the ith BS signal and γ_{T} the threshold SNR.

- Maximum Ratio Combining

  The MRC computes a weighted average of the signals arriving from the different channels. The SNR at the combiner's output is the summation of SNRs of different signals (Figure 2.7 (b)).
  An outage occurs in this case when the sum of all SNRs is lower than the threshold SNR:

  $$\mathbb{P}_{\text{out}} = \mathbb{P}(\sum_{i=1}^{M} \gamma_i < \gamma_{\text{T}}). \tag{2.2}$$

2.4 RAN evolution

Recent RAN architectures are discussed in this section. We primarily focus on the C-RAN, a centralized architecture, since it represents the main scope of our work. We then briefly discuss virtual RAN (v-RAN), a centralized and virtual

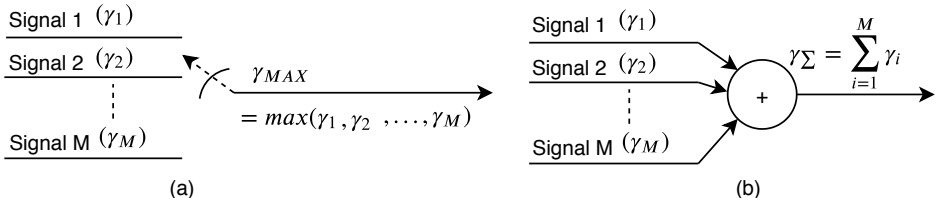


Figure 2.7: Combining diversity techniques:(a)SC (b)MRC.

architecture. Finally, we briefly explain Open-RAN (O-RAN), one of the newest network architectures, which is centralized, virtual, and open.

2.4.1 C-RAN architecture

C-RAN is an evolution of the previously known RAN architecture. It splits the functions of a legacy BS between two units: RRH (or Distributed Unit (DU)) and centralized BBU (or Centralized Unit (CU)), linked via a fronthaul. These are the three primary components of C-RAN architecture. Below is a brief description of each.

- The RRH is distributed across cell sites. It is mainly responsible for radio signal transmission and reception. Therefore, analog to digital, digital to analog, filtering, and power amplification are all performed by the RRH [Checko *et al.* 2015]. Each RRH is linked to a BBU pool.

- The BBU is centralized with other BBUs in a BBU pool. It is mainly responsible for base-band processing. The signal processing of all RRHs is combined in the BBU pool. This allows coordinated multi-point transmissions, centralized resource allocation, and joint user scheduling [Ren *et al.* 2018]. The BBU pool is linked to each RRH via the fronthaul link.

- The fronthaul is the link between each RRH and its corresponding BBU pool. Multiple protocols can be used for transmissions between RRHs and BBUs over the fronthaul. Fronthaul connections in LTE widely used the Common Public Radio Interface (CPRI) connection protocol. Fronthaul links are expected to experience increased traffic with the advent of 5G. In addition, based on the distribution of functions between the BBU and the RRH, the rate on the fronthaul is highly dependent on the configuration. There are some splits where the fronthaul bit rate scales with the number of antenna ports. In other splits, it scales with the number of MIMO layers [TR 38.801 2017]. These reasons, along with others, led to the specification of an ethernet-based CPRI (eCPRI) [ecp 2019]. CPRI uses constant bit rate transmissions while eCPRI is packet-based and varies with the actual payload. As a result, eCPRI prevents resource waste and offers greater flexibility.

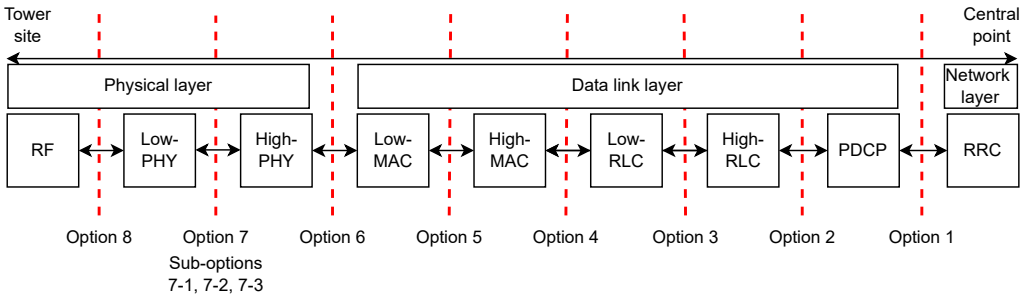


Figure 2.8: 3GPP functional splits options.

As network transport architecture evolved from LTE to 5G NR, BBU functions were split between three units (see Figure 1.1(d)). The layer 2 and 3 functions with less strict latency requirements are implemented in the CU, layer 1 and 2 functions in the DU, and the remaining layer 1 functions with very stringent latency requirements in the Radio Unit (RU) [Akbarzadeh *et al.* 2020] [ITU-T 2018].

Note that 3GPP considers a BS split into two units DU and CU. The functional split determines the number of functions implemented in the CU and the DU.

2.4.1.1 Functional splits in C-RAN

According to 3GPP, C-RAN splits the BS functions into two units. Depending on the functional split, each unit implements different functions. 3GPP has defined eight functional splits options [TR 38.801 2017]. The eight options are illustrated in Figure 2.8 [TR 38.801 2017]. What is defined by 5G is the split between RU-DU and CU. This is option 2 in Figure 2.8, where the CU is on the right side, and the RU is combined with the DU on the left side. Regarding the split between RU and DU, there are, in fact, several possibilities: option 8 is what is currently deployed, and option 7.2 is what is defined in the O-RAN architecture (described in Section 2.4.3). Among these splits, in this thesis, we consider options 1, 6, and 8 (see Figure 2.8).

As the functional split option number increases (e.g. option 8), the RAN is more centralized, while an RAN is more distributed as this option's number decreases (e.g. option 1). Each functional split has its advantages and disadvantages. In each functional split, there are some requirements—for example, the fronthaul bit rate increases with the functional split option number. A detailed survey describing all the functional splits is given in [Larsen *et al.* 2019]. Other splits options or splits naming were also given by other organizations like small cell forum [SCF 2016] and eCPRI [ecp 2019] [Larsen *et al.* 2019].

Moreover, mobile networks nowadays deal with temporally and spatially dynamic use cases, applications, and services. In addition, users with different QoS may also request to be served within the same geographic area. A flexible functional split can be implemented to accommodate network changes, and diverse requirements [Rost *et al.* 2014]. The flexible functional split enables the CU and DU to

flexibly decide which split option is right for them based on the users' requirements. By doing so, they can leverage the advantages of different functional splits. For instance, different functional splits might be used to cancel or reduce different levels of inter-cell interference. Optimal splits are dynamically switched according to the current case in each cell to achieve this [Harutyunyan & Riggio 2018].

2.4.1.2 C-RAN benefits

C-RAN offers multiple benefits for nowadays networks. Among other advantages, C-RAN helps in reducing network deployment cost and energy consumption, facilitates network expansion, and permits the coordination between different cells on distributed sites.

1. **Network deployment cost and energy consumption reduction**

   The centralization of base-band processing in a centralized pool allows the efficient utilization of available base-band resources to satisfy the users. Thus, less CUs might be used for the same network performance compared to the legacy RAN. Moreover, the simplicity of the DU makes it cheaper to deploy multiple sites. This results in CAPital EXpenses (CAPEX), OPerating EXpenses (OPEX), and energy consumption reduction [C.M.R.Institute 2010]. However, it is possible to further reduce energy consumption, as we will show in this thesis.

2. **Network flexibility and extensibility**

   In C-RAN, it is easy to assign DU to different CUs in order to respond to different network requirements. For example, during the night, some CU can take charge of additional DU in order to switch off their corresponding CU. As a result, the network can be flexible and variable in response to load fluctuations [C.M.R.Institute 2010]. As a new site can be created by implementing and connecting a simple unit to an existing CU, expanding the network is also easy and inexpensive.

3. **Multiple sites coordination**

   Due to the centralized part of the BS in the same place, different distributed sites can cooperate [Gupta *et al.* 2010]. The co-location of CU facilitates some mechanisms that aim to enhance the network performances and achieve users QoS like Coordinated Multi-Point (CoMP) and Joint Transmission (JT). Cooperation improves interference management by reducing inter-cell interference. In addition, mobility management (like handover) is also optimized and performed faster. Moreover, it allows macro-diversity and spatial multiplexing.

In our thesis, we mainly explore two of these advantages. The network flexibility and extensibility are used to adapt the network to the load. The multi-point trans-

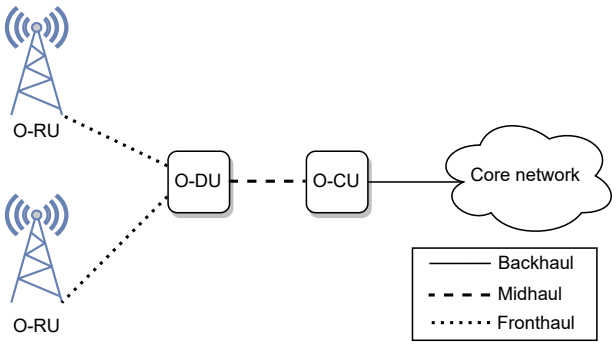


Figure 2.9: O-RAN architecture.

mission/reception is used to increase reliability using macro-diversity, i.e., receiving or transmitting by multiple DU serving different cells.

Signals are generated in the CU and transmitted to various DUs for transmission to the same user in the DL. In the UL, multiple DUs receive the signal from the user and transmit it to the CU. When receiving multiple signals, the receiver must combine them. Multiple combining techniques exist. We cite two among them illustrated in Figure 2.7.

2.4.2 v-RAN

Today, there are several approaches for automating the configuration and functions of a network in order to improve its agility. These methods are Software-Defined Networking (SDN) and Network Functions Virtualization (NFV) [Akyildiz *et al.* 2014] [Akyildiz *et al.* 2016]. As a result of applying virtualization to the DU and CU of the C-RAN, v-RAN is created. By virtualizing network functions, v-RAN eliminates the link between software and hardware. In other words, the v-RAN is no different from the C-RAN in that it is a centralized architecture with virtualized functions.

2.4.3 O-RAN

O-RAN considers the split of the BS into three parts. The O-RAN architecture consists of software-based components: an Open-RU (O-RU) connected to an Open-DU (O-DU) via a fronthaul. The O-DU is connected to the Open-CU (O-CU) via the midhaul. Then, the backhaul connects the access network to the core network. O-RAN is represented in Figure 2.9.

An essential difference between this architecture and previous ones is its openness. With the O-RAN architecture, standard BS equipment takes advantage of a modular software stack that allows base-band and radio unit modules from different vendors to interact seamlessly.

Due to its centralized model, software-based and open components, O-RAN is a virtual centralized and open network architecture.

2.5 Conclusion

In this chapter, we highlighted different topics, including technical information related to 5G NR and the evolution of RAN architecture. These elements help the reader understand the upcoming chapters.

Part I

Impact of C-RAN architecture on reliability and latency

Chapter 3

C-RAN for URLLC using HARQ: a pilot study

Contents

3.1 Introduction

Having examined macro-diversity and Hybrid Automatic Repeat reQuest (HARQ) mechanism in Chapter 2, it remains unclear whether they can achieve high reliability and low latency using Centralized-RAN (C-RAN) architecture. In the first part of our work, we combine time diversity with spatial diversity to study their impact on the reliability and the latency produced on the Radio Access Network (RAN) considered as C-RAN. We use HARQ as a time diversity mechanism and reception from multiple Radio Units (RUs) with centralized processing in C-RAN as spatial diversity. We compare this diversity combination to an architecture where the HARQ processing is performed on-site, close to the user, and only one reception is taken

into consideration. We compare HARQ close to the users without spatial diversity against HARQ far from the users with spatial diversity to see which leads to high reliability with low latency.

Before introducing our work, we highlight some relevant previous research work. We then describe the architectures considered, followed by the computation of latency produced on the RAN. The error probability evaluation in its analytic and simulation forms is given afterward. We also propose flexible switching between different architectures to save network resources while achieving the desired goals. We end this chapter with a conclusion to sum it up and introduce the next chapter.

In this chapter, we start with a pilot study where we consider a simple propagation model and a simplified delay computation. We compare two functional splits considering a different reception mode for each one: single and multiple receptions. We study only one combining technique for the multiple reception case. Afterward, we propose a flexible functional split between the two studied splits.

3.2 Related literature review

Previous studies have been conducted to respond to Ultra Reliable Low Latency Communications (URLLC) requirements. A variety of mechanisms and technologies were applied and studied. An interesting technology to explore in this area is C-RAN architecture. Authors in [Chaudhary *et al.* 2019] evaluated the delay at the switch in a packetized fronthaul using split option 7.3. The explored split is a high intra-physical layer split [Larsen *et al.* 2019]. This study showed that correct reception within a specific delay could be significantly increased by installing faster Ethernet switches on the fronthaul. In [Mountaser *et al.* 2017], the authors studied three functional split options in support of URLLC using an experimental approach. The three studied splits are options 2, 6, and 7-1 (see Figure 2.8). Option 7-1 is an intra-physical layer split where only the fast Fourier transformation is implemented locally on the cell site [Larsen *et al.* 2019]. They evaluate two packet delays. The time a packet takes from being inserted into the data link layer until being transmitted to the User Equipment (UE). The other evaluated delay is the transmission delay between the upper layer of the split to its lower layer. Evaluating these delays shows that split option 6 outperforms the other studied split options in achieving low latency for URLLC traffic. However, according to 3rd Generation Partnership Project (3GPP), one of the disadvantages of option 6 is that the centralization of the Media Access Control (MAC) layer may affect the HARQ performance due to the additional round trip fronthaul latency [TR 38.801 2017]. Nevertheless, HARQ remains the widely used error correction mechanism to increase reliability through time diversity. Therefore, it is important to investigate how localization of HARQ impacts URLLC performance while considering radio network delay.

Multiple types of diversity exist and contribute to increasing the reliability: frequency, time, and space diversity [Ohmann *et al.* 2016]. In [Swamy *et al.* 2015], researchers proved that frequency diversity alone fails to meet ultra-high reliability

in realistic channel conditions; meanwhile, multi-user diversity can make it. Time diversity is often omitted while considering URLLC because of the stringent latency requirement. It is, therefore, interesting to discover how to achieve high reliability and low latency through spatial diversity. In [Mahmood *et al.* 2019] and [Jacobsen *et al.* 2019] for example, spatial diversity through multiple receptions is studied. These studies showed the effect of spatial diversity in increasing the reliability for URLLC users. More precisely, they were interested in studying resource utilization and the network's capacity using the multi-cell reception technique. Mahmood *et al.* evaluated multi-connectivity for URLLC traffic [Mahmood *et al.* 2018]. Their study considers the reception and duplication of the packet heading to the user by a master node. The duplicate is forwarded to the secondary node. Then, both nodes independently transmit the same packet to the user. The results show that multi-connectivity achieves high reliability and low latency, especially when URLLC traffic is mixed with other types of traffic like Enhanced Mobile Broadband (eMBB). However, in this study, the duplicates are considered to be forwarded to a secondary node over an ideal Xn interface that does not experience any delays. Despite the reliability gain provided by spatial diversity, cooperating distributed Base Stations (BSs) must share information. This cooperation can be done for example through Coordinated Multi-Point (CoMP) technology [TR 36.819 2013]. Cooperation, between distant BSs, over the X2 interface is time-consuming, which breaches the stringent delay requirement in Long Term Evolution (LTE) networks [Brueck *et al.* 2010] [Artuso & Christiansen 2014]. The use of central processing units facilitates the transfer of information between cooperating BSs. It is therefore easier to collaborate across distant cell sites using C-RAN architecture [I *et al.* 2014].

As previously mentioned, time diversity is unsuitable for the service of the URLLC type. However, HARQ is an unavoidable mechanism to increase reliability in mobile networks. Various enhancements have been made to this error correction mechanism in the literature. For example, early feedback is one type of HARQ improvement. It is an approach that can provide early feedback depending on the decoder output before the decoding phase is complete. For instance, the outcome of the decoder is predicted in two steps by the authors in [Berardinelli *et al.* 2016]. First, the bit error rate is estimated based on likelihood ratios. Then, the estimated bit error rate is used to estimate the output of the block error indicator predictor based on thresholds. Their approach reduces latency and can generate more than 90% correct feedback messages. The researchers in [Imamura *et al.* 2017] predict the need for a re-transmission based on the channel state information. The drawbacks of these methods are the false feedback estimations. Later on, some research work like [Taniyama *et al.* 2019] and [Strodthoff *et al.* 2019] tried to resolve these issues.

Targeting the same objective, some researchers studied HARQ in C-RAN architecture. Several timing-critical processes exist in the lower MAC layer like HARQ. Therefore, splits options 1 to 5 have relaxed fronthaul latency requirements, while splits options 6 to 8 have highly stringent delays [Larsen *et al.* 2019] [TR 38.801 2017] (c.f. Figure 2.8). However, distributing lots of processing functions

leads to losing the C-RAN benefits like deployment cost reduction and facilitated cooperation. According to the survey in [Larsen *et al.* 2019] and references therein, CoMP is possible in both DownLink (DL) and UpLink (UL) in splits options 8, 7-1, and 7-2. In the remaining split options, CoMP advantages are limited. The more we move towards distributing the BS functions, the more the cooperation possibility is limited, and the distributed unit is complex. In [Khalili & Simeone 2017], the authors proposed to predict the decoding feedback locally by the Remote Radio Head (RRH) or the UE, while the real decoding takes place in the Baseband Unit (BBU). Doing so allows as many processing functions as possible to be centralized, and the fronthaul latency is reduced. In this study, the local feedback is considered during single and multiple UL receptions. However, the local feedback is a prediction of the actual feedback that must come from the BBU, and a feedback mismatch may occur. Moreover, when receiving from different RRHs, the combining techniques are limited due to the local feedback given by each RRH alone. In our study, we consider the actual feedback from both the RRH and the BBU, and study their impact on reliability and latency. We consider different combining techniques for signals received during multiple receptions.

Combining HARQ with macro-diversity in a C-RAN architecture has received little attention in the context of URLLC. In the first part of our work, we study the impact of C-RAN architectures with centralized and distributed processing for URLLC services.

3.3 System model

3.3.1 Network model

We consider a hexagonal cell network where the UE position is uniformly distributed. We consider the BS to be split into three units: RU, Distributed Unit (DU), and Centralized Unit (CU). The RUs are mounted with one omnidirectional antenna each and implemented in the center of each hexagonal cell of radius R_c. We consider having I RUs. The CU, connected to I RUs, is implemented at an equal distance ρ from these RUs. RU and CU are connected through a packet-based ethernet-based CPRI (eCPRI) transport network, reduced to a simple fiber optic link to ensure low latency. The propagation velocity over the fiber link is denoted by v. Therefore, the propagation delay θ between the CU and the corresponding RUs is constant. We consider different functional splits, and in each one, we precise where the DU is implemented: co-located with the RU (right side of Figure 3.1) or with the CU (left side of Figure 3.1). The network deployment is illustrated in Figure 3.1.

3.3.2 Propagation model

We consider successive transmissions of data packets in the UL direction and two reception modes: a single reception mode where only one RU receives, and a multiple reception mode where more than one RU receive the user's signal. In this

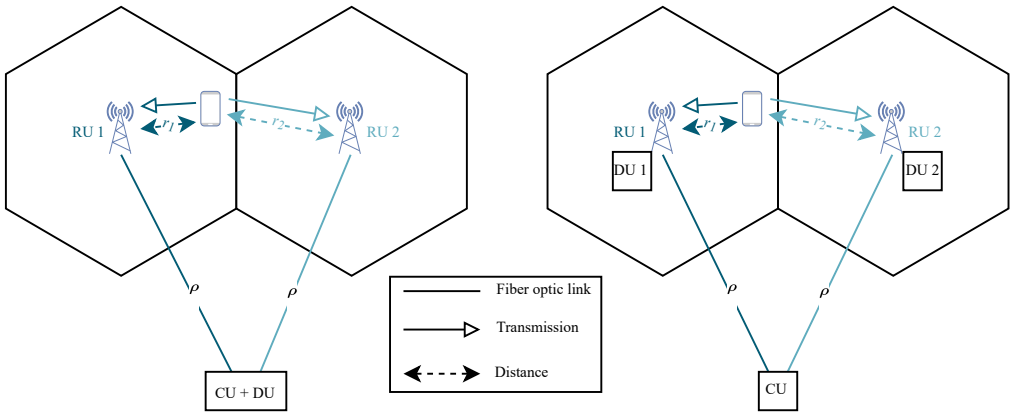


Figure 3.1: Network deployment.

chapter, a distance-based propagation model is used. The model considers path-loss attenuation as a function of distance and Rayleigh fading. The channel model is Okumura-Hata, and the received power is given by:

$$P_{\mathrm{r}} = P_{\mathrm{t}} \left(\frac{r_0}{r}\right)^{\alpha} \chi, \tag{3.1}$$

where P_{r} is the received power, P_{t} the transmitted power, r_0 a constant, r the distance between the transmitter and the receiver (the UE and the BS respectively in our case), α the path-loss exponent and χ an exponential random variable (r.v.) representing the fading whose mean is equal to 1.

We consider the background noise power N as a constant. Parameter N includes the noise and the interference. The Signal to Noise Ratio (SNR) is thus given by:

$$\gamma = \frac{P_{\mathrm{r}}}{N}. \tag{3.2}$$

With the previous considerations, we define the average SNR denoted by $\overline{\gamma}$ and given by:

$$\overline{\gamma} = \frac{P_{\mathrm{t}}}{N} \left(\frac{r_0}{r}\right)^{\alpha}. \tag{3.3}$$

3.3.3 Error model

During each transmission, the Packet Error Rate (PER) is approximated and calculated as a function of the received SNR similarly to [Liu *et al.* 2004], where the approximation was given as a curve fitting to the real PER:

$$h(\gamma) = \begin{cases} 1 & \text{if } 0 < \gamma < \gamma_M \\ ae^{-g\gamma} & \text{if } \gamma \geq \gamma_M \end{cases} \tag{3.4}$$

where a and g are parameters that depend on the Modulation and Coding Scheme (MCS) mode and $\gamma_M = \frac{\ln a}{g}$. Table 3.1 gives the variation of a and g depending on the MCS.

HARQ with Chase Combining (HARQ-CC) is the used error correction mechanism. An erroneous message leads to a re-transmission request by the receiver, and the sender re-transmits the same packet.

Table 3.1: PER model parameters.

Modulation	**Coding rate**	a	g	γ_M
BPSK [1]	1/2	274.7229	7.9932	−1.5331
QPSK [2]	1/2	90.2514	3.4998	1.0942
QPSK [2]	3/4	67.6181	1.6883	3.9722
16-QAM [3]	9/16	50.1222	0.6644	7.7021
16-QAM [3]	3/4	53.3987	0.3756	10.2488
64-QAM [3]	3/4	35.3508	0.0900	15.9784

[1] Binary Phase Shift Keying.
[2] Quadrature Phase Shift Keying.
[3] Quadrature Amplitude Modulation.

3.3.4 Architectures overview

To avoid confusion, in this chapter and the following one, we consider a CU in a central location, an RU on the tower site, and a DU. The DU can be either centralized in the central location with the CU or distributed on the cell site with the RU depending on the considered architecture (see Figure 3.1).

This chapter considers two C-RAN architectures with different functional splits: architecture A and architecture B. This section aims to describe each architecture and how it is implemented.

In **architecture A**, the MAC layer, responsible for the error correction (using HARQ), is implemented on the tower site. This corresponds, for example, to functional split option 1 in Figure 2.8, with the DU being co-located with the RU. In this architecture, data is close to the user [TR 38.801 2017]. We consider here a single-reception mode: a UE transmits, and only one RU receives it (Figure 3.2).

During each transmission, while considering architecture A, the nearest RU receives the signal, and a decoding process occurs. In the case of a failed decoding, the BS's MAC layer implemented on the tower site asks for a re-transmission by sending a Negative ACKnowledgement (NACK) to the transmitting UE. When the decoding succeeds, the data is transmitted to the CU. An ACKnowledgement (ACK) message is also sent from the BS to the UE.

In **architecture B**, the error correction is triggered from a central point. The MAC layer is, therefore, centralized. This is split option 8 in Figure 2.8, with the DU being centralized with the CU. Note that other splits with a centralized MAC layer like option 6 can also be chosen, but the delay detailed in Section 3.4 should be computed differently. Here, we consider that I RUs receive the data from the same UE (Figure 3.3). We consider a single transmission and multiple receptions:

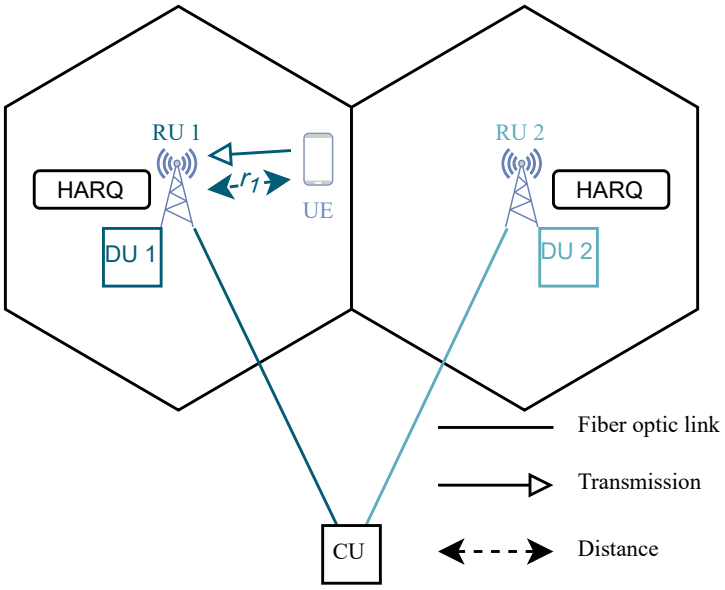


Figure 3.2: Architecture A: one receiving BS.

a UE transmits, and I RUs receive it.

The decoding is based on the I receptions in architecture B. The CU manages the coordination between the I RUs: different receiving RUs send the received signals to the CU, where the combining diversity technique takes place: the centralized combiner considers an error if all the received signals are erroneous. In case of an error, a re-transmission is requested from the CU by transmitting an NACK to the UE. When the CU succeeds in decoding the signal coming from the UE, it replies with an ACK.

3.4 Delay computation

In this section, we present the delay components and the assumptions that we consider.

The reliability is achieved by the HARQ mechanism in the first case. For the second one, both HARQ and spatial diversity are used to achieve reliability. Now, to compute the delay for each architecture, we expose its components. We assume that the processing duration is negligible. We consider that the propagation delay on the radio link is absorbed by the guard time of the slot. We also suppose we have a very high rate on the fiber link. Then, the transmission duration over the fiber is negligible.

All transmitted packets have a specific defined size L in bits. We consider that the transmission duration of a packet or an ACK/NACK on the radio link between the user and the RU is one Transmission Time Interval (TTI). Thereby, we have $T_{\mathrm{D}} = T_{\mathrm{A}}$, with T_{D} being the transmission duration of a packet of data and T_{A} the

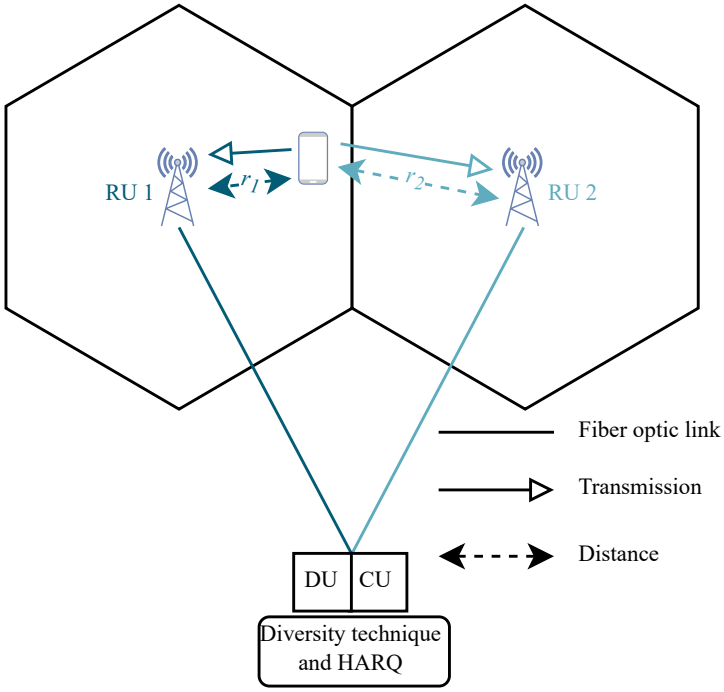


Figure 3.3: Architecture B: I receiving BSs ($I = 2$).

transmission duration of the ACK/NACK. We consider a perfect DL, i.e., ACKs and NACKs are always successfully received.

We define θ as the propagation delay between an RU and the CU that it is connected to:

$$\theta = \frac{\rho}{v}, \tag{3.5}$$

where ρ is the equal distance between the CU and the RUs connected to it, and v is the velocity over a fiber. Then, θ is a constant for all RUs connected to the same CU.

In the following, the number of transmissions is denoted by l. In other words, there are $l - 1$ failed transmissions and then a successful transmission.

For architecture A, we define the cycle duration in both cases: good and bad decoding. In Figure 3.4, $d_{\mathrm{A,f}}$ denotes the delay of one cycle in which the decoding fails:

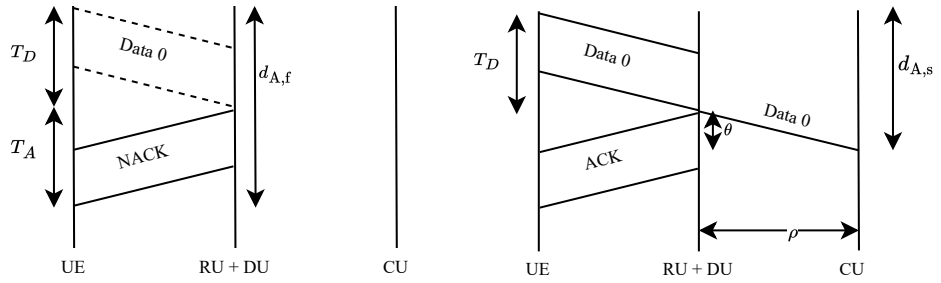
$$d_{\mathrm{A,f}} = T_{\mathrm{D}} + T_{\mathrm{A}}. \tag{3.6}$$

Let $d_{\mathrm{A,s}}$ be the delay of 1 cycle during which the decoding succeeds:

$$d_{\mathrm{A,s}} = T_{\mathrm{D}} + \theta. \tag{3.7}$$

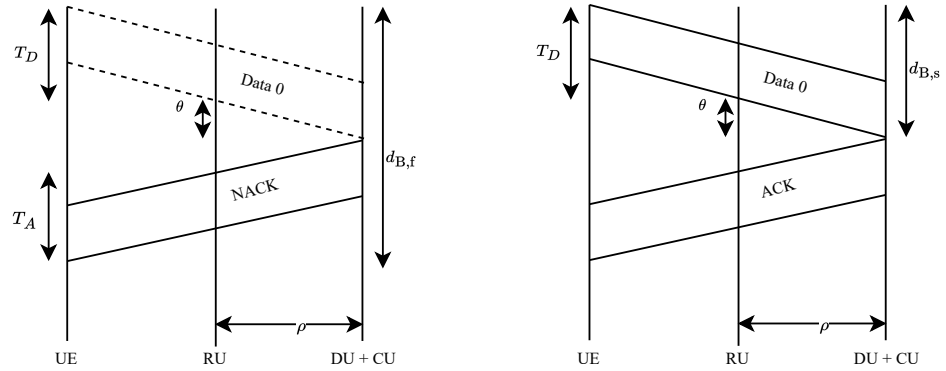
Therefore, the total delay produced by l transmissions is:

$$d_{\mathrm{A}} = (l-1)d_{\mathrm{A,f}} + d_{\mathrm{A,s}}. \tag{3.8}$$



(a) Failure case, architecture A (b) Success case, architecture A

Figure 3.4: Architecture A one cycle delay.



(a) Failure case, architecture B (b) Success case, architecture B

Figure 3.5: Architecture B one cycle delay.

The cycle delay for successful and erroneous decoding for architecture B is shown in Figure 3.5. The delays illustrated in this figure are considered between a UE and one BS (BS_i). Let $d_{B,f}$ and $d_{B,s}$ denote the delay of one cycle with erroneous and successful decoding, respectively. Then we have:

$$d_{B,f} = T_D + 2\theta + T_A, \tag{3.9}$$

and

$$d_{B,s} = T_D + \theta. \tag{3.10}$$

The total delay generated by l transmissions in architecture B is:

$$d_B = (l-1)d_{B,f} + d_{B,s}. \tag{3.11}$$

3.5 Analytic formulation

The delay detailed in Section 3.4 is a function of the number of transmissions l. Consequently, the distribution of l is needed to know the delay and its distribution.

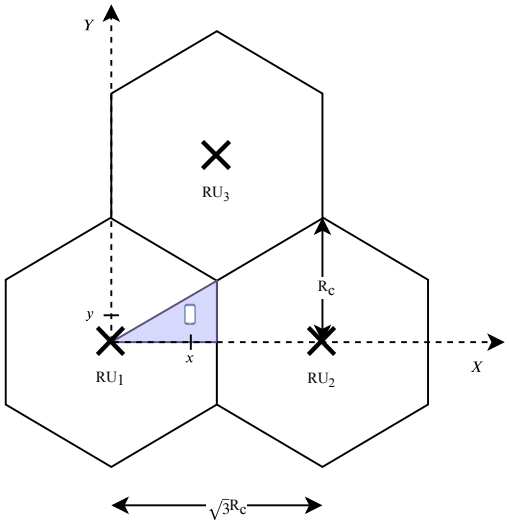


Figure 3.6: Zone of study in the hexagonal cell.

The UE position is uniformly distributed in a hexagonal cell of radius R_c. The RUs are at the centers of the cells. In [Lagrange 2010], the probability of having more than k transmissions as a function of $\overline{\gamma}$ was computed:

$$\mathbb{P}(l > k/\overline{\gamma}) = \Gamma_l(k, X) + e^{-X} \sum_{i=0}^{k-1} \frac{(X)^i}{i!} \frac{\Gamma(\frac{1}{g\overline{\gamma}})}{(g\overline{\gamma})^{k-i+1}\Gamma(\frac{1}{g\overline{\gamma}} + k - i + 1)}, \tag{3.12}$$

where $\mathbb{P}(l > k/\overline{\gamma})$ is the probability of having more than k transmissions for a given $\overline{\gamma}$ in a fading channel, $X = \frac{\gamma_M}{\overline{\gamma}}$, γ_M, a, g parameters that depend on the MCS mode (Table 3.1) [Liu *et al.* 2004], and $\Gamma_l(a, x)$ the lower incomplete normalized gamma function. The Probability Mass Function (PMF) of the number of transmissions l for a given $\overline{\gamma}$ is obtained by:

$$\mathbb{P}(l = k/\overline{\gamma}) = \mathbb{P}(l > k - 1/\overline{\gamma}) - \mathbb{P}(l > k/\overline{\gamma}). \tag{3.13}$$

For a fixed transmission power P_t and noise N, $\overline{\gamma}$ varies with the distance r between the transmitter and the receiver (c.f. (3.3)).

For symmetry reasons, our study is limited to the shadowed triangle of Figure 3.6. In a cell, we consider a UE at the position (x, y) relative to the RU in the center of the cell. For instance, in Figure 3.6, the represented user is in position (x, y) related to RU_1. The x position of the UE is between 0 and $\frac{\sqrt{3}}{2}R_c$. The y position is between 0 and $\frac{x}{\sqrt{3}}$ in the considered zone of study.

3.5.1 Architecture A

In architecture A, the BS in the same cell as the UE is the one receiving UL transmissions. The UE in position (x, y) relative to the cell's BS has thus the following

average SNR:

$$\overline{\gamma}(x,y) = \frac{P_{\mathrm{t}}}{N}\left(\frac{r_0}{\sqrt{x^2+y^2}}\right)^{\alpha}. \tag{3.14}$$

For each UE, we can get the Complementary Cumulative Distribution Function (CCDF) of the number of transmissions l from (3.12). During transmission number k, the probability of the BS erroneously receiving the packet is equal to the probability of needing more than k transmissions. In order to get a total error probability distribution, the CCDF of l, depending on the position of the user in the cell, should be averaged over all the studied surface:

$$\mathbb{P}(l>k) = \frac{8}{\sqrt{3}R_{\mathrm{c}}^2}\int_0^{\frac{\sqrt{3}}{2}R_{\mathrm{c}}}\int_0^{\frac{x}{\sqrt{3}}} \mathbb{P}\left(l>k/\overline{\gamma}(x,y)\right) dydx. \tag{3.15}$$

The PMF of the number of transmissions l is derived using (3.13).

3.5.2 Architecture B

In architecture B, the UE is received by the I nearest surrounding BSs. The index of a BS is represented by i, and it increases with the BS distance from the UE. For example, when $I=2$, the UE is received by RU_1 the nearest RU and RU_2 the second nearest RU (see Figure 3.3). The represented UE is in position (x,y) relative to RU_1, and in position $(\sqrt{3}R_{\mathrm{c}}-x,y)$ relative to RU_2. The average received SNR at RU_1 is $\overline{\gamma}(x,y)$, and at RU_2 is $\overline{\gamma}(\sqrt{3}R_{\mathrm{c}}-x,y)$, and can be computed using (3.14).

We consider that there is an error during the kth transmission if the I RUs fail to decode the packet. In this case, the probability of having more than k transmissions is the probability of having a failed kth transmission at the I RUs. So, the probability of having more than k transmissions is:

$$\mathbb{P}(l>k) = \frac{8}{\sqrt{3}R_{\mathrm{c}}^2}\int_0^{\frac{\sqrt{3}}{2}R_{\mathrm{c}}}\int_0^{\frac{x}{\sqrt{3}}} \prod_{i=1}^{I} A\left(x_i(x),y_i(x)\right) dydx, \tag{3.16}$$

where $(x_i(x),y_i(x))$ is the position of the considered UE relative to RU_i as a function of x and y, and $A\left(x_i(x),y_i(x)\right) = \mathbb{P}\left(l>k/\overline{\gamma}\left(x_i(x),y_i(x)\right)\right)$. For instance, when $I=2$ we get:

$$\mathbb{P}(l>k) = \frac{8}{\sqrt{3}R_{\mathrm{c}}^2}\int_0^{\frac{\sqrt{3}}{2}R_{\mathrm{c}}}\int_0^{\frac{x}{\sqrt{3}}} A(x,y)A(\sqrt{3}R_{\mathrm{c}}-x,y)dydx. \tag{3.17}$$

For $I=3$, the UE is in position $(\frac{\sqrt{3}R_{\mathrm{c}}}{2}-x,\frac{3R_{\mathrm{c}}}{2}-y)$ relative to the third nearest RU. From (3.16) we get:

$$\begin{aligned}\mathbb{P}(l>k) =& \frac{8}{\sqrt{3}R_{\mathrm{c}}^2}\int_0^{\frac{\sqrt{3}}{2}R_{\mathrm{c}}}\int_0^{\frac{x}{\sqrt{3}}} A(x,y)\times \\ & A(\sqrt{3}R_{\mathrm{c}}-x,y)A\left(\frac{\sqrt{3}R_{\mathrm{c}}}{2}-x,\frac{3R_{\mathrm{c}}}{2}-y\right) dydx.\end{aligned} \tag{3.18}$$

3.6 Flexible switch between architectures A and B

After presenting both architectures, we propose to dynamically switch between them. Since the propagation model is distance dependent, when the UE is too close to the first RU, there is no need to let the second RU receive from it. In this case, architecture A and the single reception mode are considered. On the other side, when the UE is near the cell edge, the other surrounding RUs can receive the UE. Thus, we switch to architecture B, and the multiple receptions mode is enabled. We define R_{th} as the radius of the limit zone outside which we allow surrounding RUs to receive along with the nearest RU using architecture B (Figure 3.7). Architecture A and single reception are enabled if the UE is inside the mentioned zone. The flexible functional split does this switch (c.f. the adaptive split in [Alfadhli *et al.* 2018]). The CU measures x and chooses the adopted architecture accordingly. For each transmission to/from the considered UE, each unit encodes/decodes the data until the corresponding layer (dashed lines in Figure 3.8) and transmits it to the other unit. The switching algorithm is the following:

Algorithm 1 Flexible switch between A and B

if $0 < x < \frac{\sqrt{3}}{2} R_{\text{th}}$ **then**
  consider architecture A
else
  if $x < \frac{\sqrt{3}}{2} R_c$ **then**
    consider architecture B
    enable multiple receptions
  end if
end if

The total error probability i.e., needing more than k transmissions during the kth one, when considering the flexible split is given by:

$$\begin{aligned}\mathbb{P}(l > k) = &\frac{8}{\sqrt{3}R_{\text{c}}^2}\left(\int_0^{\frac{\sqrt{3}}{2}R_{\text{th}}} \int_0^{\frac{x}{\sqrt{3}}} A(x,y)dydx + \right. \\ &\left. \int_{\frac{\sqrt{3}}{2}R_{\text{th}}}^{\frac{\sqrt{3}}{2}R_c} \int_0^{\frac{x}{\sqrt{3}}} \prod_{i=1}^{I} A\left(x_i(x), y_i(x)\right) dydx \right).\end{aligned} \tag{3.19}$$

For the flexible split with $I = 2$, we get:

$$\begin{aligned}\mathbb{P}(l > k) = &\frac{8}{\sqrt{3}R_{\text{c}}^2}\left(\int_0^{\frac{\sqrt{3}}{2}R_{\text{th}}} \int_0^{\frac{x}{\sqrt{3}}} A(x,y)dydx + \right. \\ &\left. \int_{\frac{\sqrt{3}}{2}R_{\text{th}}}^{\frac{\sqrt{3}}{2}R_c} \int_0^{\frac{x}{\sqrt{3}}} A(x,y)A(\sqrt{3}R_{\text{c}} - x, y)dydx \right).\end{aligned} \tag{3.20}$$

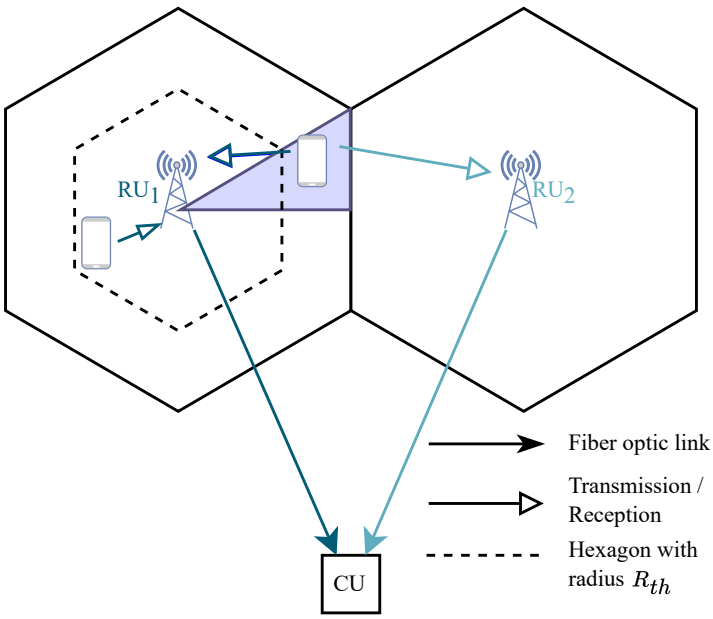


Figure 3.7: Zone delimitation to switch between A and B.

3.7 Simulations

The analytic evaluation of the error probability is accompanied by Monte-Carlo simulations using Matlab. Our simulations follow the analytic steps. The propagation and error models from (3.1) and (3.4) are used.

During each simulation, 100 000 user positions were uniformly picked in the shadowed area of Figure 3.6. The received SNR was computed according to the distance from pre-fixed BSs at the center of each cell. The SNR was then used to evaluate the PER according to (3.4).

During the first transmission, the corresponding SNR is evaluated for architec-

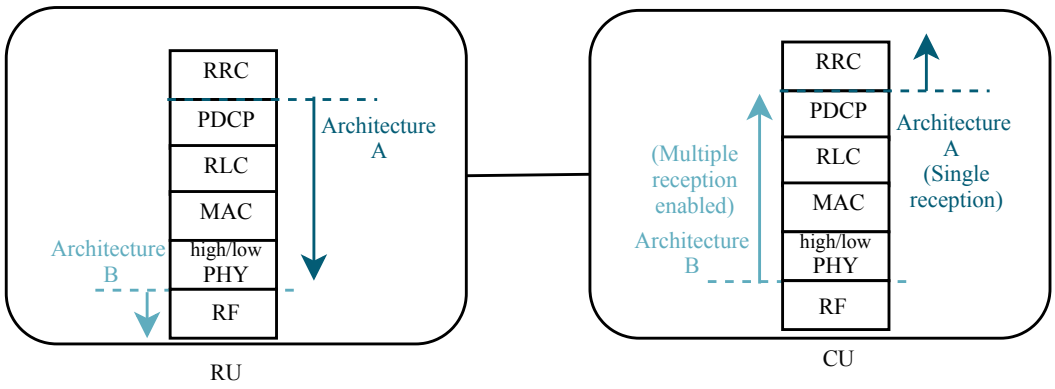


Figure 3.8: Flexible split.

Table 3.2: Parameters values.

Symbol	Parameter	Values
a [Liu *et al.* 2004]	Parameters depending	274.7
g [Liu *et al.* 2004]	on the MCS	7.993
I	Number of receiving BSs	1 for A, 2/3 for B
N (dBm)	Noise power	-116.45
P_{t} (dBm)	UE's transmission power	23
r_0 (m)	Reference distance	0.2
R_{c} (km)	Cell radius	3.2
T_{D} (ms)	Data transmission duration	1 TTI=0.25 [a]
T_{A} (ms)	ACK/NACK transmission time	1 TTI=0.25 [a]
v(m/s)	Velocity over the fiber link	2×10^8
α	Path-loss exponent	3.38
ρ (Km)	CU-RUs distance	3.5

[a] Numerology 2 of the 5G New Radio (NR)[TS 38.211 2022].

ture A, and the PER is calculated. For the second and subsequent transmissions, the SNR used to evaluate the PER on transmission l is the sum of the SNR perceived during the lth transmission, and all the $l-1$ SNRs previously perceived at the BS. For architecture B, during the first transmission, the considered SNR to evaluate the PER is the maximum received SNR at the I RUs. If the BS with the highest SNR does not well receive the UE, then none of the I BSs will do. For transmission number 2 and beyond, the maximum SNR is added to the previously received SNRs to evaluate the PER. The computed PER per UE is then averaged over the picked positions.

The process is repeated 1000 times, and the result is the average of all the simulation results.

3.8 Results and discussion

For the results, the numerical integration method is used to compute the probabilities in equations (3.15), (3.16), and (3.19). The simulation and calculation parameters are summarized in Table 3.2. The cell radius was chosen so that the perceived SNR at the cell edge for a UE in position $\left(\frac{\sqrt{3}R_{\mathrm{c}}}{2}, \frac{R_{\mathrm{c}}}{2}\right)$ is equal to γ_M.

3.8.1 Comparison of the two architectures

Figure 3.9 shows the similarity between the mathematical computation and the simulation results for the PMF distribution of the number of transmissions $\mathbb{P}(l=k)$ (see Table 3.3). It also compares the single reception with architecture A to multiple receptions with architecture B. It shows that for the case of two receiving BSs, we have a higher chance of getting successful decoding from the first transmission. For

further transmissions, the probability is lower for the case of two BSs. Thereby, in architecture B, the number of transmissions needed to get successful decoding is reduced. No significant improvement is observed for receiving from a third BS. This is because the third BS is too far. For this reason, we limit our study to $I = 2$. The

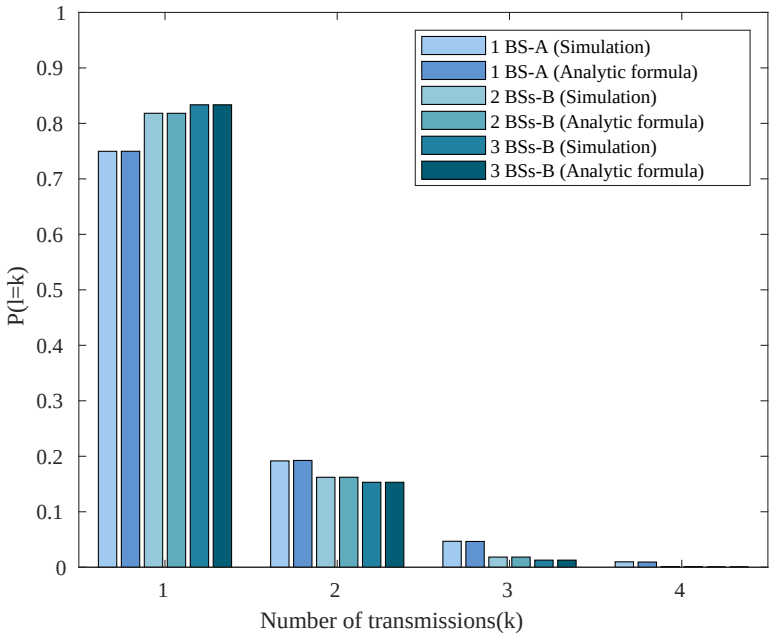


Figure 3.9: PMF of the number of transmissions for architectures A and B.

delay depends on l the number of transmissions. We can get the delay's distribution based on the distribution of l. Figure 3.10 illustrates the CCDF of the delay.

We get the latency distribution with the number of transmissions with a certain probability. If we define reliability as receiving the data successfully within a delay, this CCDF represents the outage probability. For example, if we want to assure a delay of less than 2 ms, we have higher reliability by a factor of 850 in the case of two BSs with architecture B compared to one BS with architecture A. This improvement is larger if we allow higher delays. If we desire to get an outage probability of 10^{-6}, we can see in Figure 3.10 that we can get lower latency with architecture B (2.9 ms for architecture B compared to 3.8 ms latency for architecture A).

For the case of one receiving BS with architecture A, when the data is poorly received, it is treated in the RU-DU, and a re-transmission is performed locally on the tower site. While for the second case, for each transmission, the data should go to the DU-CU. The transmission's Round Trip Time (RTT) in architecture B is longer. Nevertheless, due to the diversity gain, fewer transmissions are needed. Consequently, we can see that we get shorter delays using architecture B, even with longer RTT.

However, we can notice that distancing the CU from the RUs can add more delay. This additional delay affects architecture B more than architecture A (which

Table 3.3: PMF of the number of transmissions for architectures A and B with simulations confidence margins.

Number of transmissions (k)	PMF (Analytic)	PMF (Simulations average)	95% confidence margin
One BS-A			
1	0.7498	0.7497	$[0.7458, 0.7539]$
2	0.1925	0.1916	$[0.1888, 0.1961]$
3	0.0465	0.0468	$[0.0446, 0.0483]$
Two BSs-B			
1	0.8182	0.8182	$[0.8160, 0.8205]$
2	0.1622	0.1621	$[0.1601, 0.1643]$
3	0.0184	0.0183	$[0.0175, 0.0190]$
Three BSs-B			
1	0.8335	0.8336	$[0.8313, 0.8358]$
2	0.1532	0.1531	$[0.1510, 0.1553]$
3	0.0127	0.0128	$[0.0122, 0.0134]$

is seen in (3.8) and (3.11)). This leads the delay's CCDF of architecture B to approach the CCDF of architecture A. The difference between the two delays (between architectures A and B) at 10^{-6} PER is 0.83 ms. If we let $\rho = 20$ km, (3.8) and (3.11) give the same results for the mentioned PER. A distance higher than 20 km increases more the delay in architecture B. That way, we observe higher delays in architecture B compared to architecture A for the same PER.

3.8.2 Flexible C-RAN architecture

To evaluate the flexible split between architectures A and B, Figure 3.11 illustrates the distribution of the number of transmissions $\mathbb{P}(l = k)$ for different R_{th}. We note that $R_{\text{th}} = R_{\text{c}}$ is equivalent to the case of adopting architecture A with one receiving BS and $R_{\text{th}} = 0$ corresponds to architecture B with two receiving BSs. The threshold is chosen when we start getting the same performance as receiving from two BSs. Accordingly, in our case, we can choose $R_{\text{th}} = 0.6R_{\text{c}}$. This is also shown in Figure 3.12, where we have the same performances for the case of $R_{\text{th}} = 0.6R_{\text{c}}$ and $R_{\text{th}} = 0$. The delays achieved by the switching algorithm are slightly lower than the case where architecture B is only implemented (2.34 ms and 2.4 ms respectively for 10^{-5} PER). The maximum observed difference is 0.13 ms. By flexibly splitting the BS's functions, we save approximately 40% of the unnecessary use of the second RU. This percentage is obtained by dividing the area of the hexagon of radius R_{c} by the area of that of radius R_{th}.

Nevertheless, we can also choose $R_{\text{th}} = 0.7R_{\text{c}}$, since the difference seen between

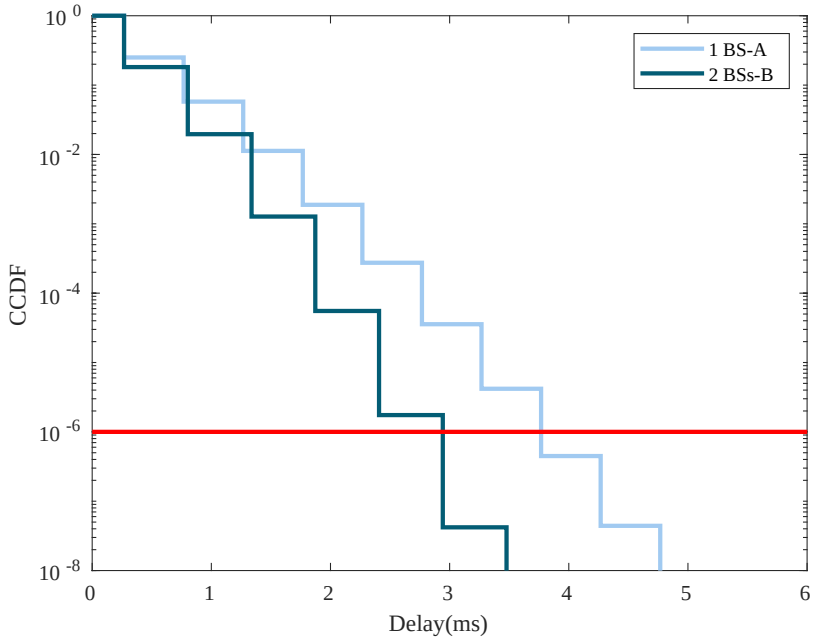


Figure 3.10: Delay CCDF for architecture A (with one BS) and architecture B (with two BSs).

the flexible split at $0.7R_c$ and $R_{th} = 0$ is not significant. We benefit in this case from an additional use reduction of the second BS that becomes approximately 50%.

3.9 Conclusion

In this chapter, we took two different C-RAN architectures A and B. In the first one, we chose the re-transmissions to be triggered in the RU-DU and we adopted a single reception. For the second one, we centralized the HARQ in the DU-CU and adopted multiple receptions.

This chapter aimed to put in evidence the importance of centralization of the error correction mechanism HARQ for URLLC applications. It was shown that the centralization of the MAC layer leads to longer transmissions and re-transmissions RTT. However, by enabling macro-diversity, allowing multiple receptions, and centralized combining technique, we got better decoding probabilities and thus lower delays.

We proposed to switch between architectures A and B dynamically. We proved that by enabling the reception from two BSs just when needed, we could reach the aimed target: lower latency and higher reliability while saving the use of far away BSs.

We consider a more realistic channel model introducing the shadowing effect in the next chapter. We add a third architecture to the comparison and the Maximum Ratio Combining (MRC) as a combining technique.

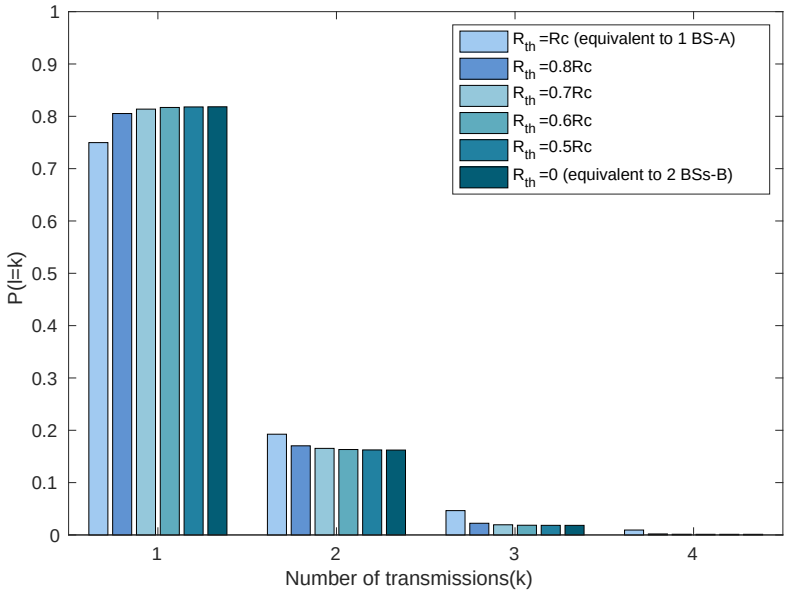


Figure 3.11: Probability distribution of the number of transmissions for different R_{th}.

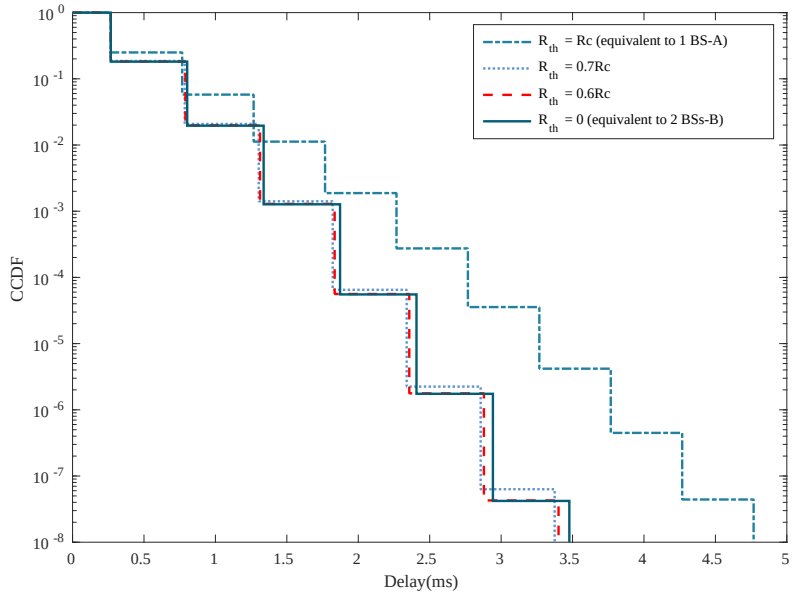


Figure 3.12: Delay CCDF with different R_{th}.

Chapter 4

Macro-diversity techniques for latency and reliability

Contents

4.1 Introduction

In the previous chapter, we did a pilot study to check the impact of two Centralized-RAN (C-RAN) architectures on latency and reliability. The results showed that centralizing the Hybrid Automatic Repeat reQuest (HARQ) mechanism and taking advantage of centralized processing for macro-diversity is beneficial for achieving high reliability and low latency. The promising results led us to investigate in the same direction by refining our work. This chapter compares three Radio Access Network (RAN) architectures, illustrated in Figure 4.1. In the previous chapter, we only assumed a distance-based path loss. This chapter considers a more realistic channel model that includes shadowing. The impact of shadowing is essential because the nearest Base Station (BS) is not always the best receiving BS. The rate on

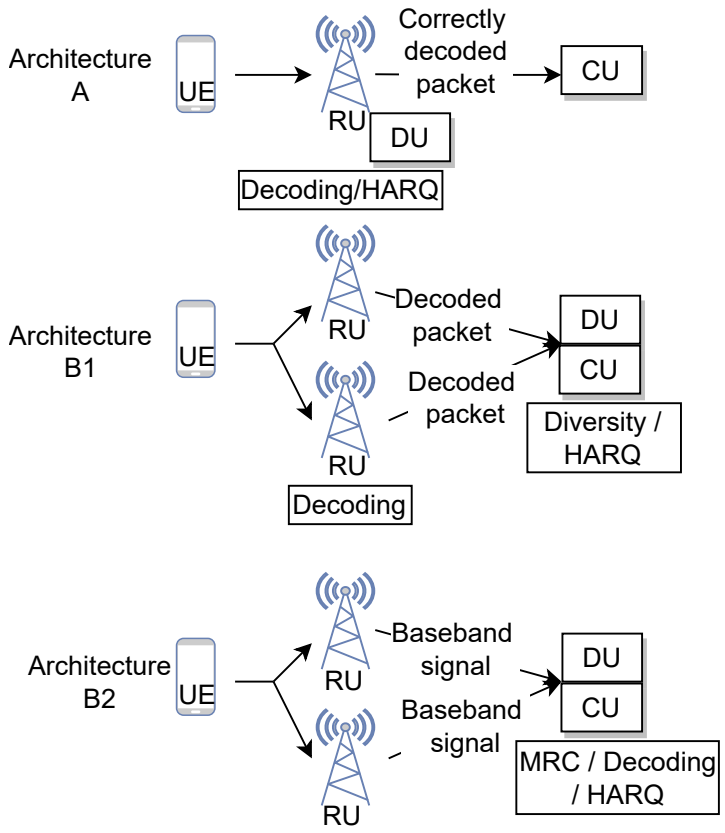


Figure 4.1: Architectures A, B1, and B2.

the fronthaul is considered finite in this study, unlike in the previous chapter, where it was considered infinite. As a matter of clarity, if the same model or equation as the previous chapter appears in this chapter, it will not be represented again. For any change, a new model or equation will be presented.

4.2 System model

Similarly to the previous chapter, we consider transmissions of data packets in the UpLink (UL) direction. We consider the same network model described in Section 3.3.1. The DownLink (DL) is considered error-free: no losses on the feedback channel. The error model is computed from (3.4) as a function of the Signal to Noise Ratio (SNR). HARQ with Chase Combining (HARQ-CC) is also used for error correction. The User Equipment (UE) transmits a packet. The receiver replies with an ACKnowledgement (ACK) if the packet is correctly decoded. Otherwise, a Negative ACKnowledgement (NACK) is sent. In the case of erroneous decoding, another round of HARQ starts again. At the receiver, all the replicas of the packets are saved and combined with the new receptions for decoding. The process is repeated until decoding is successful.

4.2.1 Propagation model

Three factors influence path loss in this chapter: distance-related path loss, multipath loss or fading, and obstacles loss or shadowing. We consider a correlated shadowing that consists of two parts: a common part and a specific part. The common part is the effect of common obstacles between the transmitter and the receiver, and the specific part counts the effect of obstacles specific to a transmitter or a receiver. When considering UL transmissions, shadowing between a UE and different BSs are correlated due to the obstacles near the UE. The common shadowing part is between a UE and all the BSs and the specific part changes between the UE and each BS. The propagation model is COST-231 Hata [TR 36.942 2020], and the received power is given by:

$$P_{\mathrm{r}} = P_{\mathrm{t}} \left(\frac{r_0}{r}\right)^{\alpha} e^{\xi_c + \xi_s} \chi, \tag{4.1}$$

where P_{r} is the received power, P_{t} the transmitted power, r_0 a constant reference distance, r the distance between the UE and the BS, α the path-loss exponent, χ an exponential random variable (r.v.) representing fading with mean = 1, ξ_c and ξ_s two normal r.v. representing the common and the specific part of the correlated shadowing, respectively: $\xi_c \sim N(0, \sigma_c^2)$, with $\sigma_c = \sigma_{c_{\mathrm{dB}}} \ln(10)/10$, and $\xi_s \sim N(0, \sigma_s^2)$, with $\sigma_s = \sigma_{s_{\mathrm{dB}}} \ln(10)/10$. The shadowing correlation coefficient is $\delta = \frac{\sigma^2_{c_{\mathrm{dB}}}}{\sigma^2_{\mathrm{dB}}}$, with $\sigma^2_{\mathrm{dB}} = \sigma^2_{c_{\mathrm{dB}}} + \sigma^2_{s\mathrm{dB}}$. We consider two connection types between the UE and the serving Radio Unit (RU). In the first one, the UE is connected to the nearest RU. This connection is not the optimal one due to the shadowing effect. In the second one, the UE is connected to the best RU, which is the case of an ideal network. In fact, due to the hysteresis effect, the UE is not always connected to the best BS.

4.2.2 Architecture overview

The three architectures laid out in this section are illustrated in Figures 4.1 and 4.2. They involve a single transmission and I receptions. For architecture A, only one RU receives the signal transmitted by the UE: $I = 1$. For architectures B1 and B2, we have multiple receptions: $I > 1$. We assume that data packets are processed at the Centralized Unit (CU), which for example includes mobile edge computing functions.

In all the architectures we focus mainly on the location of the Media Access Control (MAC) layer and on the location of the decoding process. The MAC layer contains the HARQ-CC process for error correction. Architectures A, B1, and B2 are summarized in Table 4.1.

In architecture A, all the processing is done on the tower site. The Distributed Unit (DU) is considered implemented with the RU on the cell site. This is split A in [ecp 2019]. The RU-DU receives the packet, decodes it, and stores a version. If the decoding is successful, the packet is sent to the CU. In case of error, the HARQ-CC mechanism is used. A re-transmission is triggered by the RU-DU since this is where

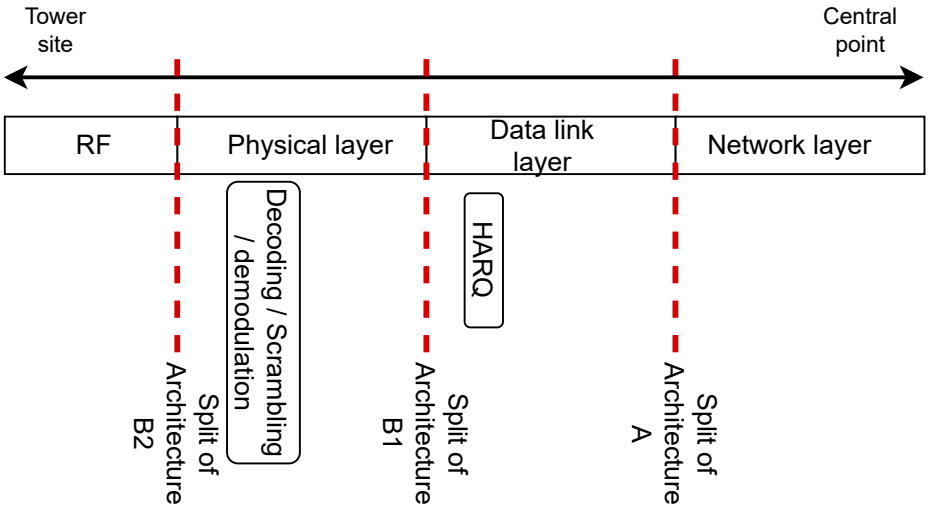


Figure 4.2: Functional splits of architectures A, B1, and B2.

the MAC layer is implemented. The re-transmitted packet is combined with its previously received and stored versions during previously failed transmissions. If a failure occurs again, another re-transmission is requested until successfully decoding the packet.

In architecture B1, each RU on the tower site receiving the transmitted packet decodes it and forwards it to the central point where the DU is implemented with the CU. A version of the received packet is stored to be combined with upcoming transmissions in case of an error. This is split D in [ecp 2019]. In the DU-CU, the redundancy of different decoded packets from different RUs is removed. An error happens if the I RUs fail to correctly decode the packet. That means if none of the received decoded packets at the DU-CU is correct. In case of error, the MAC layer in the DU-CU asks for re-transmission.

In architecture B2, each RU receiving the transmitted packet forwards it to the central point where the DU is implemented with the CU. Decoding and HARQ-CC are centralized and performed by the DU-CU in this architecture. This is the traditional Functional Split (FS) detailed in [Duan *et al.* 2016]. In the DU-CU, the signals are combined by the Maximum Ratio Combining (MRC) technique. The SNR of the I signals are summed. Then, the decoding process takes place. If an error occurs, a re-transmission is initiated from the CU. Each received version of the signal is stored to be combined with subsequent receptions in case of re-transmissions.

Table 4.1: Architectures summary.

Element/Architecture	A	B1	B2
MAC layer location	RU-DU	DU-CU	DU-CU
Decoding location	RU	RU	DU-CU
Functional split [Larsen *et al.* 2019]	Option 1	Option 6	Option 8
Transmission/reception type	Single transmission Single reception	Single transmission I receptions	
Combining technique	None	At least one good reception	MRC

4.3 Delay components

In this section, we provide the delay components for architectures A, B1, and B2. The calculated delay includes propagation and transmission duration over the radio interface and over the fronthaul until the CU. The processing delay is not taken into consideration.

4.3.1 Architecture A

We define cycle duration in both cases: good and bad decoding. When the receiving RU-DU fails to correctly decode the packet, a re-transmission is triggered locally on-site. The RU-DU transmits an NACK to the UE. In Figure 4.3, $d_{\mathrm{A,f}}$ denotes the delay of one cycle with bad decoding in architecture A:

$$d_{\mathrm{A,f}} = T_{\mathrm{D,R}} + T_{\mathrm{A,R}} + 2\frac{r}{c}, \tag{4.2}$$

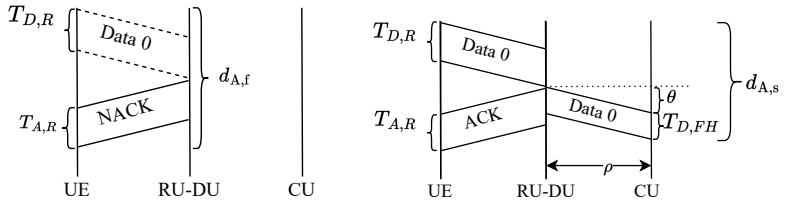
where $T_{\mathrm{D,R}}$ and $T_{\mathrm{A,R}}$ are data and ACK/NACK transmission delay, respectively, over the radio interface, and r/c the propagation delay over the radio interface. When the reception and decoding processes are successful, the successfully decoded packet is transmitted to the CU. For good reception, the delay of one cycle is:

$$d_{\mathrm{A,s}} = T_{\mathrm{D,R}} + \frac{r}{c} + T_{\mathrm{D,FH}} + \theta, \tag{4.3}$$

where $T_{\mathrm{D,FH}}$ is the data transmission duration over the fronthaul. In architecture A, a correctly decoded packet in the RU is transmitted over the fronthaul towards the CU. Thus, the transmission over the fronthaul duration is related to the packet size L_{P} and its corresponding headers size:

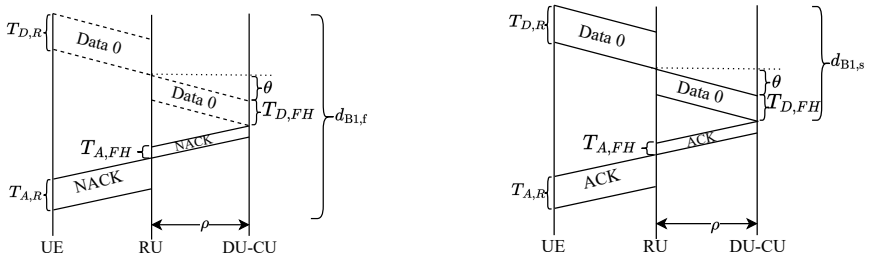
$$T_{\mathrm{D,FH}} = \frac{L_{\mathrm{P}} + L_{\mathrm{H_{eCPRI}}} + L_{\mathrm{H_{TN}}}}{C_{\mathrm{FH}}}, \tag{4.4}$$

where $L_{\mathrm{H_{eCPRI}}}$ is the ethernet-based CPRI (eCPRI) header size, $L_{\mathrm{H_{TN}}}$ the transport-network layer headers size, and C_{FH} the maximum throughput over the fronthaul.



(a) Failure case (b) Success case

Figure 4.3: Architecture A one cycle delay.



(a) Failure case (b) Success case

Figure 4.4: Architecture B1 one cycle delay.

So, the total delay produced by l transmissions ($l-1$ failed and one successful), for architecture A is:

$$d_{\mathrm{A}} = (l-1)d_{\mathrm{A,f}} + d_{\mathrm{A,s}}. \tag{4.5}$$

4.3.2 Architecture B1

The cycle delays for architecture B1 are shown in Figure 4.4. In this architecture, each received packet at the I RUs is decoded and transmitted to the DU-CU. When none of the RUs succeed to have a correctly decoded packet, the DU-CU transmits an NACK to the UE asking for a re-transmission. Parameter $d_{\mathrm{B1,f}}$ denotes the delay of one cycle with bad reception in architecture B1:

$$d_{\mathrm{B1,f}} = T_{\mathrm{D,R}} + T_{\mathrm{A,R}} + 2\frac{r}{c} + T_{\mathrm{D,FH}} + T_{\mathrm{A,FH}} + 2\theta. \tag{4.6}$$

When the DU-CU receives at least one correctly decoded packet from the I RUs, the HARQ-CC process is stopped and an ACK is transmitted to the UE. The delay of one cycle with successful decoding $d_{\mathrm{B1,s}}$:

$$d_{\mathrm{B1,s}} = T_{\mathrm{D,R}} + \frac{r}{c} + T_{\mathrm{D,FH}} + \theta. \tag{4.7}$$

Architecture B1 is split D from [ecp 2019]. In this split, each decoded packet is transmitted over the fronthaul. As a response, an ACK or NACK can also be

transmitted over the fronthaul. The transmission on this link depends on the packet size (L_{P}) or the ACK/NACK size (L_{A}) and the length of the corresponding headers. For data packet transmission we have:

$$T_{\mathrm{D,FH}} = \frac{L_{\mathrm{P}} + L_{\mathrm{H_{DLL}}} + L_{\mathrm{H_{eCPRI}}} + L_{\mathrm{H_{TN}}}}{C_{\mathrm{FH}}}, \tag{4.8}$$

where $L_{\mathrm{H_{DLL}}}$ is the data link layer headers size. For an ACK/NACK transmission:

$$T_{\mathrm{A,FH}} = \frac{L_{\mathrm{A}} + L_{\mathrm{H_{DLL}}} + L_{\mathrm{H_{eCPRI}}} + L_{\mathrm{H_{TN}}}}{C_{\mathrm{FH}}}. \tag{4.9}$$

The total propagation and transmission delay caused by $l-1$ failed transmissions and one successful for architecture B1 is:

$$d_{\mathrm{B1}} = (l-1)d_{\mathrm{B1,f}} + d_{\mathrm{B1,s}}. \tag{4.10}$$

4.3.3 Architecture B2

Architecture B2 is split E from [ecp 2019]. The radio signal received on the RU is sampled and quantized. From [Duan *et al.* 2016], we take the calculation of the source rate on the fronthaul:

$$\mathrm{R_{FH,B2}} = f_{\mathrm{s}} \times N_{\mathrm{IQ}} \times 2 \times F_{\mathrm{os}} \times N_{\mathrm{a}} \times \eta, \tag{4.11}$$

where f_{s} is the sampling frequency, N_{IQ} the number of I and Q bits (multiplied by 2 to cover both I and Q bits), F_{os} the oversampling factor, N_{a} the number of antennas, and η the Common Public Radio Interface (CPRI) Forward Error Correction (FEC) code rate. At each symbol duration, the samples are transmitted over the fronthaul in one eCPRI frame. We define $T_{\mathrm{S,FH}}$ the transmission duration of one sampled and quantized Orthogonal Frequency Division Multiplexing (OFDM) symbol over the fronthaul:

$$T_{\mathrm{S,FH}} = T_{\mathrm{symb}} + T_{\mathrm{CP}} + \frac{\mathrm{R_{FH,B2}}(T_{\mathrm{symb}} + T_{\mathrm{CP}}) + L_{\mathrm{H_{eCPRI}}} + L_{\mathrm{H_{TN}}}}{C_{\mathrm{FH}}}, \tag{4.12}$$

where T_{symb} is one symbol duration and T_{CP} the cyclic prefix duration. The cycles duration are represented in Figure 4.5. When the CU-DU receives the I signals, it combines them using MRC. If the decoding after the combining is not successful, an NACK is transmitted from the central point to the UE. The delay produced during a bad reception for architecture B2 is:

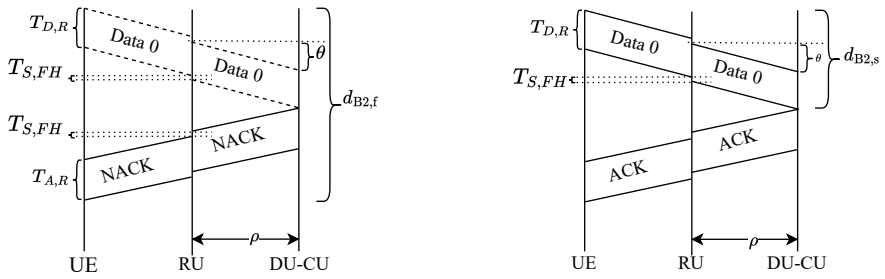
$$d_{\mathrm{B2,f}} = T_{\mathrm{D,R}} + T_{\mathrm{A,R}} + 2\frac{r}{c} + 2T_{\mathrm{S,FH}} + 2\theta. \tag{4.13}$$

When the DU-CU succeeds in decoding the packet after the MRC combining, the HARQ-CC process is stopped and an ACK is transmitted to the UE. The delay generated during a successful transmission in architecture B2 is:

$$d_{\mathrm{B2,s}} = T_{\mathrm{D,R}} + \frac{r}{c} + T_{\mathrm{S,FH}} + \theta. \tag{4.14}$$

The total delay caused by $l-1$ failed transmissions and one successful, for architecture B2 is:

$$d_{\mathrm{B2}} = (l-1)d_{\mathrm{B2,f}} + d_{\mathrm{B2,s}}. \tag{4.15}$$



(a) Failure case

(b) Success case

Figure 4.5: Architecture B2 one cycle delay.

4.4 Analytic formulation

The symmetry of the hexagonal shape allows us to restrict our study to the highlighted area in Figure 3.6. We consider that a UE is in the position (x_i, y_i) relative to an RU_i. We take (x_i, y_i) as a function of (x, y) representing the coordinates relative to RU_1 coexisting in the same cell with the UE. That means that $(x_1, y_1) = (x, y)$.

We are interested in finding the distribution of the number of transmissions which is an r.v. affecting delay. More than k transmissions are needed if the kth transmission is erroneous. Similarly to the previous chapter, from reference [Lagrange 2010], we use (3.12) the expression of the probability of having a bad kth transmission using HARQ-CC as a function of the average SNR $\overline{\gamma}(x_i, y_i, \xi_c, \xi_s)$:

$$\overline{\gamma}(x_i, y_i, \xi_c, \xi_s) = \frac{P_{\mathrm{t}}}{N} \left(\frac{r_0}{\sqrt{x_i^2 + y_i^2}} \right)^{\alpha} e^{\xi_c + \xi_s}. \tag{4.16}$$

The Probability Mass Function (PMF) of the number of transmissions is then computed using (3.13).

For simplicity, we let $A_i = \mathbb{P}(l > k/\overline{\gamma}(x_i, y_i, \xi_c, \xi_s))$ which is given by (3.12). We use A_i in the remaining equations of the chapter.

4.4.1 Architecture A, nearest BS

In architecture A, a UE is served by one and only one RU. In this section, we assume that each UE is served by the nearest RU. To get the total probability of having a bad reception at the kth try, we average (3.12) over the area in question and over the different shadowing values as well:

$$\mathbb{P}(l > k) = q \int_0^{\frac{\sqrt{3}}{2}R_{\mathrm{c}}} \int_0^{\frac{x}{\sqrt{3}}} \int_{-\infty}^{+\infty} A \frac{1}{\sqrt{\sigma_c^2 + \sigma_s^2}} e^{-\frac{w^2}{2(\sigma_c^2 + \sigma_s^2)}} \mathrm{d}w\mathrm{d}y\mathrm{d}x, \tag{4.17}$$

where $q = \frac{8}{\sqrt{2\pi}\sqrt{3}R_{\mathrm{c}}^2}$, $A = \mathbb{P}\left(l > k/\overline{\gamma}(x, y, w)\right)$, and w represents $\xi_c + \xi_s$.

4.4.2 Architecture A, best BS

When shadowing is considered, the nearest BS is not always the best serving BS. In this section, the UE is assumed to be connected to the best receiving RU among the I nearest RUs. We let RU_i be the best receiving RU with the best received SNR. Assuming constant noise, the best signal can be determined by the highest received power:

$$\mathbb{P}(\text{UE connected to RU}_i) = \mathbb{P}(P_{r,i} > P_{r,j}\ \forall j \neq i), \tag{4.18}$$

where $1 \leq i \leq I$, $1 \leq j \leq I$ and $P_{r,i}$ is the power received by RU_i from the UE when all UEs use the same transmission power. The received powers are independent because the I RUs are not co-located. Thus, we get the probability of being connected to RU_i:

$$\mathbb{P}(\text{UE connected to RU}_i) = \int_{-\infty}^{+\infty} \prod_{j\neq i} \mathbb{P}\left(\xi_{sj} < u + \alpha \ln\left(\frac{\sqrt{x_j^2+y_j^2}}{\sqrt{x_i^2+y_i^2}}\right)\right) \frac{e^{-\frac{u^2}{2\sigma_s^2}}}{\sigma_s\sqrt{2\pi}} \mathrm{d}u, \tag{4.19}$$

where u is the integration variable representing ξ_{si}. To simplify the writing of the upcoming equations, we let $\mathbb{P}_i = \prod_{j\neq i} \mathbb{P}\left(\xi_{sj} < u + \alpha \ln \frac{\sqrt{x_j^2+y_j^2}}{\sqrt{x_i^2+y_i^2}}\right)$. The probability in the previous product represents the Cumulative Distribution Function (CDF) of ξ_{sj} and is calculated as $\mathbb{P}(\xi_{sj} < h) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{h}{\sigma_s\sqrt{2}}\right)\right)$, where $h = u + \alpha \ln \frac{\sqrt{x_j^2+y_j^2}}{\sqrt{x_i^2+y_i^2}}$. The total distribution of the number of transmissions, while being connected to the best BS, is given by:

$$\mathbb{P}(l > k) = \frac{q}{\sqrt{2\pi}} \int_0^{\frac{\sqrt{3}}{2}R_c} \int_0^{\frac{x}{\sqrt{3}}} \sum_{i=1}^{I} \left[\int_{-\infty}^{+\infty} \left(\mathbb{P}_i \int_{-\infty}^{+\infty} \frac{A_i}{\sigma_c} e^{-\frac{v^2}{2\sigma_c^2}} \mathrm{d}v \right) \frac{1}{\sigma_s} e^{-\frac{u^2}{2\sigma_s^2}} \mathrm{d}u \right] \mathrm{d}y\mathrm{d}x, \tag{4.20}$$

where v is the integration variable representing ξ_c.

4.4.3 Architecture B1

For architecture B1, we consider that I RUs are receiving the signal transmitted by the UE. We consider that when all the I RUs experience a bad reception, an error is detected. The different receptions are independent and thus the probability of error is a product of errors occurring on each RU, which is then averaged over the surface in question:

$$\mathbb{P}(l > k) = \frac{q}{\sqrt{2\pi}} \int_0^{\frac{\sqrt{3}}{2}R_c} \int_0^{\frac{x}{\sqrt{3}}} \int_{-\infty}^{+\infty} \left(\prod_{i=1}^{I} \int_{-\infty}^{+\infty} \frac{A_i}{\sigma_s} e^{-\frac{u^2}{2\sigma_s^2}} \mathrm{d}u \right) \frac{1}{\sigma_c} e^{-\frac{v^2}{2\sigma_c^2}} \mathrm{d}v\mathrm{d}y\mathrm{d}x, \tag{4.21}$$

where u and v are the integration variables representing ξ_s and ξ_c, respectively.

4.4.4 Architecture B2, general case

With architecture B2, I RUs receive the signal transmitted by the UE. The different receptions are combined by the MRC technique. During each transmission, the CU processes the sum of I signals received on I RUs. An error is identified when the sum can not be decoded correctly. The total SNR experienced at the CU during the kth transmission is:

$$\gamma_{S,k} = \sum_{l=1}^{I} \gamma_{l,k}, \tag{4.22}$$

where $\gamma_{l,k}$ is the SNR perceived at RU_l during the kth transmission. Since every reception consists now of the sum of I SNRs, (3.12) is not valid anymore. We need to derive the probability of error at the kth transmission. However, repeating the calculation steps done in [Lagrange 2010] is unfeasible for this case. We thus adopt a pure simulation approach for architecture B2 in the general case. However, for a specific position of the terminal, a computation is possible as explained in the next section.

4.4.5 Architecture B2, particular case

In this section, we develop the probability of having an error at the kth transmission for the MRC case with particular considerations that are valid only for the analytic calculation. We consider only two receiving RUs: $I = 2$. We also consider $\overline{\gamma}_1 = \overline{\gamma}_2 = \overline{\gamma}$, with $\overline{\gamma}_i$ being the mean of the exponential r.v. $\gamma_{i,k}$ while communication with RU_i at the kth transmission. During each transmission k, the perceived average SNR is the sum of the received SNRs at the two RUs. Thus, $\gamma_{S,k} = \gamma_{1,k} + \gamma_{2,k}$ is an Erlang r.v. with the following distribution:

$$f_\gamma(\gamma_{S,k}) = \left(\frac{1}{\overline{\gamma}}\right)^2 \gamma_{S,k} e^{-\frac{\gamma_{S,k}}{\overline{\gamma}}}. \tag{4.23}$$

Note that a simulation approach is used for other considerations ($I > 2$ for example). We again use HARQ-CC. During the kth transmission, the SNR used to determine the Packet Error Rate (PER) in (3.4) is:

$$\gamma_{T,k} = \sum_{i=1}^{k} \gamma_{S,i}. \tag{4.24}$$

The PER during the kth transmission is, therefore, $h(\gamma_{T,k})$ (c.f. (3.4)). The probability of having more than k transmissions results from having errors during all of the first k transmissions. So, it depends on all the previous SNRs (all γ_i with $i \leq k$). We consider successive packet transmissions, so the SNR is independent and identically distributed (i.i.d.) for different transmissions. Therefore, the probability of error during the kth transmission is the following:

$$\mathbb{P}(l > k) = \int_0^\infty \ldots \int_0^\infty \prod_{i=1}^{k} h(\gamma_{T,i}) f_\gamma(\gamma_{S,1}) \ldots f_\gamma(\gamma_{S,k}) \mathrm{d}\gamma_{S,1} \ldots \mathrm{d}\gamma_{S,k}. \tag{4.25}$$

Similarly to [Lagrange 2010], we split each integral into two integrals, resulting in:

$$\mathbb{P}(l > k) = B_k + \sum_{l=1}^{k-1} C_{k,l} + D_k. \tag{4.26}$$

After several calculation steps, detailed in Appendix A, we get:

$$B_k = 1 - e^{-\frac{\gamma_M}{\overline{\gamma}}} \sum_{n=0}^{2k-1} \left(\frac{\gamma_M}{\overline{\gamma}}\right)^n \frac{1}{n!}, \tag{4.27}$$

$$\begin{aligned} C_{k,l} =& e^{-\frac{\gamma_M}{\overline{\gamma}}} \frac{1}{(2l+1)!} \left(\frac{\gamma_M}{\overline{\gamma}}\right)^{2l} \prod_{j=l+1}^{k} \frac{1}{(1+g\overline{\gamma}(k+1-j))^2} \\ & \times \left[\gamma_M \left(\frac{1}{\overline{\gamma}} + g(k-l)\right) + 2l + 1\right], \end{aligned} \tag{4.28}$$

$$D_k = e^{-\frac{\gamma_M}{\overline{\gamma}}} \prod_{j=1}^{k} \frac{1}{(1+g\overline{\gamma}(k+1-j))^2} \left[1 + \gamma_M \left(\frac{1}{\overline{\gamma}} + gk\right)\right]. \tag{4.29}$$

Finally, by substituting B_k, $C_{k,l}$, and D_k in (4.26) by (4.27), (4.28), and (4.29) respectively, we get $\mathbb{P}(l > k)$ for the MRC particular case:

$$\mathbb{P}(l > k) = 1 - e^{-G} \sum_{i=0}^{2k-1} \frac{G^i}{i!} + \sum_{l=0}^{k-1} \frac{e^{-G} G^{2l} T_{k,l}}{(2l+1)!} \prod_{j=1}^{k-l} \frac{1}{(1+\overline{\gamma} g j)^2}, \tag{4.30}$$

where $T_{k,l} = \gamma_M \left(\frac{1}{\overline{\gamma}} + g(k-l)\right) + 2l + 1$.

4.5 Results and discussions

We carried out both simulations and computations to check whether they gave the same results. In the same way as the previous chapter, we conducted Monte Carlo simulations. In each simulation, 100 000 users were uniformly distributed in the shadowed area of Figure 3.6. For each user, fading and shadowing were randomly generated. The fading changed with each transmission while the shadowing was considered the same for the same UE for all the transmissions. The PER is given by (3.4). The simulation was iterated 1000 times. The confidence margin is evaluated at 95%. For the results, the numerical integration method is used to compute the probabilities in equations (4.17), (4.20), and (4.21). The simulation and calculation parameters are summarized in Table 4.2. We consider an upper bound for the propagation delay over the radio interface: $\frac{r}{c} = \frac{R_c}{c}$.

Table 4.3 shows the similarity of the results between the simulation and our mathematical computation for the PMF of the number of transmissions. It compares the PMF for three cases: receiving from the nearest BS, receiving from the best BS among the two nearest BSs (architecture A), and receiving from the two nearest BSs (architecture B1). We can see the improvement when the best BS is selected.

Table 4.2: Parameters values.

Symbol	Parameter	Values
a [Liu *et al.* 2004]	Parameter depending on the Modulation and Coding Scheme (MCS)	274.7
c (m/s)	Light velocity	3×10^8
C_{FH} (Gbps)	Fronthaul maximum capacity	100
F_{os}	Over sampling factor	2
f_{s} (Msamples/s)	Sampling frequency	153.6
g [Liu *et al.* 2004]	Parameter depending on the MCS	7.993
L_{A} (Bits)	ACK/NACK length	1
$L_{\text{H}_{\text{DLL}}}$(Bytes)	Data link layer header (SDAP, PDCP, RLC, MAC)	11
$L_{\text{H}_{\text{eCPRI}}}$(Bytes)	eCPRI header	4
$L_{\text{H}_{\text{TN}}}$(Bytes)	Transport-network header (UDP, IP, ethernet)	62
L_{P} (Bytes)	Data packet length	32[TR 38.913 2017]
N (dBm)	Noise power	-116
N_{a}	Number of antennas	1
N_{IQ} (bits)	Number of I and Q bits	16
P_{t} (dBm)	UE's transmission power	23
r_0 (m)	Reference distance	0.2
R_{c} (km)	Cell radius	2.2
$T_{\text{A,R}}$ (ms)	ACK/NACK transmission duration over the radio interface	0.25 [a]
$T_{\text{D,R}}$ (ms)	Data transmission duration over the radio interface	0.25 [a]
$T_{\text{CP}}(\mu s)$	Cyclic prefix duration	1.17 [a]
$T_{\text{symb}}(\mu s)$	Symbol duration	16.67 [a]
α	Path-loss exponent	3.38
δ	Correlation coefficient	0.5
η	CPRI FEC code rate	$\frac{10}{8}$
ρ (Km)	CU-RUs distance	3.5
σ_{dB} (dB)	Shadowing's standard deviation	5

[a] Numerology 2 of the 5G New Radio (NR) [TS 38.211 2022].

Table 4.3: PMF of the number of transmissions for architectures A and B1 with simulations confidence margins.

Number of transmissions (k)	PMF (Analytic)	PMF (Simulations average)	95% confidence margin
Nearest BS-A			
1	0.8805	0.8803	$[0.8774, 0.8832]$
2	0.0925	0.0924	$[0.0900, 0.0950]$
3	0.0183	0.0183	$[0.0172, 0.0194]$
Best BS(2 nearest)-A			
1	0.8976	0.8977	$[0.8958, 0.8995]$
2	0.0844	0.0843	$[0.0826, 0.0860]$
3	0.0135	0.0135	$[0.0129, 0.0142]$
2 BSs-B1			
1	0.9420	0.9420	$[0.9406, 0.9434]$
2	0.0497	0.0497	$[0.0484, 0.0510]$
3	0.0063	0.0062	$[0.0058, 0.0067]$

In such cases, fewer transmissions are needed to get a correct packet. An additional improvement is observed when macro-diversity is used with architecture B1.

Knowing the distribution of the number of transmissions, we can get the distribution of the delay. The Complementary Cumulative Distribution Function (CCDF) of the delay in Figures 4.6 and 4.8 represents the probability of not receiving a good packet within a certain amount of time. In other words, this CCDF represents the PER.

The importance of macro-diversity is highlighted in Figure 4.6. When shadowing is not considered, macro-diversity does not improve as much as when decorrelated shadowing ($\delta = 0$) is considered. For instance, for a PER of 10^{-6}, going from from 2 to 4 receiving BSs reduces 0.55 ms in terms of latency with $\sigma = 0$ dB. Whereas, with 5 dB decorrelated shadowing deviation, we have a reduction of 2.2 ms. When the shadowing is correlated, moving from 2 to 4 receiving BSs does not improve a lot. The improvement is quite similar to the non-shadowed case (see 10^{-5} PER also). For a PER of 10^{-6}, this improvement is a reduction of 1.1 ms. This is due to the fact that when the shadowing is correlated, the performance of the nearest BS approaches the performance of the best one. Nevertheless, it is worth mentioning that spatial diversity, with architecture B1, improves compared to only receiving from the nearest BS, when the shadowing is correlated (Figure 4.8). But, a high diversity order is not required for additional improvement. We get similar analytic and simulation results for the MRC with architecture B2 as shown in Table 4.4. Figure 4.7 shows that receiving from two BSs, whether using the first combining technique or the second one, improves the average number of transmissions. A lower average number

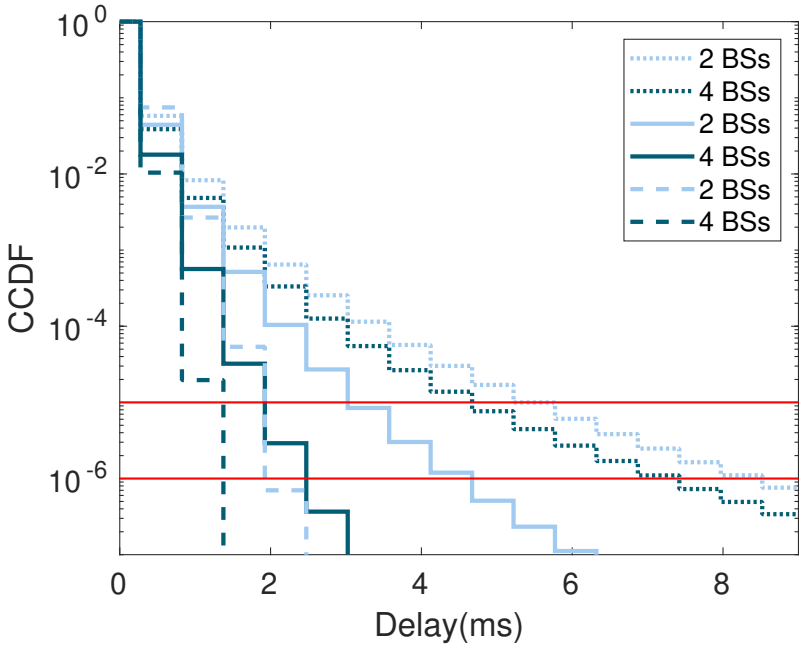


Figure 4.6: Delay CCDF for architecture B1 with $\sigma_{\text{dB}} = 5$ dB correlated shadowing (dotted lines), $\sigma_{\text{dB}} = 5$ dB decorrelated shadowing (continuous lines) and $\sigma_{\text{dB}} = 0$ dB (dashed lines).

of transmissions produces lower average latency. For $\overline{\gamma} = \overline{\gamma}_1 = \overline{\gamma}_2 = -2$ dB, we have an average delay of 0.915 ms for the nearest BS with architecture A, compared to 0.6350 ms and 0.4973 ms with architectures B1 and B2, respectively. So, we notice that MRC outperforms the combining technique used with architecture B1. Nevertheless, when both channels, between the UE and both RUs, are good (high $\overline{\gamma}_1$ and $\overline{\gamma}_2$), B1 and B2 have almost the same performance.

Figure 4.8 compares the three architectures with all the cases under study. The results of architecture B2 shown here are the simulation results. We can see the improvement produced when receiving from the best BS compared to the nearest one. An additional improvement is noticed when receiving form 4 BSs with architecture B1. The improvement rises when using MRC with architecture B2. In fact, summing the signals before decoding increases the chances of good decoding, i.e. high reliability, and decreases the number of re-transmissions, i.e. low latency. The difference appears to be significant for low PERs. If we take a PER of 10^{-2}, the difference between the delay of architecture B1 and architecture B2 is 0.06 ms, with a shorter delay for architecture B1. On the other hand, for an ultra-reliability of 10^{-6}, this difference increases to approximately 3.59 ms, with a shorter delay using architecture B2. So, for error-tolerant applications, architecture B1 is sufficient. However, for Ultra Reliable Low Latency Communications (URLLC) applications, architecture B2 with MRC is better.

Table 4.4: PMF of the number of transmissions and 95% simulation confidence margin for the MRC technique (architecture B2, particular case).

Number of transmissions (k)	**PMF (Analytic)**	**PMF (Simulations average)**	**95% confidence margin**
$\overline{\gamma} = -3$ dB			
1	0.5124	0.5124	$[0.5096, 0.5154]$
2	0.3997	0.4008	$[0.3969, 0.4024]$
3	0.0802	0.0794	$[0.0787, 0.0817]$
$\overline{\gamma} = 0$ dB			
1	0.7985	0.7970	$[0.7961, 0.8008]$
2	0.1907	0.1921	$[0.1885, 0.1930]$
3	0.0105	0.0107	$[0.0100, 0.0110]$
$\overline{\gamma} = 3$ dB			
1	0.9337	0.9338	$[0.9323, 0.9352]$
2	0.0653	0.0653	$[0.0639, 0.0667]$

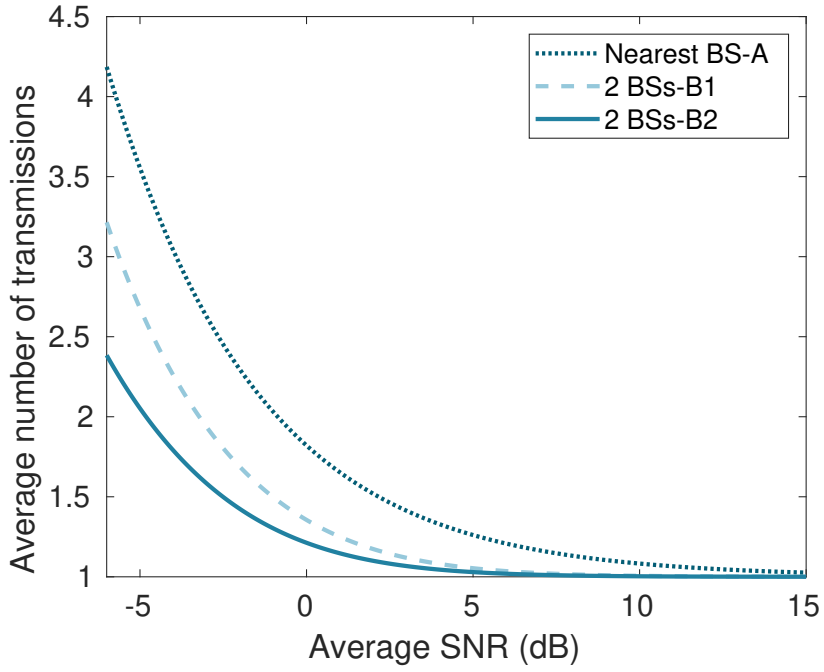


Figure 4.7: Average number of transmissions as a function of $\overline{\gamma}$ when the average SNR is the same on both sites: $\overline{\gamma}_1 = \overline{\gamma}_2 = \overline{\gamma}$.

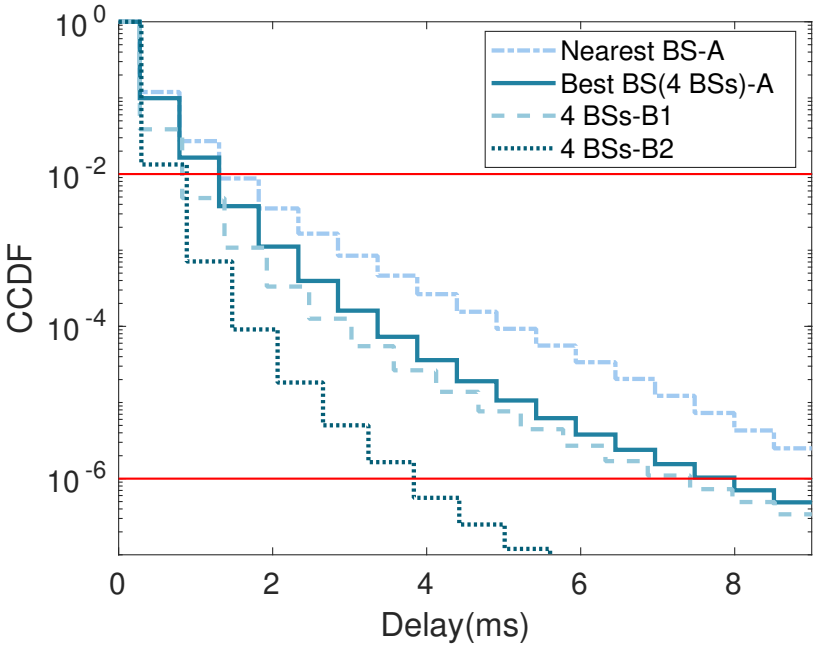


Figure 4.8: Delay CCDF for architectures A (2 cases), B1, and B2.

4.6 Conclusion

In this chapter, we studied the impact of architecture and macro-diversity on reliability and latency with three different functional splits. The comparison has been made through analytic calculations and simulations. The channel model was chosen to be realistic with a correlated shadowing effect. The evaluation of latency on the fronthaul was detailed depending on the considered split.

It was shown that receiving from the best BS induces less latency and higher reliability than receiving from the nearest BS using architecture A and single reception. Macro-diversity with architectures B1 and B2 had a better impact on reliability and latency compared to a single reception with architecture A. This chapter mainly shows the importance of the MRC technique when both high reliability and low latency are required.

In this part, we showed that the RAN architecture has a direct impact on the network performance, namely reliability and latency. C-RAN has shown to provide an additional gain that has not been highlighted before: simultaneous high reliability and low latency.

Part II

Impact of C-RAN architecture on BS energy consumption

Chapter 5

Indoor C-RAN functioning in low to medium load regimes: network coverage and capacity

Contents

5.1 Introduction

The use of wireless networks is increasing to a greater degree. According to Cisco, mobile Fourth Generation (4G) connections will grow to 6 billion by 2023 compared to 3.7 billion in 2018. Fifth Generation (5G) connections are estimated to reach 1.4 billion by 2023 compared to 13 million in 2019 [Cisco 2020]. Network operators find solutions to serve all users with stringent requirements wishing to use their networks. Consequently, this evolution of demands increases total energy consumption and indirectly contributes to large carbon footprints. 5G wireless networks are intended to offer high data rates, serve a large number of users, and achieve high reliability and low latency. Network energy consumption reduction is calling the attention of network operators.

For some 5G use cases, networks are built to provide service for many devices that might be active simultaneously. However, in specific periods (e.g. during night or weekends in business centers), only very few devices are active. This triggers network operators to find solutions to avoid wasting resources, especially energy consumption. One reasonable solution is to make the network reactive to load variation.

This part of the thesis studies the performance of an indoor Centralized-RAN (C-RAN) architecture during periods of low and medium loads.

Before introducing our work, we present the context of our study. Afterward, we present some related research relevant to our investigation, which covers this and the following two chapters. In this part, we aim to reduce the energy consumption of the network during low and medium loads without deteriorating the Quality of Service (QoS). Later in this chapter, we develop the considered network operating modes. We evaluate the network coverage and the blocking probability in each mode. This chapter aims to evaluate the network coverage and capacity for the considered operating modes with different Base Stations (BSs) positioning. In Chapter 6, we present some existing BS energy consumption models. We then detail the used energy consumption model and evaluate the C-RAN energy consumption in each considered operating mode. Finally, in Chapter 7, we introduce resource reuse on the DownLink (DL) to increase the network capacity while maintaining low energy consumption.

5.2 Context

The context of this study is related to cooperation between IMT-Atlantique and Société du Grand Paris, which is in charge of designing and building 200 km of automated metro lines with 68 stations connecting the suburbs without passing through Paris. These lines are intended to carry about 3 million passengers per day. Société du Grand Paris aims to provide a seamless experience of telecommunications services to passengers in stations and on trains.

Railway trains and subways are commonly equipped with electrical engines. This

means that their use can contribute to transportation CO2 emissions reduction, especially with new green electricity generation like wind energy generation, where zero emissions can be reached. In the Netherlands, for instance, since 2017, trains that run without CO2 emissions have been deployed [Nederlandse Spoorwegen 2017]. Using trains/subways for transportation is thus increasing, leading to higher use of train/subway stations. Network operators are concerned about satisfying their users even during peak hours, so they seek to deploy high-quality connections at these stations. Nevertheless, during the night, there are no passengers in these stations, but a network must be reachable for maintenance and security reasons. The load variation between these two extreme periods is significant. Adapting the network to this load fluctuation is essential to use the available resources optimally. The period where the load is low is not to be neglected since it lasts some hours per 24-hour day. Thus, studying the network energy consumption during low to medium loads is attractive in order to improve the use of available resources and reduce CO2 emissions as much as possible. In a subway or train station, the deployment of a macro-cell is challenging as it is usually an indoor underground environment. Deploying multiple micro-cells to achieve the desired coverage is thus unavoidable.

Although the study was suggested in the context of urban transportation, it can be applied to any case where there is no macro-cell coverage and a long period with moderate loads.

The BS is the most energy-consuming component in a network. Over 57% of the energy consumed in wireless access networks comes from the BS according to Vodafone [Han *et al.* 2011]. Network operators claim that this is a significant source of energy consumption. It is, therefore, energy-intensive to deploy multiple micro-BSs to achieve the desired coverage in an indoor underground station. To reduce the energy consumed by BSs, it is essential to identify the sources of this energy consumption. The energy-consuming elements in a BS are given hereafter:

- Processing units: the energy is consumed by the Central Processing Unit (CPU) to process received and transmitted signals.
- Radio power generation: the energy is consumed by the power amplifier to generate the power necessary for transmission on the radio link.
- Current generation: the energy is consumed by the power supply to generate electricity for the operation of the BS.
- Air conditioning: the energy is consumed by the air conditioners to cool down the heat of the BS. Note that this consumption is taken into account for macro-BSs and neglected for micro-BSs.

In macro-cells, the transmission energy cost is the highest in a BS due to the power amplifiers, followed by the base-band processing consumption and the cooling energy cost. However, in small cells, the transmission energy consumption for radio power generation is reduced due to short distances and ranges and the cooling cost

becomes negligible [Gunther *et al.* 2012]. The energy consumed by the processing units may therefore predominate. This triggers the thoughts about C-RAN, which is an improvement of the traditional Radio Access Network (RAN) and is supported by 5G wireless networks. In this centralized architecture, the Radio Units (RUs) can be distributed to achieve the desired coverage, while all the processing can be centralized in one central point.

Among the benefits of C-RAN, it is essential to mention network deployment cost reduction by expanding the network using low-cost units on different tower sites. Moreover, centralizing Centralized Units (CUs) reduces cooling and processing energy consumption. Managing the radio resources in a centralized unit can also optimize the network radio resources use.

In this part of the thesis, we study the BS performance and energy consumption of indoor micro-cells with C-RAN architecture in low and medium loads. We adopt an indoor environment where we deploy I RUs connected to a CU. We evaluate the network coverage and capacity in this chapter. Then, in Chapter 6, we study BS energy consumption under different modes. Finally, in Chapter 7, we optimize the scheduling of resources shared by different cell sites. We evaluate the BS energy consumption in low to medium load regimes. The available resources are shared between the I RUs and scheduled by the CU. The system is meant to serve a certain number of users with target UpLink (UL) and DL data rates.

5.3 Related literature review

This section reviews some research work related to C-RAN architectures' energy consumption reduction. Following that, a discussion of Coordinated Multi-Point (CoMP) in C-RAN is presented, along with optimizing energy consumption.

5.3.1 C-RAN energy consumption reduction

Much research has been done to reduce the energy consumption of BS. The architecture C-RAN has proven to be an energy-efficient architecture [Chen *et al.* 2014], [Sigwele *et al.* 2015]. Centralizing baseband processing can lead to a reduction in the number of processing units or better cooperation between sites, for example, to reduce energy consumption. Despite this, it can still consume more energy when radio units are deployed densely. Research interest in lower energy consumption in C-RANs tends to increase. Previous studies optimized the energy consumption of the three components of a C-RAN: the fronthaul, the central point (Baseband Unit (BBU) pool), and the distributed unit on sites (the Remote Radio Head (RRH)). Some studies considered one component to optimize its energy consumption. In comparison, other studies optimized the energy consumption of more than one component. Some studies considered user-to-RRH association and RRH-to-BBU association to reduce C-RAN energy consumption. For instance, to reduce the amount of energy produced on the fronthaul, Zuo *et al.* optimized user-RRH

association to turn off unused RRHs [Zuo *et al.* 2016]. Optimizing BBU computation resources is one method for reducing BBU pool energy consumption. Lyazidi *et al.* formulated an integer linear programming problem for BBU selection with three objectives including minimization of BBU processing power [Lyazidi *et al.* 2017]. Researchers in [Aqeeli *et al.* 2018] optimized computational resources allocation between BBUs and RRHs. Their main goal was to reduce the number of used BBUs, reducing thus the BBU pool energy consumption while maintaining the users QoS requirements in terms of data rate. The researchers in [Zhang *et al.* 2016b] proposed aggregating sparsely used BBUs to reduce energy consumption in C-RAN. In this study, underutilized BBUs are sent to sleep mode and their corresponding RRHs are connected to other BBUs. Nevertheless, switching off some BBUs results in scheduling problems due to RRH-BBU or user-RRH re-association. To reduce RRH's energy consumption, one proposed solution is to allow some RRHs to be switched off. The RRHs chosen to be turned off are those that cover a low-loaded area. Authors in [Saxena *et al.* 2016] showed that it is possible to reduce energy consumption by dynamically switching a subset of RRHs on or off according to cellular traffic knowledge. However, switching some RRHs off can lead to coverage holes. The QoS of users originally covered by the switched-off RRH is degraded. These users might perform a handover to other RRHs. The new serving RRHs might provide the re-associated users with a bad service especially if the RRHs are performing on high frequencies where the signal attenuation is high [Feng *et al.* 2017].

In our work, during low-load periods, we take advantage of BSs cooperation offered by C-RAN instead of switching off the C-RAN entities, namely the BBU and RRH.

5.3.2 CoMP for energy consumption reduction

Coordination between different cell sites appeared to enhance users' experience. CoMP refers to techniques that allow coordinated transmissions/receptions between different cell sites. A DL CoMP example is the joint transmission technique when multiple BSs serve a user. Joint reception is when multiple BSs receive a user in the UL [Dahlman *et al.* 2011]. CoMP is one of the best techniques to improve the received Signal to Interference and Noise Ratio (SINR). By using CoMP, authors in [Marsch & Fettweis 2011] maximized the average SINR and reduced the chance of SINR outage. An increased SINR increases the throughput. CoMP is thus essential to achieve high 5G throughput [Li *et al.* 2014] [Irram *et al.* 2020]. CoMP is also used to reduce the BS energy consumption. Authors in [Jahid *et al.* 2018] used Dynamic Point Selection (DPS)-CoMP [1] to determine the user's best serving BS and associate them together. They considered three UE-BS association CoMP metrics: distance, SINR, and distance-SINR. Their results show that UE-BS association based on SINR gains the best energy efficiency by maximizing system throughput. Another

[1] In DPS-CoMP, data is transmitted to a User Equipment (UE) by one BS with the best channel conditions [TR 36.819 2013].

CoMP mechanism, the Joint Transmission (JT)-CoMP [2], is also explored for better energy efficiency in [Landou & Barreto 2015]. In this study, authors consider users with bad network conditions to be served simultaneously by more than one BS. They consider cell zooming [3] and BS sleep techniques to reduce BS power consumption. To keep an acceptable QoS they use the aforementioned CoMP technique. Although this cooperation enhanced the energy efficiency, its use was limited in this study: i) they only consider it for users with bad conditions, ii) they consider a maximum of two serving BSs simultaneously, and iii) they consider it in addition to other energy reduction techniques.

Since C-RAN has centralized units at a central location, it presents a better opportunity for collaboration between cell sites. Therefore, C-RAN architecture eases the implementation of CoMP techniques. C-RAN is therefore an enabler for CoMP techniques, and researchers were interested in optimizing energy consumption and efficiency in C-RAN architectures with CoMP techniques. For example, using CoMP techniques, authors in [Luo *et al.* 2015] examined how joint DL and UL user-BS association and beamforming optimize overall power consumption in C-RANs.

Using the C-RAN architecture, we further explore CoMP techniques to minimize the BS energy consumption and avoid QoS degradation.

5.4 Load-dependent system considerations

In the context mentioned above, we consider small-cell indoor cases such as train railway stations, subway stations, indoor factories, or shopping malls. We consider I RUs connected to a Distributed Unit (DU)-CU pool. In this part of the thesis, we consider that the DU is implemented at the central point with the CU. For simplicity, we refer to these two entities located at the same central point as CU.

As previously explained in Section 2.3.1, a user receives the Synchronization Signal Block (SSB) periodically broadcasted by the network before asking to be connected. Thanks to these blocks, the user recognizes the identity of the serving cell and synchronizes with it. The user initiates random access during a Physical Random Access CHannel (PRACH) or transmission occasion, periodically available at the network. After a successful random access, resource allocation and signal exchange take place. Radio resource allocation consists of allocating resources in time and frequency to users. In 5G New Radio (NR), one Resource Block (RB) is defined in the frequency domain (it consists of 12 subcarriers) [4]. This task is done by the BS. After these steps, data exchange between the user and the network is possible. The energy consumption of this communication, between the user and

[2] JT-CoMP improves the reception quality by transmitting information simultaneously from different coordinated BSs to the UE [TR 36.819 2013].

[3] Cell zooming is a technique where the cell coverage is adapted to the cell load [Niu *et al.* 2010].

[4] In Long Term Evolution (LTE), an RB is defined in frequency and time domains: it spans 12 subcarriers and one time slot, respectively in each domain. While due to the flexibility of the transmission time in 5G NR, an RB is only defined in the frequency domain where it is a block of 12 subcarriers [Dahlman *et al.* 2018].

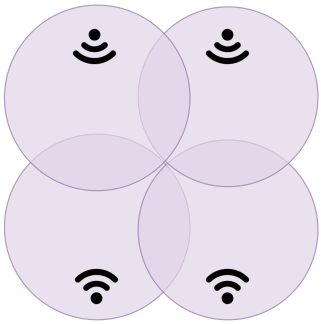


Figure 5.1: Low-load profile system, all BSs performing as one big cell.

the network, has a part that is independent of the network load and another that depends on the network load. At zero network load, the consumption is only due to SSB transmission and reception activation during the PRACH occasions. At low or medium load, we have SSB transmissions, PRACH opportunities, and data exchange on the DL and UL. However, in this case, the stream load is moderate: Multiple Input Multiple Output (MIMO) or beamforming is not required. At high load, in addition to SSB transmissions and PRACH opportunities, there is a large load to carry. Some transmission techniques such as MIMO and beamforming are essential. We focus on low and medium loads in this part. We define two modes of operation of the network depending on the load: low and medium load. Although we are not covering the high-load regime in this thesis, we include it in this section to make the presentation complete.

5.4.1 Low-load system

When the system is not heavily loaded, we consider that all RUs simultaneously broadcast the same information in the SSB. We consider here very few or almost no active users. All the RUs have the same identity and perform as one big cell, which is illustrated in Figure 5.1. The UE hears the network to synchronize and connect to it. Reciprocally, there will be one PRACH occasion where the user can request access to the network. The CoMP technique considered here is the JT technique. This mode is referred to as mode 1.

In this low-load profile, there is at most one RU transmission/reception on each RB at each time slot in the system. This transmission/reception comes from the simultaneous transmissions/receptions of the same information by all RUs. In this mode, the system is one big cell and the user is unaware of the presence of multiple RUs.

5.4.2 Medium-load system

In a medium-load system, serving multiple users in parallel is important to flow the needed load. To this end, we consider different SSBs to be sent consecutively by each RU. Each SSB is different from the others. This is illustrated in Figure 5.2.

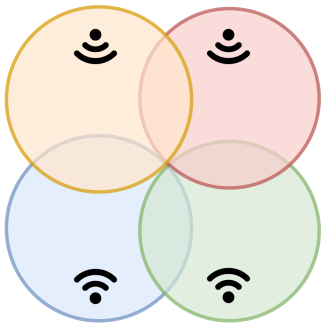


Figure 5.2: Medium-load profile system, BSs performing independently.

The difference between the SSBs allows the UE receiving an SSB to know which RU is serving. Then, there will be as many PRACH occasions as the number of RUs. The RU that receives/transmits best from/to the UE is selected to serve it. When needed, multiple RUs can transmit on the same RB in the absence of interference constraints. There are at most as many parallel transmissions in this mode as the number of RUs. The CoMP technique considered here is the DPS. This mode is referred to as mode 2.

5.4.3 High-load system

In a crowded system, to flow the high load, specific techniques might be needed like Multi-User MIMO (MU-MIMO) with beamforming. Consequently, multi-antenna RUs must be deployed. In that case, multiple SSBs are time multiplexed per RU and form a Synchronization Signal Burst (SS-Burst). This is called beam-sweeping SSBs. Each beam has its own SSB and performs as an independent cell. This is illustrated in Figure 5.3. This beamforming can be used to improve the coverage in each beam direction. It also allows beam-sweeping from the receiver side to go through the access process (preamble transmission during PRACH occasions). During this process, the network is able to determine the DL beam used to transmit each SSB. Considering beam correspondence, this beam can be used for the upcoming UL and DL transmissions. There are, at most, as many parallel transmissions in this mode as the number of beams. This mode is referred to as mode 3.

Table 5.1 summarizes the different load levels and the corresponding network behavior.

5.5 Presentation of the model

To evaluate the performances of the modes introduced in Section 5.4, we present the considered model in this section. We recall that we consider a C-RAN architecture where the BSs are split into RUs distributed on cell sites and a centralized DU-CU entity referred to as CU for simplicity.

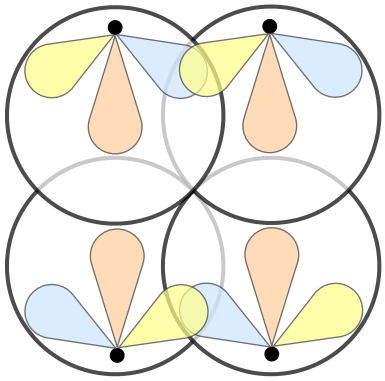


Figure 5.3: High-load profile system, BSs with beamforming.

Load level	Mode name	Main points
Low load	Mode 1	• All RUs perform as one big cell • Same SSBs and PRACH occasions for all RUs • One user served on the same RBs
Medium load	Mode 2	• Each RU performs as a cell • SSBs and PRACH occasions specific to each RU • Up to the number of RUs users served on the same RBs (treated in Chapter 7)
High load	Mode 3	• Multiple beams per RU each performing as a cell • SSBs and PRACH occasions specific to each beam (beam sweeping) • Up to the number of beams users served on the same RBs

Table 5.1: Different system loads summary.

5.5.1 Network deployment

The service zone is a rectangular indoor area of dimensions $A \times B$ m^2. The position of UE$_j$, given by its coordinates (x_j, y_j), is assumed to be uniformly distributed in the rectangular service area.

We consider I RUs to cover the considered rectangular area. Each RU$_i$ position is fixed and located by its coordinates (x_i, y_i). The considered RU placements are represented in Figure 5.4 for $I = 4$. The RUs can be in parallel positions where on one side of the rectangular, $I/2$ RUs are implemented, and the other half faces these implemented RUs on the opposite side. The side can be either the length or the width of the rectangular as shown in Figures 5.4(a) and 5.4(b), respectively. On one side, the RUs are separated by $\frac{A}{I}$ m or $\frac{B}{I}$ m from the corner and by $\frac{2A}{I}$ m or $\frac{2B}{I}$ m from each others. The RUs can also be implemented on the length or width of the rectangle in staggered positions like in figures 5.4(c) and 5.4(d), respectively. On one side, the RUs are separated by $\frac{A}{I+1}$ m or $\frac{B}{I+1}$ m from one corner and by $\frac{2A}{I+1}$ m or $\frac{2B}{I+1}$ m from the other corner and from each other. The last RU placement is represented in Figure 5.4(e). Here the RUs are placed in the centers of I rectangular areas that form the considered area. On the length, they are separated by $\frac{A}{I}$ m from the corner and by $\frac{2A}{I}$ m from each other. On the width, they are separated by $\frac{B}{I}$ m from the corner and by $\frac{2B}{I}$ m from each other.

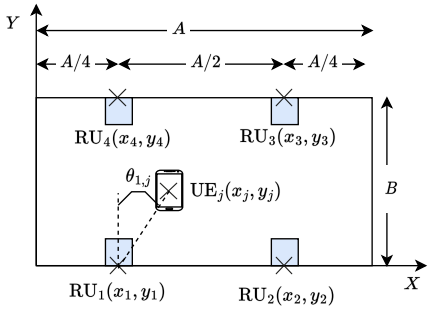
The RUs are connected to a CU implemented in a central point.

5.5.2 Traffic model

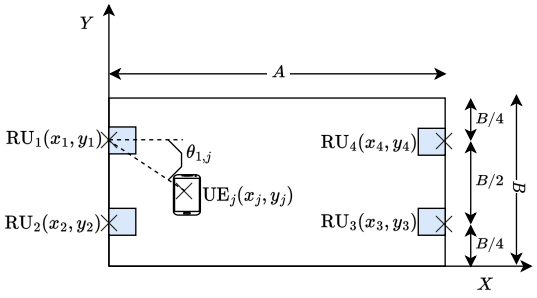
Many previous studies have considered a full-buffer traffic model. This model considers that users always have a full buffer. In other words, there is an infinite amount of data to transmit. All users have something to transmit at all times. In our study, we consider a user-requested service that needs a certain target data rate to be performed well [Ezzaouia *et al.* 2018]. The target data rate R_{T}^{n} to serve a user in direction $n = \{\mathrm{DL,UL}\}$ depends on the service the user is asking for (cf. Table 5.2, where the Federal Communications Commission recommends the values). Note that $R_{\mathrm{T}}^{\mathrm{UL}}$ is given by $R_{\mathrm{T}}^{\mathrm{UL}} = \frac{R_{\mathrm{T}}^{\mathrm{DL}}}{4}$.

Service	Target DL data rate ($R_{\mathrm{T}}^{\mathrm{DL}}$) (Mbps)
Browsing/e-mail	1
Social media	1
Streaming standard-high definition video	3 - 8
Streaming ultra high definition 4K video	25
Personal video call	1 - 1.5
High definition video teleconferencing	6
Online gaming	3 - 4

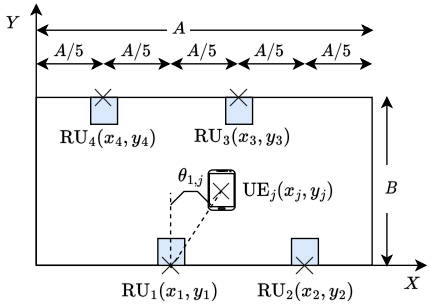
Table 5.2: Target DL data rate $R_{\mathrm{T}}^{\mathrm{DL}}$ for different services [FCC 2022].



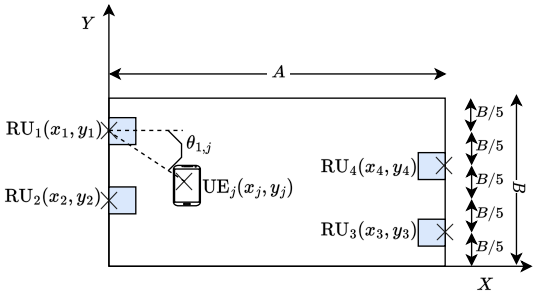
(a) Parallel RU placement (Length)



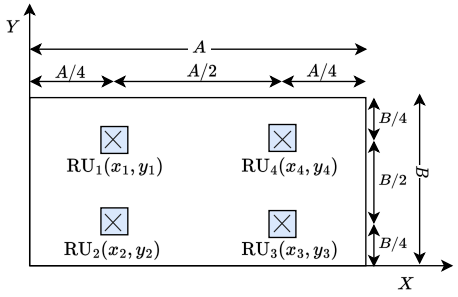
(b) Parallel RU placement (Width)



(c) Staggered RU placement (Length)



(d) Staggered RU placement (Width)



(e) Centered RU placement

Figure 5.4: RU placement for $I = 4$.

5.5.3 Propagation model

We consider an indoor factory environment where the heights of the receiving and transmitting antennas are lower than the average clutter height. The clutter density is less than 40% [Jiang *et al.* 2021]. From Table 7.4.1-1 in [TR 38.901 2019] (Appendix B), we take the path-loss model given by the 3rd Generation Partnership Project (3GPP) in dB for an Indoor factory Sparse clutter, Low base station height (InF-SL) with:

$$\mathrm{PL}_{\text{InF-SL,dB}} = 25.5\log(d_{i,j}) + 33 + 20\log(f_c), \tag{5.1}$$

where $\log(u)$ is the log base 10 of u, f_c the central frequency in GHz ($0.5 \leq f_c \leq 100$ GHz), $d_{i,j}$ the 2D distance between RU_i in position (x_i, y_i) and the jth user in position (x_j, y_j) having the same height, measured in m ($1 \leq d \leq 600$ m):

$$d_{i,j} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}. \tag{5.2}$$

The path-loss can also be expressed linearly as $\mathrm{PL} = \left(\frac{d_{i,j}}{r_0}\right)^{\alpha}$, where α is the path-loss exponent and r_0 is the reference distance for which the path-loss is equal to one. Parameters α and r_0 are derived, for a given f_c, from the following form of the path-loss equation: $\mathrm{PL}_{\text{dB}} = 10\alpha\log(d_{i,j}) - 10\alpha log(r_0)$.

A log-normal shadowing is considered by 3GPP. It is modeled by $e^{\sigma\xi}$ with ξ being a normal random variable (r.v.): $\xi \sim N(0,1)$, $\sigma = \frac{\ln(10)}{10}\sigma_{\text{dB}}$, and $\ln(u)$ the natural logarithm of u.

We also consider the RUs to be equipped with directive antennas with the following gain in dB [TR 36.942 2020]:

$$G_{\text{dB}}(\theta_{i,j}) = G_{\text{A}} - \min\left[12\left(\frac{\theta_{i,j}}{\theta_{\text{3dB}}}\right)^2, A_{\text{m}}\right], \tag{5.3}$$

where G_{A} is the antenna gain in the boresight direction, θ_{3dB} the 3 dB beam width, A_{m} the maximum attenuation, and $\theta_{i,j}$ the angle between RU_i and UE_j, measured from the antenna boresight:

$$\theta_{i,j} = \arctan\left(\frac{x_j - x_i}{y_j - y_i}\right). \tag{5.4}$$

The angle $\theta_{i,j}$ is illustrated in Figure 5.4 for $i = 1$, between user j and RU_1. Note that for the RU placement represented in Figure 5.4(e), each RU is mounted with an omnidirectional antenna with gain $G_{\text{dB}} = G_{\text{A}} = 2.15$ dBi. The UE is considered equipped with an omnidirectional antenna.

With all the above considerations, the received power by a receiver from a transmitter is computed as:

$$P^n_{\text{r},i,j,q}(x_j, y_j) = P^n_{\text{t},q}\left(\frac{r_0}{d_{i,j}}\right)^{\alpha} G(\theta_{i,j})e^{\sigma\xi_{i,j}}, \tag{5.5}$$

where $P^n_{\mathrm{r},i,\,j,\mathrm{q}}(x_j, y_j)$ is the received power in mode $q = \{1, 2\}$ and direction $n = \{\mathrm{DL},\mathrm{UL}\}$. The received power depends on the position (x_j, y_j) of UE_j. Parameter $P^n_{\mathrm{t},q}$ is the transmitted power per RB in mode q and in direction n. Let us remind that $q = 1$ designates the low-load mode such that we have a big cell with a multi-point transmission/reception (the RUs perform simultaneously), and $q = 2$ the medium-load mode in which each RU is a single cell. For instance, $P^n_{\mathrm{t},q} = P^{\mathrm{DL}}_{\mathrm{t},1}$ is the transmission power, per RB, of the RU DL transmissions in the low load or mode 1. These two modes will be detailed later in this chapter.

5.5.4 Reception model

5.5.4.1 Noise power

Each receiver is characterized by its noise figure N^n_{NF} in direction $n = \{\mathrm{DL},\mathrm{UL}\}$ expressed in dB and the thermal noise over an occupied bandwidth w for a receiver temperature T_{K}. The receiver's noise power in direction n is expressed as:

$$N^n_{\mathrm{p}} = N^n_0 w, \tag{5.6}$$

where $N^n_0 = 10^{\frac{N^n_{\mathrm{NF}}}{10}} K_b T_{\mathrm{K}}$ is the noise density, with K_b being the Boltzmann constant.

5.5.4.2 Receiver sensitivity

The receiver's sensitivity is the minimum power the receiver can detect. It is computed using the noise power and the receiver minimum Signal to Noise Ratio (SNR) [Penttinen 2019]:

$$P^n_{\mathrm{min}} = 10^{\frac{\gamma^n_{\mathrm{min}}}{10}} N^n_0 w_{\mathrm{RB}} m_{\mathrm{RB}}, \tag{5.7}$$

where γ^n_{min} is the minimum SNR in direction n, w_{RB} the bandwidth occupied by one RB, and m_{RB} the number of RBs used during the transmission.

5.5.5 Achievable data rate

For a given user j in position (x_j, y_j) the reachable data rate, per RB, using Shannon's equation is:

$$R^n_{j,q}(x_j, y_j) = w_{\mathrm{RB}} \log_2 \left(1 + \gamma^n_{j,q}(x_j, y_j)\right), \tag{5.8}$$

where w_{RB} is the bandwidth occupied by one RB, and $\gamma^n_{j,q}(x_j, y_j)$ the SNR or the SINR perceived by the receiver. However, this formula assumes a Gaussian distribution source, which is not feasible in practice. Thus, the rate in (5.8) can not be reached. In [Mogensen *et al.* 2007], a modified Shannon formula was proposed:

$$R^n_{j,q}(x_j, y_j) = w_{\mathrm{RB}} \delta_{\mathrm{BW}} \log_2 \left(1 + \frac{\gamma^n_{j,q}(x_j, y_j)}{\delta_{\mathrm{SNR}}}\right), \tag{5.9}$$

where δ_{BW} and δ_{SNR} are correction factors of the bandwidth and the SNR, respectively. The values of these parameters are taken from [Mogensen *et al.* 2007] for the single input single output case.

5.6 Resource usage and system capacity

In this section, we formulate the evaluation of resource usage along with the blocking probability. In the considered system, M_{RB} RBs are available per slot. We evaluate the amount of RBs used among the available resource blocks to serve user j requesting a service with a target data rate R_{T}^{n} in mode q and direction n using:

$$\tau_{j,q}^{n} = \frac{R_{\mathrm{T}}^{n}}{M_{\mathrm{RB}} R_{j,q}^{n}(x_j, y_j)}, \tag{5.10}$$

The number of RBs needed for user j is given by:

$$m_{\mathrm{RB},j,\,q}^{n} = \tau_{j,q}^{n} M_{\mathrm{RB}}. \tag{5.11}$$

The Cumulative Distribution Function (CDF) of the resource usage of user j in position (x_j, y_j) is given by the probability of not reaching the target data rate with certain resource usage. It is given by:

$$F_{\tau_{j,q}^{n}}(u) = \mathbb{P}\left(\tau_{j,q}^{n} \leq u | (x_j, y_j)\right). \tag{5.12}$$

The system capacity is evaluated by the ability to serve a certain number of users using the number of available RBs (M_{RB}). User j is blocked if more than M_{RB} RBs are needed to reach the target data rate: $m_{\mathrm{RB},j,\,q}^{n} > M_{\mathrm{RB}}$. Thus, the blocking probability for one user is given by:

$$\mathbb{P}_{\mathrm{blocking},1,\,q}^{n} = \mathbb{P}\left(m_{\mathrm{RB},j,\,q}^{n} > M_{\mathrm{RB}}\right). \tag{5.13}$$

For J users, the blocking probability is given by:

$$\mathbb{P}_{\mathrm{blocking},J,\,q}^{n} = \mathbb{P}\left(\sum_{j=1}^{J} m_{\mathrm{RB},j,\,q}^{n} > M_{\mathrm{RB}}\right). \tag{5.14}$$

5.7 System performance evaluation metrics

In this section, we detail the operations of both modes: mode 1 for a low-load system and mode 2 for a medium-load system.

We then evaluate the network coverage during SSB transmission over the beacon channel. A UE is considered covered if it receives a power higher than its sensitivity power $P_{\mathrm{min}}^{\mathrm{DL}}$ obtained from (5.7) with $m_{\mathrm{RB}} = m_{\mathrm{RB}}^{\mathrm{SSB}}$ the number of RBs occupied by the SSB. This is used to determine the BS transmission power $P_{\mathrm{t},q}^{\mathrm{DL}}$ needed in order to achieve a certain coverage. When a UE is uncovered, it is considered in an outage, and the probability of failure is expressed as $\mathbb{P}_{\mathrm{out},q}^{\mathrm{DL}}$.

Afterward, we study whether the user is well-received by the BS or not during PRACH occasions. A user is considered well received, if the BS is able to detect its signal with a power higher than its sensitivity $P_{\mathrm{min}}^{\mathrm{UL}}$ obtained from (5.7) with $m_{\mathrm{RB}} = m_{\mathrm{RB}}^{\mathrm{PRACH}}$ the number of RBs occupied by the PRACH. The probability of a

UE being not well-received is $\mathbb{P}^{\mathrm{UL}}_{\mathrm{out},q}$. The evaluation is conducted to determine $P^{\mathrm{UL}}_{\mathrm{t},q}$ the transmission power of the UE needed to guarantee a certain percentage of UE reception by the BS.

Next, we evaluate both modes' resource usage in the DL and UL directions. In mode 1, all RUs transmit the same data information to the user and receive the transmitted data signals from the user. This follows the simultaneous SSB transmissions and PRACH occasions. In mode 2, for DL data transmission, the best RU transmits DL data. For UL data transmissions, the RU with the best receiving conditions is considered. This mode follows the successive SSB transmissions over the beacon channel and PRACH occasions. Here, we determine the number of RBs required per time slot to achieve a target user data rate in the DL and UL. This amount is then used to determine how many users can be served during a time slot interval before exceeding the system's available capacity.

Note that in mode 2, resource reuse is possible since each RU performs as an independent cell. During data transmission, UE_j associated to RU_i can be served on the same resources with $\mathrm{UE}_{j'}$ associated to $\mathrm{RU}_{i'}$. Further details about the multi-cell multi-user concept are given in Chapter 7. In this chapter, we do not consider resource reuse. We only evaluate the performance of the two modes: one user is served by all RUs on some RBs in mode 1, and one user is served by the best RU on some RBs in mode 2.

5.7.1 Low-load system, mode 1

In this mode, the same SSB is transmitted by all RUs. One common PRACH occasion is available simultaneously on the same RBs for all RUs. All RUs transmit the same DL data to the user, and all of them receive the user in the UL direction.

5.7.1.1 Low load network coverage

As mentioned earlier, the I RUs perform as one big cell in this mode. All the RUs broadcast, simultaneously, the same SSB. The user receives all these transmissions and combines them using the Maximum Ratio Combining (MRC) technique. In this case, the I similar signals are summed up on the user side, then decoded. The combination of these signals achieves a higher probability of good decoding. The user then knows the cell's identity and is synchronized with it. It should be noted that the UE is unaware of the existence of different RUs. The UE detects a single cell consisting of the existing RUs. A UE is considered covered if it can detect an SSB. Thus, a UE is uncovered when the power received from the sum of the I SSB signals is below the minimum power it can detect $P^{\mathrm{DL}}_{\mathrm{min}}$. The outage probability $\mathbb{P}^{\mathrm{DL}}_{\mathrm{out},1}$ can be translated by the following equation for UE_j receiving from all the RUs:

$$\mathbb{P}^{\mathrm{DL}}_{\mathrm{out},1} = \mathbb{P}\left(\sum_{i=1}^{I} P^{\mathrm{SSB}}_{\mathrm{r},i,j,1}(x_j, y_j) < P^{\mathrm{DL}}_{\mathrm{min}}\right), \tag{5.15}$$

where $P^{\mathrm{SSB}}_{\mathrm{r},i,j,1}(x_j, y_j) = m^{\mathrm{SSB}}_{\mathrm{RB}} P^{\mathrm{DL}}_{\mathrm{r},i,j,1}$, and $P^{\mathrm{DL}}_{\mathrm{r},i,j,1}$ is defined in (5.5).

5.7.1.2 Low load PRACH occasions

After receiving the SSB, the UE initiates a random access process during the available PRACH opportunity. The transmitted preamble is received by the I RUs having the same PRACH occasion simultaneously. MRC combining technique of the same request received by all RUs is performed in the CU. In this mode, we consider the I RUs to simultaneously receive the UE. The achieved received power is the sum of all received signals by the I RUs from the UE. UE_j is badly received when this sum is lower than the BS sensitivity $P_{\text{min}}^{\text{UL}}$. The outage probability in the UL $\mathbb{P}_{\text{out},1}^{\text{UL}}$ is given by:

$$\mathbb{P}_{\text{out},1}^{\text{UL}} = \mathbb{P}\left(\sum_{i=1}^{I} P_{\text{r},i,j,1}^{\text{PRACH}}(x_j, y_j) < P_{\text{min}}^{\text{UL}}\right), \tag{5.16}$$

where $P_{\text{r},i,j,1}^{\text{PRACH}}(x_j, y_j) = m_{\text{RB}}^{\text{PRACH}} P_{\text{r},i,j,1}^{\text{UL}}$, and $P_{\text{r},i,j,1}^{\text{UL}}$ is defined in (5.5).

After successful random access, the UE can exchange data.

5.7.1.3 Low load resource usage for a target user data rate

For DL data transmission, the CU generates a user data signal and forwards it to all RUs, which transmit it to the user at the same time. Received at the user, the I data signals are combined using MRC. When all the RUs are transmitting simultaneously, the received SNR at the jth UE is the sum of all the received SNRs. During data transmission on the UL, the user transmits a signal that is received on the same RBs by all RUs. The MRC technique then takes place in the CU. When the user is received by all the RUs simultaneously, the received SNR at the BS is the sum of all the received SNRs. Then, using the modified Shannon rate equation, we get the received rate per RB:

$$R_{j,1}^{n}(x_j, y_j) = w_{\text{RB}} \delta_{\text{BW}} \log_2\left(1 + \frac{1}{\delta_{\text{SNR}}} \sum_{i=1}^{I} \frac{P_{\text{r},i,j,1}^{n}}{N_0^n w_{\text{RB}}}\right). \tag{5.17}$$

From (5.10), we have $R_{j,1}^{n} = \frac{R_{\text{T}}^{n}}{M_{\text{RB}} \tau_{j,1}^{n}}$, and by replacing $R_{j,1}^{n}$ in (5.17) by $\frac{R_{\text{T}}^{n}}{M_{\text{RB}} \tau_{j,1}^{n}}$, we get $\tau_{j,1}^{n}$ the resource usage of one user for a target DL/UL data rate in the low-load mode. The blocking probability for J users is calculated using (5.14).

5.7.2 Medium-load system, mode 2

In this mode, different SSBs are transmitted consecutively by different RUs. Distinct PRACH occasions are also available on successive RBs for every RU. Each user is associated with the best serving RU. DL data is transmitted by the best transmitting RU. The best transmitting RU is the one that provides the user with the highest SNR. UL data is received by the best receiving RU. The best receiving RU is the one that receives the highest SNR from the user.

5.7.2.1 Medium load network coverage

Each RU performs as one cell when the system has a medium load. RUs broadcast, one after the other, different SSB each. A UE is considered to be served by the best RU among the I RUs. Therefore, the user receives the SSB corresponding to the best SSB signal received. The UE tries to decode this signal to recognize its serving cell and synchronize with it. A UE is considered uncovered if none of the I RUs can provide a received power higher than its sensitivity. The outage probability $\mathbb{P}^{\mathrm{DL}}_{\mathrm{out},2}$ is given by:

$$\mathbb{P}^{\mathrm{DL}}_{\mathrm{out},2} = \prod_{i=1}^{I} \mathbb{P}\left(P^{\mathrm{SSB}}_{\mathrm{r},i,j,2}(x_j, y_j) < P^{\mathrm{DL}}_{\mathrm{min}}\right), \tag{5.18}$$

where $P^{\mathrm{SSB}}_{\mathrm{r},i,j,2}(x_j, y_j) = m^{\mathrm{SSB}}_{\mathrm{RB}} P^{\mathrm{DL}}_{\mathrm{r},i,j,2}$, and $P^{\mathrm{DL}}_{\mathrm{r},i,j,2}$ is defined in (5.5).

5.7.2.2 Medium load PRACH occasions

A UE requests random access after receiving an SSB. This is done during the PRACH occasion, available at the serving RU. This RU receives the transmitted preamble. A UE is badly heard by the system if none of the RUs can receive the transmitted preamble. The outage probability $\mathbb{P}^{\mathrm{UL}}_{\mathrm{out},2}$ in this mode is given by:

$$\mathbb{P}^{\mathrm{UL}}_{\mathrm{out},2} = \prod_{i=1}^{I} \mathbb{P}\left(P^{\mathrm{PRACH}}_{\mathrm{r},j,i,2}(x_j, y_j) < P^{\mathrm{UL}}_{\mathrm{min}}\right), \tag{5.19}$$

where $P^{\mathrm{PRACH}}_{\mathrm{r},i,j,2}(x_j, y_j) = m^{\mathrm{PRACH}}_{\mathrm{RB}} P^{\mathrm{UL}}_{\mathrm{r},i,j,2}$, and $P^{\mathrm{UL}}_{\mathrm{r},i,j,2}$ is defined in (5.5).

After successful random access, the UE can start exchanging data.

5.7.2.3 Medium load resource usage for a target user data rate

For DL data transmission, we consider only one RU transmitting data at a time. The transmitting RU is the RU that gives the user the highest received power among the I RUs. For UL data transmission, we consider only one user transmitting data at a time. When a UE is transmitting data in the UL direction, it is received by the best RU. In other words, the RU that receives this signal is the one that perceives the highest received power among the I RUs. Then, using the modified Shannon rate equation, the received power by RU_i the best receiving RU, from UE_j is:

$$R^n_{j,2} = w_{\mathrm{RB}} \delta_{\mathrm{BW}} \log_2 \left(1 + \frac{1}{\delta_{\mathrm{SNR}}} \frac{P^n_{\mathrm{t},2} r_0^{\alpha} e^{\sigma \xi_{i,j}} G(\theta_{i,j})}{N^n_0 w_{\mathrm{RB}} d^{\alpha}_{i,j}}\right). \tag{5.20}$$

We derive $\tau^n_{j,2}$ the resource usage of one user for a target data rate in the medium-load mode. The blocking probability for J users is given by (5.14).

5.8 Simulations

To evaluate the performance of the mentioned modes, we conduct Monte Carlo simulations using Matlab. The network coverage in the DL is evaluated during SSB transmissions. One thousand users are randomly positioned in the studied area. The RUs are positioned depending on the scenario. The received power at each UE is evaluated, and if it is less than the user's sensitivity, the user is considered in an outage. This is repeated for all UEs and iterated 10000 times. The number of users in an outage is then averaged over the number of trials to get the average outage probability. Similarly to DL coverage, a user is considered heard by the network in the UL if the network can receive a power higher than its sensitivity from the user during PRACH occasions.

The resource usage is then evaluated to determine whether we exceed the number of available resources or not. For a certain number of users J, we evaluated the needed resources to achieve the target data rate in DL and UL after randomly distributing J users and randomly generating their corresponding shadowing. The total number of RBs needed to serve the J users is determined during each simulation and compared to the number of available resources. If the number of needed RBs is higher than the number of available resources, the system is considered blocked and not blocked otherwise. For the same number of users, the simulation is iterated 10000 times. The number of trials then averages the blocking probability.

The simulation parameters are given in Table 5.3.

5.9 Results and discussions

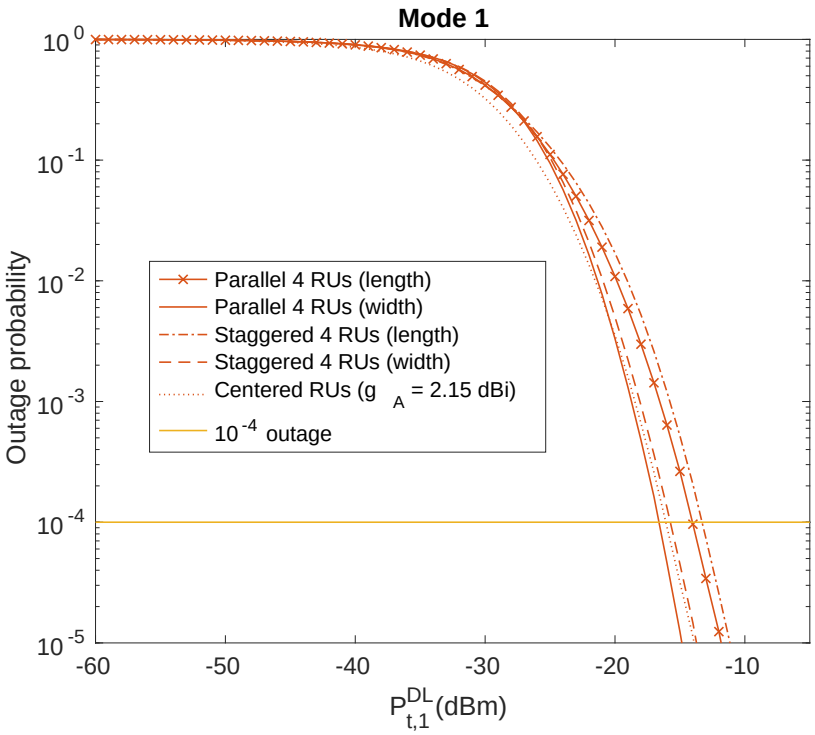
5.9.1 Network coverage

We evaluate the network coverage in all the RU placements illustrated in Figure 5.4. First, we start by evaluating the network coverage during SSB transmissions on the DL. Figure 5.5 illustrates the DL network outage occurring when a user is not able to receive an SSB. The network outage is represented as a function of the RU transmission power per RB. Modes 1 and 2 performances are illustrated in Figures 5.5(a) and 5.5(b), respectively. These performances are discussed below.

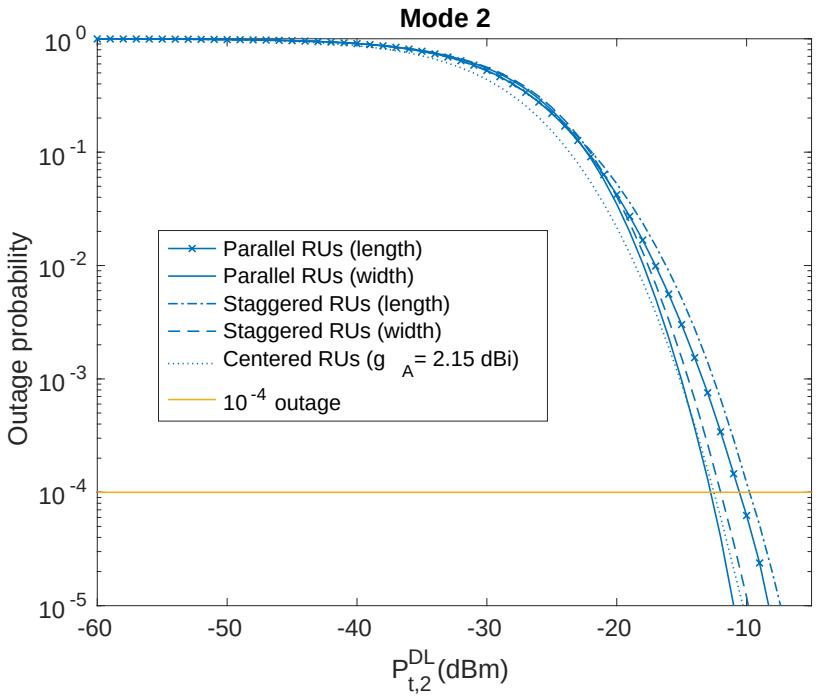
The outage probability with all RU placements is high for low transmission powers. The outage is higher than 30% for a transmission power lower than -28 dBm and -26 dBm in modes 1 and 2, respectively. For higher transmission powers, the outage probability decreases. At low transmission powers, the performance with the staggered and parallel positions is almost the same for both length and width positioning. When $P_{\mathrm{t},q}^{\mathrm{DL}}$ increases, the parallel positions appear to achieve a higher coverage compared to the staggering positions. With transmission powers higher than $-28/-26$ dBm, it becomes clear that the lowest outage probability is achieved by placing the RUs in parallel on the width of the rectangular area (see Figure 5.4(b)). This is due to the position of the RUs with respect to the corners and each other. When the RUs are positioned on the length of the rectangular area, their

Symbol	Parameter	Value
A (m)	Indoor area length	100
A_{m}	Antenna maximum attenuation	18
B (m)	Indoor area width	50
f_c (GHz)	Central frequency	26
G_{A} (dBi)	Antenna gain	7 [Kathrein Inc.]
M_{RB}	Available RBs in one time slot	135
$m_{\mathrm{RB}}^{\mathrm{PRACH}}$	RBs used for preamble reception	12
$m_{\mathrm{RB}}^{\mathrm{SSB}}$	RBs used for SSB transmission	20
$N_{\mathrm{NF}}^{\mathrm{DL}}$ (dB)	UE noise figure	7 [Penttinen 2019]
$N_{\mathrm{NF}}^{\mathrm{UL}}$ (dB)	BS noise figure	2 [Penttinen 2019]
r_0 (m)	Path-loss reference distance	0.0039
$R_{\mathrm{T}}^{\mathrm{DL}}$ (Mbps)	Target DL data rate	2
T (ms)	Reference period of study	20
T_{K} (K)	Receiver temperature	290 [Penttinen 2019]
w_{RB} (kHz)	RB bandwidth	720 [TS 38.211 2022]
W_{total} (MHz)	System bandwidth	100
α	Path-loss exponent	2.55
$\gamma_{\mathrm{min}}^{\mathrm{DL}}$ (dB)	Minimum DL SNR	-10 [Penttinen 2019]
$\gamma_{\mathrm{min}}^{\mathrm{UL}}$ (dB)	Minimum UL SNR	-7 [Penttinen 2019]
δ_{BW}	Bandwidth correction factor	0.56 [Mogensen *et al.* 2007]
δ_{SNR}	SNR correction factor	2 [Mogensen *et al.* 2007]
θ_{3dB}	Half power beam width	90° [Kathrein Inc.]
μ	NR numerology	2
σ_{dB} (dB)	Shadowing standard deviation	5.7 [TR 38.901 2019]

Table 5.3: Parameters values for the system model.



(a) Mode 1



(b) Mode 2

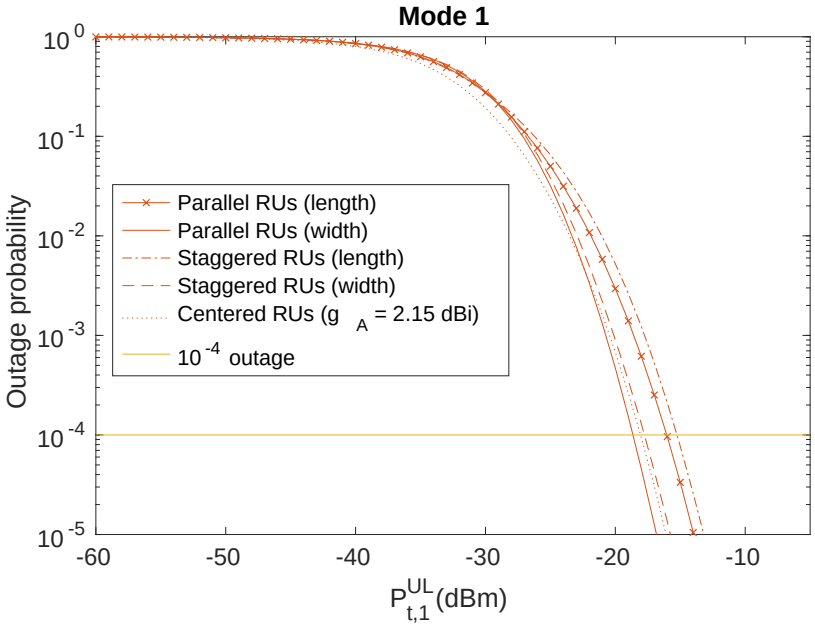
Figure 5.5: DL network outage with different RU placements for $I = 4$.

positions are farther from the corners and each other compared to implementing them on the width. It is also the case between staggered and parallel positions. With the chosen directive antennas, when the RU position is far from the corner, more users in the corner are not covered, for example. This is also the case when the distance between two consecutive RUs increases, keeping some users between these consecutive RUs uncovered.

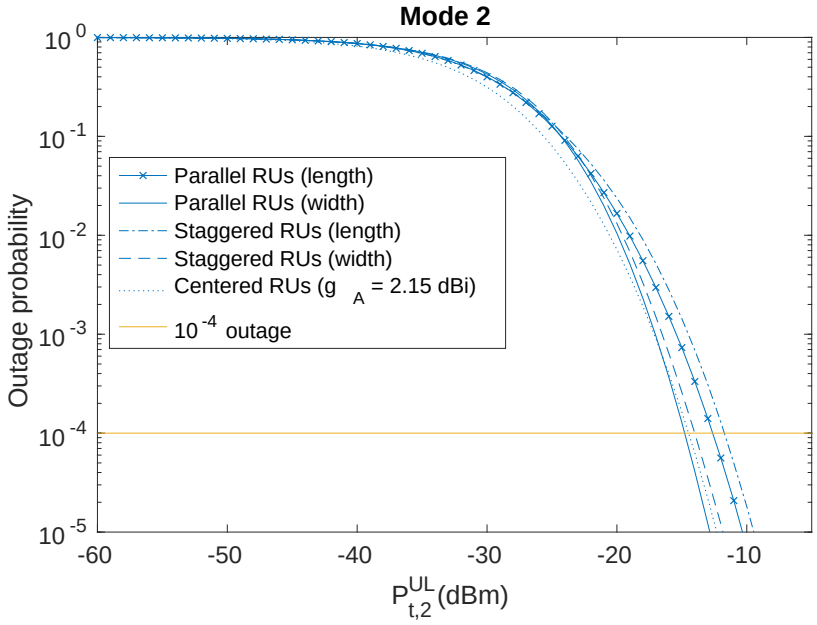
As for the central position with omnidirectional antennas, the outage probability is the lowest compared to other RU placements for low transmission powers. Even though this position appears to be equitable for users in all positions (in the corners or not), the directivity of the antennas gives an extra gain that enhances the network coverage. With higher transmission powers, the performance of the central placement is deteriorated (after being the best one) to be close to the performance when the RUs are on the width of the rectangular area in staggered positions (dashed lines) in mode 1 on the DL and UL. In mode 2, with high transmission powers, the coverage provided by the centered position is between the coverage when the RUs are on the width of the rectangular area in staggered positions (dashed lines) and the coverage with parallel positioning on the width of the rectangular area (continuous lines).

Comparing mode 1 and mode 2, we see that mode 1 requires less transmission power to achieve the same coverage as mode 2 for all RU placements. For example, for an outage of 10^{-4} or less, a transmission power of minimum -17 dBm is required in mode 1 with the parallel placement of RUs across the width, compared to -13 dBm in mode 2. In fact, the simultaneous transmissions of SSBs, increase the power received by the user who sums all the received signals. Thus, the probability of receiving an SSB with a power higher than the user's sensitivity increases. Note that similar performance appears on the UL as shown in Figure 5.6. The lower transmit powers observed in this figure are due to the lower noise at the BS (see Table 5.3). The lowest user transmission power, i.e., when the RUs are parallel across the width, is equal to -19 dBm and -15 dBm in modes 1 and 2, respectively.

Now, to evaluate the impact of the number of RUs on network coverage, we take the placement that gives the best coverage. This RUs placement is illustrated in Figure 5.4(b): the RUs are placed on the width of the rectangular area in a parallel way. If we look at each mode alone in Figure 5.7, we can see that increasing the number of RUs enhances the coverage. For mode 2, this is due to the greater chance of finding an RU that can cover a user as the number of RUs increases. This is also the case for mode 1. In addition, the increase in the number of simultaneous signals in mode 1 leads to an increase in the received power and, consequently, in the quality of the network coverage. Nevertheless, the improvement observed between 2 and 4 RUs is greater than that observed between 4 and 6 RUs. The position of the RUs as the number of RUs increases does not change much since the spacing between them decreases. This is clearly observed with mode 2 in Figure 5.7(b) where we observe almost the same performance between 4 and 6 RUs in mode 2. When increasing the number of RUs from 2 to 4, we notice better network coverage



(a) Mode 1



(b) Mode 2

Figure 5.6: UL network outage with different RU placements for $I = 4$.

represented by a lower outage for both modes. However, when increasing from 4 to 6 RUs, we observe, in Figures 5.7(a) and 5.7(b), that mode 1 with 4 RUs outperforms mode 2 with 6 RUs. For those reasons, we continue our study with 4 RUs with the placement represented in Figure 5.4(b).

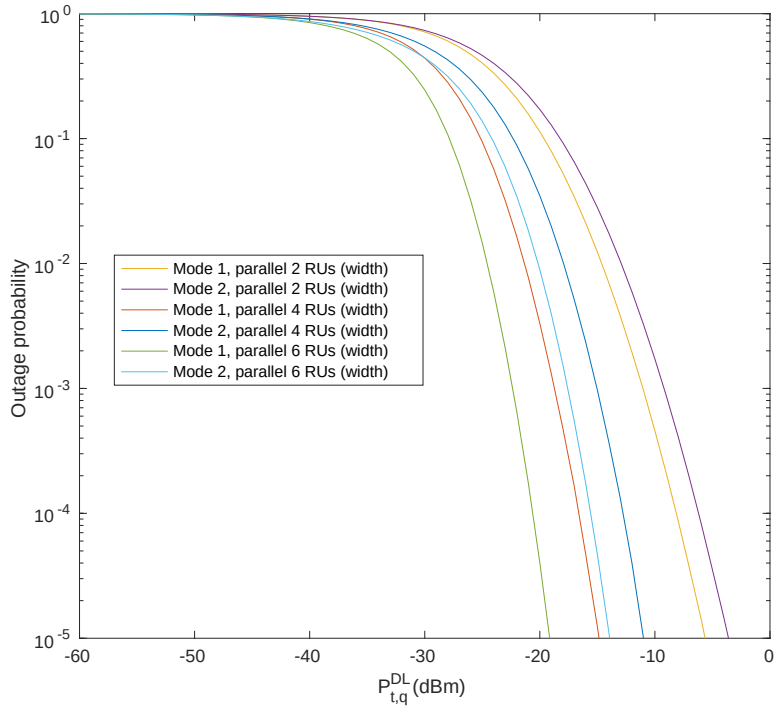
5.9.2 Resource usage and blocking probability

We evaluate the resource usage and blocking probability in modes 1 and 2 in the DL and UL directions. For the DL direction we choose $P_{t,1}^{DL} = P_{t,2}^{DL} = -13$ dBm. This transmission power guarantees an outage of 10^{-4} or less in modes 1 and 2 (see Figure 5.5). In the UL direction, the chosen transmission power is $P_{t,1}^{UL} = P_{t,2}^{UL} = -15$ dBm (see Figure 5.6). The distribution of RBs usage is illustrated in Figure 5.8. It is given by the probability of not reaching the target data rate for a user based on the number of RBs allocated to that user out of the total number of RBs available per slot. For the same resource usage, we can see that mode 1 outperforms mode 2. This means we have lower risks of not reaching the target data rate with the simultaneous transmissions by all RUs compared to the best RU transmission. In Figure 5.8(a), we see that to reach the target DL data rate with a probability of 99%, 8% of the available resources are needed in mode 1 while 18% are needed in mode 2. This difference increases for higher success rates. The perceived data rate is calculated using (5.9). So, if we increase the number of allocated RBs, we have higher chances of achieving the target data rate in both modes. However, this increase does not affect both modes equally due to the SNR increase by using simultaneous transmissions in mode 1. The same difference between modes 1 and 2 is observed in the UL as seen in Figure 5.8(b). A lower RB usage is noted in the UL direction. This is mainly due to the lower target UL data rate. For example, less than 1% of the available RBs are needed to guarantee the target UL data rate with a probability of 99% in mode 1 compared to 8% in the DL (mode 1).

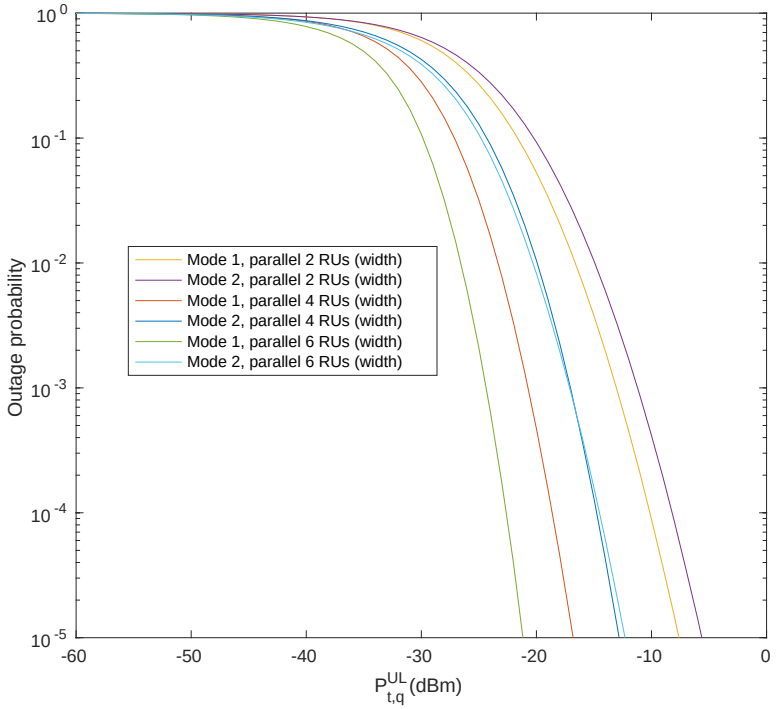
The blocking probability as a function of the number of users is shown in Figure 5.9. Blocking occurs when the number of RBs needed to serve the total number of users is greater than the number of available RBs. In these simulations, we consider that the number of RBs is $M_{RB} = 135$ for each transmission direction (DL and UL). The results show that the same blocking probability is reached for a larger number of users in mode 1 compared to mode 2 in both DL and UL directions. Thus, more users can be served in mode 1 compared to mode 2. For example, for 10^{-2} blocking probability on the DL, 32 users can be served in mode 1 compared to 17 users in mode 2.

5.10 Conclusion

In this chapter, we presented an indoor C-RAN network functioning in two modes. Multiple RUs are deployed to achieve the desired coverage and are linked to a CU. In mode 1, all RUs perform as one cell and serve the user simultaneously, while in mode 2 each RU performs as one independent cell. In mode 2, the best RU for each

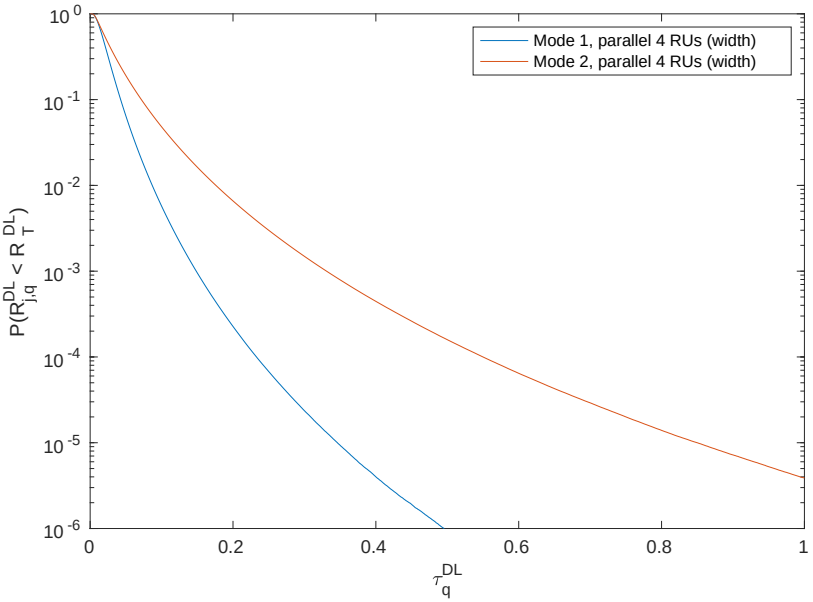


(a) Downlink

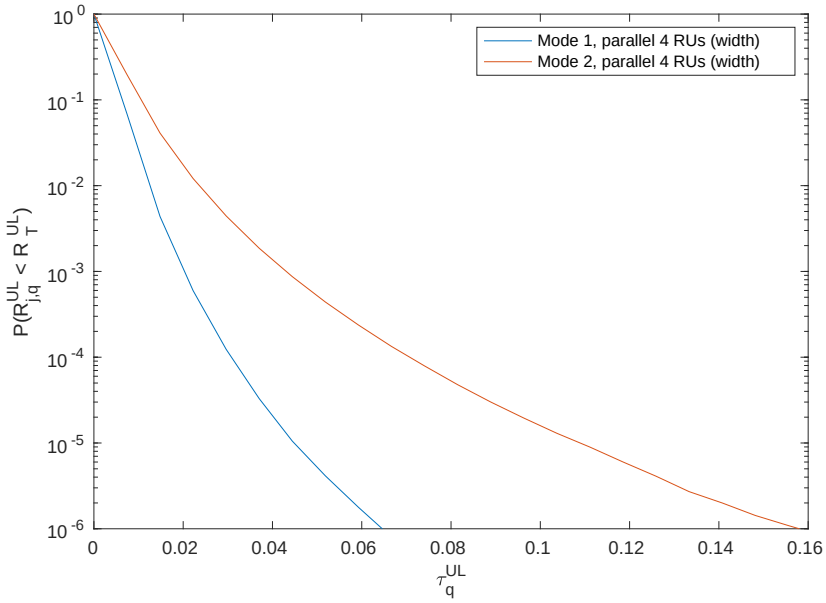


(b) Uplink

Figure 5.7: Network outage for $I = 2,\ 4,$ and 6.

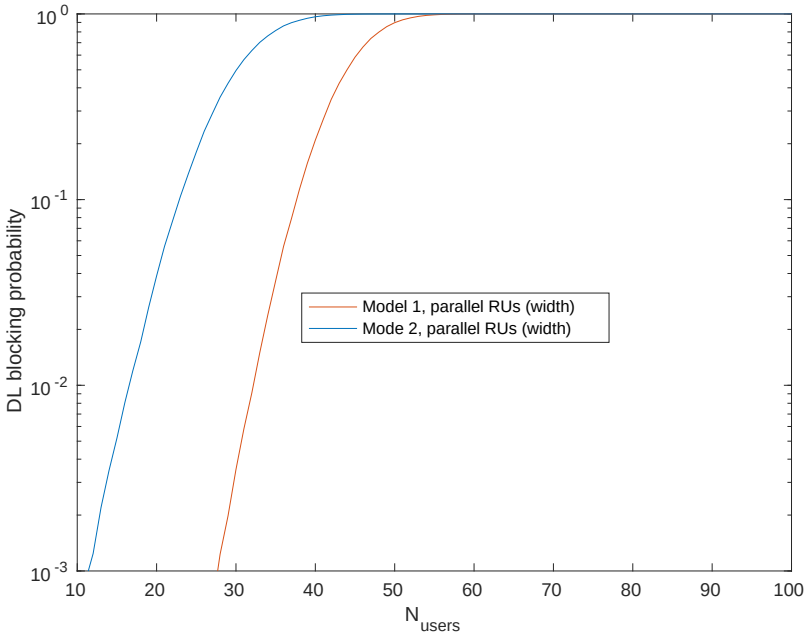


(a) Downlink

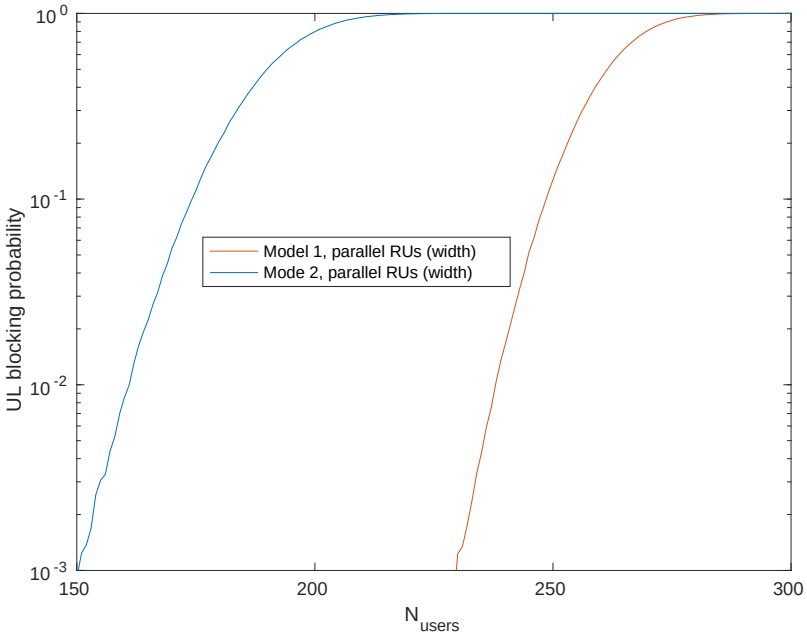


(b) Uplink

Figure 5.8: Resource usage distribution per user per slot.



(a) Downlink



(b) Uplink

Figure 5.9: Blocking probability as a function of the total number of users.

user is selected to be the serving RU. We evaluated the network coverage for five RU placements in the rectangular area. Placing the RUs in parallel on the widths of the rectangle with directive antennas achieved the best coverage. We evaluated the impact of the number of RUs on the network coverage. We compared 2, 4, and 6 RUs. Based on the results, we decided to continue our study with 4 RUs.

Serving the user in mode 1 ends up by increasing the received SNR that increases the received data rate and decreases the number of RBs needed to achieve the user's target data rate. This led to better performances for mode 1 compared to mode 2. Better coverage and higher chances of reaching the target data rate in mode 1 compared to mode 2 are noticed. We also noticed that due to the less RBs needed, we can serve more users in mode 1 compared to mode 2 before we reach the maximum number of available resources. Nevertheless, the independent performance of each RU in mode 2 allows resource reuse to increase the number of served users. The resource reuse is later treated in Chapter 7. Before this, we evaluate the BS energy consumption with the proposed model and modes in the next chapter.

Chapter 6

Indoor C-RAN energy consumption

Contents

6.1 Introduction

Our study's purpose is to reduce energy consumption during low-load periods while maintaining good coverage and service. In the previous chapter, we evaluated the network coverage and blocking probability in two modes. In this chapter, we evaluate the energy consumption of the Base Station (BS) while operating in these two modes. The described model and modes are the same as in Chapter 5.

The energy consumption of the BS with a centralized Radio Access Network (RAN) is evaluated in this chapter. First, we review some existing BS energy consumption models. We give the diagram of a two-unit BS with a brief description of its electronic components. Then, a general energy consumption model is given

based on the chosen model. Later, we give the energy consumption equations for each studied transmission/reception case in modes 1 and 2.

6.2 Existing BS energy consumption models

Multiple BS energy consumption models were proposed in the literature for the one-unit traditional BS and for the disaggregated BS.

For the traditional BS, some models consider the size of the BS. For instance, authors in [Arnold *et al.* 2010] developed power consumption models for macro and micro BSs. They also differentiated between static and dynamic consumption. The static power is consumed when the cell is empty, and the dynamic power depends on the cell load. The macro BS accounts for only static consumption, whereas the micro BS accounts for both static and dynamic consumption. This is because a small cell's power consumption is more dynamic than a large cell. After all, the number of users is statistically more variable in a small cell. Other power consumption models considered the load variation in a cell, for example, the model proposed by Energy Aware Radio and neTwork tecHnologies (EARTH) project [Gunther *et al.* 2012]. The proposed approximated linear energy consumption model is a function of a fixed and a load-dependent power. Static power is the consumption of electronic components when the output radio power is zero. The static part considers cooling devices, signal processing, and other non-variable consumption. The dynamic part is the product of a load-dependent slope and the output radio power. This model also considers the power consumption of a BS in sleep mode. When in sleep mode, the dynamic consumption is zero; however, there is a fixed power cost. It is worth mentioning that the fixed power consumed during sleep mode is lower than that consumed during active mode. In [Liu *et al.* 2016], authors proposed a BS power consumption model based on hardware components. The model was given as the sum of fixed power values of different BS components.

The energy consumption of a centralized architecture BS is of our interest. EARTH project power consumption model (previously presented) was then extended to evaluate the energy consumption of a two-unit BS [Khan *et al.* 2015]. The model's output radio power is correlated with its supplied power. However, the model does not provide the detailed consumption of each component. It is used with the proposed parameters without the opportunity to analyze each of the BS components' power consumption. In [Jung *et al.* 2014] a more detailed BS power consumption model was given where the power consumption model of each BS component was given. A correlation between BS components' power consumption is considered. The authors took into account the load variation in the model. They also provided BS power consumption models based on the cell size (micro or macro cell) and the RAN architecture. They proposed a power consumption model of a BS in Centralized-RAN (C-RAN) architecture.

6.3 Reference period of study

We evaluate the energy consumption of the BS taking into account the consumption of the signal processing, the electronic components, and the radio power generation. A certain amount of energy is consumed to perform the signal processing and this depends on the number of allocated Resource Blocks (RBs) to serve the user. A fixed power is consumed by electronic components, and this power is evaluated over the BS activation period to calculate the consumed energy. Radio power generation is mainly the power consumption of the power amplifier that is evaluated over the transmission duration to calculate the radio power generation energy consumption.

To evaluate energy consumption, we choose a reference period in which we determine the number of allocated resources, the activation period, and the transmission duration. The reference period of study is chosen to be equal to the Synchronization Signal Block (SSB) and Physical Random Access CHannel (PRACH) occasions periodicity. Thus, during that period, there is one SSB transmission and one PRACH occasion per cell, followed by user data exchange.

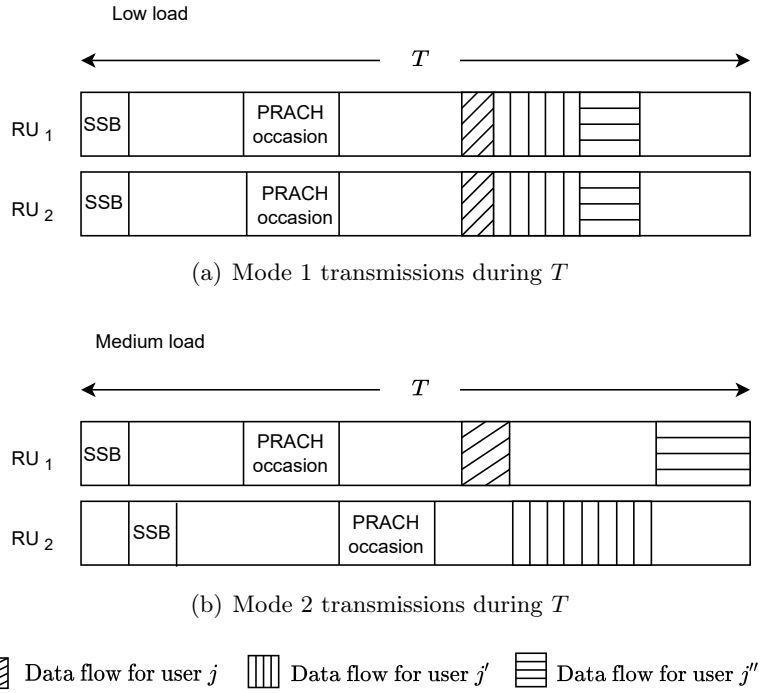
We consider Time Division Duplex (TDD) transmissions where a time slot is used either for an UpLink (UL) transmission or for a DownLink (DL) transmission. We define T as the period over which we compute the energy consumption of the BS. During this reference period, the SSBs must be transmitted. A PRACH occasion has to be also available for the User Equipment (UE), even if the user does not transmit during every PRACH occasion. Then, we have data flows if there are active users to serve. The flow of user data depends on the number of active users. We consider that each active user asks for a service with a target DL/UL data rate R^n_{T}. Although all users ask for the same service with the same target data rate, each of them requires a different number of RBs to be served. If we look at equations (5.10) and (5.11), we see that the number of RBs required to serve a user with the requested data rate depends on the perceived rate. The perceived data rate depends on the network conditions between the user and the BS represented by the Signal to Interference and Noise Ratio (SINR) $\gamma^n_{j,q}(x_j, y_j)$ in (5.9).

The simultaneous transmissions and receptions during T in low load, and the successive transmissions and receptions during medium load are illustrated in Figures 6.1 (a) and (b), respectively.

6.4 C-RAN based BS energy consumption model

To evaluate the energy consumption of the BS, we start by choosing a BS energy consumption model. After briefly discussing some of the existing models in Section 6.2, we detail the model that we consider from [Jung *et al.* 2014].

The BS diagram considering a C-RAN architecture is represented in Figure 6.2. When the BS is split into two entities, the energy consumption is due to the energy consumed by both entities. In the Radio Unit (RU), the energy is consumed by the radio power generation, Radio Frequency (RF) transceiver chain, and other



(a) Mode 1 transmissions during T

(b) Mode 2 transmissions during T

Figure 6.1: SSB, PRACH, and data transmissions/receptions over T.

electronic components. To exchange signals between the RU and the Centralized Unit (CU) over the fiber optic fronthaul, an electric to optical conversion takes place or vice versa. The CU is responsible for signal processing that has an energy cost. Electronic components in the CU also contribute to this consumption.

6.4.1 BS energy consuming components

The BS energy cost is due to multiple elements: the RF transceiver chain components, electronic components, radio power generation, and signal processing. The BS is considered split into two units. The main energy-consuming components of each one of these units are given hereafter.

The main electronic components, including transceiver chain ones in the RU are:

- The filters responsible for filtering the signal after reception: Band Pass Filter (BPF) and before transmission: Low Pass Filter (LPF). Both filters consume a fixed power P_{BPF} and P_{LPF}, respectively.

- The power amplifier unit is responsible for amplifying transmitted and received signals. A High Power Amplifier (HPA) is responsible for amplifying transmitted signals and is characterized by its efficiency η_{PAE}. A Low-Noise Amplifier (LNA) is responsible for amplifying received signals and consumes a fixed power P_{LNA}.

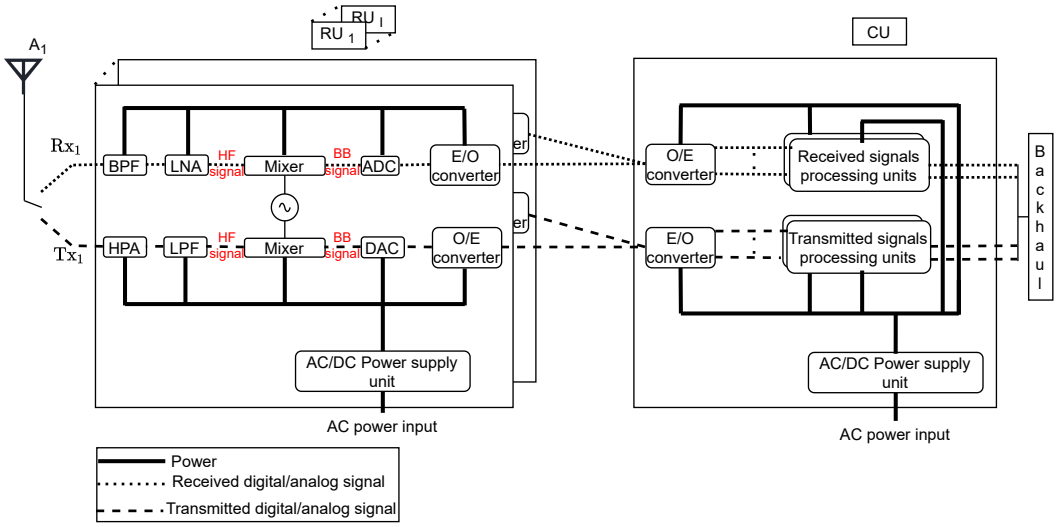


Figure 6.2: Power-consuming units in a C-RAN architecture.

- The mixer shifts a received high-frequency signal to a baseband signal and shifts a baseband signal to a high-frequency one for transmission. The mixer consumes a fixed power denoted by P_{Mixer}.
- The Analog to Digital Converter (ADC) and Digital to Analog Converter (DAC) convert an analog received signal into a digital signal and a digital signal into an analog one for transmission, respectively. These converters consume fixed powers denoted by P_{ADC} and P_{DAC}, respectively.

In the CU, we have:

- Received and transmitted signal processing units. Their energy consumption accounts for a fixed part related to electronic components and a variable part dependent on the packet length. The computation of signal processing energy consumption is detailed in Section 6.4.2.

In the RU and CU, we have:

- The O/E converter (or E/O converter) converts the optical signal received from the optical link into an electrical signal when the BS is in transmitting mode and conversely when BS is receiving. The power consumption of this unit $P_{\text{O/E}}$ (or $P_{\text{E/O}}$) is considered constant.
- The AC/DC power supply unit that converts the Alternating Current (AC) into Direct Current (DC) for all the components consuming power in the RU or the CU. The power supply is characterized by its efficiency $\eta_{\text{AC/DC}}$. This efficiency is a measure of how much actual power is delivered to the components relative to the real input power.

The energy consumption of a transmitting RU consists of the power consumption of the transceiver chain during transmission and reception and the radio power generation when the BS is transmitting over a period in which the RU is active for transmission.

The energy consumption in the CU includes a fixed and variable consumption that depends on the packet length. The fixed part is the power consumed by components like the E/O and O/E converters over a period of time. It also includes the fixed consumption of the signal processing unit. The variable energy consumption is the consumption of the signal processing units. It has a part dependent on the packet length and another part independent of the packet length.

6.4.2 Packet processing energy consumption

The energy consumption of the signal processing unit consists of two main parts. A fixed part of this energy consumption is consumed by the signal processing unit which includes the consumption of the motherboard, peripherals, and fans. These components consume a fixed power denoted by $P_{\mathrm{SP,f}}$ and evaluated over the activity period to get the consumed energy. Each time a packet is transmitted or received, some energy is needed to code or decode it. Thus, another part of this energy consumption is variable. It has a part dependent on the packet length $E_{\mathrm{SP,v}}$ and accounts for the energy consumed by layer 1 processes like coding/decoding and modulation/demodulation and is consumed per bit. The other part is consumed per packet independently of the packet length. It calculates the energy consumed by upper layers processes and is denoted by $E_{\mathrm{SP,f}}$. We define T_n, the time parameter that determines the activation duration of an RU or CU in direction $n = \{\mathrm{DL,UL}\}$ during T. To encode (in the DL) or decode (in the UL) packets to serve J users, the consumed energy in direction n is given by:

$$E^n_{\mathrm{SP}} = \underbrace{\underbrace{T_{\mathrm{n}}\underbrace{P_{\mathrm{SP,f}}}_{\text{Power}}}_{\substack{\text{Energy proportional}\\ \text{to } T_n \text{ the activity}\\ \text{duration during } T}}}_{} + \underbrace{E^n_{\mathrm{SP,v}} L^n_{\mathrm{P,T}} + E_{\mathrm{SP,f}} N^n_{\mathrm{Pkt}}}_{\substack{\text{Processing energy}\\ \text{proportional to the}\\ \text{packet length and}\\ \text{number of processed}\\ \text{packets during } T}}, \tag{6.1}$$

where $L^n_{\mathrm{P,T}} = JR^n_{\mathrm{T}}T$ is the length of all transmitted ($n = \mathrm{DL}$) or received ($n = \mathrm{UL}$) packets during T, and N^n_{Pkt} the number of transmitted or received packets during T.

6.4.3 General energy consumption model

We consider the energy consumption model from [Jung *et al.* 2014] of a BS represented in Figure 6.2. The energy consumption is computed during the activation period of the BS T_n. The BS is considered inactive when it has nothing to transmit/receive. We consider that the BS deactivation and activation consume negligible energy that is not taken into account in our study.

1. CU energy consumption

   The consumption of the CU has two main parts. The first part is a fixed part $P^{n}_{\mathrm{CU,f}}$ that accounts for the consumption of some electronic equipment (O/E and E/O converters, fans, motherboards, peripherals) of the CU in direction n. This part is consumed as soon as the CU is active and is computed as shown in Table 6.1. Then, the other part is the consumption of the signal processing of a received or transmitted packet. Its computation is given in Section 6.4.2. Note that $P^{n}_{\mathrm{CU,f}}$ includes $P_{\mathrm{SP,f}}$ from (6.1) (see Table 6.1). In the remaining equations, we consider that variables $P^{n}_{\mathrm{CU,f}}$, $E^{n}_{\mathrm{SP,v}}$, and $E^{n}_{\mathrm{SP,f}}$ include the AC/DC power supply efficiency ($\eta_{\mathrm{AC/DC}}$).

   The energy consumption of a transmitting/receiving CU to serve J users in direction n is:

$$E^{n}_{\mathrm{CU}} = T_n P^{n}_{\mathrm{CU,f}} + E^{n}_{\mathrm{SP,v}} L^{n}_{\mathrm{P,T}} + E^{n}_{\mathrm{SP,f}} N^{n}_{\mathrm{Pkt}}. \tag{6.2}$$

2. RU energy consumption

   An active RU at a certain time is considered either transmitting or receiving. The consumption of an RU is divided into two parts: the first part $P^{\mathrm{n}}_{\mathrm{RU,f}}$ is fixed and it accounts for the consumption of the RF transceiver chain (LPF/BPF, mixers, DAC/ADC, and O/E or E/O converters). Its computation is given in Table 6.1. The second part is a function of the transmission power, and it is accounted only when the RU is transmitting. In the upcoming equations, we consider that $P^{\mathrm{n}}_{\mathrm{RU,f}}$ includes the AC/DC power supply efficiency ($\eta_{\mathrm{AC/DC}}$).

   The energy consumed by a transmitting RU is given by:

$$E^{\mathrm{DL}}_{\mathrm{RU}} = T_{\mathrm{DL}} \left(P^{\mathrm{DL}}_{\mathrm{RU,f}} + \frac{m^{\mathrm{DL}}_{\mathrm{RB}} P^{\mathrm{DL}}_{\mathrm{t},q}}{\eta_{\mathrm{T}}} \right), \tag{6.3}$$

   where T_{DL} is the duration where the RU is active for transmission during T, $m^{\mathrm{DL}}_{\mathrm{RB}}$ the number of RBs used for the transmission, $P^{\mathrm{DL}}_{\mathrm{t},q}$ the DL transmission power per RB when the system is in mode $q = \{1,2\}$ (cf. Chapter 5), $\eta_{\mathrm{T}} = \eta_{\mathrm{PAE}}\eta_{\mathrm{AC/DC}}$, and the power consumption of the HPA is taken from [FOURIKIS 2000]:

$$P_{\mathrm{HPA}} = \left(\frac{1}{\eta_{\mathrm{PAE}}} - 1 \right) P_{\mathrm{PA,out}}, \tag{6.4}$$

   with $P_{\mathrm{PA,out}}$ being the output power of the power amplifier and is the transmission power in our case.

   The energy consumed by a receiving RU is:

$$E^{\mathrm{UL}}_{\mathrm{RU}} = T_{\mathrm{UL}} P^{\mathrm{UL}}_{\mathrm{RU,f}}, \tag{6.5}$$

   where T_{UL} is the duration where the RU is active for reception during T.

The fixed power consumptions of different units are summarized in Table 6.1. Parameter T_n (including T_{DL} and T_{DL}) depends on the transmission/reception phase and conditions. In Sections 6.5 - 6.7, we provide the energy consumption of the CU and RU during different transmission and reception phases. We thus replace T_n with the corresponding duration of transmission or reception.

Symbol	Computation
$P^{\mathrm{DL}}_{\mathrm{RU,f}}$	$\frac{P_{\mathrm{O/E}}+P_{\mathrm{LPF}}+P_{\mathrm{Mixer}}+P_{\mathrm{DAC}}}{\eta_{\mathrm{AC/DC}}}$
$P^{\mathrm{UL}}_{\mathrm{RU,f}}$	$\frac{P_{\mathrm{E/O}}+P_{\mathrm{BPF}}+P_{\mathrm{LNA}}+P_{\mathrm{Mixer}}+P_{\mathrm{ADC}}}{\eta_{\mathrm{AC/DC}}}$
$P^{\mathrm{DL}}_{\mathrm{CU,f}}$	$\frac{P_{\mathrm{E/O}}+P_{\mathrm{SP,F}}}{\eta_{\mathrm{AC/DC}}}$
$P^{\mathrm{UL}}_{\mathrm{CU,f}}$	$\frac{P_{\mathrm{O/E}}+P_{\mathrm{SP,F}}}{\eta_{\mathrm{AC/DC}}}$

Table 6.1: CU and RU fixed power consumption due to electronic components.

6.5 Beacon channel energy consumption

Each SSB occupies four Orthogonal Frequency Division Multiplexing (OFDM) symbols. This section provides the energy consumption during SSB transmissions.

6.5.1 Simultaneous transmissions, mode 1

In this mode, the same SSB is transmitted by the I RUs concurrently. Thus, one SSB has to be processed in the CU, and the RUs are active altogether to transmit the SSB.

1. CU energy consumption

   The energy consumption of the CU for the simultaneous SSB transmissions is:

   $$E^{\mathrm{DL},1}_{\mathrm{CU,SSB},T} = T_{\mathrm{SSB}} P^{\mathrm{DL}}_{\mathrm{CU,f}} + E^{\mathrm{DL}}_{\mathrm{SP,f}}, \tag{6.6}$$

   where $T_{\mathrm{SSB}} = 4\frac{T_{\mathrm{s}}}{14}$ is one SSB duration, and T_{s} the slot duration. The network configuration does not change for each SSB transmission. Thus, the SSB is not encoded each time it has to be transmitted. For this reason, we neglect the packet length-dependent part of the processing energy consumption $E_{\mathrm{SP,v}}$ during SSB transmissions.

2. RU energy consumption

   The energy consumption of one RU is:

   $$E^{\mathrm{DL},1}_{\mathrm{RU,SSB},T} = T_{\mathrm{SSB}} \left(P^{\mathrm{DL}}_{\mathrm{RU,f}} + \frac{m^{\mathrm{SSB}}_{\mathrm{RB}} P^{\mathrm{DL}}_{\mathrm{t},1}}{\eta_{\mathrm{T}}} \right), \tag{6.7}$$

   where $m^{\mathrm{SSB}}_{\mathrm{RB}}$ is the number of RBs used for SSB transmission.

3. Total energy consumption

   The total energy consumed for the SSB transmission over the beacon channel in mode 1 is:

$$E^{\mathrm{DL},1}_{\mathrm{Total,SSB},T} = E^{\mathrm{DL},1}_{\mathrm{CU,SSB},T} + IE^{\mathrm{DL},1}_{\mathrm{RU,SSB},T}. \tag{6.8}$$

6.5.2 Successive transmissions, mode 2

During the SSB transmissions in mode 2, each RU is supposed to transmit different information in the SSB at different times. Therefore, for I RUs, I SSBs need to be processed in the CU and transmitted successively by the I RUs.

1. CU energy consumption

   The energy consumption of the CU for the successive SSB transmissions in mode 2 is:

$$E^{\mathrm{DL},2}_{\mathrm{CU,SSB},T} = IT_{\mathrm{SSB}}P^{\mathrm{DL}}_{\mathrm{CU,f}} + IE^{\mathrm{DL}}_{\mathrm{SP,f}}. \tag{6.9}$$

2. RU energy consumption

   The energy consumed by one RU for the successive SSB transmissions is:

$$E^{\mathrm{DL},2}_{\mathrm{RU,SSB},T} = T_{\mathrm{SSB}}\left(P^{\mathrm{DL}}_{\mathrm{RU,f}} + \frac{m^{\mathrm{SSB}}_{\mathrm{RB}}P^{\mathrm{DL}}_{\mathrm{t},2}}{\eta_{\mathrm{T}}}\right). \tag{6.10}$$

3. Total energy consumption

   The total energy consumed for the SSB transmission over the beacon channel in mode 2 is:

$$\begin{aligned} E^{\mathrm{DL},2}_{\mathrm{Total,SSB},T} &= E^{\mathrm{DL},2}_{\mathrm{CU,SSB},T} + IE^{\mathrm{DL},2}_{\mathrm{RU,SSB},T} \\ &= I\left(T_{\mathrm{SSB}}P^{\mathrm{DL}}_{\mathrm{CU,f}} + E^{\mathrm{DL}}_{\mathrm{SP,f}} + T_{\mathrm{SSB}}\left(P^{\mathrm{DL}}_{\mathrm{RU,f}} + \frac{m^{\mathrm{SSB}}_{\mathrm{RB}}P^{\mathrm{DL}}_{\mathrm{t},2}}{\eta_{\mathrm{T}}}\right)\right). \end{aligned} \tag{6.11}$$

6.6 PRACH occasion energy consumption

During the studied reference period, at least one PRACH occasion has to be available. In order to give this opportunity to the user, the BS must be on and ready to receive the user's preamble. That's why the BS consumes a certain amount of energy while being active for the preamble reception. For a low and medium load system, the average number of preamble transmissions is negligible. Thus, we only consider a constant energy consumption of the BS that accounts for the active electronic equipment consumption.

6.6.1 Simultaneous receptions, mode 1

When all the BSs are performing as one big cell, there will be one PRACH occasion in which all the RUs are active for reception and the CU performs one processing.

1. CU energy consumption

   The energy consumed by the CU is:

   $$E^{\mathrm{UL},1}_{\mathrm{CU,PRACH},T} = T_{\mathrm{s}} P^{\mathrm{UL}}_{\mathrm{CU,f}}, \tag{6.12}$$

   where T_{s} is the slot duration and is considered the duration of the PRACH occasion. No signal processing is considered here. We consider that the BS has to be active to be able to receive probable random accesses.

2. RU energy consumption

   The energy consumed by one RU is:

   $$E^{\mathrm{UL},1}_{\mathrm{RU,PRACH},T} = T_{\mathrm{s}} P^{\mathrm{UL}}_{\mathrm{RU,f}}. \tag{6.13}$$

3. Total energy consumption

   The total energy consumed for the PRACH occasion in mode 1 is:

   $$\begin{aligned} E^{\mathrm{UL},1}_{\mathrm{Total,PRACH},T} &= E^{\mathrm{UL},1}_{\mathrm{CU,PRACH},T} + I E^{\mathrm{UL},1}_{\mathrm{RU,PRACH},T} \\ &= T_{\mathrm{s}} P^{\mathrm{UL}}_{\mathrm{CU,f}} + I T_{\mathrm{s}} P^{\mathrm{UL}}_{\mathrm{RU,f}}. \end{aligned} \tag{6.14}$$

6.6.2 Successive receptions, mode 2

When each RU is performing independently, there will be as many PRACH occasions as the number of RUs.

1. CU energy consumption

   The energy consumed by the CU in this case is:

   $$E^{\mathrm{UL},2}_{\mathrm{CU,PRACH},T} = I T_{\mathrm{s}} P^{\mathrm{UL}}_{\mathrm{CU,f}}. \tag{6.15}$$

   As mentioned earlier, the number of random accesses in a relatively low-medium load system is negligible. Therefore, we don't compute any signal processing energy consumption here.

2. RU energy consumption

   For one RU, the energy consumption is given by:

   $$E^{\mathrm{UL},2}_{\mathrm{RU,PRACH},T} = T_{\mathrm{s}} P^{\mathrm{UL}}_{\mathrm{RU,f}}. \tag{6.16}$$

3. Total energy consumption

   The total energy consumed for the PRACH reception in the mode 2 is:

   $$\begin{aligned} E^{\mathrm{UL},2}_{\mathrm{Total,PRACH},T} &= E^{\mathrm{UL},2}_{\mathrm{CU,PRACH},T} + I E^{\mathrm{UL},2}_{\mathrm{RU,PRACH},T} \\ &= I \left(T_{\mathrm{s}} P^{\mathrm{UL}}_{\mathrm{CU,f}} + T_{\mathrm{s}} P^{\mathrm{UL}}_{\mathrm{RU,f}} \right). \end{aligned} \tag{6.17}$$

6.7 Data transmission and reception energy consumption

During data transmission and reception, the CU and RUs are active for an average duration of $\bar{\tau}_q^n T$ seconds. This activity duration is used to evaluate the energy consumption of the BS during T seconds. Parameter $\bar{\tau}_q^n$ represents the average resource usage to transmit data for J users, and is given by:

$$\bar{\tau}_q^n = \frac{\mathbb{E}\left(\sum_{j=1}^{J} m_{\mathrm{RB},j,q}^n\right)}{M_{\mathrm{RB}}} \tag{6.18}$$

$$= \mathbb{E}\left(\sum_{j=1}^{J} \tau_{j,q}^n\right), \tag{6.19}$$

where $m_{\mathrm{RB},j,q}^n$ is the number of RBs needed to serve user j in mode q in direction n and can be determined using equations (5.17) and (5.20), and $\tau_{j,q}^n = \frac{m_{\mathrm{RB},j,q}^n}{M_{\mathrm{RB}}}$ from (5.11). We recall that M_{RB} is the total number of available RBs per slot.

1. CU energy consumption

   The energy consumption of the CU in mode q and direction n is given by:

   $$E_{\mathrm{CU,data},T}^{n,q} = \bar{\tau}_q^n T P_{\mathrm{CU,f}}^n + J E_{\mathrm{SP,v}}^n L_{\mathrm{P}} + E_{\mathrm{SP,f}}^n \bar{\tau}_q^n \frac{T}{T_{\mathrm{s}}}. \tag{6.20}$$

2. RU energy consumption

   The DL energy consumption of one transmitting RU in mode q consists of the electronic components and the transmission power generation energy consumption:

   $$E_{\mathrm{RU,data},T}^{\mathrm{DL},q} = \bar{\tau}_q^{\mathrm{DL}} T \left(P_{\mathrm{RU,f}}^{\mathrm{DL}} + \frac{M_{\mathrm{RB}} P_{\mathrm{t},q}^{\mathrm{DL}}}{\eta_{\mathrm{T}}}\right). \tag{6.21}$$

   The UL energy consumption of one receiving RU in mode q consists of the energy consumption of electronic components:

   $$E_{\mathrm{RU,data},T}^{\mathrm{UL},q} = \bar{\tau}_q^{\mathrm{UL}} T P_{\mathrm{RU,f}}^{\mathrm{UL}}. \tag{6.22}$$

3. Total energy consumption in mode 1

   The total energy consumption in mode 1 when all RUs receive and transmit simultaneously is:

   $$E_{\mathrm{Total,data},T}^{n,1} = E_{\mathrm{CU,data},T}^{n,1} + I E_{\mathrm{RU,data},T}^{n,1}. \tag{6.23}$$

4. Total energy consumption in mode 2

   The total energy consumption in mode 2 when one (the best) RU receives and transmits at a time is:

$$E^{n,2}_{\text{Total,data},T} = E^{n,2}_{\text{CU,data},T} + E^{n,2}_{\text{RU,data},T}. \tag{6.24}$$

6.8 Total DL and UL energy consumption

As already mentioned in Section 6.3, we evaluate the energy consumption during T the reference period of study. This period is equal to the periodicity of SSB transmissions and PRACH occasions. Therefore, during T, there is one SSB transmission, the BS should be active to ensure one PRACH occasion, and then there is the energy consumption for data transmission and reception. The total energy consumed during T when the network is operating in mode q is given by:

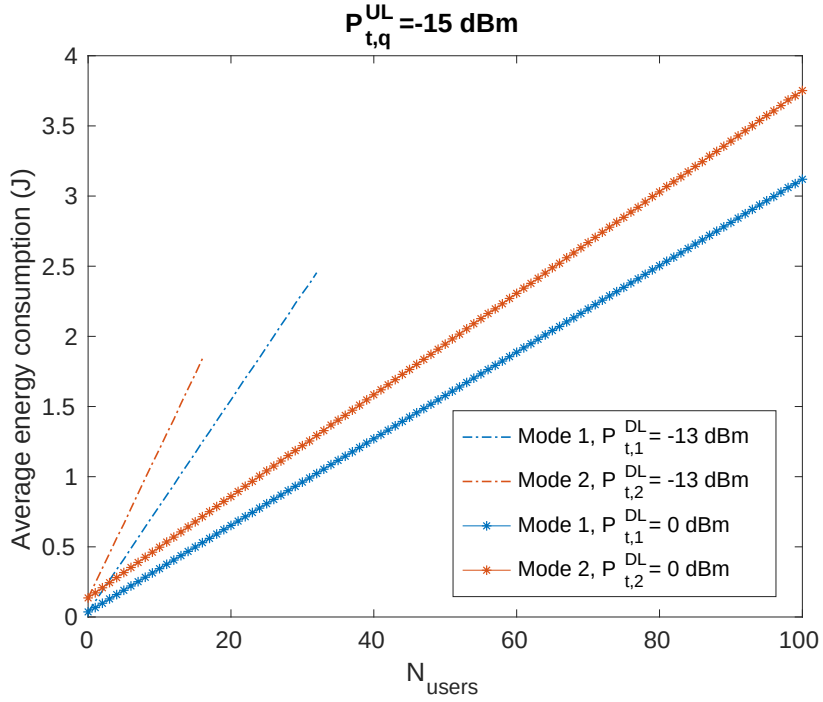
$$E^{q}_{\text{Total},T} = E^{\text{DL},q}_{\text{Total,SSB},T} + E^{\text{UL},q}_{\text{Total,PRACH},T} + E^{\text{DL},q}_{\text{Total,data},T} + E^{\text{UL},q}_{\text{Total,data},T}. \tag{6.25}$$

6.9 Results and discussions

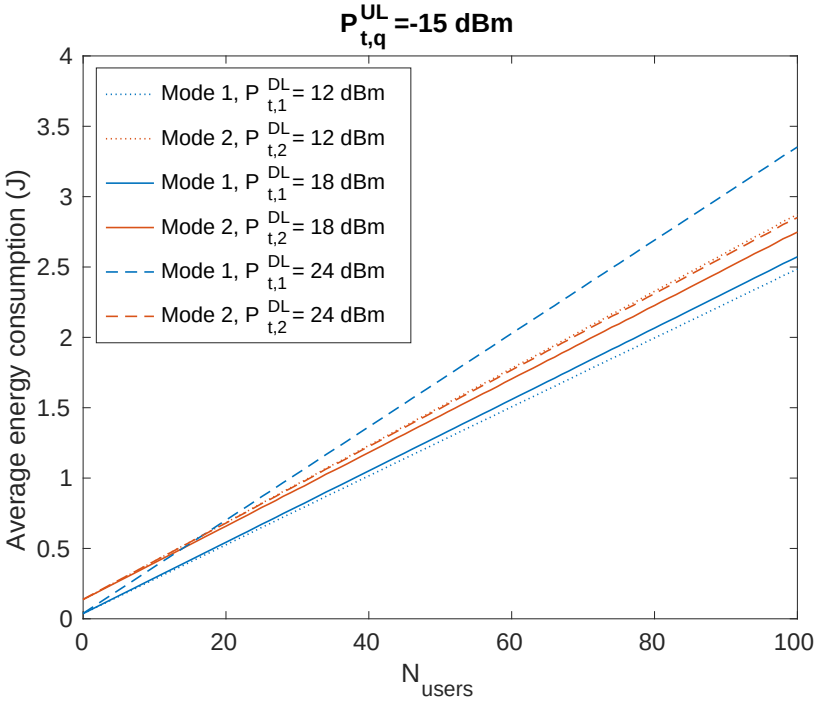
The used energy and power consumption parameters values for different electronic components and processing are available in Table 6.2.

In Figure 6.3, we represent the energy consumption as a function of the number of users for different DL transmission powers $P^{\text{DL}}_{\text{t}} = P^{\text{DL}}_{\text{t},1} = P^{\text{DL}}_{\text{t},2}$. Before starting our discussion we give some indications on the illustrated curves. Figure 6.3 represents the average total energy consumption of the BS computed using (6.25) during T. The average resource usage $\bar{\tau}^{n}_{q}$ in this equation is computed from (6.18). For a target data rate per user in direction n, Monte-Carlo simulations are conducted for J users. During each simulation, the number of needed RBs is calculated for each user using equations (5.10) and (5.11). The average of the total number of needed RBs to serve the J users is then determined based on 10000 simulations.

In Figure 6.3 we maintain $P^{\text{UL}}_{\text{t}} = P^{\text{UL}}_{\text{t},1} = P^{\text{UL}}_{\text{t},2} = -15$ dBm for all plots. Note that we don't change the transmission power in the UL direction to avoid impacting the UE energy consumption knowing that it is an energy-limited device. In Figure 6.3(a), the dash-dotted lines represent the energy consumption for $P^{\text{DL}}_{\text{t}} = -13$ dBm and $P^{\text{UL}}_{\text{t}} = -15$ dBm. These transmission powers are determined for a coverage outage of 10^{-4} (cf. Figures 5.5 and 5.6 in Section 5.9.1). Using these transmission powers, for a 10^{-2} blocking probability, up to 32 and 17 users can be served on the DL in modes 1 and 2, respectively (cf. Figure 5.9(a)). In the UL direction more users can be served before reaching this blocking probability, but we take 32 and 16 users to make sure that the blocking probability is lower than 10^{-2} in both directions (DL and UL). Thus, the energy consumption for these transmission powers, in Figure 6.3(a), is only evaluated with 32 and 17 users for modes 1 and 2, respectively. A closer look is given only on the DL energy consumption for data transmission in



(a)



(b)

Figure 6.3: BS total energy consumption for $I = 4$.

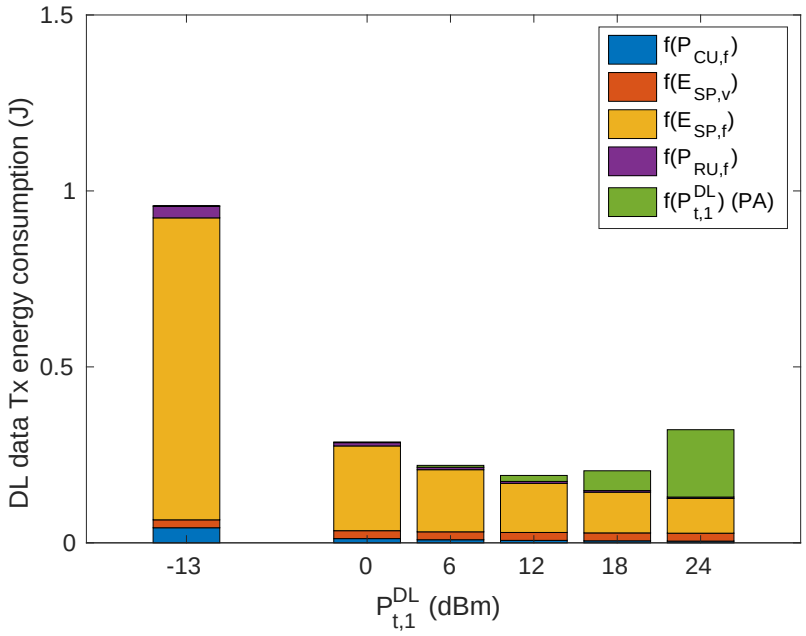
Symbol	Parameter	Value
$E^{\mathrm{DL}}_{\mathrm{SP,f}}$ (J)	Packet length independent DL signal processing EC	$\frac{25\times10^{-3}}{0.8} = 0.0312$ (d)
$E^{\mathrm{DL}}_{\mathrm{SP,v}}$ (J/bit)	Packet length dependent DL signal processing EC(a)	$\frac{30\times10^{-9}}{0.8} = 3.75\times10^{-8}$ (d)
$E^{\mathrm{UL}}_{\mathrm{SP,f}}$ (J)	Packet length independent UL signal processing EC	$\frac{25\times10^{-3}}{0.8} = 0.0312$ (d)
$E^{\mathrm{UL}}_{\mathrm{SP,v}}$ (J/bit)	Packet length dependent UL signal processing EC	$\frac{70\times10^{-9}}{0.8} = 8.75\times10^{-8}$ (d)
P_{ADC} (W)	ADC PC	0.8 (c)
P_{BPF} (W)	BPF PC	0.3 (c)
P_{DAC} (W)	DAC PC	0.135 (c)
$P_{\mathrm{E/0}}$ (W)	E/O converter PC	0.5×10^{-9} (d)
P_{LNA} (W)	LNA PC (b)	0.425 (c)
P_{LPF} (W)	LPF PC	0.3 (c)
P_{Mixer} (W)	Mixer PC	0.54 (c)
$P_{\mathrm{O/E}}$ (W)	O/E converter PC	0.5×10^{-9} (d)
$P_{\mathrm{SP,f}}$ (W)	Fixed signal processing PC	5(d)
T (ms)	Reference period of study	20
T_{s} (ms)	Time slot duration	0.25
$\eta_{\mathrm{AC/DC}}$	AC/DC power supply gain	0.8
η_{PAE}	Power Amplifier (PA) gain	0.7

(a) Energy consumption (EC) (including $\eta_{\mathrm{AC/DC}}$)
(b) Power consumption (PC)
(c) [Kumar & Gurugubelli 2011]
(d) [Jung *et al.* 2014]

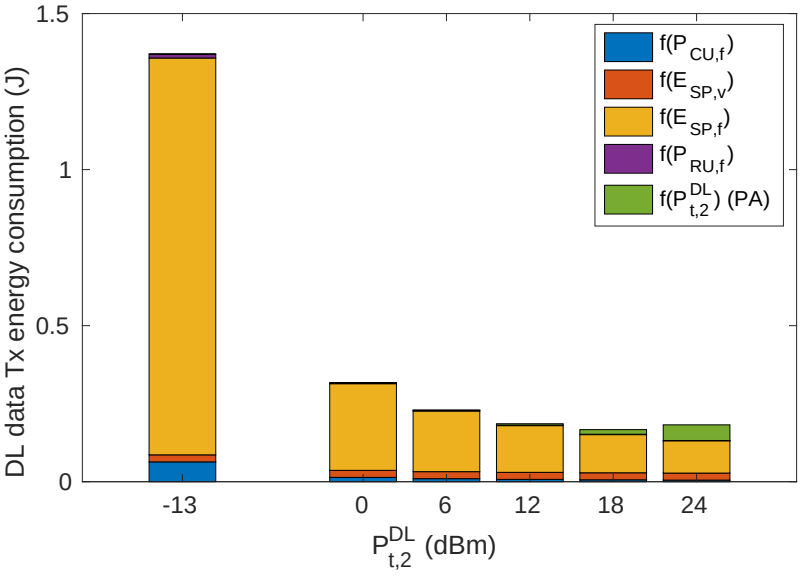
Table 6.2: Parameters values for energy consumption.

Figure 6.4. Data transmission energy consumption is obtained from (6.23) for mode 1, and (6.24) for mode 2.

Now, looking at Figure 6.3(a), we see that mode 1 consumes less energy than mode 2. In mode 1, the simultaneous transmissions increase the perceived rate at the receiver, leading to the need for less RBs. Thus the BS is active for a shorter period of time. This puts in evidence that the low load mode (mode 1) that we consider can save energy while maintaining the desired coverage. What is interesting is that increasing the DL transmission power from -13 dBm to 0 dBm reduces the total energy consumption of the BS. A higher DL transmission power ends up with less RBs needed to serve the users on the DL. Therefore, the BS is active for a shorter period in this direction. When $P^{\mathrm{DL}}_{\mathrm{t}}$ is further increased to 12 dBm, the same effects appear in Figure 6.3(b). The total energy consumption is reduced and mode 1 is still less energy-consuming compared to mode 2. Nevertheless, when $P^{\mathrm{DL}}_{\mathrm{t},q}$ is increased to



(a)



(b)

Figure 6.4: BS DL data transmission energy consumption for $I = 4$ and $J = 15$.

18 dBm, the total energy consumption in mode 2 decreases but it starts to increase in mode 1. Note that for 18 dBm, mode 1 is still less consuming than mode 2. Further increase of the transmission power, to 24 dBm for instance, results in an increase in the total energy consumption in both modes. Mode 1 also becomes more energy-consuming compared to mode 2 after 20 users.

To explain these results we take a look at the DL energy consumption during data transmission. Figure 6.4 represents the energy consumption during DL data transmission for different transmission powers per RB for 15 users. The BS energy consumption on the DL is due to multiple factors. In the CU, there is the electronic components energy consumption (function of $P_{\mathrm{CU,f}}$) and the signal processing energy consumption (function of $E_{\mathrm{SP,v}}$ and $E_{\mathrm{SP,f}}$). In the RU, there is the energy consumption of the electronic components (function of $P_{\mathrm{RU,f}}$) and the energy consumption due to radio power generation (function of the transmission power $P_{\mathrm{t},q}^{\mathrm{DL}}$).

When transmitting with low power, the generation of the radio power is not considerable. The DL BS energy consumption is thus reduced when the transmission power on the DL is increased from -13 to 12 dBm in mode 1 and from -13 to 18 dBm in mode 2 (Figure 6.4). Here, the CU is predominant in terms of energy consumption. Thus, reducing the number of RBs where the BS is active to perform the signal processing, leads to reduced energy consumption on the DL. This explains also why mode 1 is less energy-consuming compared to mode 2 even though in mode 1 the four RUs are transmitting simultaneously, while in mode 2 one RU transmits at a time.

However, when the transmission power increases a lot, its production starts to be costly in terms of energy. The energy consumption of the RU is predominant in this case leading to mode 1 being more energy consuming compared to mode 2 when $P_{\mathrm{t},q}^{\mathrm{DL}} \geq 12$ dBm. For instance when $P_{\mathrm{t},q}^{\mathrm{DL}} = 18$ dBm, the energy consumed to transmit data on the DL for 15 users is 0.2 J in mode 2 compared to 0.17 J in mode 1 (Figure 6.4). We remark that when the transmission power production starts to predominate the energy consumption, mode 1 is more affected than mode 2 because, in mode 1, all the RUs contribute to data transmission while in mode 2 only one RU transmits at a time. For instance, when increasing from 12 to 18 dBm, the DL energy consumption of mode 1 increases while it still decreases with mode 2. Also, when the DL transmission power is increased from 18 to 24 dBm, the DL energy consumption in mode 1 increases much more than in mode 2. It increases 0.12 J in mode 1 compared to only 0.015 J in mode 2. When the transmission power increases, the number of RBs where the BS is active for transmission is always reduced and the BS is active for a shorter period. However, the energy consumption for data transmission increases when the transmission power is considerable, even if the BS is active for a shorter period because its generation is energy hungry.

6.10 Conclusion

In this chapter, we evaluated the BS energy consumption for two operating modes: mode 1 and mode 2. In mode 1, all the RUs serve the user simultaneously. In mode 2, the user is served by the best RU. We presented the considered BS energy consumption model. Then, we provided the computation of the energy consumption during SSB transmissions, PRACH occasions, DL data transmission, and UL data reception.

After presenting the results, we realized that if we are using a low transmission power, mode 1 can be considered as an energy-saving mode in low load periods. We also showed that an increase in transmission power can reduce total energy consumption. However, beyond a transmission power value, its generation requires a considerable amount of energy. In that case, increasing the transmission power leads to an increase in the total energy consumption, and mode 1 becomes more energy-consuming compared to mode 2. There is a minimum energy consumption that can be achieved by varying the transmission power. The trends of the results can be generalized. However, the minimum energy consumption may be found outside the operating range of the parameters (e.g. transmission power higher than the allowed values).

If we merge this chapter's results with Chapter 5's results, we notice that using mode 1, the energy consumption can be reduced while maintaining a good coverage, without coverage holes. Also, increasing the transmission power when it is too low, not only permits receiving more users before the system is saturated but also reduces the total energy consumption.

Mode 1 is considered for low load because only one user can be served at a time. In mode 2, more users can be served at once to increase the capacity of the network (medium or high load). This is done through resource reuse and is the subject of the next chapter.

Chapter 7

DL scheduling in C-RAN for BS energy minimization

Contents

7.1 Introduction

We retake the network model and energy consumption model detailed in Chapters 5 and 6, respectively. We recall that we have two network operating modes: mode 1 where all Radio Units (RUs) cooperate to serve simultaneously one user at a time, and mode 2 where each RU performs independently and one user can be served by one RU, considered the best, at a time.

In Chapter 5, we saw that mode 1 can serve more users compared to mode 2. We also saw in Chapter 6 that mode 1 consumes less energy than mode 2 provided that the transmission power per RU is below a given value. Nevertheless, when each RU is performing independently like in mode 2, multiple users can be served on the same resources. Resource reuse can increase the number of served users. However, this may come at the cost of increased energy consumption if the interference is not well managed. Interference produced when serving different users on the same resources decreases the received Signal to Interference and Noise Ratio (SINR), which contributes to a reduction in the perceived data rate per Resource Block (RB). Thus, the number of needed RBs to achieve the target data rate may increase. The Base Station (BS) may need to be active for a longer period of time and consumes more energy. In this chapter, we aim to increase the number of served users while avoiding the total energy consumption increase. The goal is to select the serving RU as well as the number of needed RBs for each User Equipment (UE), allowing resource reuse and taking into account the potential multiple-access interference. For these reasons, we formulate a joint RB and RU allocation problem whose objective is the total BS energy consumption minimization under the same constraints and considerations as detailed in Chapter 5, but with potential resource reuse.

7.2 Related literature review

Resource allocation is an important issue to handle in wireless networks. It ensures fairness and improved performance, such as guaranteed low latency, increased throughput, and reduced energy consumption. Allocations of resources include computing resources, bandwidth, and user-BS association allocations [Nguyen 2018].

Multiple algorithms were proposed in the literature to assign available time and frequency resources to users. The authors in [Huang & Kadoch 2020] developed an approach for scheduling resources for high-load mobile networks that guarantees low latency using machine learning methods considering traditional BSs. In [Ferdouse *et al.* 2017], in a Centralized-RAN (C-RAN) context, resource allocation through an optimization problem with the objective to minimize latency was proposed. To resolve the problem, they adopted distributed scheduling in order to minimize the response time. In [Schwarz *et al.* 2010], an optimization problem has been formulated and linearized to allocate resources with the objective of maximizing the network throughput. This study handled the traditional BS architecture. The total active users' throughput in C-RAN was maximized in [Lyazidi *et al.* 2016]. Resource allocation and admission control were studied with the objective of maximizing throughput with constraints on data rate, fronthaul capacity, and transmission power.

Due to obvious environmental concerns, improving system energy efficiency has recently become one major issue and many studies focus on the optimization of scheduling algorithms with this criterion as an objective. An algorithm with global throughput maximization and Quality of Service (QoS) constraints is considered in

[Kaddour *et al.* 2015]. The work was conducted with a traditional BS consisting of one unit in Long Term Evolution (LTE) networks. The authors show that their algorithm increases energy efficiency. Korrai *et al.* proposed a resource allocation solution that reduces energy consumption using mixed numerologies [Korrai *et al.* 2020]. They considered a traditional one-unit BS in a Fifth Generation (5G) wireless network.

Energy consumption reduction with C-RAN architectures also gained the researchers' attention. Among other solutions, resource allocation was considered with C-RAN architectures to reduce this consumption. Three main resource allocation problems were tackled: the Centralized Unit (CU) computational resource allocation, the radio resource allocation, and the user-RU association. In a lot of studies, authors decided to optimize the CU computational resources in order to reduce the number of active CUs. Matching the number of computing servers with the network load was proposed in [Sigwele *et al.* 2015]. Researchers in this study considered three bin packing schemes to allocate processing tasks to CUs. The unused processing units are switched off, thereby reducing energy consumption. Also in [Tang *et al.* 2015], the authors formulated an optimization problem for computation CU capacity and power allocation. Their solution appeared to provide energy-efficient CU capacity allocation.

Energy minimization with radio resource allocation studies has also been conducted. In [Peng *et al.* 2015], RB allocation with energy efficiency maximization was proposed. Authors considered hybrid C-RAN architecture consisting of one macro cell served by a high power node and multiple small cells served by RUs. The high power node and the RUs are all connected to a CU. The RBs were divided into two groups: one is shared between the high power node and the RUs and another one shared only among the RUs like in a C-RAN. When shared among RUs, an RB can be assigned to more than one user served by different RUs. Their simulations showed a high energy efficiency with the considered resource allocation. However, in their study, one user can be connected to at most one RU. Authors in [Huang *et al.* 2016] proposed scheduling along with beamforming coordination in C-RAN. Based on their position, users are merged into groups and each group is served by a set of RUs. Resources can be reused by all the RUs from the same or different groups. Cooperative beamforming between the RUs of the same group is considered to combat interference. They were able to demonstrate good performance, such as energy efficiency, with their proposal. Nevertheless, in their study, they consider that the number of serving RUs in a cluster must be at least equal to the number of UEs in the group.

User-RU association has also gained particular attention especially to manage the produced interference in a multi-cell scenario. In [Luo *et al.* 2015], DownLink (DL) and UpLink (UL) power consumption in a C-RAN was minimized and joint user association and beamforming solution was used to manage interference. In this study, one user can be connected to one RU in each direction and can be served through beamforming. More recently, Taleb *et al.* proposed a distributed user

association and RU clustering in C-RAN [Taleb *et al.* 2020]. They divided their problem into two sub-problems: user association and RU clustering. In this study, a user can be connected to only one RU. RUs connected to the same CU share the radio resources available on this CU and can reuse these resources orthogonally, i.e. without interference. RUs connected to different CUs can cause interference if they use the same resources. Their results demonstrated reduced energy consumption, increased network throughput, and rapid adaptation to traffic variations. Energy minimization through bandwidth allocation and UE-RU association have been discussed independently. However, it is crucial to address these schemes together because the association problem has a direct impact on resource allocation and both schemes have an impact on energy consumption.

All the mentioned studies focused on fully loaded systems to optimize resource scheduling. Few studies have proposed scheduling solutions in low-load systems where some network resources are available and can be exploited to improve network performance, such as reducing power consumption for example. Authors in [Videv & Haas 2011], proposed a bandwidth trade-off method during low load with traditional BSs. This method consists of allocating more bandwidth to the user since a part of the bandwidth is free and available due to the low load. Although the user is allocated more resources, energy consumption is reduced. This is due to the ability to lower the order of the modulation scheme and adjust other link parameters while maintaining the rate achieved by the user when more bandwidth is allocated.

In this chapter, we're interested in optimizing the radio resource allocation since we consider one CU. We consider joint RU selection (or association to UE) and time-frequency resource allocation. We propose an optimization problem to schedule radio resources for users on the DL during low to medium loads considering a C-RAN architecture with an RU and CU energy consumption minimization objective. In our study, a user can be served by one RU with or without resource reuse where another RU can be serving another UE on the same resources. We also take advantage of the BS cooperation provided by C-RAN to provide a multi-point service where one user can be served by all RUs simultaneously.

7.3 Resource scheduling in modes 1 and 2

The network model considered is the same as in Chapter 5. We consider an indoor rectangular area in which we deploy I RUs to obtain the desired coverage. These RUs are connected to a CU and operate in two modes. Mode 1 is where all the RUs operate as one large cell: they have the same Synchronization Signal Block (SSB), a common opportunity for Physical Random Access CHannel (PRACH), and they transmit or receive simultaneously. In mode 2, each RU functions as an independent cell: each RU has a specific SSB and PRACH opportunity, and only one RU transmits or receives at a time. The UE position is uniformly distributed in the rectangular area.

In mode 1, to serve a user, all RUs use the same RBs simultaneously. One occupied resource engages all the RUs to transmit or receive at the same time. In mode 2, to serve one user, one RU (the best one) uses the needed RBs. When an RB is occupied, one RU is transmitting or receiving and the other $I-1$ RUs are inactive. In modes 1 and 2, an RB is fully dedicated to a user, whether it is served by all RUs (in mode 1) or by the best one (mode 2). Nevertheless, the independent operation of each RU (like in mode 2) allows the resource reuse among different RUs.

7.4 DL resource reuse scheduling

All the RUs share the same resources available in the system. Resource reuse is possible if the RUs are independent (i.e. like in mode 2). It is possible to allocate the same resources to two different users, each served by an RU, if the interference conditions allow it. This is the reuse of resources that we develop in this chapter. We consider a resource reuse scheduling on the DL.

Let us refer to the jth UE by UE_j where $j \in \mathcal{J} = \{j | 1 \leq j \leq J\}$ and to the ith RU by RU_i where $i \in \mathcal{I} = \{i | 1 \leq i \leq I\}$. We introduce an activity state for each RU denoted by s_i for RU_i and defined by:

$$s_i = \begin{cases} 1 & \text{if } \text{RU}_i \text{ is active} \\ 0 & \text{otherwise} \end{cases} \tag{7.1}$$

An activation configuration ν determines which RUs are active and which ones are idle in the system. An active RU can serve a maximum of one UE per RB. Each configuration is determined by the I states s_i of the I RUs: $\nu = (s_1, s_2, ..., s_I)$. For I RUs we have $2^I - 1$ system configurations that include the configurations where only one or more RUs are active and exclude the configuration where all the RUs are idle. The configurations are numbered from 1 to $2^I - 1$. The first I configurations are those that have only one RU in use. Then we have the configurations where more than one (two to I) RU are active and each serves different users. Configuration number $2^I - 1$ is the configuration where all RUs are active and each one is serving different users. The indicator $d(\nu, i)$ determines if RU_i is active in configuration ν or not:

$$d(\nu, i) = \begin{cases} 1 & \text{if } \text{RU}_i \text{ is active in configuration } \nu \\ 0 & \text{otherwise} \end{cases} \tag{7.2}$$

In a configuration ν, if only one RU's state is set to one, it means that this RU is operating alone on its allocated resources. In other words, this performance is similar to mode 2 in the previous chapters. If multiple RUs are active in a configuration, it means that these active RUs are using the same resources to serve different users.

We introduce a virtual RU that operates as if all RUs were active to serve one user simultaneously. When the virtual RU is not enabled, the system performs as described earlier with the $2^I - 1$ possible configurations. The virtual RU is denoted

by RU_{I+1} and the received power at user j by this virtual RU is denoted and defined by:

$$P_{\text{r},I+1,j} = \sum_{i=1}^{I} P_{\text{r},i,j}, \tag{7.3}$$

where $P_{\text{r},i,j} = P^{\text{DL}}_{\text{r},i,j}(x_j, y_j)$ is the received power at user j from RU_i and is computed using (5.5). Note that in this chapter, parameter q is omitted because we do not consider modes 1 and 2 separately any longer.

When the virtual RU is enabled we have 2^I possible configurations. To the $2^I - 1$ already described configurations, we add the configuration where the virtual RU is active and the I RUs are idle. With the virtual RU enabled, the configurations are also numbered from 1 to 2^I as mentioned above, with configuration number $\nu = 2^I$ being the one where the virtual RU is active.

When the virtual RU is not considered in the system, each RU has a specific SSB generated and transmitted periodically. An extra SSB that is common to all RUs is added to the I different SSBs for the virtual RU when considered. When a user is served by the virtual RU, it is technically served by all the RUs simultaneously. In other words, when all users are served in the configuration where only the virtual RU is active, the performance is similar to mode 1 described in the previous chapters: all RUs perform as one large cell.

7.5 Decision variable: RU-UE-configuration assignment

The objective is to minimize energy consumption while serving the users on the available RBs. The optimization problem determines the user's serving RU and the number of needed RBs for this service (depending on the chosen service configuration by the optimization problem). To schedule the resources on the DL, we need to determine which RU serves the user and in which configuration. This is determined by an order-3 tensor $\boldsymbol{x} = (x_{i,j,\nu})$, where:

$$x_{i,j,\nu} = \begin{cases} 1 & \text{if } \text{RU}_i \text{ is serving } \text{UE}_j \text{ with configuration } \nu \\ 0 & \text{otherwise} \end{cases} \tag{7.4}$$

This decision variable is the output of the optimization problem that schedules the service of the users.

The system's constraint on this decision variable ensures that each UE is served and served only once: definitely served by one RU in one configuration only. This constraint is formulated as follows:

$$\sum_{i,\nu} x_{i,j,\nu} d(\nu, i) = 1 \ \ \forall\, j. \tag{7.5}$$

7.6 DL transmission model

We study resource reuse in the DL direction. In this section, we provide the SINR and the calculation of the needed RBs to serve one user by an RU_i in a configuration ν.

7.6.1 SINR computation

With resource reuse, if more than one RU transmit on the same RBs, they interfere with each other. Thus, the quality of the wireless connection between the BS and the UE has to include the signal of interest, the interfering signal, and the receiver's noise. When the virtual RU is not considered, the SINR for a user j served by RU_i in configuration ν is given by:

$$\gamma_{i,j,\nu} = \begin{cases} \frac{P_{\mathrm{r},i,j}}{\sum_{i' \neq i} P_{\mathrm{r},i',j} d(\nu,i') + N_{\mathrm{p}}^{\mathrm{DL}}} & \text{if } d(\nu,i) = 1 \ \ \forall i \leq I \\ \text{n.a.} & \text{otherwise} \end{cases} \tag{7.6}$$

where $N_{\mathrm{p}}^{\mathrm{DL}}$ is the noise power at the UE side and is given by (5.6).

When the virtual RU is enabled, the perceived SINR is given by:

$$\gamma_{i,j,\nu} = \begin{cases} \frac{P_{\mathrm{r},i,j}}{\sum_{i' \neq i} P_{\mathrm{r},i',j} d(\nu,i') + N_{\mathrm{p}}^{\mathrm{DL}}} & \text{if } \nu \neq 2^I,\ d(\nu,i) = 1 \ \ \forall i \neq I+1 \\ \frac{\sum_{i=1}^{I} P_{\mathrm{r},i,j}}{N_{\mathrm{p}}^{\mathrm{DL}}} & \text{if } \nu = 2^I,\ d(\nu, I+1) = 1 \\ \text{n.a.} & \text{otherwise} \end{cases} \tag{7.7}$$

We use the SINR to evaluate the number of needed RBs.

7.6.2 Number of needed resource blocks computation

We aim to provide each user with the desired target data rate $R_{\mathrm{T}}^{\mathrm{DL}}$. Based on the configuration ν where the user is served and the RU_i that is serving a user j, the perceived data rate per RB is computed using (5.9):

$$R_{(i,j,\nu)}^{\mathrm{DL}} = w_{\mathrm{RB}} \delta_{\mathrm{BW}} \log_2 \left(1 + \frac{\gamma_{i,j,\nu}}{\delta_{\mathrm{SNR}}} \right), \tag{7.8}$$

where w_{RB} is the bandwidth occupied by one RB, and δ_{BW} and δ_{SNR} are correction factors of the bandwidth and the Signal to Noise Ratio (SNR), respectively [Mogensen *et al.* 2007].

The number of needed RBs to achieve the target data rate on the DL $R_{\mathrm{T}}^{\mathrm{DL}}$ is then determined using (5.10) and (5.11):

$$m(i,j,\nu) = \frac{R_{\mathrm{T}}^{\mathrm{DL}}}{R_{(i,j,\nu)}^{\mathrm{DL}}}. \tag{7.9}$$

7.6.3 Occupied resources on the system level

Each UE_j is served by an RU_i in a configuration ν. To achieve the user's target data rate in the mentioned conditions, a certain number of RBs is needed. To serve a set of users in a configuration ν, the active RUs occupy a number of RBs. In each configuration, each active RU requires a certain number of RBs to serve the users assigned to it. Each RU occupies the sum of the needed RBs to serve its assigned users in the considered configuration. The number of occupied RBs on the system level y_ν during each configuration ν is the maximum of the number of RBs occupied by all the active RUs in this configuration:

$$y_\nu = \max_i \left(\sum_j m(i,j,\nu) x_{i,j,\nu} \right). \tag{7.10}$$

7.7 Objective function: DL data transmission energy consumption

The optimization problem objective is to minimize the BS energy consumption. In this section, we provide the energy consumption computation. The energy consumption model is taken from Chapter 6. We adapt hereafter the equations to the network functioning using the decision variables.

1. CU energy consumption

   The energy consumption of the CU has three parts. The first is the fixed power consumption of the electronic components $P^{\text{DL}}_{\text{CU,f}}$ which is counted during the CU activity period. It is consumed as soon as the system is active: independent of the number of active RUs or served UEs. The second part is the fixed signal processing energy consumption $E^{\text{DL}}_{\text{SP,f}}$ that is consumed per processed packet per user. We consider one packet per slot. Thus, to evaluate this energy consumption, the number of occupied slots to serve all the users is counted. Note that here each RB is counted as many times as used to serve different users. For instance if one RB is used to serve UE_j and $\text{UE}_{j'}$ by RU_i and $\text{RU}_{i'}$, respectively, it is counted twice. The third part is the energy consumption that depends on the packet length $E^{\text{DL}}_{\text{SP,v}}$. This is multiplied by the total number of served users. Note that the constants $P^{\text{DL}}_{\text{CU,f}}$, $E^{\text{DL}}_{\text{SP,f}}$, and $E^{\text{DL}}_{\text{SP,v}}$ include $\eta_{\text{AC/DC}}$ the AC/DC gain. The energy consumption of the CU is therefore given by:

   $$E^{\text{DL}}_{\text{CU}}(\boldsymbol{x}) = \frac{\sum_\nu y_\nu}{M_{\text{RB}}} P^{\text{DL}}_{\text{CU,f}} T + \frac{\sum_i \sum_j \sum_\nu m(i,j,\nu) x_{i,j,\nu}}{M_{\text{RB}}} E^{\text{DL}}_{\text{SP,f}} \frac{T}{T_{\text{s}}} + J E^{\text{DL}}_{\text{SP,v}} L^{\text{DL}}_{\text{p}}, \tag{7.11}$$

   where $\sum_\nu y_\nu$ is the total number of RBs occupied by used configurations on the system level, M_{RB} the total number of available RBs per slot, T the reference period of study (c.f. Section 6.3), $\sum_i \sum_j \sum_\nu m(i,j,\nu) x_{i,j,\nu}$ the number of RBs used to serve all the users, T_{s} the slot duration, and L^{DL}_{p} the transmitted packet length during T, and is given by $L^{\text{DL}}_{\text{p}} = R^{\text{DL}}_{\text{T}} T$.

2. Energy consumption of I real RUs

   Here, we provide the energy consumption of the I RUs when the virtual RU is not enabled. With or without resource reuse, an RU must be active to serve each user for the duration of the RBs it needs to reach the user's target data rate. The energy consumption of the I RUs is computed by counting the transmission and fixed power consumption during all the RBs needed to serve the J users:

$$E_{\mathrm{RU}}^{\mathrm{DL}}(\boldsymbol{x}) = \sum_{i,\nu} \frac{\sum_j m(i,j,\nu) x_{i,j,\nu}}{M_{\mathrm{RB}}} \left(P_{\mathrm{RU,f}}^{\mathrm{DL}} + \frac{M_{\mathrm{RB}} P_{\mathrm{t}}^{\mathrm{DL}}}{\eta_{\mathrm{T}}} \right) T, \tag{7.12}$$

   where $P_{\mathrm{t}}^{\mathrm{DL}}$ is the BS transmission power, and $\eta_{\mathrm{T}} = \eta_{\mathrm{PA}} \eta_{\mathrm{AC/DC}}$ includes the power amplifier and AC/DC gains.

3. Energy consumption of I real and one virtual RU

   The virtual RU performs as if the I RUs were active to serve the same user simultaneously. The energy consumption of this RU is thus multiplied by I the number of real RUs in the system. The energy consumption of I real and one virtual RU to serve J users is given by:

$$E_{\mathrm{RU}}^{\mathrm{DL}}(\boldsymbol{x}) = \left(\frac{\sum_{i=1}^{I} \sum_{\nu} \sum_j m(i,j,\nu) x_{i,j,\nu} + I \sum_j m(I+1,j,2^I) x_{I+1,j,2^I}}{M_{\mathrm{RB}}} \right) \times \left(P_{\mathrm{RU,f}}^{\mathrm{DL}} + \frac{M_{\mathrm{RB}} P_{\mathrm{t}}^{\mathrm{DL}}}{\eta_{\mathrm{T}}} \right) T, \tag{7.13}$$

   where I is the number of deployed (real) RUs, $i = I + 1$ the index referring to the virtual RU, and $\nu = 2^I$ the configuration in which the virtual RU is active.

7.8 SSB transmission energy consumption

The SSBs are transmitted periodically, regardless of the served UEs in the system. When the virtual RU is not considered, I different SSBs are generated in the CU. Each RU then broadcasts its corresponding SSB. The reference period of study T is chosen equal to the SSB transmission periodicity. Thus, during T, the SSB generation and transmission is done once. The energy consumption of the SSB generation and transmission during T is given by:

$$E_{\mathrm{Total,SSB},T}^{\mathrm{DL}} = I \left(T_{\mathrm{SSB}} P_{\mathrm{CU,f}}^{\mathrm{DL}} + E_{\mathrm{SP,f}}^{\mathrm{DL}} + T_{\mathrm{SSB}} \left(P_{\mathrm{RU,f}}^{\mathrm{DL}} + \frac{m_{\mathrm{RB}}^{\mathrm{SSB}} P_{\mathrm{t}}^{\mathrm{DL}}}{\eta_{\mathrm{T}}} \right) \right), \tag{7.14}$$

where $m_{\mathrm{RB}}^{\mathrm{SSB}}$ is the number of RBs occupied during an SSB transmission.

The virtual RU functions as if the I RUs are functioning as one large cell. A common SSB is generated in the CU and transmitted simultaneously by all RUs. This SSB is useful to the user to connect simultaneously to the I RUs in configuration number 2^I where only the virtual RU is active. In addition, each RU needs a specific SSB to be generated and transmitted independently of the others. These SSBs are used by the user to connect to each of the I RUs separately when $\nu \leq 2^I - 1$. In the CU, $I+1$ SSBs are generated (I for the real RUs and one for the virtual RU). Each RU has to transmit two SSBs: one specific to it and one common with other RUs for the simultaneous transmissions. Thus, the energy consumption to generate and transmit SSBs when the virtual RU is enabled is given by:

$$E^{\mathrm{DL}}_{\mathrm{Total,SSB},T} = (I+1)\left(T_{\mathrm{SSB}} P^{\mathrm{DL}}_{\mathrm{CU,f}} + E^{\mathrm{DL}}_{\mathrm{SP,f}}\right) + 2IT_{\mathrm{SSB}}\left(P^{\mathrm{DL}}_{\mathrm{RU,f}} + \frac{m^{\mathrm{SSB}}_{\mathrm{RB}} P^{\mathrm{DL}}_{\mathrm{t}}}{\eta_{\mathrm{T}}}\right). \quad (7.15)$$

The SSB energy consumption is constant and independent of the number of served users and resource scheduling during data transmission. We thus do not include it in the objective function but we include it in the BS energy consumption evaluation.

7.9 Optimization problem formulation

As mentioned earlier, we aim to minimize the BS energy consumption during DL data transmission. To properly schedule the service of available users, we have some constraints to guarantee the good functioning of the system.

The optimization problem is formulated as follows:

$$\begin{aligned}
&\text{minimize } E^{\mathrm{DL}}_{\mathrm{CU}}(\boldsymbol{x}) + E^{\mathrm{DL}}_{\mathrm{RU}}(\boldsymbol{x}) && (7.16a)\\
&\text{s.t. } \mathbf{C1:} \sum_{i,\nu} x_{i,j,\nu} d(\nu,i) = 1 \ \ \forall\, j && (7.16b)\\
&\mathbf{C2:} \sum_{\nu} y_\nu \leq M_{\mathrm{RB}} && (7.16c)\\
&\text{with } y_\nu = \max_i \left(\sum_j m(i,j,\nu) x_{i,j,\nu}\right)\\
&\mathbf{C3:} \ x_{i,j,\nu} \in \{0,1\} \ \ \forall\, i,j,\nu && (7.16d)
\end{aligned}$$

The decision variable is $\boldsymbol{x} = (x_{i,j,\nu})$.

The objective function in (7.16a) is the minimization of the DL data transmission energy consumption provided by (7.11) and (7.12) or (7.13). The SSB transmission energy consumption is independent of the decision variable $\boldsymbol{x} = (x_{i,j,\nu})$. Thus, when we evaluate the energy consumption including the SSB transmissions, we add the energy consumed for SSB transmission to the optimal energy consumed by the BS after the resolution of the optimization problem.

Constraint **C1** in (7.16b) is defined in Section 7.5. It guarantees that all users are served and that each user is served by one RU in one configuration.

Constraint **C2** in (7.16c) is the blocking constraint that forces the algorithm to find a solution without exceeding the number of RBs available in the system (M_{RB}). The number of occupied RBs on the system level in configuration ν is determined by y_ν, and this is what counts to evaluate the system blocking.

Constraint **C3** in (7.16d) defines the set of $x_{i,j,\nu}$ as binary variables.

The objective function and constraint **C1** are linear. Constraint **C2** contains y_ν that is a maximum and function of $m(i,j,\nu)$ that is not linear. With the binary variables in **C3**, the problem is Mixed Integer NonLinear Programming (MINLP).

7.9.1 MILP optimization problem formulation

In our problem resolution, the preliminary task is the computation of $m(i,j,\nu)$ for all triplets (i,j,ν). Then, these values are used in the optimization resolution. Also, to replace the max function in the computation of y_ν in (7.10), we add the constraint **C4** detailed hereinafter to the previously formulated problem. The modified optimization problem becomes:

$$\text{minimize } E_{\mathrm{CU}}^{\mathrm{DL}}(\boldsymbol{x}, y_\nu) + E_{\mathrm{RU}}^{\mathrm{DL}}(\boldsymbol{x}, y_\nu) \tag{7.17a}$$

$$\text{s.t. } \mathbf{C1:} \sum_{i,\nu} x_{i,j,\nu} d(\nu,i) = 1 \ \ \forall\, j \tag{7.17b}$$

$$\mathbf{C2:} \sum_{\nu} y_\nu \leq M_{\mathrm{RB}} \tag{7.17c}$$

$$\mathbf{C3:}\ x_{i,j,\nu} \in \{0,1\} \ \ \forall\, i,j,\nu \tag{7.17d}$$

$$\mathbf{C4:}\ y_\nu \geq \sum_{j} m(i,j,\nu) x_{i,j,\nu} \ \ \forall\, i,\nu \tag{7.17e}$$

Constraint **C4** in (7.17e) is the new constraint that rewrites (7.10) linearly with all $m(i,j,\nu)$ being pre-calculated. Variable y_ν also becomes a new decision variable. This decision variable is a vector whose length equals 2^I or $2^I - 1$ depending on whether the virtual RU is considered or not. The new constraint imposes restrictions on the number of occupied system-level RBs in configuration ν. In each configuration, each active RU serves a certain number of users. The RBs occupied by one RU_i are the sum of the RBs necessary to serve all the users served by this RU_i ($\sum_j m(i,j,\nu) x_{i,j,\nu}$). The resources occupied by each RU are evaluated for all active RUs in the considered configuration ($\forall\, i$). The number of occupied RBs at the system level in configuration ν (y_ν) must be greater than the RBs occupied by each active RU in this configuration. This must be true for all the considered configurations ($\forall\, \nu$). The other elements of the optimization problem remain unchanged. The formulated problem in (7.17a)-(7.17e) is thus Mixed Integer Linear Programming (MILP).

7.10 Results and discussions

In this section, we evaluate the performance of the optimization problem in (7.17). We first evaluate the execution time to solve this problem. We compute the blocking probability that is defined by the probability that the required RBs number exceeds the number of available RBs. For more details on the definition of the blocking probability, please see Section 5.6. We also measure the energy consumption produced on the DL.

7.10.1 Simulations

We consider four RUs in a rectangular area implemented as in Figure 5.4(b). The charge in the network is represented by J the number of available users. For a certain charge, J users' positions are randomly chosen with random shadowing. For all the users, $m(i, j, \nu)$ in all possible configurations is calculated with parameters from Table 5.3. The energy consumption parameters of the objective function are taken from Table 6.2.

The optimization problem is run with the random selection using Cplex with Python. These steps are done 10000 times for each charge. The energy consumption is then evaluated by averaging the objective function value after the problem resolution over the trials where the solver can find a solution. If we observe the different constraints of the optimization problem in (7.17), we remark that the only one that can prevent the solver from finding a solution is constraint **C2** in (7.17c). Thus, when the solver can't find a solution, i.e. can't respect constraint **C2** in (7.17c), the system is blocked. So, when the optimization problem solver is run, if a solution is found, it is taken into account to evaluate the energy consumption and the usage of the configurations. If not, the system is considered blocked. The number of times when the solver can't find a solution is counted among the 10000 trials to evaluate the blocking probability.

7.10.2 Optimization execution time

The simulations were performed on a remote server with 792 GB of RAM and 96 cores with simultaneous multi-threading. The optimization problem resolution execution time is given in Figure 7.1. In the Y axis, we have the execution time per simulation in seconds. In the X axis, we have the network load (given by J the number of users). The curve represents the execution time when the optimization problem is considered without the virtual RU as a function of the network load. The given time is averaged on 10000 simulations for each value of J. The execution time increases with the network load. Increasing the number of users increases the number of variables and constraints in the optimization problem. The execution time increase is relatively slow for the low loads (for $J \leq 30$ users) and increases faster for high loads when the maximum system capacity starts to be reached. Nevertheless, even with this increase, the execution time remains acceptable. For 30 users, one

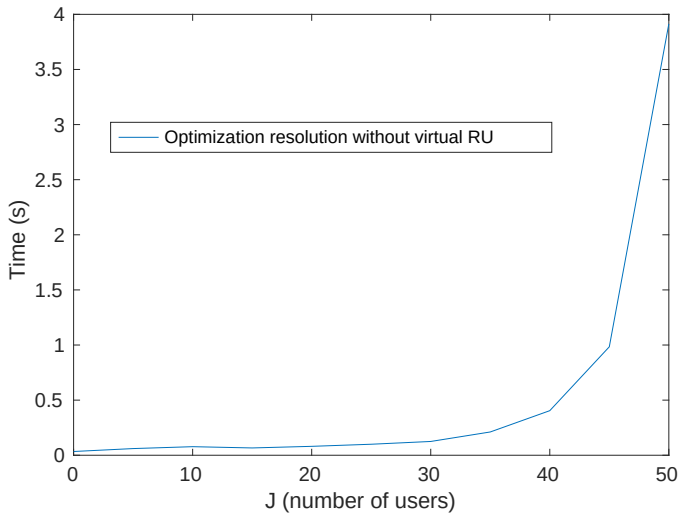


Figure 7.1: Optimization resolution execution time per simulation.

simulation takes 0.125 second while for 45 users, it takes 0.983 second and 3.91 seconds for $J = 50$.

7.10.3 Blocking probability

Now, let us evaluate the blocking probability of the optimization problem provided in (7.17). Figure 7.2 shows the blocking probability as a function of the number of users available in the network for mode 1, mode 2 (see Figure 5.9(a)), and the output of the optimization problem. The transmission power is considered equal to -13 dBm. We note that considering different configurations with four RUs with and without virtual RU increases the system capacity. The configurations without virtual RU include transmission from a single RU (like mode 2) and configurations with resource reuse. When the virtual RU is considered, in addition to the mentioned configurations, we have the transmission from all RUs simultaneously (using the virtual RU, similar to mode 1). For a blocking probability of 10^{-2}, in mode 2, 17 users can be served compared to 32 in mode 1, 48 without virtual RU, and 50 with virtual RU. In comparison to mode 1, the proposed solution increases the system capacity by 50% and 56% without and with virtual RU, respectively. To further explain this increase in capacity, we look at the configurations used as a function of the network load (or J the number of users).

In Figure 7.3, we evaluate the usage of the configurations. On the X axis of Figures 7.3(a) and 7.3(b), we have the 15 and 16 configurations, respectively. Each configuration is represented by the active RUs it contains. For example, configuration 5 is denoted by RU(3,4), which means that in this configuration, RU_3 and RU_4 are active and serve different users on the same RBs. On the Y axis, we count the number of RBs where each configuration is active. For example, for $J = 50$ in Figure 7.3(b), the configuration where RU_1 and RU_4 are active is used on average

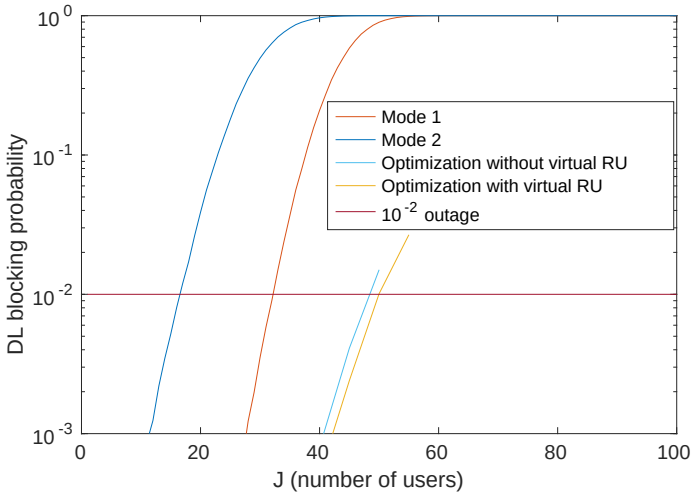


Figure 7.2: DL blocking probability.

on 2.7 RBs. In other words, RU_1 and RU_4 are active and reuse 2.7 RBs to serve different users on these resources. A zoom-in in Figure 7.3(b) is given in the box for a better view of the configurations that are used on less than 6 RBs.

For a low load like $J = 10$ in Figure 7.3, we can see that the solver of the optimization problem tends to serve users with configurations where only one RU is active: configurations number 1 to 4 (transmission by a single RU, similar to mode 2) and configuration number 16 where only the virtual RU is active (simultaneous transmissions, similar to mode 1) when the virtual RU is considered. The objective of the optimization problem is to minimize energy consumption. Simultaneous service increases the perceived SNR and reduces thus the number of needed RBs to reach the target data rate. Since the energy consumed by the BS majorly depends on the activity duration of the CU with a transmission power of -13 dBm, serving users with the virtual RU, when considered, reduces energy consumption (mode 1 is less energy consuming than mode 2, see Chapter 6). However, even when the virtual RU is considered, the configurations with a single RU serving are still used (configurations 1 to 4). In these specific cases, it is possible that serving these users with one or four RUs simultaneously requires almost the same number of RBs. For example, when channel conditions are very good between a user and RU_1 and very bad between that user and the other three RUs. Here, serving this user by RU_1 requires a certain number of RBs. The simultaneous service does not offer a big enhancement in the reduction of the number of needed RBs (bad channel conditions with RUs 2, 3, and 4). Thus, the solver of the optimization problem chooses to serve users like this with one RU instead of four simultaneously (even if virtual RU is considered) to save the energy of three active RUs that are not decreasing the number of RBs. This is why we have some RBs where the first four configurations are used in Figure 7.3(b).

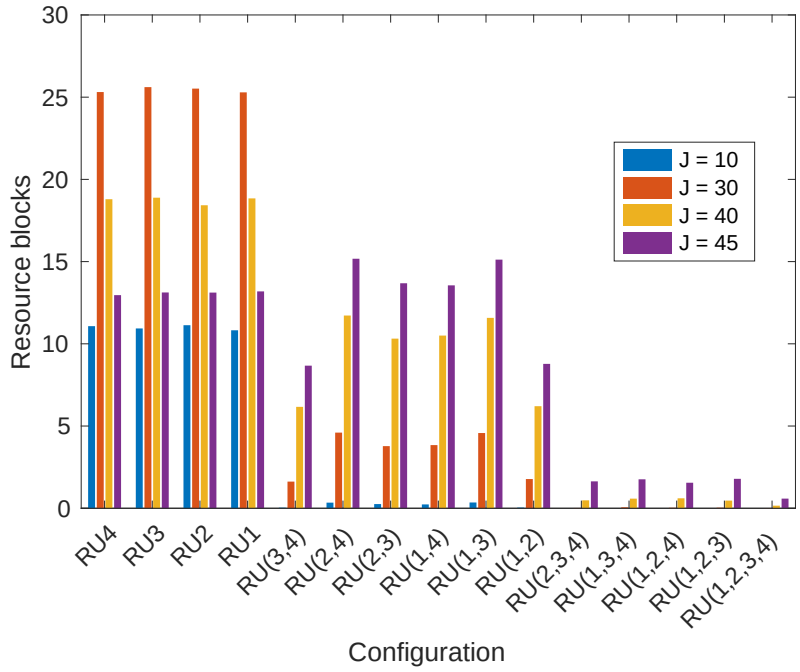
When the load increases and the maximum system capacity starts to be reached

(see Figure 7.4), the configurations with resource reuse are used as shown in Figure 7.3. It can be seen that when the virtual RU is not considered, resource reuse is more important compared to the case with virtual RU. The resource reuse starts to appear on more than one RB per configuration with lower loads (less than 30 users, see Appendix C) without the virtual RU, while there is no significant resource reuse for this load with the virtual RU. In fact, when the virtual RU is active, most users are assigned to that virtual RU. With the virtual RU in Figure 7.3(b), for $J < 40$ users, resource reuse by only two RUs is used on less than one RB per configuration. For $J \geq 40$ users, more configurations with resource reuse are used (on more than one RB per configuration) (see Appendix C). The configurations with three and four RUs performing resource reuse are increasingly used in these cases. Without virtual RU, with $J = 45$, on average, 83 RBs are used by configurations with resource reuse (last 11 configurations) compared to 52 RBs without resource reuse (first four configurations). While the use of configurations with resource reuse does not exceed that with service by one RU when considering the virtual RU even with 55 users (40 RBs with resource reuse compared to 95 without resource reuse). This behavior decreases the use of the first four configurations and the last one when considered. For instance, with virtual RU in Figure 7.3(b), for 45, 50, and 55 users, the use of the first four configurations is reduced and with 50 and 55 users, the use of the last configuration is also reduced. A more complete figure that contains the use of configurations for more network load cases is given in Appendix C.

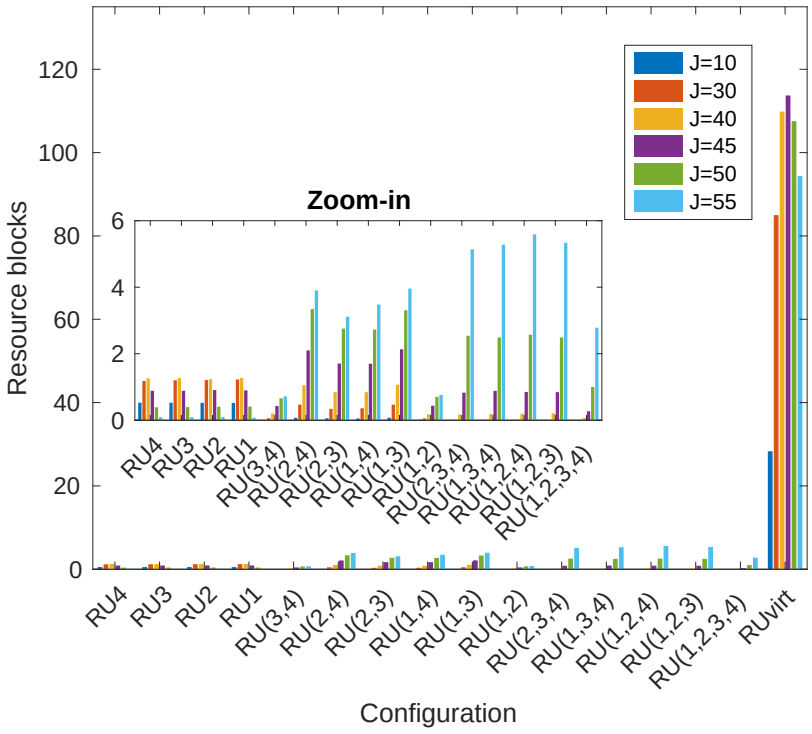
The mean and standard deviation of the total number of required RBs for each load are given in Figure 7.4. The maximum capacity of the system is represented by the horizontal line for $M_{\text{RB}} = 135$. The average number of needed RBs increases linearly with the system load up to $J = 20$ and $J = 40$ without virtual RU and with virtual RU, respectively. During the linear phase, about 4.5 RBs are required per user on average without virtual RU compared to 3 RBs with virtual RU. For higher loads, a slower increase is observed since the constraint **C2** in (7.17c) prevents exceeding M_{RB}. This also explains the decrease in standard deviation for more than 35 users after it increased with the network load from zero to 30 users in Figure 7.4(b) with virtual RU, for example.

7.10.4 Energy consumption

The energy consumed by the BSs for data transmission is given in Figure 7.5. In Figure 7.5(a), we compare the energy consumed with the optimization solution with and without virtual RU against the energy consumed by modes 1 and 2 considered alone (see Chapter 6). The energy consumption with the optimization problem with virtual RU and without virtual RU is similar to the energy consumption of mode 1 and mode 2, respectively. The energy consumed increases linearly with the number of users up to 40 users without virtual RU and 45 users with virtual RU. For these loads and beyond, resource reuse becomes important (as seen in Figure 7.3) and energy consumption increases faster than the previous linear increase. This means that with resource reuse, the interference generated increases the number of

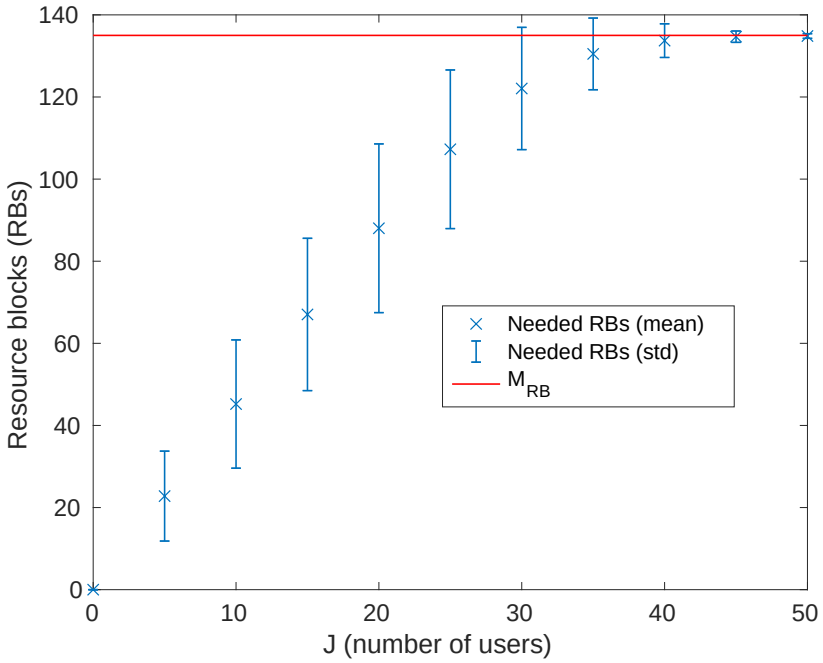


(a) Without virtual RU

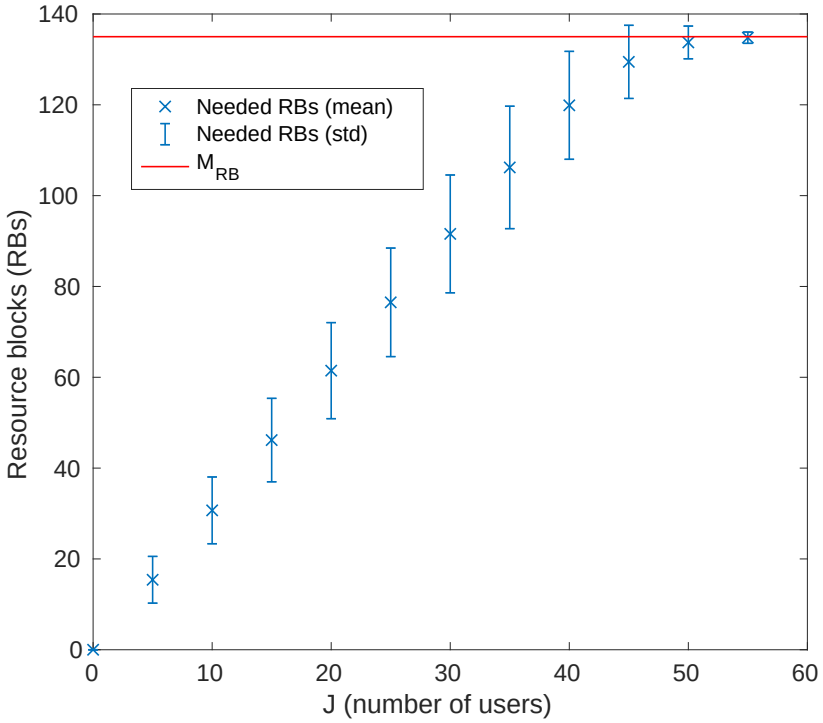


(b) With virtual RU

Figure 7.3: Configurations usage for $I = 4$ without and with virtual RU.



(a) Without virtual RU



(b) With virtual RU

Figure 7.4: Number of RBs needed as a function of J the number of users.

RBs needed to serve users on the reused RBs. Thus, energy consumption becomes significant. This explains the results of the previous section where resource reuse is almost not used unless it is really needed (when the maximum system capacity is reached without resource reuse).

The energy consumed by the BSs to transmit the data with the virtual RUs is lower than that without the virtual RUs. However, simultaneous transmissions by the virtual RU require an additional SSB transmission which consumes some energy to be transmitted. In Figure 7.5(b), we show the energy consumption for data and SSB transmission with and without the virtual RU. The additional energy to transmit the additional SSB with the virtual RU appears when there is no load ($J = 0$). This increase is so small that its impact disappears quickly and the energy consumption with the virtual RU is lower than the energy consumption without it, even with the additional SSB.

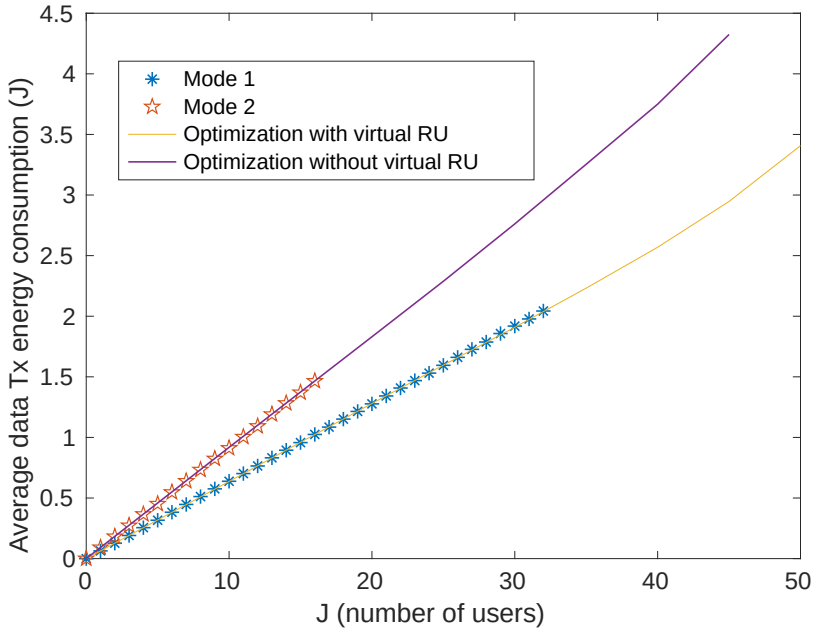
The optimization problem with the virtual RU consumes less energy than in the case without the virtual RU. The simultaneous transmissions provided by the virtual RU improve the user-perceived SNR and decrease the number of RBs where the BS is active. Since the power consumption of the CU is predominant with $P_{\mathrm{t}}^{\mathrm{DL}} = -13$ dBm (see Figure 6.4), even if all four RUs are transmitting, the power consumption is reduced.

7.10.5 Impact of BS parameters variation

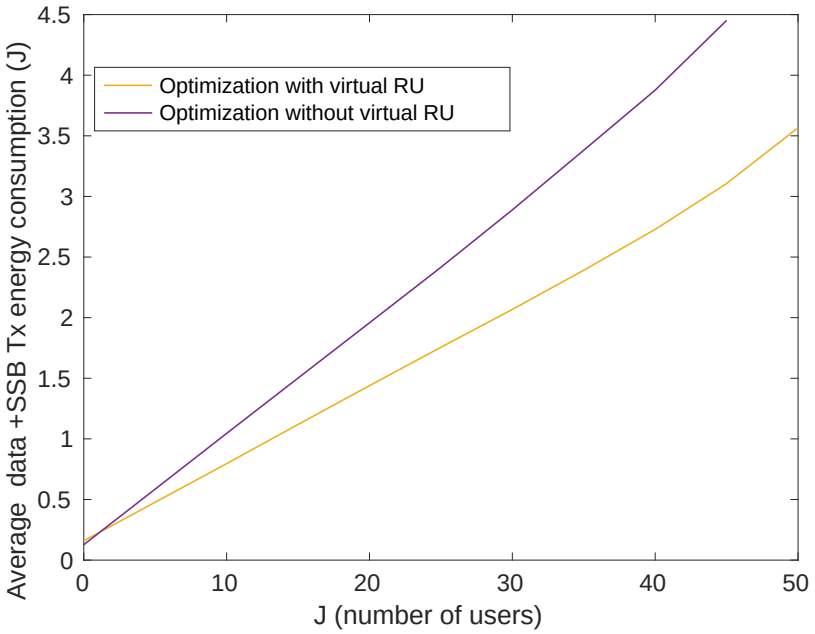
The results of our study depend on the values of several parameters. We have taken these values from some scientific articles. However, there are uncertainties as they may depend on the type of hardware or the software implementation. Therefore, we vary some parameters in this section and analyse their influence on the performance.

In this section, for each parameter variation, the optimization problem is run 10000 times, each time with random UE positions and shadowing. The results are then averaged over these trials.

When considering only modes 1 and 2 as in Chapters 5 and 6, varying the fixed energy consumption parameters has no impact on resource allocation. The impact is only on energy consumption. However, with the proposed optimization problem, the variation of the fixed parameters has a direct impact on resource allocation and energy consumption since the allocation is the output of the optimization problem that aims to minimize energy consumption. Referring back to Figure 6.4, we see that the energy consumption due to the fixed signal processing energy ($E_{\mathrm{SP,f}}$) predominates when $P_{\mathrm{t}}^{\mathrm{DL}} = -13$ dBm. We consider here the decrease of $E_{\mathrm{SP,f}}$ and the increase of $P_{\mathrm{t}}^{\mathrm{DL}}$ the transmission power. Parameter $E_{\mathrm{SP,f}}$ can be decreased if some optimized signal processing techniques are provided for example. Note that increasing the transmission power has an impact on the number of needed RBs. For example, it improves the perceived SNR and reduces the number of RBs needed in the case of a single transmission without interference. When we decrease $E_{\mathrm{SP,f}}$, the number of required RBs remains unchanged unless the assignment of the RU-UE-configuration is changed. The results of these variations are shown in Figures 7.6



(a) Data transmission energy consumption.



(b) Data and SSB transmission energy consumption.

Figure 7.5: DL energy consumption.

and 7.7 and are given for a load with $J = 40$ users.

When the transmission power is increased, the use of virtual RU is reduced and the use of configurations with transmission by a single RU is favored as shown in Figure 7.6(a). In this case, the generation of the transmission power in the RU consumes more energy, and the transmission with all RUs simultaneously to serve a user is strongly disadvantaged. For high transmission power such as $P_{\mathrm{t}}^{\mathrm{DL}} = 24$ dBm, only the first four configurations are used.

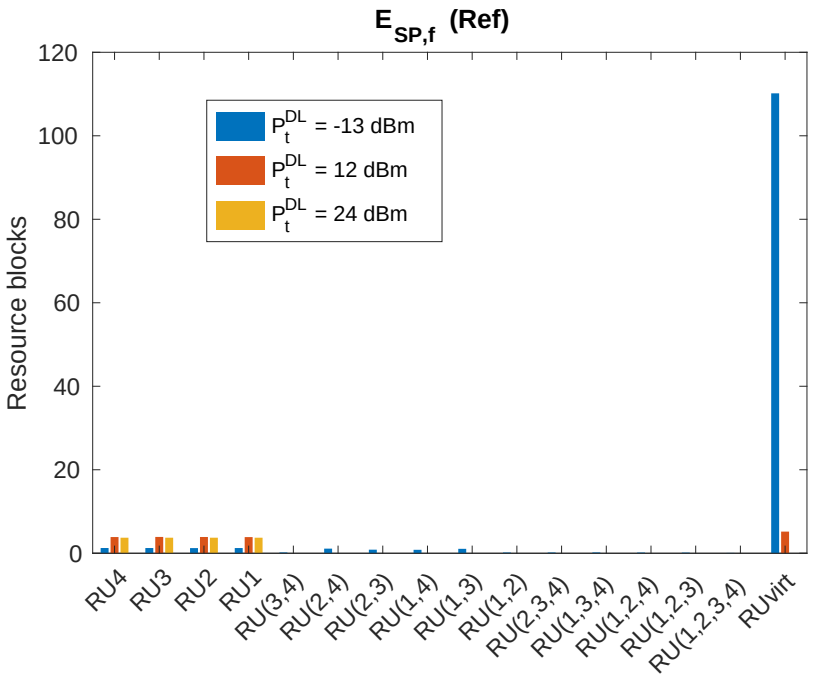
The reduction in $E_{\mathrm{SP,f}}$ privileged resource reuse. We can see from Figure 7.6(b) that configurations with resource reuse are more used with lower $E_{\mathrm{SP,f}}$. Note that $E_{\mathrm{SP,f}}$ (Ref) refers to the original value of this parameter, previously used for the evaluation of energy consumption (see Table 6.2). In this figure, the transmission power is low and when we decrease $E_{\mathrm{SP,f}}$, the use of the virtual RU is a bit disadvantaged (less than when we decrease the transmission power).

To explain this behavior, we examine in detail the energy consumption function of different parameters with variations of $P_{\mathrm{t}}^{\mathrm{DL}}$ and $E_{\mathrm{SP,f}}$ in Figure 7.7. This figure represents the energy consumed for data transmission as a function of $P_{\mathrm{t}}^{\mathrm{DL}}$ the transmission power for different values of $E_{\mathrm{SP,f}}$. The energy consumed is divided into five parts according to the different BS consuming components, including the fixed consumption of the CU ($f(P_{\mathrm{CU,f}})$), signal processing in the CU ($f(E_{\mathrm{SP,v}})$ and $f(E_{\mathrm{SP,f}})$), the fixed consumption of the RU ($f(P_{\mathrm{RU,f}})$), and the consumption of the Power Amplifier (PA) for radio power generation ($f(P_{\mathrm{t}}^{\mathrm{DL}})$).

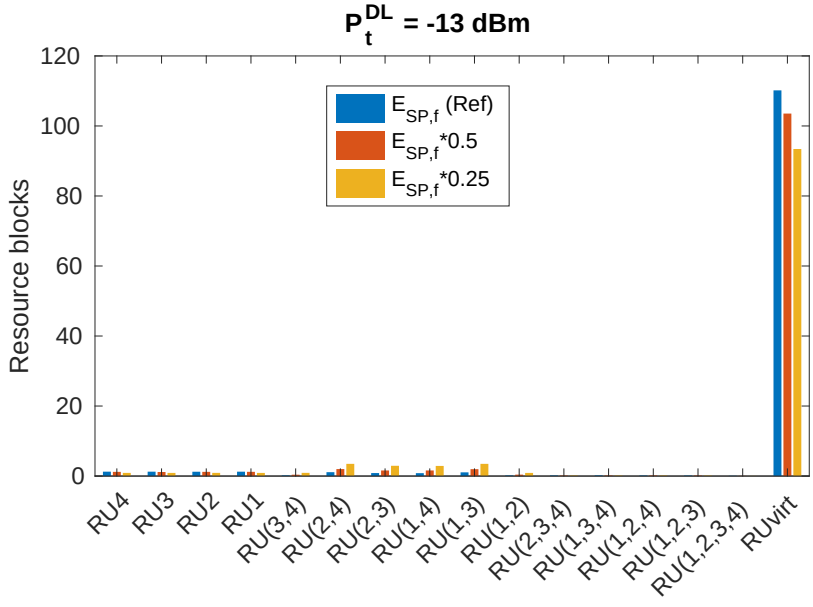
In Figure 7.7(a) we use the original value of $E_{\mathrm{SP,f}}$ ($E_{\mathrm{SP,f}}$ (Ref)). This value dominates the energy consumption for all transmission powers, even when the transmission power is high and its generation consumes a lot of energy. Nevertheless, for 24 dBm transmission power, the part consumed for radio power generation is not to be neglected, which explains the disadvantage of the virtual RU in Figure 7.6(a). We can also see that increasing the transmission power from -13 to 12 dBm decreases the total energy consumption. Further increasing the transmission power to 24 dBm has almost no impact on the total energy consumption.

However, when the original value of $E_{\mathrm{SP,f}}$ is halved ($0.5E_{\mathrm{SP,f}}$, Figure 7.7(b)), the portion of the energy consumption that depends on it, becomes comparable to the energy consumed for the generation of the transmission power (a function of $P_{\mathrm{t}}^{\mathrm{DL}}$) for $P_{\mathrm{t}}^{\mathrm{DL}} = 24$ dBm. An increase in total energy consumption is observed in this case when increasing the transmission power from 12 to 24 dBm. A similar behavior is observed with $0.25E_{\mathrm{SP,f}}$ in Figure 7.7(c). Here, for $P_{\mathrm{t}}^{\mathrm{DL}} = 24$ dBm, the transmission power generation predominates the other parts of the energy consumption.

When we look at the three cumulative bars in Figure 7.7 for $P_{\mathrm{t}}^{\mathrm{DL}} = -13$ dBm, we realize that the energy consumed decreases with the reduction of $E_{\mathrm{SP,f}}$. The transmission power generation part remains low for this low transmission power. But, since what is predominant (function part of $E_{\mathrm{SP,f}}$) is reduced, the use of the virtual RU is somewhat decreased (see Figure 7.6(b)). This also explains the increase in utilization for configurations with resource reuse where, even though the BS is active for a longer period of time due to the decrease in channel quality (from

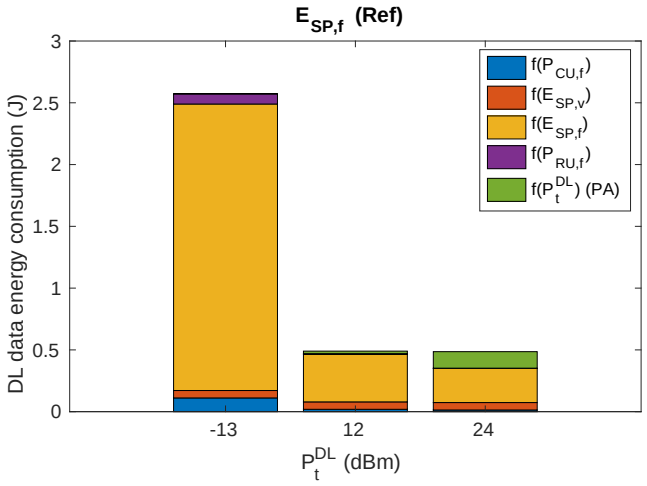


(a) $P_{\mathrm{t}}^{\mathrm{DL}}$ variation.

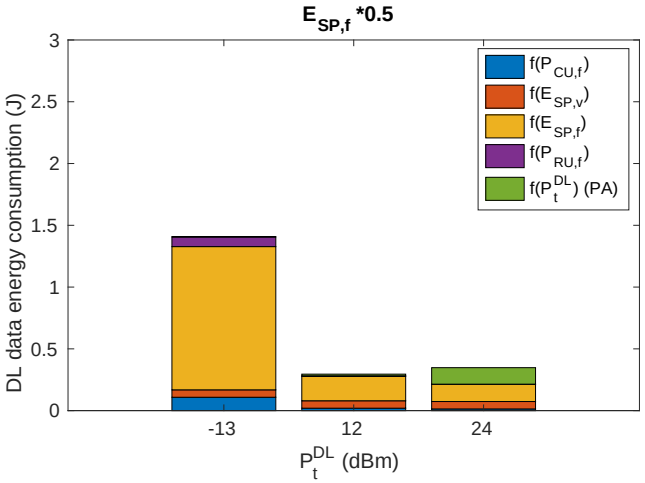


(b) $E_{\mathrm{SP,f}}$ variation.

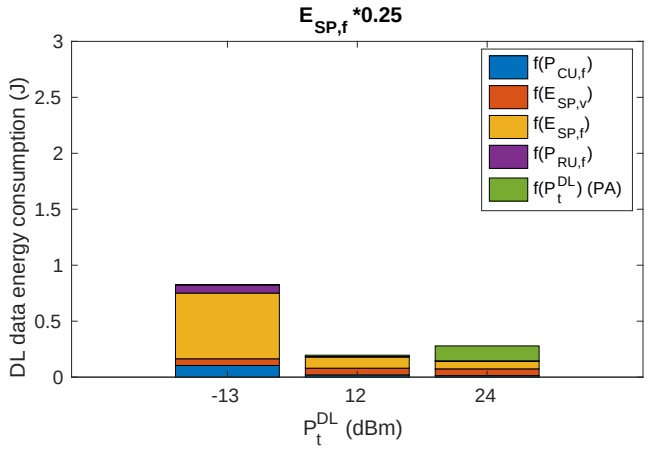
Figure 7.6: Energy consumption parameters variation.



(a)



(b)



(c)

Figure 7.7: DL data transmission energy consumption.

interference when resource reuse is taken into account), the power consumption is still acceptable. This is due to the part that depends on $P_{\text{CU,f}}$ because this part is consumed with the same amount as soon as the CU is active, even if it is active to serve more than one user on the same resources. When the reference value of $E_{\text{SP,f}}$ is considered, the energy consumption function of $P_{\text{CU,f}}$ is negligible with a ratio of 4% compared to the part depending on $E_{\text{SP,f}}$. When $E_{\text{SP,f}}$ is divided by four, the consumption function of $P_{\text{CU,f}}$ remains small compared to that of $E_{\text{SP,f}}$, but the ratio becomes 18%. This ratio, which remains low but is higher than the previous one (4%), explains the small advantage given to resource reuse in Figure 7.6(b).

7.11 Conclusion

In this chapter, we took the system and energy consumption models from Chapters 5 and 6, respectively. We aimed to minimize the BS energy consumption while increasing the system capacity. We formulated an optimization problem that is a mixture of modes 1 and 2 (without resource reuse) explained in the previous chapters, and mode 2 with resource reuse detailed in this chapter. We provided the computation of the BS energy consumption with the considered model. This consumption is the objective function to be minimized by the optimization problem. A user can be served by any of the available RUs and in any configuration defined by the active RUs. In a configuration, an RU can be active and serving alone, active and serving with other active RUs (resource reuse with interference), or active and cooperating with other RUs (simultaneous transmissions). After the resolution of this optimization problem, each user is assigned to be served by an RU in a certain configuration so that the consumed energy by the BS is minimized.

The main finding of this chapter appears after the resolution of the optimization problem. We found that the system capacity is increased while maintaining a minimum amount of energy consumed equal to the consumption of modes 1 and 2 considered alone when the use of the configurations with resource reuse is not so large. The configurations with resource reuse seem to consume more energy compared to modes 1 and 2 (without resource reuse) alone. Their use is therefore limited to the real need: when the maximum system capacity is reached.

The obtained performances highly depend on the parameter values taken. We thus vary two of these parameters and observe their impact on the optimization problem solution. We realized that increasing the transmission power advantages the configurations where only one RU is active and disadvantages the simultaneous transmissions. This is due to the increase in the RU energy consumption that is high with simultaneous transmissions (involving all RUs to transmit). We also found that decreasing the fixed signal processing energy, advantages configurations with resource reuse.

This study showed interesting results, however, it has some limitations:

- The computation of the needed RBs is a bit pessimistic when the interference is considered. When two RUs are interfering with each other, the interference

is considered for the whole occupied RBs on the system level even if one of the RUs finishes serving before the other. Taking into account the interference only when it really exists complicates the resolution of the problem but gives results that are closer to reality.

- The performance may highly depend on the considered parameters and thus can not be generalized. A look at other parameters references may be interesting to take in order to validate this work.

- The computation time increases with the number of variables and constraints. Thus, increasing the number of RUs increases the number of configurations that can quickly increase the number of variables and constraints with the load increase. The number of tested RUs is limited by the solver and the formulation of the problem. There may be other solvers and other formulations of the problem (with possibly a suboptimality of the solution) that would allow increasing the number of RUs.

In summary, in this part of the thesis, we have shown that the choice of only one of the modes (1 or 2) does not allow to have both a low energy consumption and a reasonable capacity (up to 50 users). The combination of the two modes is interesting even if we vary the main parameters (transmission power and computation energy).

Chapter 8

Conclusion

Contents

8.1 Thesis summary

This thesis evaluated the impact of Radio Access Network (RAN) architecture on reliability, latency, and energy consumption. The evolution of the RAN architecture led to a centralized architecture where some Base Station (BS) functions can be centralized in a central point and other functions are distributed on cell sites.

In Fifth Generation (5G) new use cases with stringent requirements like Ultra Reliable Low Latency Communications (URLLC) appeared. URLLC requires very low latency and ultra-high reliability simultaneously. In the first part of this thesis, we studied the impact of the centralization or distribution of some BS functions on reliability and latency.

Moreover, with the increased use of wireless networks and the increased CO2 emissions, it's becoming mandatory to find solutions to reduce the energy consumption of the networks. However, this reduction should not come at the cost of a reduced Quality of Service (QoS). The BS being the most energy-consuming component, in the second part its energy consumption during low and medium loads is minimized while maintaining the users' QoS.

Throughout this thesis, the contributions of this work were explained and justified using scientific interpretation and/or simulations and can be summarized as follows:

- Study of the impact of two RAN architectures with Hybrid Automatic Repeat reQuest (HARQ) mechanism on UpLink (UL) reliability and latency:

  Each architecture implements the Media Access Control (MAC) layer in a location. In the first architecture, the HARQ re-transmissions are performed locally on the cell site with single reception (distributed MAC layer, short Round Trip Time (RTT)). In the second one, multiple UL receptions are combined in a central point where the re-transmissions are performed (centralized MAC layer, long RTT). The delay on the radio link and the error

probability were provided. Analytic computations and simulations showed that by enabling macro-diversity, allowing multiple receptions, and centralized combining technique, fewer re-transmissions are needed and the latency is reduced with high reliability even with longer RTT. A dynamic switch between the two architectures is then proposed. It was shown that the same performances as in the second architecture can be reached with this flexible switch.

- Extended study of the impact of architecture and macro-diversity on reliability and latency with three RAN architectures:

  Compared to the first contribution, a more realistic channel model and latency computation were considered along with an additional architecture and combining technique. Through analytical calculations and simulations, a comparison has been made. We have demonstrated the importance of the Maximum Ratio Combining (MRC) technique if providing simultaneous high reliability and low latency is needed.

- Evaluation of two network operating modes with Centralized-RAN (C-RAN) architecture:

  Radio Units (RUs) are deployed to provide a certain coverage and are linked to a Centralized Unit (CU). While mode 1 operates with all RUs acting as one cell, mode 2 operates with each RU acting independently. In Mode 1, all RUs serve one user simultaneously on some Resource Blocks (RBs). In mode 2, only one RU is active at a time and serves one user at a time. Monte Carlo simulations are performed to evaluate the network coverage, resource usage, and blocking probability. As a result of serving the user in mode 1, increased Signal to Noise Ratio (SNR) is received, resulting in a higher received data rate, and a lower number of necessary RBs. Better coverage and higher capacity are thus achieved with mode 1 compared to mode 2.

- Evaluation of the DownLink (DL) and UL energy consumption of two-unit BSs in the previously proposed operating modes (modes 1 and 2):

  The BS energy consumption calculation is provided for different transmissions/receptions. The simulation results showed that mode 1 is less energy-consuming compared to mode 2 with low BS transmission power. Also, an increase in the transmission power decreases the total energy consumption. This is due to the relatively low transmission power which generation does not consume a lot of energy (processing consumption predominating). This remains true with the transmission power increase until a certain value where the generation of this power in the RU becomes important and mode 1 is disadvantaged. An increase in such high transmission powers also leads to an increase in total energy consumption. We proved that energy consumption can be reduced in low loads without negatively affecting the network coverage and capacity.

- Provide a resource scheduling solution with the goal of minimizing the BS energy consumption:

  We considered multiple configurations for the set of RUs in the system. The configurations are a mix of mode 1, mode 2 without resource reuse (as previously considered), and mode 2 with resource reuse. The independent performance of the RUs in mode 2 makes resource reuse possible. We formulated an optimization problem to schedule the available resources with the objective to minimize the BS energy consumption. The resolution of the problem assigns each user to be served by an RU in a certain configuration. By using configurations with resource reuse, the system capacity is increased while maintaining a minimum amount of energy consumed equal to modes 1 and 2 considered individually. Nevertheless, when the use of configurations with resource reuse is important, energy consumption performs faster increase. The configurations with resource reuse are energy-consuming and are thus only used to avoid reaching the maximum system capacity.

8.2 Perspectives

Future work to extend the present work and overcome its limitations are summarized as follows:

- When a distance-based simplified channel model was considered, the flexible switch between architectures was proposed. The flexible switch aims to achieve high reliability and low latency while saving the use of multiple receptions when not necessary, i.e. when the requirements can be reached with a single reception. The proposed switching algorithm was simple and based on distance thresholds. To extend the work in the first part of the thesis, we propose to provide a flexible split when the realistic channel model is considered. In this case, the switching algorithm must depend not only on the distance but also on the shadowing. A more complex switching algorithm is needed in that case.

- The evaluation of the error probability using HARQ with MRC was only provided for a specific case. It is complicated to provide a generalized analytic formulation for this error probability but it remains interesting. The MRC is a technique that reduces latency, increases reliability, increases system capacity, and reduces energy consumption with low transmission powers. An approximated yet accurate analytic computation of this probability is thus interesting to find.

- In this thesis, we showed that C-RAN along with multiple receptions are able to achieve high reliability and low latency for URLLC applications. Nevertheless, the study of the security aspect of this solution is unavoidable [Tian *et al.* 2017] [Yoshizawa *et al.* 2019]. First, the required low latency limits the computational complexity of strict security, second, multiple receptions

increase the number of paths that can be attacked and thus the vulnerability, and third, the centralization of the HARQ process makes it critical in case of denial of service in the central point.

- In the second part, we provided three operating modes but only evaluated two of them. As an extension, multiple antenna RUs can be deployed and beamforming can be used to increase the system capacity. The evaluation of the network performances like network coverage, capacity, and energy consumption may be evaluated in this case.

- The proposed optimization problem does not consider all the possible configurations. For example, a configuration where RU_1 andRU_2 are simultaneously serving a user, and RU_3 and RU_4 are inactive is not considered. Adding some of the omitted configurations is a bit complicated to deal with. However, these configurations might enhance the system's performance.

- The obtained results come from a pessimistic evaluation of the number of needed RBs. For example, sometimes an RU interfering on a user in a certain configuration finishes transmission before the interfered user finishes being served by the serving RU. In our case, we consider that the interference continues until the end of the interfered user service (even if it is not the real case). It is therefore also interesting to compute the number of needed RBs in each configuration in a more realistic (less pessimistic) way.

- The proposed optimization problem is limited by the number of RUs. Nevertheless, it is interesting to have a solution that can be extended to larger use cases with more RUs. It is therefore possible to use other solvers or to propose optimization methods or heuristic algorithms to prevent the computation time from exploding.

- In addition to reducing energy consumption through resource allocation and cooperative transmissions presented in this thesis, the deployment of BSs powered by renewable energy has been widely considered to ensure sustainability [Hassan *et al.* 2013]. It would be interesting to consider this solution in the context of the C-RAN as in [Zhang *et al.* 2016a] and [Temesgene *et al.* 2018]. The idea may be to optimize the distribution of renewable energy sources: in the Baseband Unit (BBU), the Remote Radio Head (RRH), or all units depending on the functional split, cooperation techniques, and other network parameters to minimize brown energy consumption and deployment cost.

Appendices

Appendix A

PER using HARQ with macro-diversity combined by MRC

We consider transmissions in the UpLink (UL) direction. We consider that one User Equipment (UE) is transmitting and two Base Stations (BSs) are receiving. The channel suffers fast fading. Thus, the Signal to Noise Ratio (SNR) is an exponential random variable with mean the average SNR: $\overline{\gamma}$. In [Liu *et al.* 2004], the Packet Error Rate (PER) as a function of the SNR was provided:

$$h(\gamma) = \begin{cases} 1 & \text{if } 0 < \gamma < \gamma_M \\ ae^{-g\gamma} & \text{if } \gamma \geq \gamma_M \end{cases} \tag{A.1}$$

where a and g are parameters that are Modulation and Coding Scheme (MCS) mode dependent and $\gamma_M = \ln(a)/g$.

A.1 Analytic derivation

For the Maximum Ratio Combining (MRC), during each transmission, the received SNR is the sum of the two SNRs received at each BS as shown in section 13.4-1 in [Proakis & Salehi 2008]:

$$\gamma_{S,i} = \gamma_{1,i} + \gamma_{2,i} \tag{A.2}$$

with $\gamma_{m,i}$ being the SNR received from the UE at BS_m during the ith transmission.

If $\overline{\gamma_1} << \overline{\gamma_2}$, then $\gamma_{S,i} \approx \gamma_{2,i}$. Similarly, if $\overline{\gamma_2} << \overline{\gamma}_1$, then $\gamma_{S,i} \approx \gamma_{1,i}$. We thus take the case where $\overline{\gamma}_1 = \overline{\gamma}_2 = \overline{\gamma}$. We have $\gamma_{m,i}$ an exponential random variable: $\gamma_{m,i} \sim exp(\frac{1}{\overline{\gamma}_m})$. Then, $\gamma_{S,i} = \gamma_{1,i} + \gamma_{2,i}$ follows the Erlang law with the following distribution:

$$f_\gamma(\gamma_{S,i}) = \left(\frac{1}{\overline{\gamma}}\right)^2 \gamma_{S,i} e^{-\frac{1}{\overline{\gamma}}\gamma_{S,i}} \tag{A.3}$$

Using HARQ with Chase Combining (HARQ-CC), the probability of error during the kth transmission depends on the SNR of all the k transmissions: it is $h(\gamma_{T,k})$ where $\gamma_{T,k} = \sum_{i=1}^{k} \gamma_{S,i}$.

Needing more than k transmissions is based on all the SNRs perceived by the receiver from the first transmission until the kth one. Consequently, the probability

of needing more than k transmissions has to be averaged over all $\gamma_{S,i}$ with $i : 1 \rightarrow k$. We consider successive packet transmissions, so the SNR is independent and identically distributed (i.i.d) for different transmissions. Thereby, the probability of error during the kth transmission is the following:

$$\mathbb{P}(l > k) = \int_0^{\infty} ... \int_0^{\infty} \prod_{i=1}^{k} h(\gamma_{T,i}) f_\gamma(\gamma_{S,1}) ... f_\gamma(\gamma_{S,k}) \mathrm{d}\gamma_{S,1} ... \mathrm{d}\gamma_{S,k}. \tag{A.4}$$

Similarly to [Lagrange 2010], we split (A.4) into the sum of three series:

$$\mathbb{P}(l > k) = B_k + \sum_{l=1}^{k-1} C_{k,l} + D_k \tag{A.5}$$

where:

$$B_k = \int_0^{\gamma_M} \int_0^{\gamma_M - \gamma_{S,1}} ... \int_0^{\gamma_M - \gamma_{T,k-1}} f_\gamma(\gamma_{S,1}) f_\gamma(\gamma_{S,2}) ... f_\gamma(\gamma_{S,k}) \mathrm{d}\gamma_{S,1} \mathrm{d}\gamma_{S,2} ... \mathrm{d}\gamma_{S,k} \tag{A.6}$$

$$C_{k,l} = \int_0^{\gamma_M} ... \int_0^{\gamma_M - \gamma_{T,l-1}} \int_{\gamma_M - \gamma_{T,l}}^{\infty} \int_0^{\infty} ... \int_0^{\infty} f_\gamma(\gamma_{S,1}) ... f_\gamma(\gamma_{S,k}) a e^{-g\gamma_{T,l+1}} ... a e^{-g\gamma_{T,k}} \mathrm{d}\gamma_{S,1} ... \mathrm{d}\gamma_{S,k} \tag{A.7}$$

$$D_k = \int_{\gamma_M}^{\infty} \int_0^{\infty} ... \int_0^{\infty} f_\gamma(\gamma_{S,1}) ... f_\gamma(\gamma_{S,k}) a e^{-g\gamma_{T,1}} ... a e^{-g\gamma_{T,k}} \mathrm{d}\gamma_{S,1} ... \mathrm{d}\gamma_{S,k}. \tag{A.8}$$

The calculation steps in Section A.2 lead to the following expressions:

$$B_k = 1 - e^{-\frac{\gamma_M}{\overline{\gamma}}} \sum_{n=0}^{2k-1} \left(\frac{\gamma_M}{\overline{\gamma}} \right)^n \frac{1}{n!} \tag{A.9}$$

$$\begin{aligned} C_{k,l} =& e^{-\frac{\gamma_M}{\overline{\gamma}}} \frac{1}{(2l+1)!} \left(\frac{\gamma_M}{\overline{\gamma}} \right)^{2l} \times \\ & \prod_{j=l+1}^{k} \frac{1}{(1 + g\overline{\gamma}(k+1-j))^2} \left[\gamma_M \left(\frac{1}{\overline{\gamma}} + g(k-l) \right) + 2l + 1 \right] \end{aligned} \tag{A.10}$$

$$D_k = e^{-\frac{\gamma_M}{\overline{\gamma}}} \prod_{j=1}^{k} \frac{1}{(1 + g\overline{\gamma}(k+1-j))^2} \left[1 + \gamma_M \left(\frac{1}{\overline{\gamma}} + gk \right) \right]. \tag{A.11}$$

Finally, the probability of having an error at the kth transmission while using HARQ-CC and MRC based macro-diversity is:

$$\boxed{\mathbb{P}(l > k) = 1 - e^{-Y} \sum_{i=0}^{2k-1} \frac{Y^i}{i!} + \sum_{l=0}^{k-1} \frac{e^{-Y} Y^{2l}}{(2l+1)!} \prod_{j=1}^{k-l} \frac{1}{(1 + \overline{\gamma} g j)^2} E_{k,l}} \tag{A.12}$$

where $Y = \frac{\gamma_M}{\overline{\gamma}}$ and $E_{k,l} = \gamma_M \left(\frac{1}{\overline{\gamma}} + g(k-l) \right) + 2l + 1$.

A.2 Detailed derivation of B_k, $C_{k,l}$, and D_k

A.2.1 Computation of B_k

$$B_k = \int_0^{\gamma_M} \int_0^{\gamma_M-\gamma_{S,1}} ... \int_0^{\gamma_M-\gamma_{T,k-1}} f_\gamma(\gamma_{S,1}) f_\gamma(\gamma_{S,2}) ... f_\gamma(\gamma_{S,k}) \mathrm{d}\gamma_{S,1} \mathrm{d}\gamma_{S,2} ... \mathrm{d}\gamma_{S,k}.$$

Now, we make a variable change by dividing $\gamma_{S,i}$ by γ_M:

$$\begin{aligned} B_k = \left(\frac{\gamma_M}{\overline{\gamma}}\right)^{2k} \int_0^1 \int_0^{1-\gamma_{S,1}} ... \int_0^{1-\gamma_{T,k-1}} \gamma_{S,1}\gamma_{S,2} ... \gamma_{S,k} \times \\ \left(e^{-\frac{\gamma_{S,1}\gamma_M}{\overline{\gamma}}} e^{-\frac{\gamma_{S,2}\gamma_M}{\overline{\gamma}}} ... e^{-\frac{\gamma_{S,k}\gamma_M}{\overline{\gamma}}}\right) \mathrm{d}\gamma_{S,1} \mathrm{d}\gamma_{S,2} ... \mathrm{d}\gamma_{S,k}. \end{aligned} \tag{A.13}$$

Using the multiple integral expression from [Gradshteyn & Ryzhik 2007, eq. 2 pp. 614], we get:

$$B_k = \frac{1}{\Gamma(2k)} \left(\frac{\gamma_M}{\overline{\gamma}}\right)^{2k} \int_0^1 e^{-x\frac{\gamma_M}{\overline{\gamma}}} x^{2k-1} \mathrm{d}x.$$

The previous integral can be solved using [Gradshteyn & Ryzhik 2007, eq. 11 §2.33, pp. 108]. We finally get:

$$B_k = 1 - e^{-\frac{\gamma_M}{\overline{\gamma}}} \sum_{n=0}^{2k-1} \left(\frac{\gamma_M}{\overline{\gamma}}\right)^n \frac{1}{n!}. \tag{A.14}$$

A.2.2 Computation of $C_{k,l}$

$$\begin{aligned} C_{k,l} = \int_0^{\gamma_M} ... \int_0^{\gamma_M-\gamma_{T,l-1}} \int_{\gamma_M-\gamma_{T,l}}^{\infty} \int_0^{\infty} ... \int_0^{\infty} f_\gamma(\gamma_{S,1}) ... f_\gamma(\gamma_{S,k}) \times \\ a e^{-g\gamma_{T,l+1}} ... a e^{-g\gamma_{T,k}} \mathrm{d}\gamma_{S,1} ... \mathrm{d}\gamma_{S,k} \end{aligned} \tag{A.15}$$

where $\gamma_{T,k} = \sum_{i=1}^{k} \gamma_{S,i}$.
For $i > l+1$:

$$\int_0^{\infty} \frac{\gamma_{S,i}}{\overline{\gamma}^2} e^{-\frac{\gamma_{S,i}}{\overline{\gamma}}} a e^{-g(k-i+1)\gamma_{S,i}} \mathrm{d}\gamma_{S,i} = \frac{a}{(1+g\overline{\gamma}(k+1-i))^2}. \tag{A.16}$$

So we have:

$$\begin{aligned} &\int_0^{\infty} ... \int_0^{\infty} f_\gamma(\gamma_{S,l+2}) ... f_\gamma(\gamma_{S,k}) a e^{-g\gamma_{T,l+2}} ... a e^{-g\gamma_{T,k}} \mathrm{d}\gamma_{S,l+2} ... \mathrm{d}\gamma_{S,k} = \\ &\prod_{i=l+2}^{k} \frac{a}{(1+g\overline{\gamma}(k+1-i))^2}. \end{aligned} \tag{A.17}$$

Then, for $i > l$:

$$U_{k,l} = \int_{\gamma_M-\gamma_{T,l}}^{\infty} f_\gamma(\gamma_{S,l+1}) a e^{-g(k-l)\gamma_{S,l+1}} \prod_{i=l+2}^{k} \frac{a}{(1+g\overline{\gamma}(k+1-i))^2} \mathrm{d}\gamma_{S,l+1}. \tag{A.18}$$

After several elementary computation steps we get:

$$U_{k,l} = V_{k,l} e^{\gamma_{T,l}\left(g(k-l)+\frac{1}{\overline{\gamma}}\right)} \left[1 + \left(\frac{1}{\overline{\gamma}} + g(k-l)\right)(\gamma_M - \gamma_{T,l})\right]$$

with $V_{k,l} = \prod_{i=l+1}^{k} \frac{1}{(1+g\overline{\gamma}(k+1-i))^2} e^{-\frac{\gamma_M}{\overline{\gamma}}}$.

So,

$$\begin{aligned} C_{k,l} =& \frac{1}{\overline{\gamma}^{2l}} V_{k,l} \int_0^{\gamma_M} ... \int_0^{\gamma_M - \gamma_{T,l-1}} \gamma_{S,1}...\gamma_{S,l} \mathrm{d}\gamma_{S,1}...\mathrm{d}\gamma_{S,l} \\ &+ \frac{1}{\overline{\gamma}^{2l}} \left(\frac{1}{\overline{\gamma}} + g(k-l)\right) V_{k,l} \int_0^{\gamma_M} ... \int_0^{\gamma_M - \gamma_{T,l-1}} \gamma_{S,1}...\gamma_{S,l}(\gamma_M - \gamma_{S,1} - ... - \gamma_{S,l}) \mathrm{d}\gamma_{S,1}...\mathrm{d}\gamma_{S,l} \end{aligned}$$

$$C_{k,l} = \frac{1}{\overline{\gamma}^{2l}} V_{k,l} I_l \left[1 + \gamma_M \left(\frac{1}{\overline{\gamma}} + g(k-l)\right)\right] + \frac{1}{\overline{\gamma}^{2l}} V_{k,l} J_l \left(\frac{1}{\overline{\gamma}} + g(k-l)\right)$$

where

$$I_l = \int_0^{\gamma_M} \int_0^{\gamma_M - \gamma_{S,1}} ... \int_0^{\gamma_M - \gamma_{T,l-1}} \gamma_{S,1}\gamma_{S,2}...\gamma_{S,l} \mathrm{d}\gamma_{S,1}\mathrm{d}\gamma_{S,2}...\mathrm{d}\gamma_{S,l} = \frac{{\gamma_M}^{2l}}{(2l)!} \quad \text{(A.19)}$$

and

$$\begin{aligned} J_l =& \int_0^{\gamma_M} \int_0^{\gamma_M - \gamma_{S,1}} ... \int_0^{\gamma_M - \gamma_{T,l-1}} \gamma_{S,1}\gamma_{S,2}...\gamma_{S,l}(-\gamma_{S,1} - \gamma_{S,2} - ... - \gamma_{S,l}) \mathrm{d}\gamma_{S,1}\mathrm{d}\gamma_{S,2}...\mathrm{d}\gamma_{S,l} \\ =& \frac{-{\gamma_M}^{2l+1}}{(2l+1)(2l-1)!}. \end{aligned} \quad \text{(A.20)}$$

Finally,

$$C_{k,l} = e^{-\frac{\gamma_M}{\overline{\gamma}}} \frac{1}{(2l+1)!} \left(\frac{\gamma_M}{\overline{\gamma}}\right)^{2l} \prod_{j=l+1}^{k} \frac{1}{(1+g\overline{\gamma}(k+1-j))^2} \left[\gamma_M \left(\frac{1}{\overline{\gamma}} + g(k-l)\right) + 2l + 1\right] \quad \text{(A.21)}$$

A.2.2.1 Proof of (A.19)

We prove (A.19) by induction. By definition of I_l, I_1 is:

$$I_1 = \int_0^{\gamma_M} \gamma_{S,1} \mathrm{d}\gamma_{S,1} = \frac{{\gamma_M}^2}{2}.$$

Thus, (A.19) is valid for $l = 1$. We suppose that:

$$I_{l-1} = \frac{{\gamma_M}^{2(l-1)}}{(2(l-1))!}$$

and we denote it as $I_{l-1}(\gamma_M)$ for the sake of clarity in the next steps. Now, we check if it remains true for l:

$$I_l = \int_0^{\gamma_M} \gamma_{S,1} \int_0^{\gamma_M-\gamma_{S,1}} ... \int_0^{\gamma_M-\gamma_{T,l}} \gamma_{S,2}...\gamma_{S,l} \mathrm{d}\gamma_{S,1}\mathrm{d}\gamma_{S,2}...\mathrm{d}\gamma_{S,l}$$

$$I_l = \int_0^{\gamma_M} \gamma_{S,1} I_{l-1}(\gamma_M - \gamma_{S,1})\mathrm{d}\gamma_{S,1}$$

$$I_l = \int_0^{\gamma_M} \gamma_{S,1} \frac{(\gamma_M - \gamma_{S,1})^{2l-2}}{(2l-2)!}\mathrm{d}\gamma_{S,1}$$

$$I_l = \frac{{\gamma_M}^{2l}}{(2l)!}. \tag{A.22}$$

A.2.2.2 Proof of (A.20)

$$J_l = \int_0^{\gamma_M} \int_0^{\gamma_M-\gamma_{S,1}} ... \int_0^{\gamma_M-\gamma_{T,l-1}} \gamma_{S,1}\gamma_{S,2}...\gamma_{S,l}(-\gamma_{S,1}-\gamma_{S,2}-...-\gamma_{S,l})\mathrm{d}\gamma_{S,1}\mathrm{d}\gamma_{S,2}...\mathrm{d}\gamma_{S,l}.$$

Like for (A.19), we make the proof by induction for (A.20). By definition of J_l, J_1 is:

$$J_1 = \int_0^{\gamma_M} \gamma_{S,1}(-\gamma_{S,1})\mathrm{d}\gamma_{S,1} = -\frac{{\gamma_M}^3}{3}.$$

So, (A.20) is valid for $l = 1$. We suppose that

$$J_{l-1} = -\frac{{\gamma_M}^{2l-1}}{(2l-1)(2l-3)!}$$

and we denote it as $J_{l-1}(\gamma_M)$. Now, we check if it is still true for l:

$$\begin{aligned} J_l = &\int_0^{\gamma_M} -{\gamma_{S,1}}^2 \int_0^{\gamma_M-\gamma_{S,1}} ... \int_0^{\gamma_M-\gamma_{T,l}} \gamma_{S,2}...\gamma_{S,l}\mathrm{d}\gamma_{S,1}\mathrm{d}\gamma_{S,2}...\mathrm{d}\gamma_{S,l} \\ &+ \int_0^{\gamma_M} \gamma_{S,1} \int_0^{\gamma_M-\gamma_{S,1}} ... \int_0^{\gamma_M-\gamma_{T,l}} \gamma_{S,2}...\gamma_{S,l}(-\gamma_{S,2}... - \gamma_{S,l})\mathrm{d}\gamma_{S,1}\mathrm{d}\gamma_{S,2}...\mathrm{d}\gamma_{S,l} \end{aligned}$$

$$J_l = \int_0^{\gamma_M} -{\gamma_{S,1}}^2 I_{l-1}(\gamma_M - \gamma_{S,1})\mathrm{d}\gamma_{S,1} + \int_0^{\gamma_M} \gamma_{S,1} J_{l-1}(\gamma_M - \gamma_{S,1})\mathrm{d}\gamma_{S,1}$$

$$J_l = \int_0^{\gamma_M} -{\gamma_{S,1}}^2 \frac{(\gamma_M - \gamma_{S,1})^{2(l-1)}}{(2(l-1))!}\mathrm{d}\gamma_{S,1} + \int_0^{\gamma_M} \gamma_{S,1} \left(-\frac{(\gamma_M - \gamma_{S,1})^{2l-1}}{(2l-1)(2l-3)!} \right) \mathrm{d}\gamma_{S,1}$$

$$J_l = \frac{-{\gamma_M}^{2l+1}}{(2l+1)(2l-1)!}. \tag{A.23}$$

A.2.3 Computation of D_k

$$D_k = C_{k,0} = U_{k,0}$$

$$D_k = \prod_{j=2}^{k} \frac{a}{(1+g\overline{\gamma}(k+1-j))^2} \int_{\gamma_M}^{\infty} \frac{\gamma_{S,1}}{\overline{\gamma}^2} e^{-\frac{\gamma_{S,1}}{\overline{\gamma}}} a e^{-g(k-l)\gamma_{S,1}} \mathrm{d}\gamma_{S,1}$$

$$D_k = e^{-\frac{\gamma_M}{\overline{\gamma}}} \prod_{j=1}^{k} \frac{1}{(1+g\overline{\gamma}(k+1-j))^2} \left[1 + \gamma_M \left(\frac{1}{\overline{\gamma}} + gk\right)\right]. \tag{A.24}$$

Appendix B

Indoor path-loss model

B.1 Path-loss models

From Table 7.4.1-1 in [TR 38.901 2019] we present the path-loss models considering indoor offices and different scenarios in indoor factories:

- Indoor Offices (InOf):
  - $\mathrm{PL}_{\mathrm{InOf}} = 38.3\log(d) + 17.30 + 24.9\log(f_c)$
- Indoor Factories (InF):
  - Indoor factory Sparse clutter, Low base station height (InF-SL): $\mathrm{PL}_{\mathrm{InF\text{-}SL}} = 25.5\log(d) + 33 + 20\log(f_c)$
  - Indoor factory Dense clutter, Low base station height (InF-DL): $\mathrm{PL}_{\mathrm{InF\text{-}DL}} = 35.7\log(d) + 18.6 + 20\log(f_c)$
  - Indoor factory Sparse clutter, High base station height (InF-SH): $\mathrm{PL}_{\mathrm{InF\text{-}SH}} = 23\log(d) + 32.4 + 20\log(f_c)$
  - Indoor factory Dense clutter, High base station height (InF-DH): $\mathrm{PL}_{\mathrm{InF\text{-}DH}} = 21.9\log(d) + 33.63 + 20\log(f_c)$

In indoor factories, the scenarios depend on the density of the clutter and its position relative to the transmitters and receivers. The clutter consists of machines or storage shelves, for example. Clutter density is calculated as the percentage of the area that the clutter occupies and scenarios are classified as sparse or dense based on a threshold of 40%. Sparse clutter scenarios correspond to densities below 40%, and dense scenarios correspond to densities above 40%. In the low Base Station (BS) scenarios, transmitters and receivers are located below the clutter, while in the high BS scenarios, transmitters and receivers are located above the clutter.

In our work, we used the first scenario (InF-SL), but the study could be carried out with the three other scenarios.

Appendix C

Configurations use for different network loads

In Figures C.1 and C.2, we see a more complete representation of configurations used for different network loads. This is presented in Figure 7.3 in a simplified way: with less values of J.

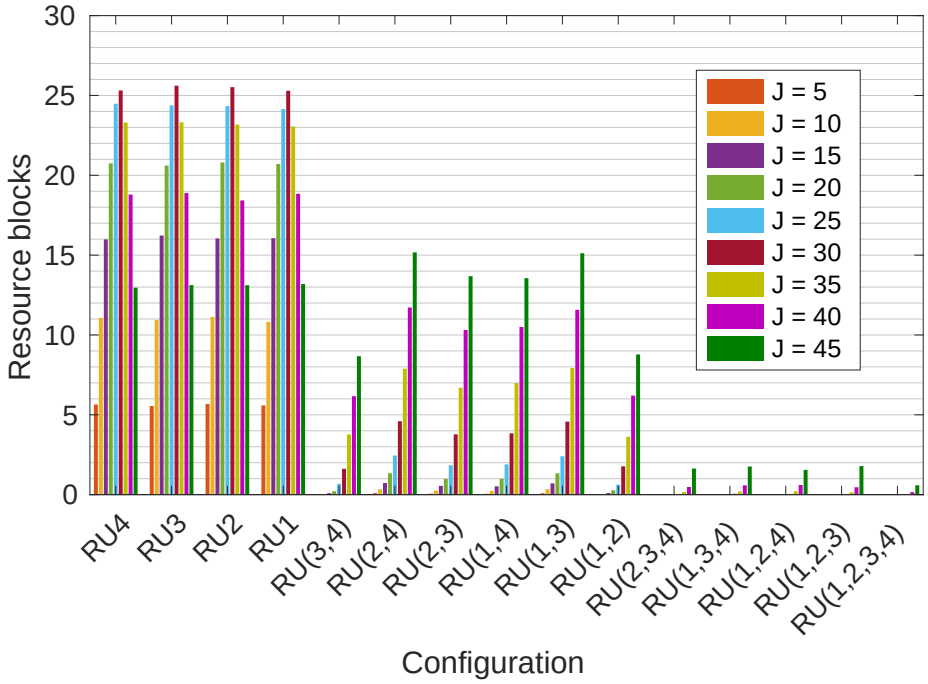


Figure C.1: Configurations usage for $I = 4$ without virtual RU.

On the one hand, in Figure C.1, we can see that some configurations with resource reuse start being used on more than one Resource Block (RB) for $J = 25$. On the other hand, in Figure C.2, this effect appears from $J = 40$ users. When the virtual Radio Unit (RU) is considered, resource reuse is less advantageous compared to the case without virtual RU. This is due to the use of the last configuration (with virtual RU active) capable of reducing energy consumption while still being able to serve higher loads before the system is blocked.

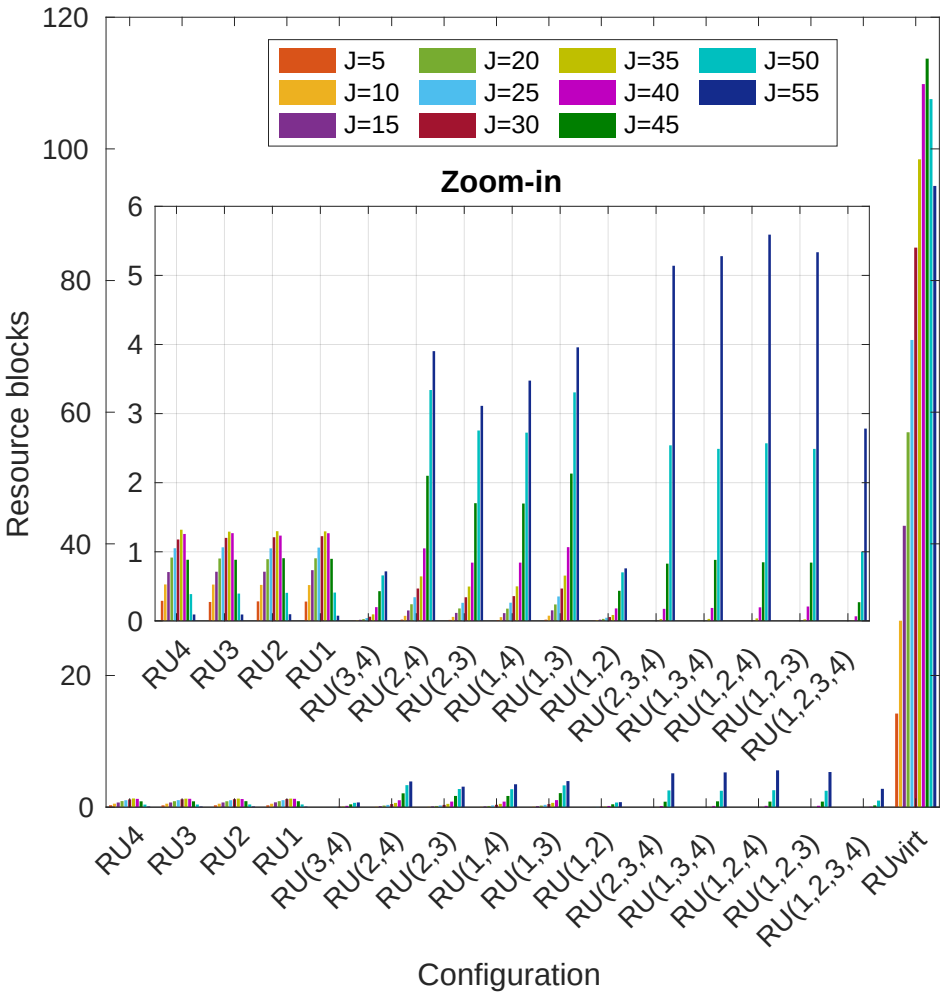


Figure C.2: Configurations usage for $I = 4$ with virtual RU.

Bibliography

[5G-Americas 2018] 5G-Americas. *New services and applications with 5G ultra-reliable low latency communications*. White paper, 5G Americas, November 2018. (Cited on pages 9 and 22.)

[Akbarzadeh *et al.* 2020] S. Akbarzadeh, J. Schwoerer, B. Bailly and W. Labidi. Les réseaux 5g: Architectures système, radio et cœur, coexistence 4g, mise en oeuvre opérationnelle. Blanche. Eyrolles, 2020. (Cited on page 40.)

[Akyildiz *et al.* 2014] Ian F. Akyildiz, Ahyoung Lee, Pu Wang, Min Luo and Wu Chou. *A roadmap for traffic engineering in SDN-OpenFlow networks*. Computer Networks, vol. 71, pages 1–30, 2014. (Cited on page 42.)

[Akyildiz *et al.* 2016] Ian F. Akyildiz, Shuai Nie, Shih-Chun Lin and Manoj Chandrasekaran. *5G roadmap: 10 key enabling technologies*. Computer Networks, vol. 106, pages 17–48, 2016. (Cited on page 42.)

[Alfadhli *et al.* 2018] Y. Alfadhli, M. Xu, S. Liu, F. Lu, P. Peng and G. Chang. *Real-Time Demonstration of Adaptive Functional Split in 5G Flexible Mobile Fronthaul Networks*. In 2018 Optical Fiber Communications Conference and Exposition (OFC), pages 1–3, March 2018. (Cited on page 58.)

[Aqeeli *et al.* 2018] Emad Aqeeli, Abdallah Moubayed and Abdallah Shami. *Power-Aware Optimized RRH to BBU Allocation in C-RAN*. IEEE Transactions on Wireless Communications, vol. 17, no. 2, pages 1311–1322, 2018. (Cited on page 87.)

[Arnold *et al.* 2010] Oliver Arnold, Fred Richter, Gerhard Fettweis and Oliver Blume. *Power consumption modeling of different base station types in heterogeneous cellular networks*. In 2010 Future Network & Mobile Summit, pages 1–8, 2010. (Cited on page 112.)

[Artuso & Christiansen 2014] Matteo Artuso and Henrik Christiansen. *Discrete-event simulation of coordinated multi-point joint transmission in LTE-Advanced with constrained backhaul*. In 2014 11th International Symposium on Wireless Communications Systems (ISWCS), pages 106–110, 2014. (Cited on page 49.)

[Ateya *et al.* 2018] Abdelhamied A. Ateya, Ammar Muthanna, Maria Makolkina and Andrey Koucheryavy. *Study of 5G Services Standardization: Specifications and Requirements*. In 2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), pages 1–6, 2018. (Cited on page 28.)

[Bassoli *et al.* 2017] Riccardo Bassoli, Marco Di Renzo and Fabrizio Granelli. *Analytical energy-efficient planning of 5G cloud radio access network.* In 2017 IEEE International Conference on Communications (ICC), pages 1–4, 2017. (Cited on page 22.)

[Benelli 1985] G. Benelli. *An ARQ Scheme with Memory and Soft Error Detectors.* IEEE Transactions on Communications, vol. 33, no. 3, pages 285–288, 1985. (Cited on page 35.)

[Berardinelli *et al.* 2016] G. Berardinelli, S. R. Khosravirad, K. I. Pedersen, F. Frederiksen and P. Mogensen. *Enabling Early HARQ Feedback in 5G Networks.* In 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), pages 1–5, May 2016. (Cited on page 49.)

[Brueck *et al.* 2010] Stefan Brueck, Lu Zhao, Jochen Giese and M. Awais Amin. *Centralized scheduling for joint transmission coordinated multi-point in LTE-Advanced.* In 2010 International ITG Workshop on Smart Antennas (WSA), pages 177–184, 2010. (Cited on page 49.)

[Chase 1985] D. Chase. *Code Combining - A Maximum-Likelihood Decoding Approach for Combining an Arbitrary Number of Noisy Packets.* IEEE Transactions on Communications, vol. 33, no. 5, pages 385–393, 1985. (Cited on page 35.)

[Chaudhary *et al.* 2019] Jay Kant Chaudhary, Jobin Francis, André Noll Barreto and Gerhard Fettweis. *Packet Loss in Latency-constrained Ethernet-based Packetized C-RAN Fronthaul.* In 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), pages 1–6, 2019. (Cited on page 48.)

[Checko *et al.* 2015] Aleksandra Checko, Henrik L. Christiansen, Ying Yan, Lara Scolari, Georgios Kardaras, Michael S. Berger and Lars Dittmann. *Cloud RAN for Mobile Networks—A Technology Overview.* IEEE Communications Surveys & Tutorials, vol. 17, no. 1, pages 405–426, 2015. (Cited on page 39.)

[Chen *et al.* 2014] Liming Chen, Hu Jin, Haoming Li, Jun-Bae Seo, Qing Guo and Victor Leung. *An Energy Efficient Implementation of C-RAN in HetNet.* In 2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall), pages 1–5, 2014. (Cited on page 86.)

[Chen *et al.* 2018] He Chen, Rana Abbas, Peng Cheng, Mahyar Shirvanimoghaddam, Wibowo Hardjawana, Wei Bao, Yonghui Li and Branka Vucetic. *Ultra-Reliable Low Latency Cellular Networks: Use Cases, Challenges and Approaches.* IEEE Communications Magazine, vol. 56, no. 12, pages 119–125, 2018. (Cited on page 31.)

[Cisco 2020] Cisco. *Cisco Annual Internet Report (2018–2023).* White paper, Cisco, 2020. (Cited on page 84.)

[C.M.R.Institute 2010] C.M.R.Institute. *C-RAN The Road Towards Green RAN*. Technical report, China Mobile Research institute, Version 1.0, April 2010. (Cited on pages 10, 22 and 41.)

[Dahlman *et al.* 2011] Erik Dahlman, Stefan Parkvall and Johan Sköld. 4g lte/lte-advanced for mobile broadband. Academic Press, 2011. (Cited on page 87.)

[Dahlman *et al.* 2018] Erik Dahlman, Stefan Parkvall and Johan Sköld. 5g nr: The next generation wireless access technology. Academic Press, 2018. (Cited on pages 31, 33 and 88.)

[Duan *et al.* 2016] J. Duan, X. Lagrange and F. Guilloud. *Performance Analysis of Several Functional Splits in C-RAN*. In IEEE 83rd VTC Spring, 2016. (Cited on pages 13, 68 and 71.)

[ecp 2019] *eCPRI Specification V2.0*. Interface specification, May 2019. (Cited on pages 13, 39, 40, 67, 68, 70 and 71.)

[Elbamby *et al.* 2018] Mohammed S. Elbamby, Cristina Perfecto, Mehdi Bennis and Klaus Doppler. *Toward Low-Latency and Ultra-Reliable Virtual Reality*. IEEE Network, vol. 32, no. 2, pages 78–84, 2018. (Cited on page 31.)

[Ezzaouia *et al.* 2018] Mahdi Ezzaouia, Cédric Gueguen, Melhem El Helou, Mahmoud Ammar, Xavier Lagrange and Ammar Bouallegue. *A dynamic transmission strategy based on network slicing for cloud radio access networks*. In 2018 Wireless Days (WD), pages 40–45, 2018. (Cited on page 92.)

[FCC 2022] FCC. *Broadband Speed Guide*, 2022. https://www.fcc.gov/consumers/guides/broadband-speed-guide. (Cited on pages ix and 92.)

[Feng *et al.* 2017] Mingjie Feng, Shiwen Mao and Tao Jiang. *Base Station ON-OFF Switching in 5G Wireless Networks: Approaches and Challenges*. IEEE Wireless Communications, vol. 24, no. 4, pages 46–54, 2017. (Cited on page 87.)

[Ferdouse *et al.* 2017] Lilatul Ferdouse, Olivia Das and Alagan Anpalagan. *Auction Based Distributed Resource Allocation for Delay Aware OFDM Based Cloud-RAN System*. In GLOBECOM 2017 - 2017 IEEE Global Communications Conference, pages 1–6, 2017. (Cited on page 130.)

[FOURIKIS 2000] NICHOLAS FOURIKIS. *2 - From Array Theory to Shared Aperture Arrays*. In NICHOLAS FOURIKIS, editor, Advanced Array Systems, Applications and RF Technologies, Signal Processing and its Applications, pages 111–217. Academic Press, London, 2000. (Cited on page 117.)

[Goldsmith 2005] Andrea Goldsmith. Wireless communications. Cambridge University Press, New York, NY, USA, 2005. (Cited on page 38.)

[Gradshteyn & Ryzhik 2007] I. S. Gradshteyn and I. M. Ryzhik. Table of integrals, series, and products. Elsevier/Academic Press, Amsterdam, seventh édition, 2007. Translated from the Russian, Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger. (Cited on page 161.)

[Gunther *et al.* 2012] Auer Gunther, Blume Oliver, Giannini Vito, Godor Istvan, Imran Muhammad Ali, Jading Ylva, Katranaras Efstathios, Olsson Magnus, Sabella Dario, Skillermark Per and Wajda Wieslawa. *Energy efficiency analysis of the reference systems, areas of improvements and target breakdown.* Deliverable D2.3, January 2012. (Cited on pages 86 and 112.)

[Gupta *et al.* 2010] Parul Gupta, Arun Vishwanath, Shivkumar Kalyanaraman and Yong Hua Lin. *Unlocking wireless performance with co-operation in co-located base station pools.* In 2010 Second International Conference on COMmunication Systems and NETworks (COMSNETS 2010), pages 1–8, 2010. (Cited on page 41.)

[Han *et al.* 2011] Congzheng Han, Tim Harrold, Simon Armour, Ioannis Krikidis, Stefan Videv, Peter M. Grant, Harald Haas, John S. Thompson, Ivan Ku, Cheng-Xiang Wang, Tuan Anh Le, M. Reza Nakhai, Jiayi Zhang and Lajos Hanzo. *Green radio: radio techniques to enable energy-efficient wireless networks.* IEEE Communications Magazine, vol. 49, no. 6, pages 46–54, 2011. (Cited on page 85.)

[Harutyunyan & Riggio 2018] Davit Harutyunyan and Roberto Riggio. *Flex5G: Flexible Functional Split in 5G Networks.* IEEE Transactions on Network and Service Management, vol. 15, no. 3, pages 961–975, 2018. (Cited on page 41.)

[Hassan *et al.* 2013] Hussein Al Haj Hassan, Loutfi Nuaymi and Alexander Pelov. *Renewable energy in cellular networks: A survey.* In 2013 IEEE Online Conference on Green Communications (OnlineGreenComm), pages 1–7, 2013. (Cited on page 156.)

[Huang & Kadoch 2020] Qian Huang and Michel Kadoch. *5G Resource Scheduling for Low-latency Communication: A Reinforcement Learning Approach.* In 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall), pages 1–5, 2020. (Cited on page 130.)

[Huang *et al.* 2016] Xiaoyan Huang, Guoliang Xue, Ruozhou Yu and Supeng Leng. *Joint Scheduling and Beamforming Coordination in Cloud Radio Access Networks With QoS Guarantees.* IEEE Transactions on Vehicular Technology, vol. 65, no. 7, pages 5449–5460, 2016. (Cited on page 131.)

[I *et al.* 2014] C. I, J. Huang, R. Duan, C. Cui, J. Jiang and L. Li. *Recent Progress on C-RAN Centralization and Cloudification.* IEEE Access, vol. 2, pages 1030–1039, August 2014. (Cited on page 49.)

[Imamura *et al.* 2017] Yuta Imamura, Dairoku Muramatsu, Yoshihisa Kishiyama and Kenichi Higuchi. *Low latency hybrid ARQ method using channel state information before channel decoding.* In 2017 23rd Asia-Pacific Conference on Communications (APCC), pages 1–6, 2017. (Cited on page 49.)

[Irram *et al.* 2020] Fauzia Irram, Mudassar Ali, Zubdah Maqbool, Farhan Qamar and Joel JPC Rodrigues. *Coordinated Multi-Point Transmission in 5G and Beyond Heterogeneous Networks.* In 2020 IEEE 23rd International Multitopic Conference (INMIC), pages 1–6, 2020. (Cited on page 87.)

[ITU-R 2015] ITU-R. *IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond.* Technical report, September 2015. (Cited on page 28.)

[ITU-T 2014] ITU-T. *The Tactile Internet.* Technical report, August 2014. Available online https://www.itu.int/oth/T2301000023/en. (Cited on page 29.)

[ITU-T 2018] ITU-T. *Transport network support of IMT-2020/5G.* Technical report, February 2018. Available online https://www.itu.int/hub/publication/t-tut-home-2018/. (Cited on page 40.)

[Jacobsen *et al.* 2019] T. H. Jacobsen, R. Abreu, G. Berardinelli, K. I. Pedersen, I. Z. Kovács and P. Mogensen. *Multi-Cell reception for uplink grant-free ultra-reliable low-latency communications.* IEEE Access, vol. 7, pages 80208–80218, 2019. (Cited on page 49.)

[Jahid *et al.* 2018] Abu Jahid, Abdullah Bin Shams and Md. Farhad Hossain. *Green energy driven cellular networks with JT CoMP technique.* Physical Communication, vol. 28, pages 58–68, 2018. (Cited on page 87.)

[Jiang *et al.* 2021] Tao Jiang, Jianhua Zhang, Pan Tang, Lei Tian, Yi Zheng, Jianwu Dou, Henrik Asplund, Leszek Raschkowski, Raffaele D'Errico and Tommi Jämsä. *3GPP Standardized 5G Channel Model for IIoT Scenarios: A Survey.* IEEE Internet of Things Journal, vol. 8, no. 11, pages 8799–8815, 2021. (Cited on page 94.)

[Jung *et al.* 2014] Byoung Hoon Jung, Hansung Leem and Dan Keun Sung. *Modeling of Power Consumption for Macro-, Micro-, and RRH-Based Base Station Architectures.* In 2014 IEEE 79th Vehicular Technology Conference (VTC Spring), pages 1–5, 2014. (Cited on pages 112, 113, 116 and 124.)

[Kaddour *et al.* 2015] Fatima Zohra Kaddour, Emmanuelle Vivier, Lina Mroueh, Mylene Pischella and Philippe Martins. *Green Opportunistic and Efficient Resource Block Allocation Algorithm for LTE Uplink Networks.* IEEE Transactions on Vehicular Technology, vol. 64, no. 10, pages 4537–4550, 2015. (Cited on page 131.)

[Kallel 1990] S. Kallel. *Analysis of a type II hybrid ARQ scheme with code combining.* IEEE Transactions on Communications, vol. 38, no. 8, pages 1133–1137, 1990. (Cited on page 35.)

[Kathrein Inc.] Kathrein Inc. *800 10465 Dual Band Indoor Directional Antenna Integrated Combiner.* original document from Kathrein Inc., Scala Division, available at https://manualzz.com/doc/13065566/datenblatt-80010465. (Cited on page 101.)

[Khalili & Simeone 2017] Shahrouz Khalili and Osvaldo Simeone. *Uplink HARQ for Cloud RAN via Separation of Control and Data Planes.* IEEE Transactions on Vehicular Technology, vol. 66, no. 5, pages 4005–4016, 2017. (Cited on page 50.)

[Khan *et al.* 2015] M. Khan, R. S. Alhumaima and H. S. Al-Raweshidy. *Reducing energy consumption by dynamic resource allocation in C-RAN.* In 2015 European Conference on Networks and Communications (EuCNC), pages 169–174, 2015. (Cited on page 112.)

[Korrai *et al.* 2020] Praveen Kumar Korrai, Eva Lagunas, Ashok Bandi, Shree Krishna Sharma and Symeon Chatzinotas. *Joint Power and Resource Block Allocation for Mixed-Numerology-Based 5G Downlink Under Imperfect CSI.* IEEE Open Journal of the Communications Society, vol. 1, pages 1583–1601, 2020. (Cited on page 131.)

[Kumar & Gurugubelli 2011] R. V. Raja Kumar and Jagadeesh Gurugubelli. *How green the LTE technology can be?* In 2011 2nd International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace Electronic Systems Technology (Wireless VITAE), pages 1–5, 2011. (Cited on page 124.)

[Lagrange 2010] X. Lagrange. *Throughput of HARQ protocols on a block fading channel.* Communications Letters, IEEE, vol. 14, pages 257 – 259, April 2010. (Cited on pages 56, 72, 74, 75 and 160.)

[Landou & Barreto 2015] Samir Kolawolé Akanni Landou and André Noll Barreto. *Use of CoMP in 4G cellular networks for increased network energy efficiency.* In 2015 International Workshop on Telecommunications (IWT), pages 1–6, 2015. (Cited on page 88.)

[Larsen *et al.* 2019] L. M. P. Larsen, A. Checko and H. L. Christiansen. *A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks.* IEEE Communications Surveys Tutorials, vol. 21, no. 1, pages 146–172, Firstquarter 2019. (Cited on pages 40, 48, 49, 50 and 69.)

[Li *et al.* 2014] Qian Clara Li, Huaning Niu, Apostolos Tolis Papathanassiou and Geng Wu. *5G Network Capacity: Key Elements and Technologies.* IEEE

Vehicular Technology Magazine, vol. 9, no. 1, pages 71–78, 2014. (Cited on page 87.)

[Lin *et al.* 2010] Y. Lin, L. Shao, Z. Zhu, Q. Wang and R. K. Sabhikhi. *Wireless network cloud: Architecture and system requirements.* IBM Journal of Research and Development, vol. 54, no. 1, pages 4:1–4:12, 2010. (Cited on pages 10 and 22.)

[Liu *et al.* 2004] Q. Liu, S. Zhou and G.B. Giannakis. *Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links.* Wireless Communications, IEEE Transactions on, vol. 3, pages 1746 – 1755, October 2004. (Cited on pages 51, 56, 60, 76 and 159.)

[Liu *et al.* 2016] Chang Liu, Balasubramaniam Natarajan and Hongxing Xia. *Small Cell Base Station Sleep Strategies for Energy Efficiency.* IEEE Transactions on Vehicular Technology, vol. 65, no. 3, pages 1652–1661, 2016. (Cited on page 112.)

[Luo *et al.* 2015] Shixin Luo, Rui Zhang and Teng Joon Lim. *Downlink and Uplink Energy Minimization Through User Association and Beamforming in C-RAN.* IEEE Transactions on Wireless Communications, vol. 14, no. 1, pages 494–508, 2015. (Cited on pages 88 and 131.)

[Lyazidi *et al.* 2016] Mohammed Yazid Lyazidi, Nadjib Aitsaadi and Rami Langar. *Resource Allocation and Admission Control in OFDMA-Based Cloud-RAN.* In 2016 IEEE Global Communications Conference (GLOBECOM), pages 1–6, 2016. (Cited on page 130.)

[Lyazidi *et al.* 2017] Mohammed Yazid Lyazidi, Lorenza Giupponi, Josep Mangues-Bafalluy, Nadjib Aitsaadi and Rami Langar. *A Novel Optimization Framework for C-RAN BBU Selection Based on Resiliency and Price.* In 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), pages 1–6, 2017. (Cited on page 87.)

[Mahmood *et al.* 2018] Nurul Huda Mahmood, Melisa Lopez, Daniela Laselva, Klaus Pedersen and Gilberto Berardinelli. *Reliability Oriented Dual Connectivity for URLLC services in 5G New Radio.* In 2018 15th International Symposium on Wireless Communication Systems (ISWCS), pages 1–6, 2018. (Cited on page 49.)

[Mahmood *et al.* 2019] N. H. Mahmood, A. Karimi, G. Berardinelli, K. I. Pedersen and D. Laselva. *On the resource utilization of multi-connectivity transmission for URLLC services in 5G new radio.* In 2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW), pages 1–6, April 2019. (Cited on page 49.)

[Mandelbaum 1974] D. Mandelbaum. *An adaptive-feedback coding scheme using incremental redundancy (Corresp.)*. IEEE Transactions on Information Theory, vol. 20, no. 3, pages 388–389, 1974. (Cited on page 35.)

[Marsch & Fettweis 2011] Patrick Marsch and Gerhard Fettweis. *Static Clustering for Cooperative Multi-Point (CoMP) in Mobile Communications*. In 2011 IEEE International Conference on Communications (ICC), pages 1–6, 2011. (Cited on page 87.)

[Mogensen *et al.* 2007] Preben Mogensen, Wei Na, Istvan Z. Kovacs, Frank Frederiksen, Akhilesh Pokhariyal, Klaus I. Pedersen, Troels Kolding, Klaus Hugl and Markku Kuusela. *LTE Capacity Compared to the Shannon Bound*. In 2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring, pages 1234–1238, 2007. (Cited on pages 95, 101 and 135.)

[Mountaser *et al.* 2017] G. Mountaser, M. Condoluci, T. Mahmoodi, M. Dohler and I. Mings. *Cloud-RAN in Support of URLLC*. In 2017 IEEE GC Workshops, 2017. (Cited on page 48.)

[Nederlandse Spoorwegen 2017] Nederlandse Spoorwegen. *Green energy for train, bus and station*. ns.nl/en/about-ns/sustainability/climate-neutral/green-energy-for-train-bus-and-station.html, 2017. (Cited on page 85.)

[Nguyen 2018] Long D. Nguyen. *Resource Allocation for Energy Efficiency in 5G Wireless Networks*. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, vol. 5, no. 14, 6 2018. (Cited on page 130.)

[Niu *et al.* 2010] Zhisheng Niu, Yiqun Wu, Jie Gong and Zexi Yang. *Cell zooming for cost-efficient green cellular networks*. IEEE Communications Magazine, vol. 48, no. 11, pages 74–79, 2010. (Cited on page 88.)

[Ohmann *et al.* 2016] David Ohmann, Ahmad Awada, Ingo Viering, Meryem Simsek and Gerhard P. Fettweis. *Diversity Trade-Offs and Joint Coding Schemes for Highly Reliable Wireless Transmissions*. In 2016 IEEE 84th Vehicular Technology Conference (VTC-Fall), pages 1–6, 2016. (Cited on page 48.)

[Parvez *et al.* 2018] Imtiaz Parvez, Ali Rahmati, Ismail Guvenc, Arif I. Sarwat and Huaiyu Dai. *A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions*. IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pages 3098–3130, 2018. (Cited on page 29.)

[Peng *et al.* 2015] Mugen Peng, Kecheng Zhang, Jiamo Jiang, Jiaheng Wang and Wenbo Wang. *Energy-Efficient Resource Assignment and Power Allocation in Heterogeneous Cloud Radio Access Networks*. IEEE Transactions on Vehicular Technology, vol. 64, no. 11, pages 5275–5287, 2015. (Cited on page 131.)

[Penttinen 2019] J T Penttinen. 5G explained - security and deployment of advanced mobile communications. Wiley-Blackwell, Hoboken, NJ, 2019. (Cited on pages 95 and 101.)

[Proakis & Salehi 2008] J.G. Proakis and M. Salehi. Digital communications, 5th edition. McGraw-Hill Higher Education, 2008. (Cited on page 159.)

[Ren *et al.* 2018] Hong Ren, Nan Liu, Cunhua Pan, Maged Elkashlan, Arumugam Nallanathan, Xiaohu You and Lajos Hanzo. *Low-Latency C-RAN: An Next-Generation Wireless Approach.* IEEE Vehicular Technology Magazine, vol. 13, no. 2, pages 48–56, 2018. (Cited on page 39.)

[Rost *et al.* 2014] Peter Rost, Carlos J. Bernardos, Antonio De Domenico, Marco Di Girolamo, Massinissa Lalam, Andreas Maeder, Dario Sabella and Dirk Wübben. *Cloud technologies for flexible 5G radio access networks.* IEEE Communications Magazine, vol. 52, no. 5, pages 68–76, 2014. (Cited on page 40.)

[Saxena *et al.* 2016] Navrati Saxena, Abhishek Roy and HanSeok Kim. *Traffic-Aware Cloud RAN: A Key for Green 5G Networks.* IEEE Journal on Selected Areas in Communications, vol. 34, no. 4, pages 1010–1021, 2016. (Cited on page 87.)

[SCF 2016] SCF. *Small Cell Virtualization Functional Splits and Use Cases, .* Technical report Release 7.0, Small Cell Forum, January 2016. (Cited on page 40.)

[Schwarz *et al.* 2010] Stefan Schwarz, Christian Mehlführer and Markus Rupp. *Low complexity approximate maximum throughput scheduling for LTE.* In 2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers, pages 1563–1569, 2010. (Cited on page 130.)

[Shiozaki 1996] A. Shiozaki. *Adaptive type-II hybrid broadcast ARQ system.* IEEE Transactions on Communications, vol. 44, no. 4, pages 420–422, 1996. (Cited on page 35.)

[Siddiqi *et al.* 2019] Murtaza Ahmed Siddiqi, Heejung Yu and Jingon Joung. *5G Ultra-Reliable Low-Latency Communication Implementation Challenges and Operational Issues with IoT Devices.* Electronics, vol. 8, no. 9, 2019. (Cited on page 29.)

[Sigwele *et al.* 2015] Tshiamo Sigwele, Atm Shafiul Alam, Prashant Pillai and Y. Fun Hu. *Evaluating Energy-Efficient Cloud Radio Access Networks for 5G.* In 2015 IEEE International Conference on Data Science and Data Intensive Systems, pages 362–367, 2015. (Cited on pages 86 and 131.)

[Soret *et al.* 2014] Beatriz Soret, Preben Mogensen, Klaus I. Pedersen and Mari Carmen Aguayo-Torres. *Fundamental tradeoffs among reliability, latency and throughput in cellular networks.* In 2014 IEEE Globecom Workshops (GC Wkshps), pages 1391–1396, 2014. (Cited on pages 9 and 22.)

[Strodthoff *et al.* 2019] N. Strodthoff, B. Göktepe, T. Schierl, C. Hellge and W. Samek. *Enhanced Machine Learning Techniques for Early HARQ Feedback Prediction in 5G.* IEEE Journal on Selected Areas in Communications, vol. 37, no. 11, pages 2573–2587, 2019. (Cited on page 49.)

[Swamy *et al.* 2015] Vasuki Narasimha Swamy, Sahaana Suri, Paul Rigge, Matthew Weiner, Gireeja Ranade, Anant Sahai and Borivoje Nikolić. *Cooperative communication for high-reliability low-latency wireless control.* In 2015 IEEE International Conference on Communications (ICC), pages 4380–4386, 2015. (Cited on page 48.)

[Taleb *et al.* 2020] Hussein Taleb, Kinda Khawam, Samer Lahoud, Melhem El Helou and Steven Martin. *A fully distributed approach for joint user association and RRH clustering in cloud radio access networks.* Computer Networks, vol. 182, page 107445, 2020. (Cited on page 132.)

[Tang *et al.* 2015] Jianhua Tang, Wee Peng Tay and Tony Q. S. Quek. *Cross-Layer Resource Allocation With Elastic Service Scaling in Cloud Radio Access Network.* IEEE Transactions on Wireless Communications, vol. 14, no. 9, pages 5068–5081, 2015. (Cited on page 131.)

[Tang *et al.* 2019] Jianhua Tang, Byonghyo Shim and Tony Q. S. Quek. *Service Multiplexing and Revenue Maximization in Sliced C-RAN Incorporated With URLLC and Multicast eMBB.* IEEE Journal on Selected Areas in Communications, vol. 37, no. 4, pages 881–895, 2019. (Cited on page 22.)

[Taniyama *et al.* 2019] Kentarou Taniyama, Yoshihisa Kishiyama and Kenichi Higuchi. *Low Latency HARQ Method Using Early Retransmission Prior to Channel Decoding with Multistage Decision.* In 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), pages 1–5, 2019. (Cited on page 49.)

[Temesgene *et al.* 2018] Dagnachew A. Temesgene, Nicola Piovesan, Marco Miozzo and Paolo Dini. *Optimal Placement of Baseband Functions for Energy Harvesting Virtual Small Cells.* In 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), pages 1–6, 2018. (Cited on page 156.)

[Tian *et al.* 2017] Fengyu Tian, Peng Zhang and Zheng Yan. *A Survey on C-RAN Security.* IEEE Access, vol. 5, pages 13372–13386, 2017. (Cited on page 155.)

[TR 36.819 2013] TR 36.819. *Coordinated multi-point operation for LTE physical layer aspects.* Technical report, 3GPP (3rd Generation Partnership Project), September 2013. V11.2.0. (Cited on pages 49, 87 and 88.)

[TR 36.942 2020] TR 36.942. *Technical Specification Group Radio Access Network; Evolved universal terrestrial radio access (E-UTRA); radio frequency (RF) system scenarios.* Technical report, 3GPP (3rd Generation Partnership Project), June 2020. V14.0.0. (Cited on pages 67 and 94.)

[TR 38.801 2017] TR 38.801. *Study on new radio access technology; radio access architecture and interfaces.* Technical report, 3GPP (3rd Generation Partnership Project), March 2017. V14.0.0. (Cited on pages 39, 40, 48, 49 and 52.)

[TR 38.802 2017] TR 38.802. *Study on new radio access technology physical layer aspects.* Technical report, 3GPP (3rd Generation Partnership Project), September 2017. V14.2.0. (Cited on page 30.)

[TR 38.901 2019] TR 38.901. *Technical Specification Group Radio Access Network; Study on channel model for frequencies from 0.5 to 100 GHz.* Technical report, 3GPP (3rd Generation Partnership Project), October 2019. V16.0.0. (Cited on pages 94, 101 and 165.)

[TR 38.913 2017] TR 38.913. *5G; Study on Scenarios and Requirements for Next Generation Access Technologies.* Technical report, 3GPP (3rd Generation Partnership Project), June 2017. V14.3.0. (Cited on pages 29 and 76.)

[TS 22.261 2020] TS 22.261. *5G; Service requirements for next generation new services and markets.* Technical report, 3GPP (3rd Generation Partnership Project), Septeember 2020. V18.0.0. (Cited on pages 29 and 31.)

[TS 38.211 2022] TS 38.211. *5G NR physical channels and modulation.* Technical report, 3GPP (3rd Generation Partnership Project), March 2022. V17.1.0. (Cited on pages 37, 60, 76 and 101.)

[TS 38.401 2018] TS 38.401. *NG-RAN; Architecture description.* Technical report, 3GPP (3rd Generation Partnership Project), September 2018. V15.3.0. (Cited on pages 31 and 34.)

[Videv & Haas 2011] Stefan Videv and Harald Haas. *Energy-Efficient Scheduling and Bandwidth-Energy Efficiency Trade-Off with Low Load.* In 2011 IEEE International Conference on Communications (ICC), pages 1–5, 2011. (Cited on page 132.)

[Yoshizawa *et al.* 2019] Takahito Yoshizawa, Sheeba Backia Mary Baskaran and Andreas Kunz. *Overview of 5G URLLC System and Security Aspects in 3GPP.* In 2019 IEEE Conference on Standards for Communications and Networking (CSCN), pages 1–5, 2019. (Cited on page 155.)

[Zhang *et al.* 2016a] Deyu Zhang, Zhigang Chen, Lin X. Cai, Haibo Zhou, Ju Ren and Xuemin Shen. *Resource Allocation for Green Cloud Radio Access Networks Powered by Renewable Energy.* In 2016 IEEE Global Communications Conference (GLOBECOM), pages 1–6, 2016. (Cited on page 156.)

[Zhang *et al.* 2016b] Jiawei Zhang, Yuefeng Ji, Xiangzi Xu, Hui Li, Yongli Zhao and Jie Zhang. *Energy efficient baseband unit aggregation in cloud radio and*

optical access networks. Journal of Optical Communications and Networking, vol. 8, no. 11, pages 893–901, 2016. (Cited on page 87.)

[Zheng *et al.* 2014] Feng Zheng, Turner Whitted, Anselmo Lastra, Peter Lincoln, Andrei State, Andrew Maimone and Henry Fuchs. *Minimizing latency for augmented reality displays: Frames considered harmful*. In 2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pages 195–200, 2014. (Cited on page 31.)

[Zuo *et al.* 2016] Jun Zuo, Jun Zhang, Chau Yuen, Wei Jiang and Wu Luo. *Energy Efficient User Association for Cloud Radio Access Networks*. IEEE Access, vol. 4, pages 2429–2438, 2016. (Cited on page 87.)

List of publications

International conferences

- T. Alhajj and X. Lagrange, "Reliability and Low Latency: Impact of The Architecture," 2020 IEEE Symposium on Computers and Communications (ISCC), Rennes, France, 2020, pp. 1-6

  doi: 10.1109/ISCC50000.2020.9219636.

- T. Alhajj and X. Lagrange, "Impact of the RAN Architecture and Macro-diversity Techniques on Latency," 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall), Norman, OK, USA, 2021, pp. 1-7

  doi: 10.1109/VTC2021-Fall52928.2021.9625386.

Technical report

- T. Alhajj, X. Lagrange, "Computation of the probability of error for HARQ-CC with macro-diversity based on MRC," [Research Report] RR-2020-01-SRCD, IMT ATLANTIQUE. 2020. ⟨hal-02949519⟩

COLLEGE SCIENCES
DOCTORAL POUR L'INGENIEUR
BRETAGNE ET LE NUMERIQUE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Titre : Impact du réseau d'accès radio centralisé sur les performances de la 5G

Mot clés : 5G, C-RAN, HARQ, latence, fiabilité, consommation d'énergie, ressources radio.

Résumé : Les réseaux mobiles de cinquième génération (5G) ouvrent la voie à une nouvelle architecture de réseau d'accès radio (RAN). Il s'agit du RAN centralisé (C-RAN) qui regroupe certaines des fonctions de la station de base (BS) dans une unité centrale (CU) connectée à des unités radio (RU) distribuées sur différents sites où sont implémentées les fonctions restantes de la BS. Cette thèse étudie comment tirer parti du C-RAN pour concilier haute fiabilité et faible latence d'une part et pour minimiser la consommation d'énergie de la BS d'autre part. Les transmissions mono et multi-RU, ainsi qu'un mélange des deux, sont évaluées. Nous comparons une approche mono-RU où le mécanisme de retransmission HARQ, qui combine toutes les retransmissions ARQ, est situé dans le RU à une transmission multi-RU avec un HARQ centralisé dans le CU. Nous montrons que les transmissions multi-RU offrent une fiabilité élevée et une faible latence grâce à la diversité spatiale même si la centralisation augmente le temps aller-retour. Nous considérons ensuite l'allocation de ressources radios. Nous évaluons l'énergie consommée par les BSs par un modèle de consommation qui intègre la puissance de transmission et l'énergie consommée pour le traitement. À faible charge, les transmissions multi-RU dans un C-RAN, où toutes les RUs servent un utilisateur, économisent la consommation d'énergie des BSs, sans dégrader la couverture, par rapport à la desserte d'un utilisateur avec une seule RU. Nous posons et résolvons un problème d'optimisation pour minimiser la consommation d'énergie et augmenter la capacité du système à une charge modérée en réutilisant les ressources radio entre les RUs.

Title: Impact of centralized-radio access network architecture on 5G performance

Keywords: 5G, C-RAN, HARQ, latency, reliability, energy consumption, radio resources.

Abstract: Fifth Generation (5G) mobile networks are paving the way for a new Radio Access Network (RAN) architecture. This is the Centralized-RAN (C-RAN) which groups some of the Base Station (BS) functions in a Central Unit (CU) connected to Radio Units (RU) distributed in different sites where the other BS functions are implemented. This thesis studies how to take advantage of C-RAN to combine high reliability and low latency on the one hand and to minimize the power consumption of the BS on the other hand. Both single and multi-RU transmissions, as well as a mixture of both, are evaluated. We compare a single-RU approach where the Hybrid Automatic Repeat reQuest (HARQ) mechanism, which combines all the ARQ re-transmissions, is located in the RU to a multi-RU transmission with an HARQ centralized in the CU. We show that multi-RU transmissions provide high reliability and low latency due to spatial diversity, even though centralization increases the round trip time. We then consider radio resource allocation. We evaluate the energy consumed by the BSs using a consumption model that integrates both the transmission power and the energy consumed for processing. At low load, multi-RU transmissions in a C-RAN, where all RUs serve one user, save BS energy consumption, without coverage degradation, compared to serving a user with a single RU. We pose and solve an optimization problem to minimize the energy consumption and increase the capacity to moderate load by reusing radio resources between RUs.